

# Scaling cloud for LSST catalog at IN2P3

**Fabrice Jammes**

Scalable Data Systems Expert  
IN2P3

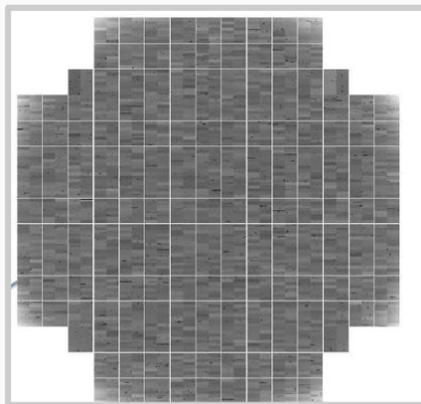
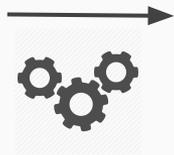
**Frédéric Gaudet**

Openstack Cloud architect  
CNRS

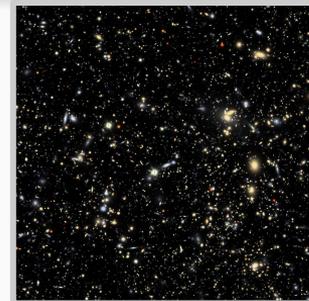
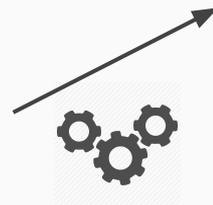
**Nicolas Chotard**

Researcher  
IN2P3

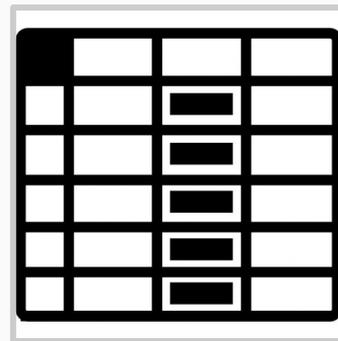
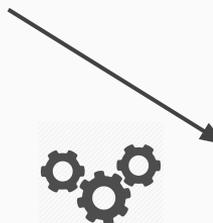
# 80+ PB of astronomical catalog



Raw data



Processed image



**Catalog** (stars, galaxies, objects, sources, transients, exposures, etc.)

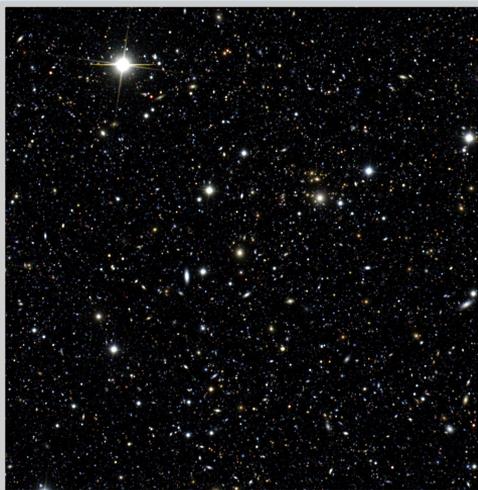
LSST will build a catalog of 20 billion galaxies and 17 billion stars and their associated physical properties

# Data

## Images

Persisted: **~38 PB**

Temporary: **~½ 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 schema



**LSST Database Schema Browser** *alpha*

Schema versions available for browsing: [baseline](#) | [DC3a](#) | [PT1\\_1](#) | [PT1\\_2](#) | [ImSim](#) | [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 CcdVisit CcdVisitMetadata DiaForcedSource DiaObject DiaObject_To_Object_Match DiaSource ForcedSource LeapSeconds Object 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	<p>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.</p> <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 ProcessingHistory 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 ProcessingHistory 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 ProcessingHistory 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.																																																																																																														

