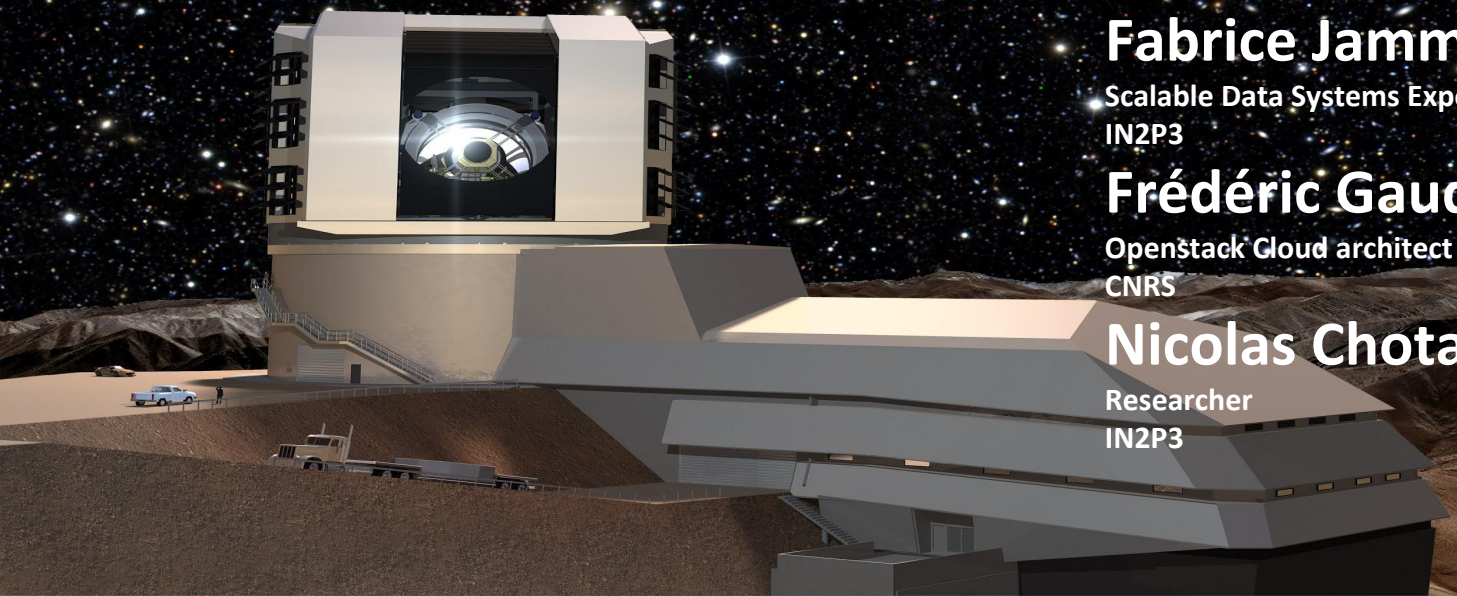
Scaling cloud for LSST catalog at IN2P3

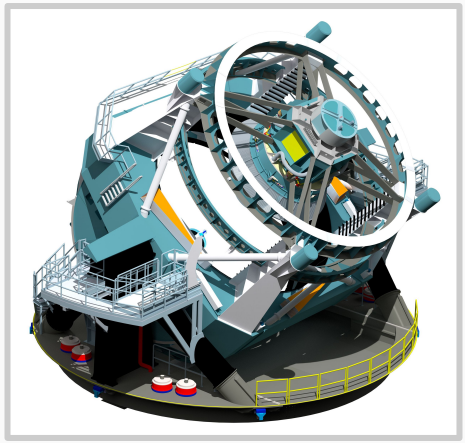


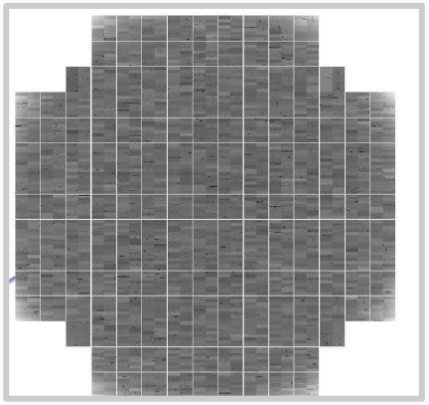
Fabrice Jammes
Scalable Data Systems Expert
IN2P3

Frédéric Gaudet
Openstack Cloud architect
CNRS

Nicolas Chotard
Researcher
IN2P3

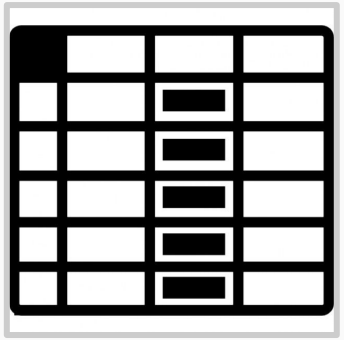
80+ PB of astronomical catalog





Raw data

Processed image

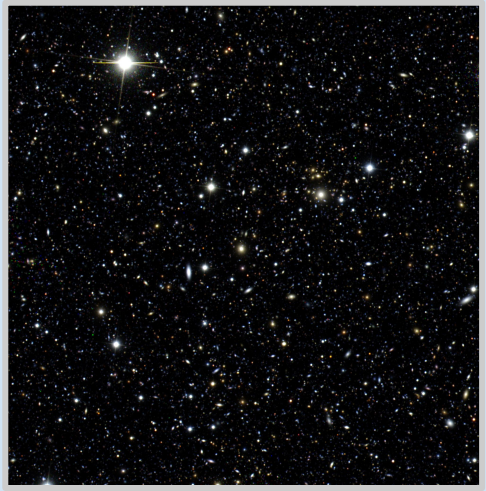


Catalog (stars, galaxies, objects, sources,transients, exposures, etc.)