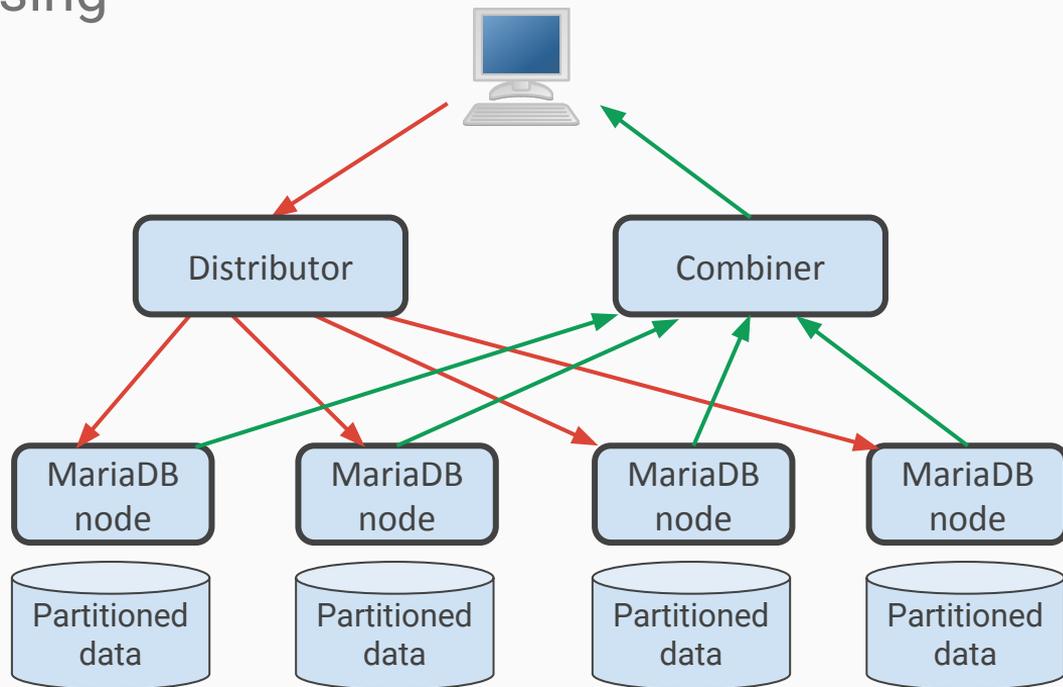
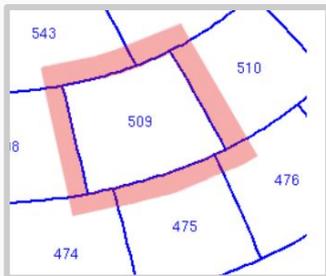
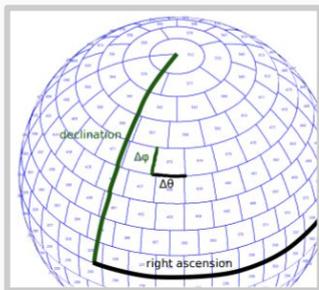
<http://ls.st/s91>

# 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



On the french side

# Who we are: French side

## Research and Engineering

- ★ Oualid Achbal: Cloud-Computing, CI
- ★ Christian Arnault: Alternate solutions (MongoDB, Spark)
- ★ Sébastien Binet: Containers, Orchestration
- ★ Nicolas Chotard: Data loading
- ★ Vincent Gatignol: Openstack, CI
- ★ Frédéric Gaudet: Openstack, CEPH
- ★ Fabrice Jammes: Qserv development
- ★ Amine Mesmoudi: Alternate solutions (Hadoop)
- ★ Bogdan Vulpescu: Data loading



# Who we are: French side, CC-IN2P3

## Operation Team @ CC-IN2P3

- ★ Fabio Hernandez: Coordinator
- ★ Osman Aidel: Database, Spark
- ★ Yvan Calas: 50 nodes cluster management
- ★ Mathieu Puel: System Administration
- ★ Loïc Tortay: Shared Storage
- ★ Fabien Wernli: Monitoring