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**

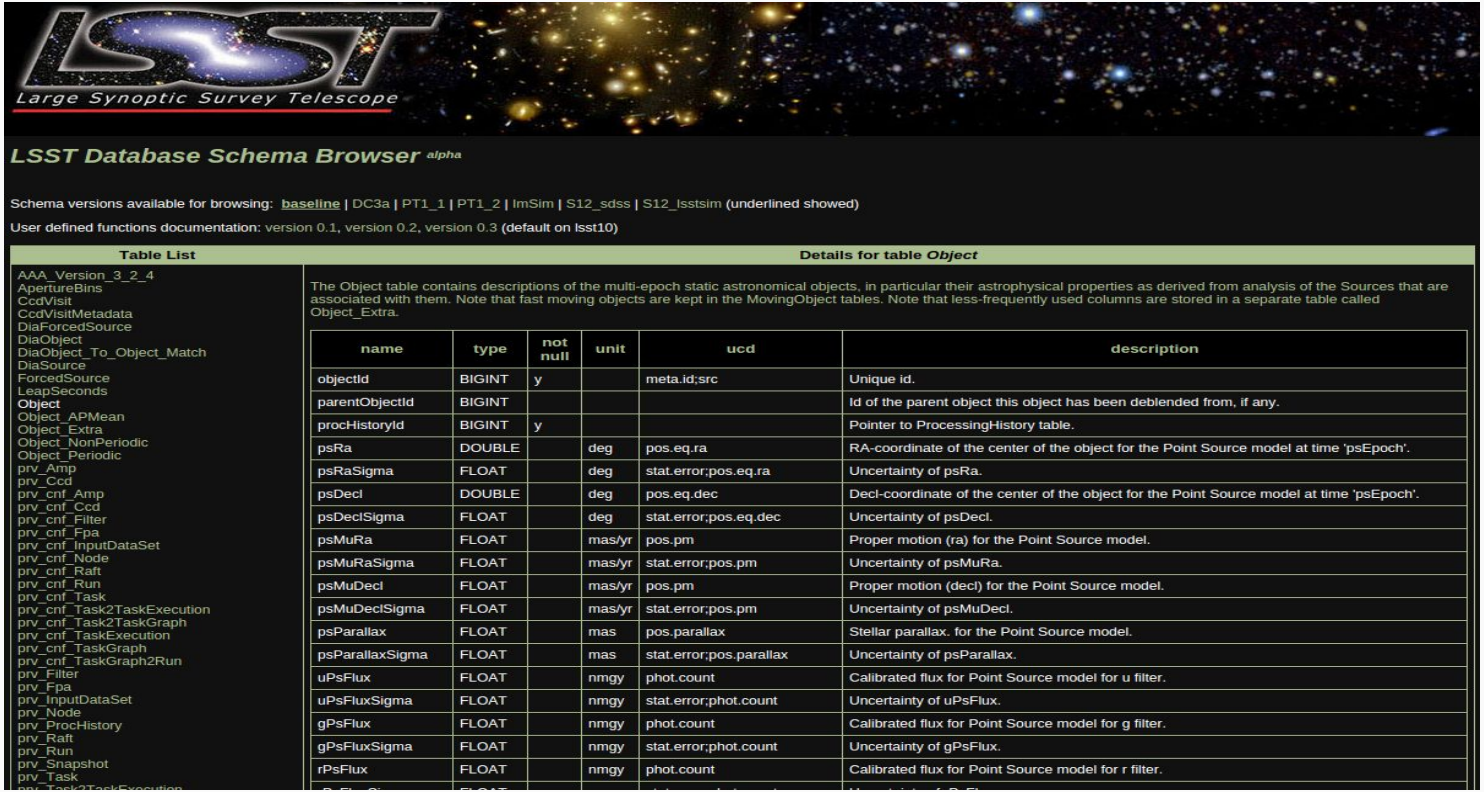


★ ~3 million “visits”
★ ~47 billion“objects”
★ ~9 trillion “detections”

★ 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



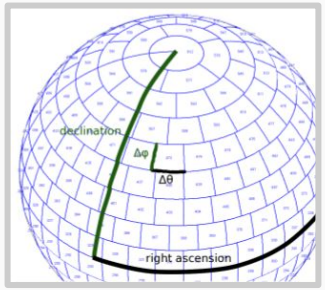
http://ls.st/s91

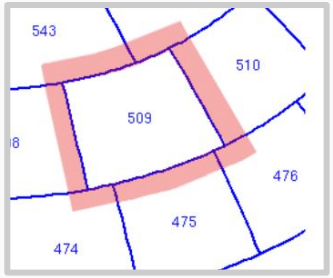
Implementation Strategy

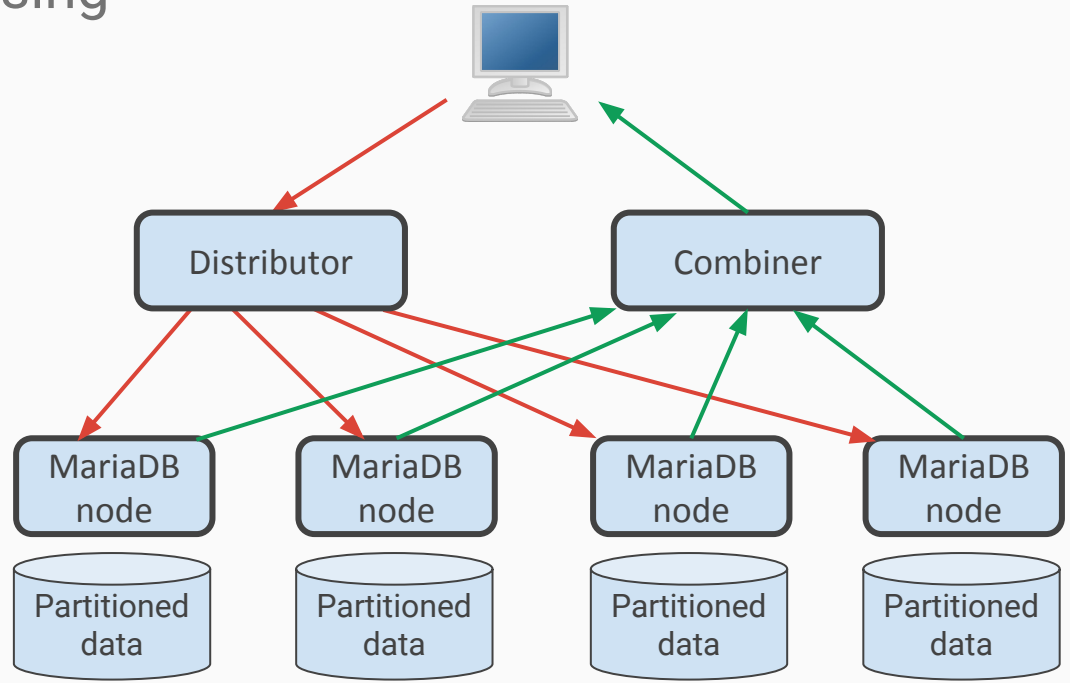
- ★ 100% Open source
- ★ Keep it flexible
- ★ Hide complexity
- ★ Reuse existing components:
  - MariaDB, MySQL Proxy, XRootD, Google protobuf, Flask
- ★ Plus custom glue
  - C++, a bit of python, some ANTLR
  - Lots of multithreading, callbacks, mutexes and sockets
- ★ And custom UDFs

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

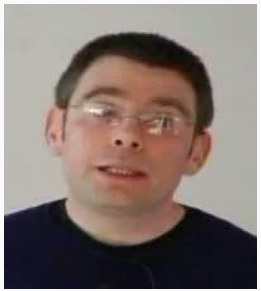
Operation Team @ CC-IN2P3

- ★ 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

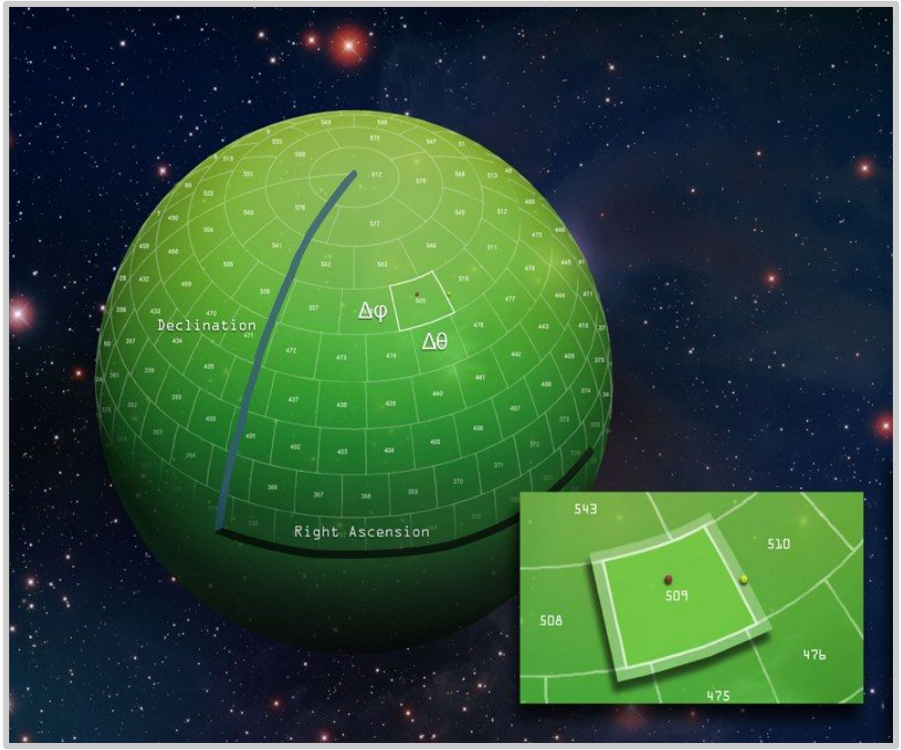
Mathieu Puel

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)