Fabio  
Hernandez



Fabien Wernli



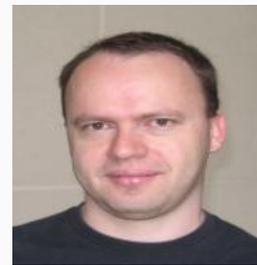
Osman Aidel



Loïc Tortay



Mathieu Puel

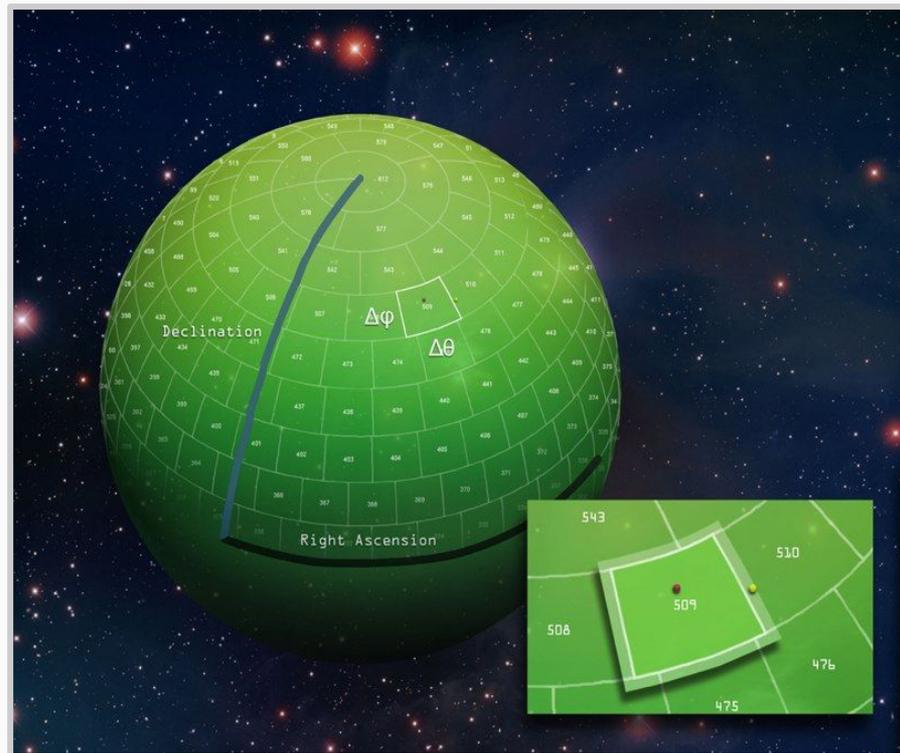


Yvan Calas

# What we do

## Data Access and Database

- ★ Qserv scientific validation
- ★ Preparing production at Large Scale:
  - Continuous integration (SQUARE)
  - Integration and Large Scale tests
  - Deployment
  - Monitoring
  - Orchestration
- ★ Study and design deployment on modern infrastructure
  - Distributed storage: CEPH
  - Containerization
  - Kubernetes
  - Openstack
- ★ Alternate solution testing (Christian Arnault, A. Mesmoudi)