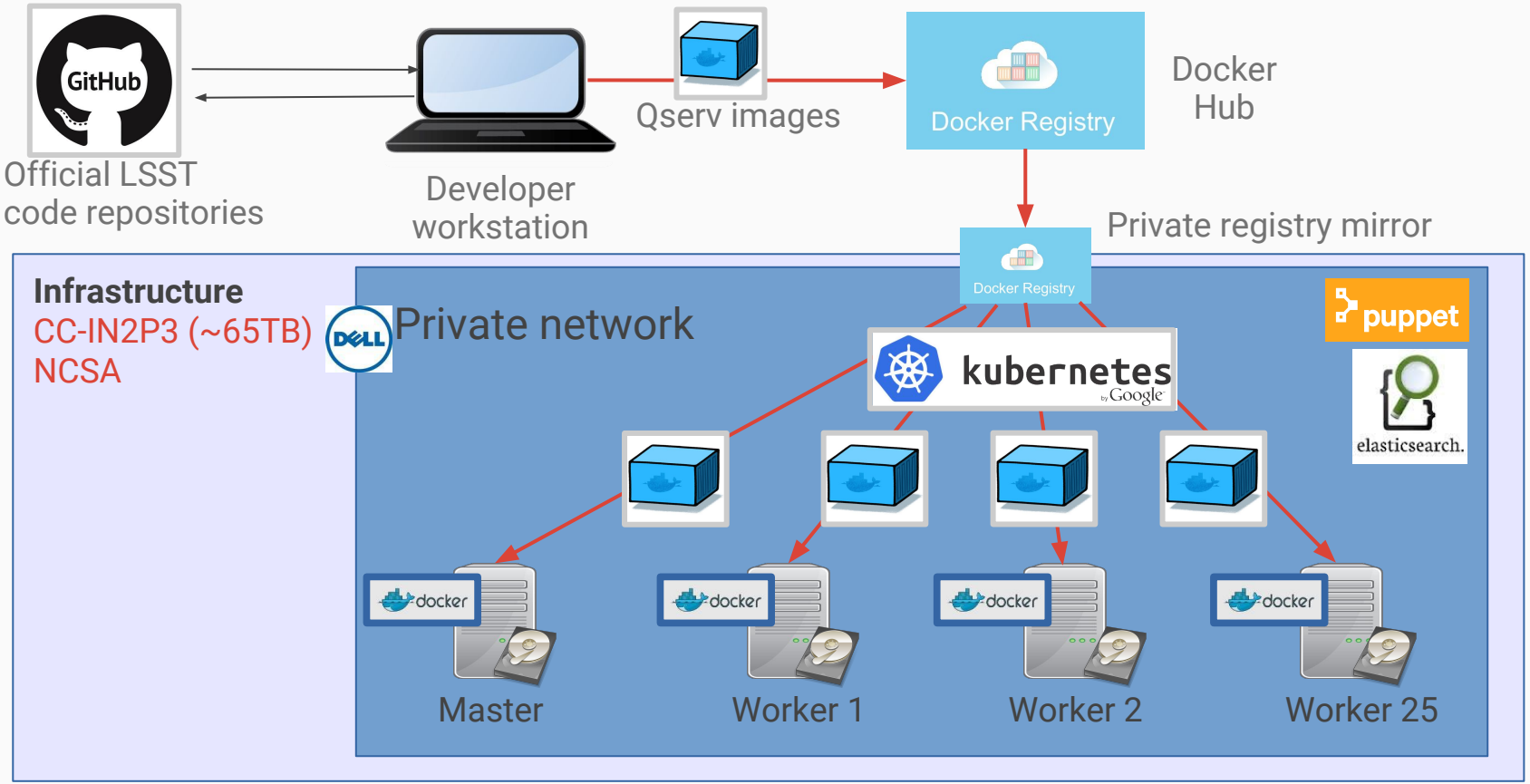
500 cores, 4 TB memory
700 TB storage,
WISE data loaded

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

- ★ 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

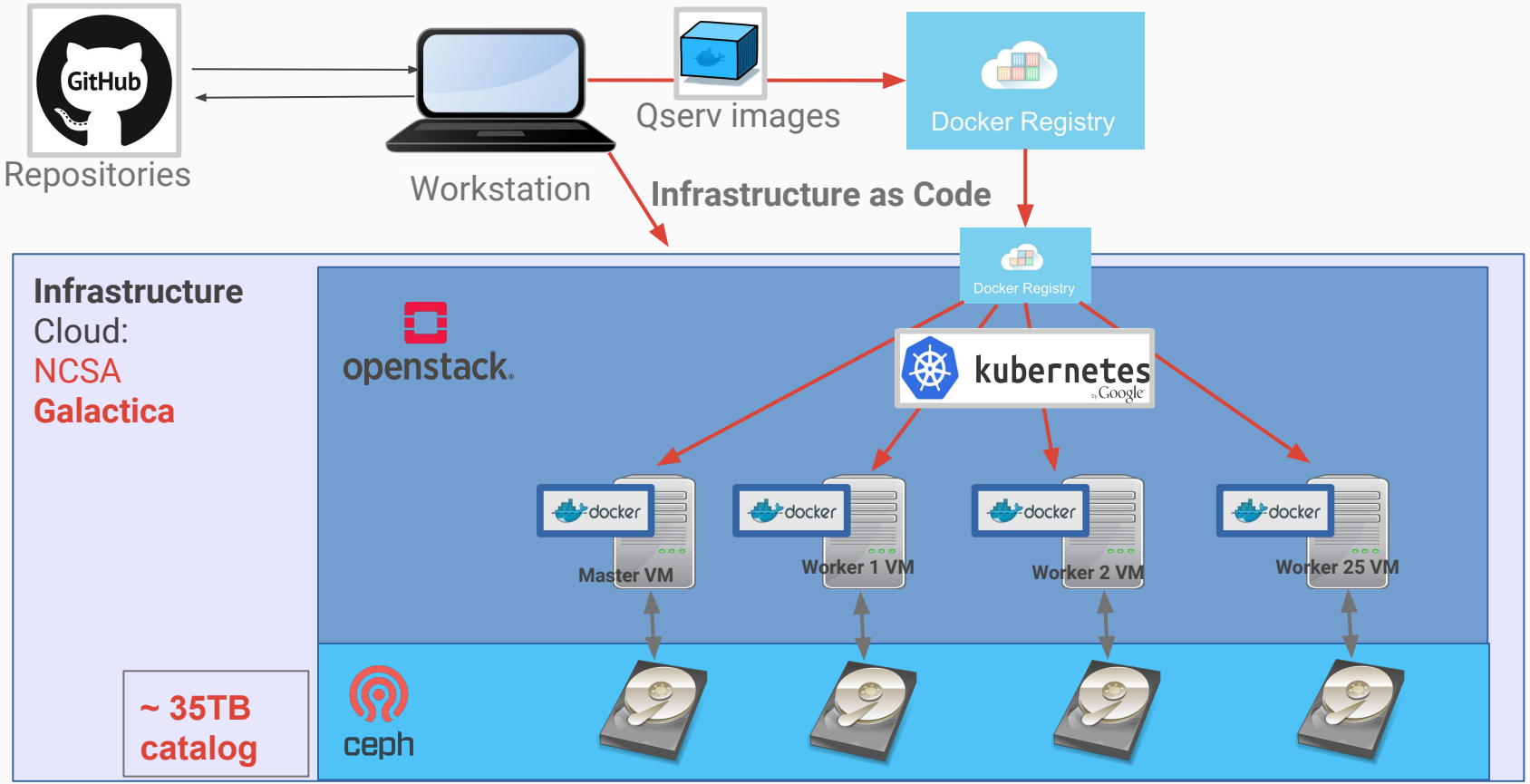
Galactica: experiment objectives

- ★ 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

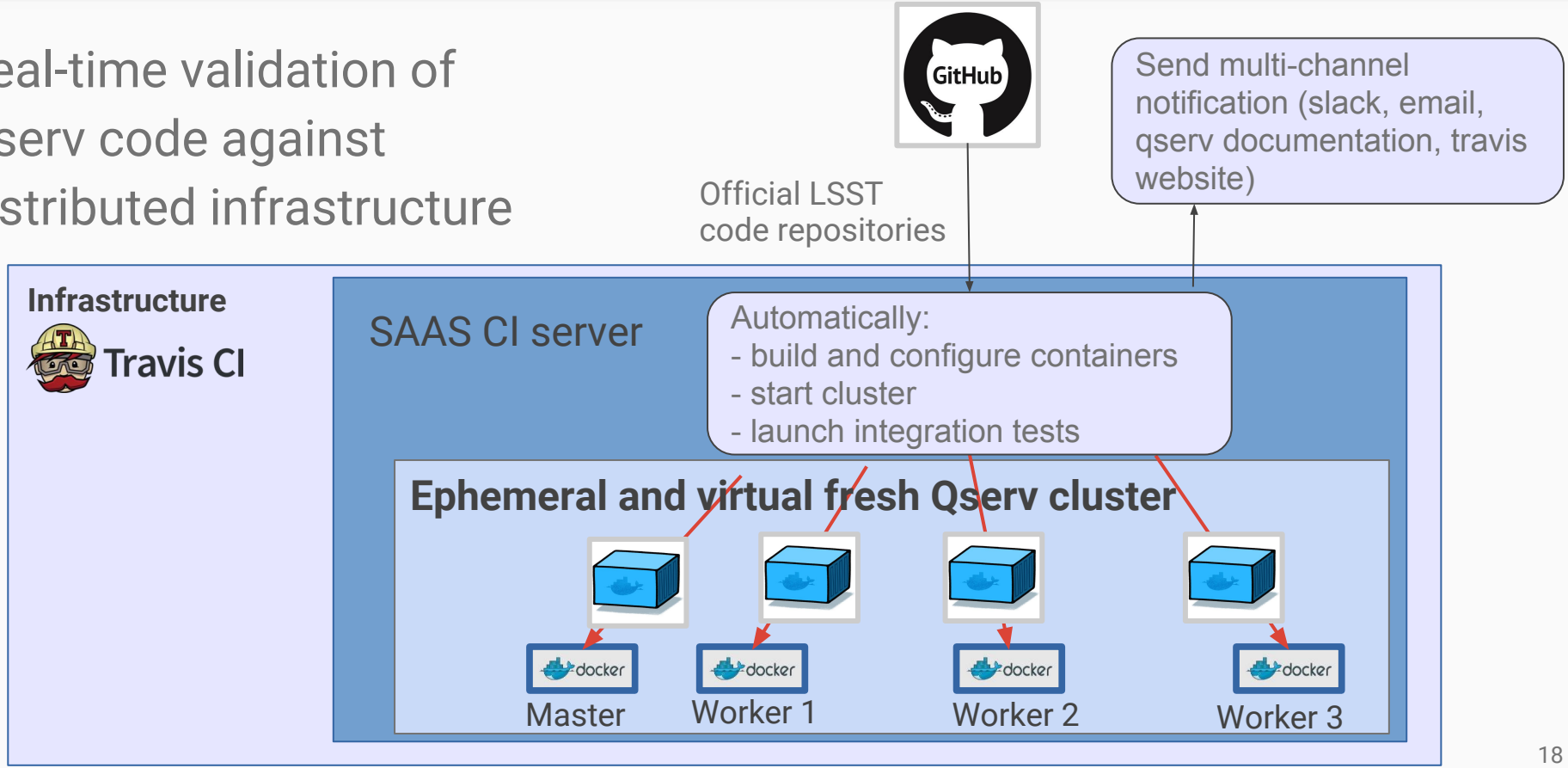
- ★ CEPH side: 24*1,5 TB virtual disks —> 35TB of data
- ★ Openstack: 25 VM using 16GB RAM and 2 vCPU each