# Tests and demonstrations

Target for production

~500 nodes clusters in 2 international data-centers

Running now

**Development platform (CC-IN2P3)**

*400 cores, 800 GB memory*

*500 TB storage,*

**=> ~65 TB data set on 2\*25 nodes**

**Prototype Data Access Center (NCSA)**

*500 cores, 4 TB memory*

*700 TB storage,*

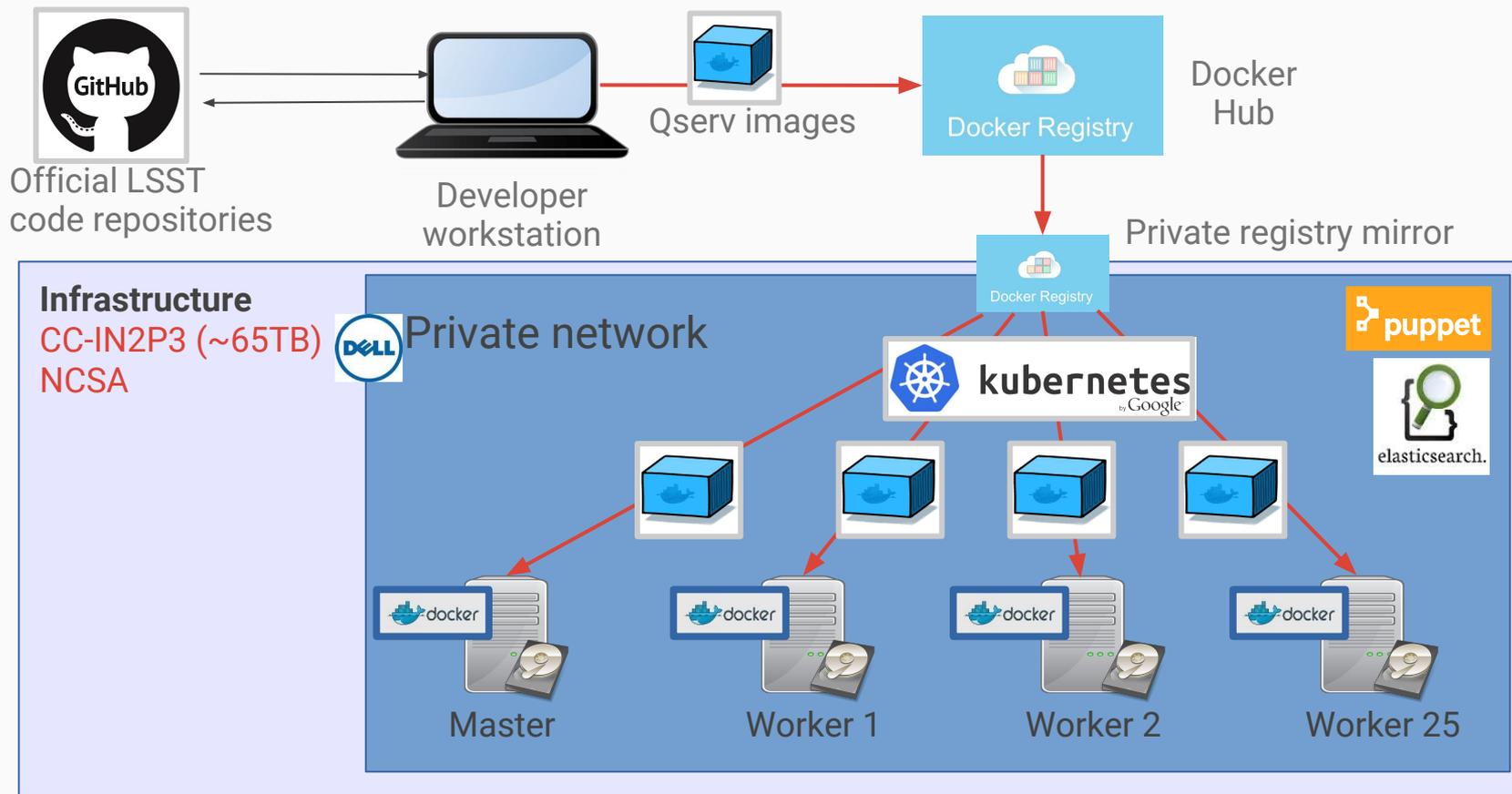
*WISE data loaded*



# Large Scale Tests: All about the data

- ★ The data set comes from the Stripe 82 released by the SDSS telescope. It has been processed by the Qserv team and duplicated to reach 35 TB in 2015 and 65 TB in 2017.
- ★ No real scientific meaning (yet !)

# Automated deployment: bare-metal



- ★ PetaSky : « Gestion et exploration des grandes masses de données scientifiques issues d'observations astronomiques grand champ »
- ★ PetaSky uses test dataset from both LSST and Euclid
- ★ Involved labs : LIMOS, LIRIS, LPC, APC, LAL, LaBRI, LIF , LIRMM, LAM, CC-IN2P3

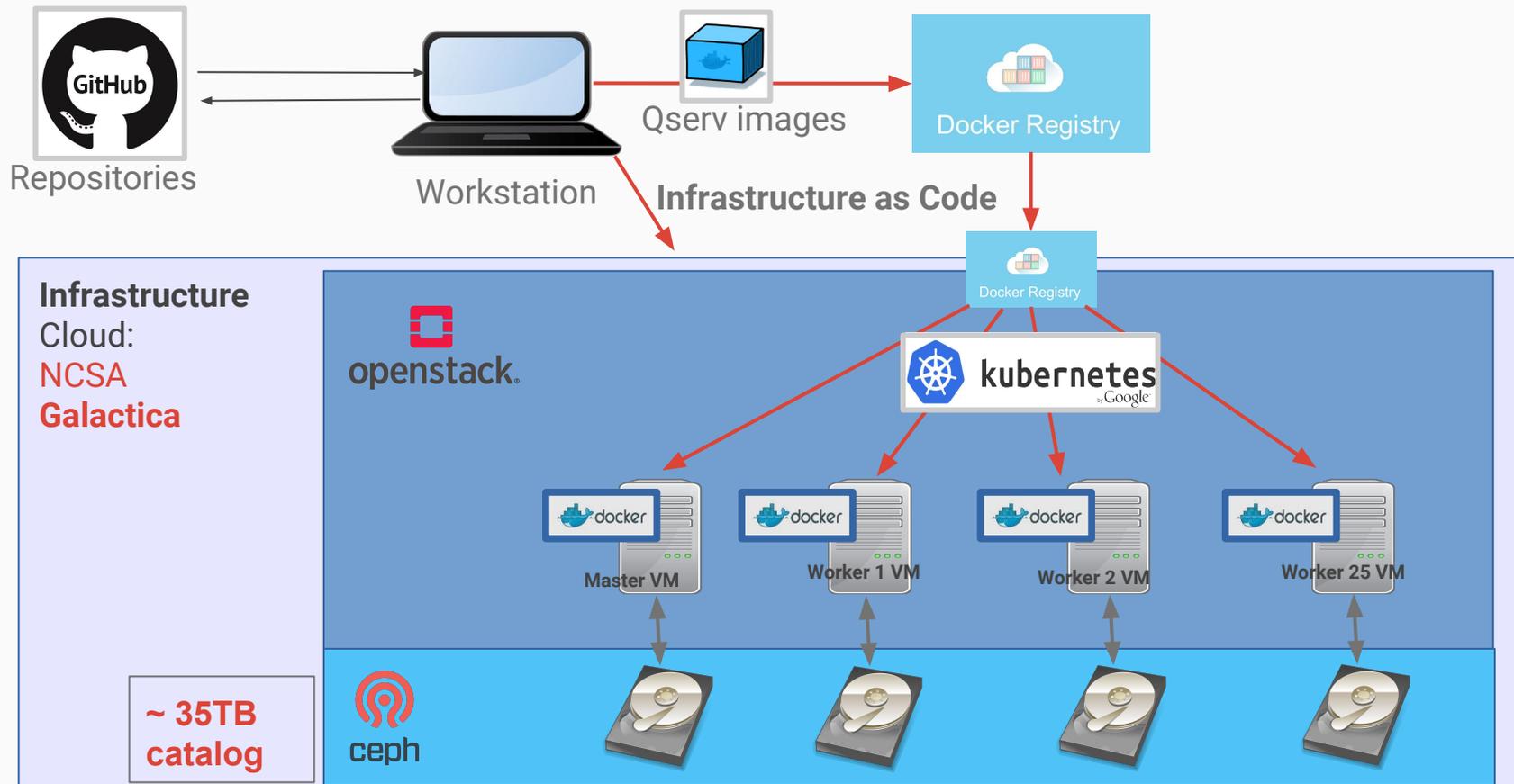
# Galactica: experiment objectives

- ★ Set up a **Large Scale Continuous Integration** platform:  
benchmark QServ releases against a 35TB data set
- ★ Prototype **Qserv deployment/orchestration over the cloud**  
QServ will land in production datacenter (CC-IN2P3/NCSA)
- ★ Experiment how **Cloud can scale to Big-Data**  
OpenStack compute nodes + Ceph storage

# Key figures

- ★ CEPH side: 24\*1,5 TB virtual disks → 35TB of data
- ★ Openstack: 25 VM using 16GB RAM and 2 vCPU each

# Automated deployment: Openstack+CEPH



# Automated deployment: advanced CI

Real-time validation of Qserv code against distributed infrastructure



Official LSST code repositories

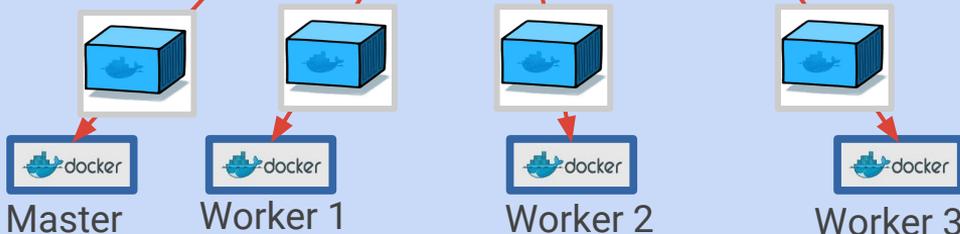
Send multi-channel notification (slack, email, qserv documentation, travis website)

Infrastructure  
 Travis CI

SAAS CI server

Automatically:  
- build and configure containers  
- start cluster  
- launch integration tests

Ephemeral and virtual fresh Qserv cluster



# What does bring an elastic infrastructure?

## Openstack+containers

In a few seconds:

- Provision a Qserv cluster
- Package and deploy development version

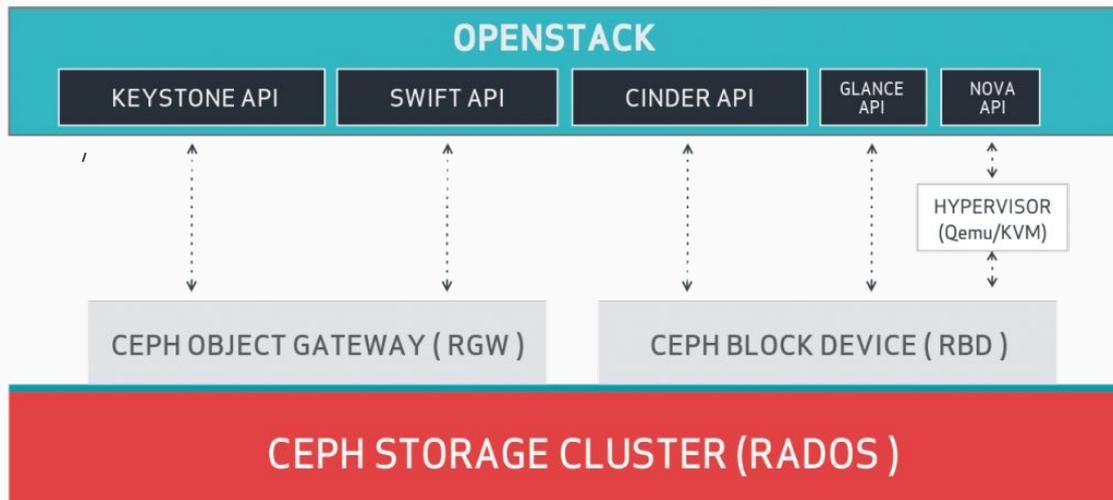
## CEPH

Enable large storage access via Openstack

Provide:

- Data replication
- Data high availability
- Data reconstruction

=> Currently trying to make it scale



# What does bring an elastic infrastructure?

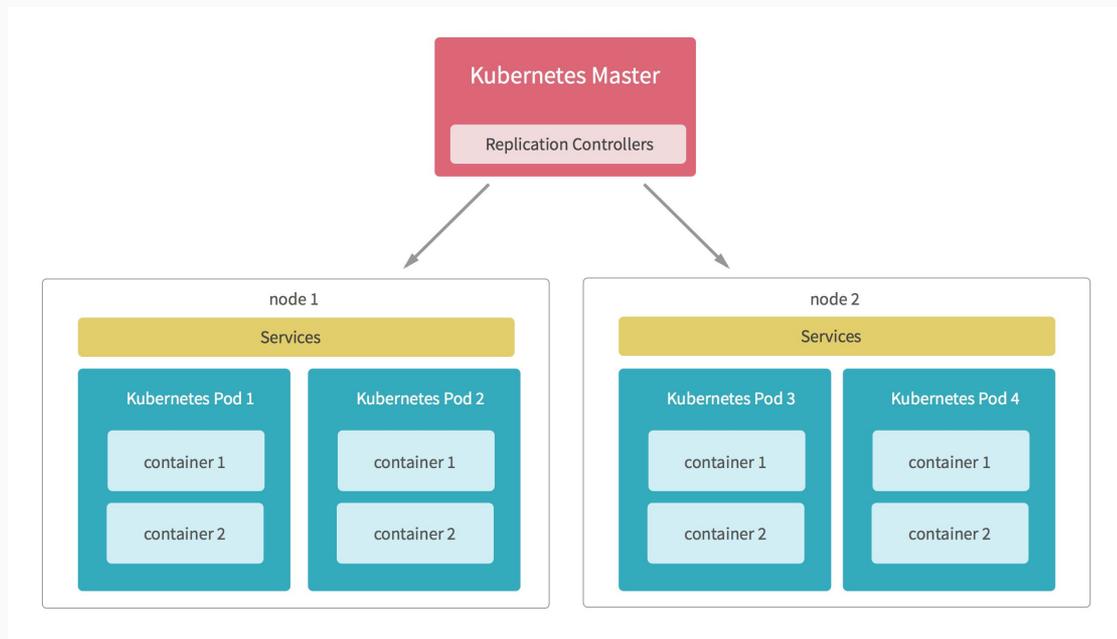
## Kubernetes

### Orchestrate Qserv processes

*Can scale to hundred of nodes*

Provide:

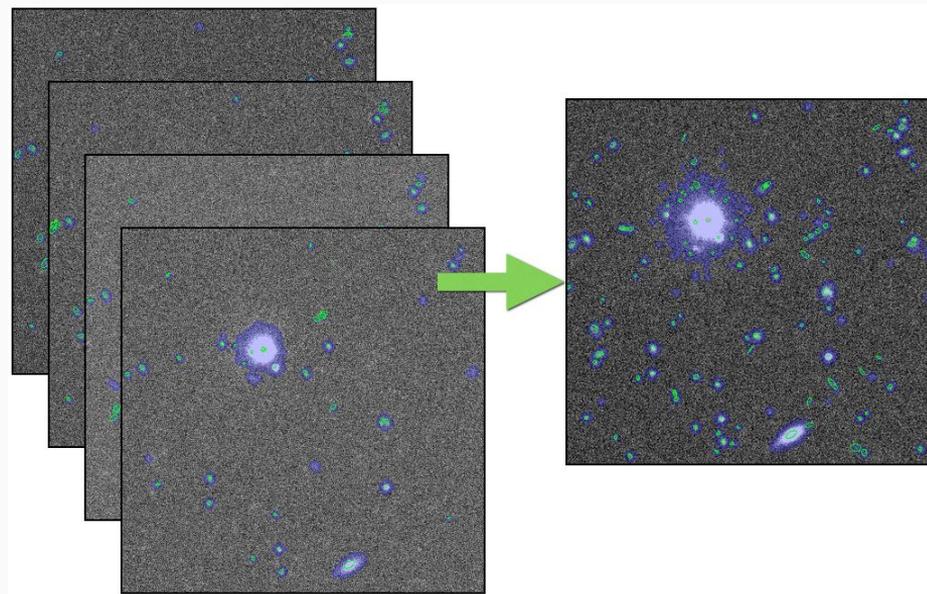
- Container placement
- Auto-scaling
- Auto-healing
- Volume management (storage)
- Resource usage monitoring
- Health checks
- Rolling update



# Qserv integration into science pipelines

Test Qserv on real data processed through the LSST stack

- ★ Process a dataset, and produce the catalogs using LSST Stack
- ★ Create a Qserv instance ready to ingest data
- ★ Transform the catalogs into a data format that Qserv can ingest
- ★ Load the catalogs into Qserv
- ★ Create a set of queries to test basic Qserv features
- ★ Run these queries using python
- ★ Implement into Clusters pipeline
- ★ Extend to other analysis



<https://github.com/nicolaschotard/qserv>

Conceptual data model for stack output ?

# Alternate solutions

Hive/HadoopDB (A. Mesmoudi & al)

**Benchmarking SQL on MapReduce systems using large astronomy databases**

DOI: [10.1007/s10619-014-7172-8](https://doi.org/10.1007/s10619-014-7172-8)

MongoDB (C. Arnault)

Test using Qserv S15 dataset

- Data for one worker node (1.3 TB)
- Using Openstack/Galactica

*Very promising results on simple queries (i.e. selection on indexed fields)*

*Investigating on:*

- ★ *Joins between large tables*
- ★ *Near-neighbors*
- ★ *Non-indexed selections*
- ★ *Data-distribution (chunk overlap distribution)*

*Moving toward Spark/Dataframe/GeoSpark*



# XLDB 2017 in Europe

Session and chairs:

1.5 days for main conference  
~1 day for Hackaton

## Polystores

- **Patrick Valduriez, INRIA:** Senior researcher, head of Zenith research team.

## Applications: earth and astronomy, neuroscience

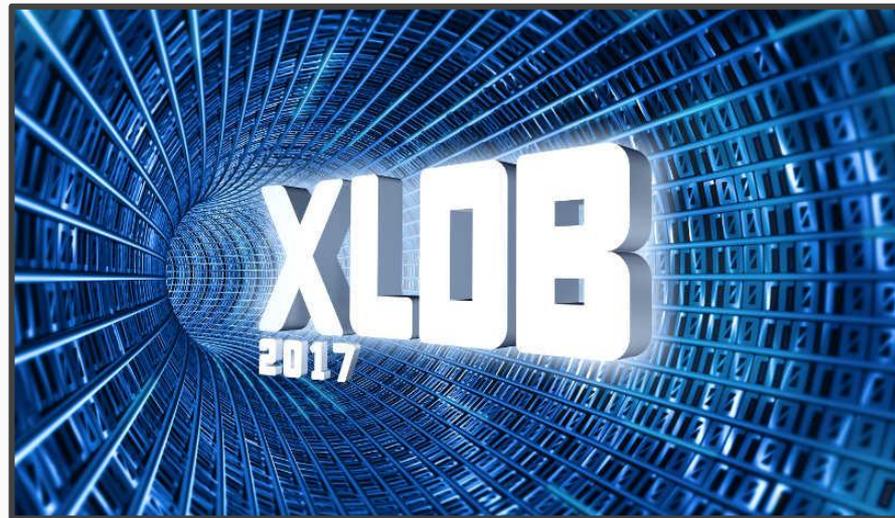
- **Peter Baumann, Jacobs University:** Professor and head of the Large-Scale Scientific Information Systems research group.
- **Romulo Goncalves, Nederland eScience Center:** Expert in Databases, Data Structures, Distributed Computing.

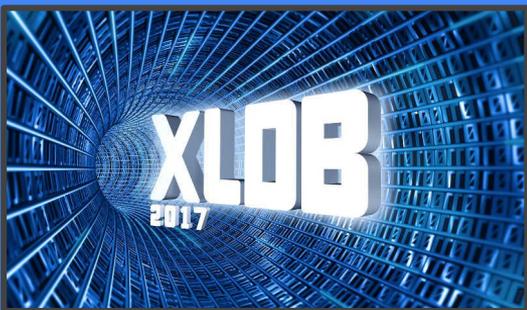
## Modern data management

- **Anastasia Ailamaki, EPFL:** Professor and Lab Director.
- **Mohand-Said Hacid, LIRIS:** Professor and Lab Director.

## Scaling Cloud to Big Data

- **Dirk Duellmann, CERN:** Deputy leader of the data and storage services group in CERN's IT.
- **Yannick Legré, EGI:** Managing director.





See you in autumn 2017 Credits: CRDTA  
in Clermont-Ferrand  
<http://xldb2017.uca.fr>



# Summary

- ★ French scientists are very interested in LSST catalog challenge
  - Alternate solution testing
  - Elastic deployment, prepare production
  - Integration with LSST stack
  
- ★ 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)

# Thanks!

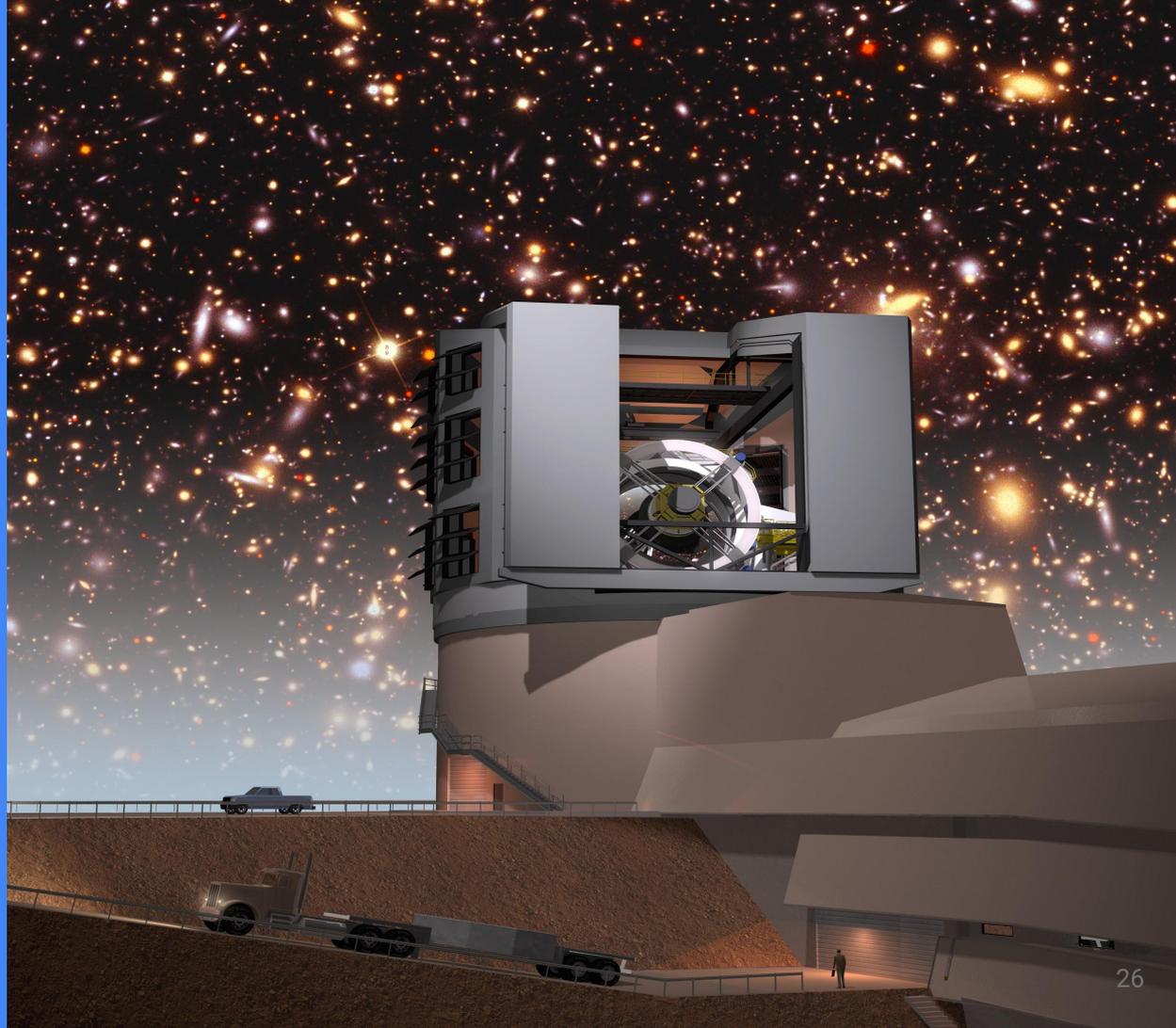
Contact:

Fabrice JAMMES

LPC

Clermont-Ferrand

[fabrice.jammes@in2p3.fr](mailto:fabrice.jammes@in2p3.fr)



# Implementation Details