Automated deployment: Openstack+CEPH



Automated deployment: advanced CI

Real-time validation of Qserv code against distributed infrastructure



What does bring an elastic infrastructure?

Openstack+containers

In a few seconds:

- Provision a Qserv cluster
- Package and deploy development version

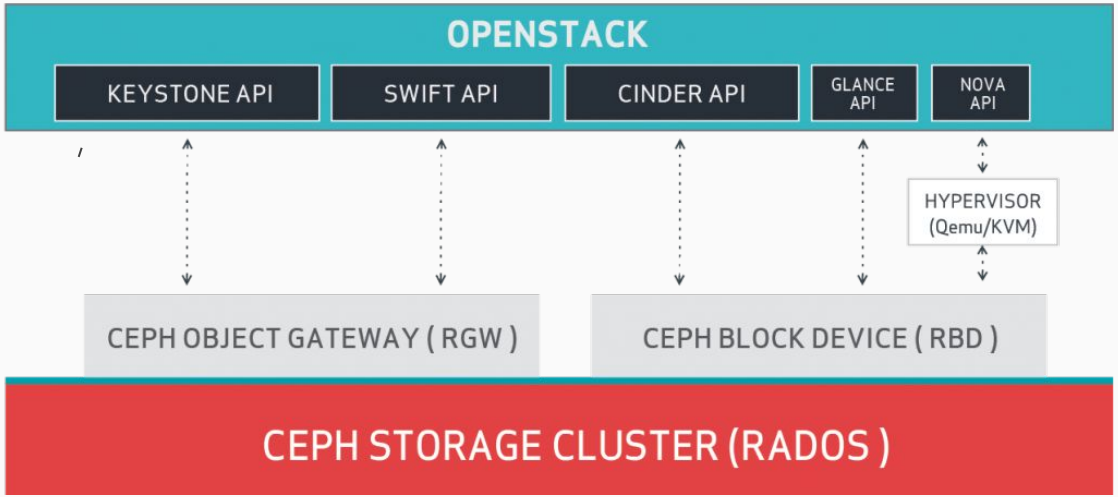
CEPH

Enable large storage access via Openstack

Provide:

- Data replication
- Data high availability
- Data reconstruction

=> Currently trying to make it scale



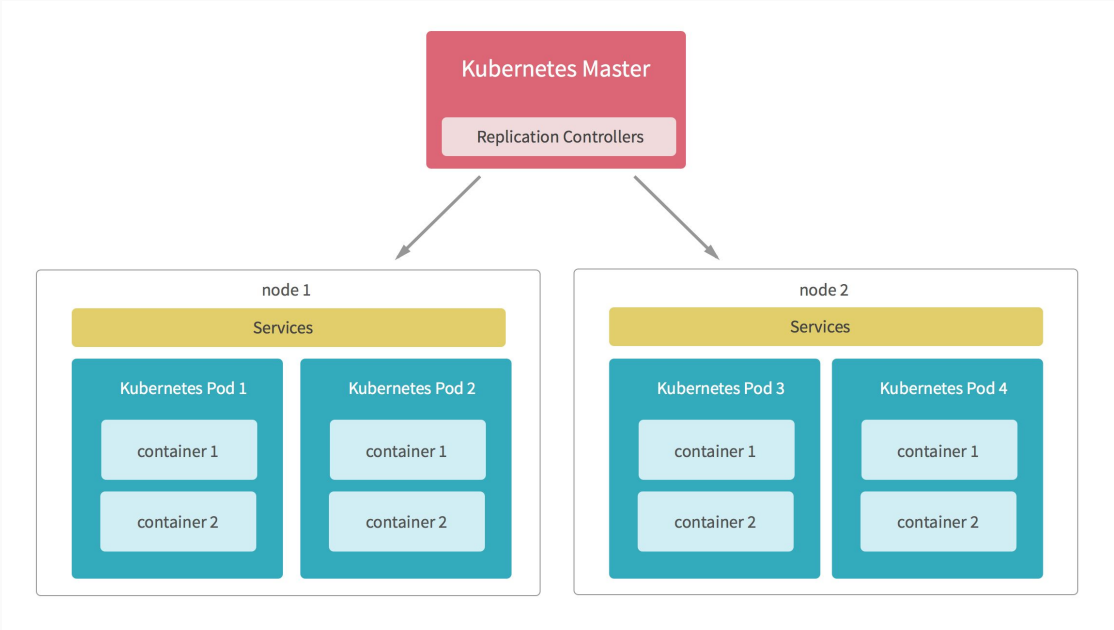
What does bring an elastic infrastructure?

-

Kubernetes

Orchestrate Qserv processes
Can scale to hundred of nodes
Provide:

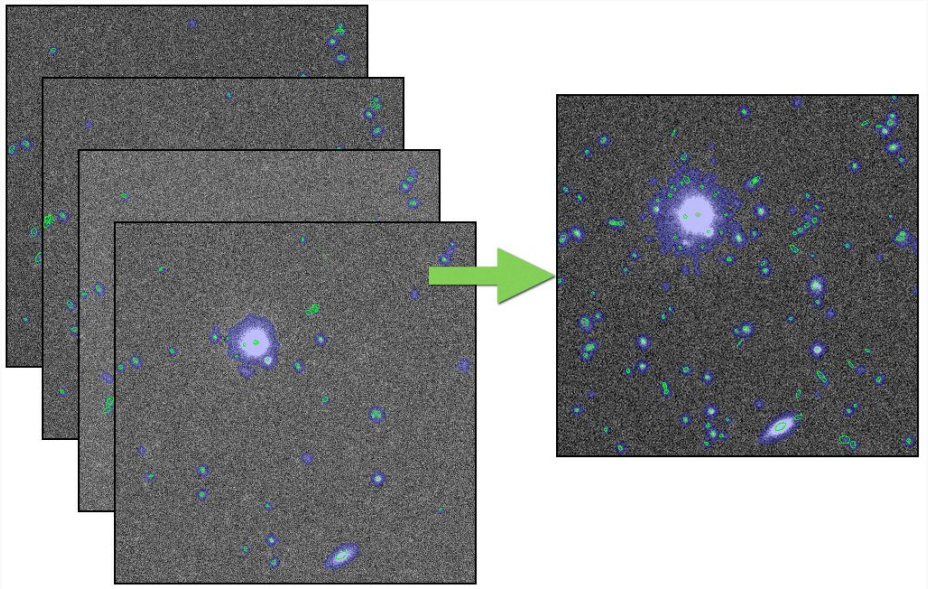
- Container placement
- Auto-scaling
- Auto-healing
- Volume management (storage)
- , Resource usage monitoring
- Health checks
- Rolling update



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/qservi

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

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:

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.

1.5 days for main conference
~1 day for Hackaton

databricks

APACHE Spark™

Flink

See you in autumn 2017

in Clermont-Ferrand

http://xldb2017.uca.fr

Credits : CRDTA

Summary

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

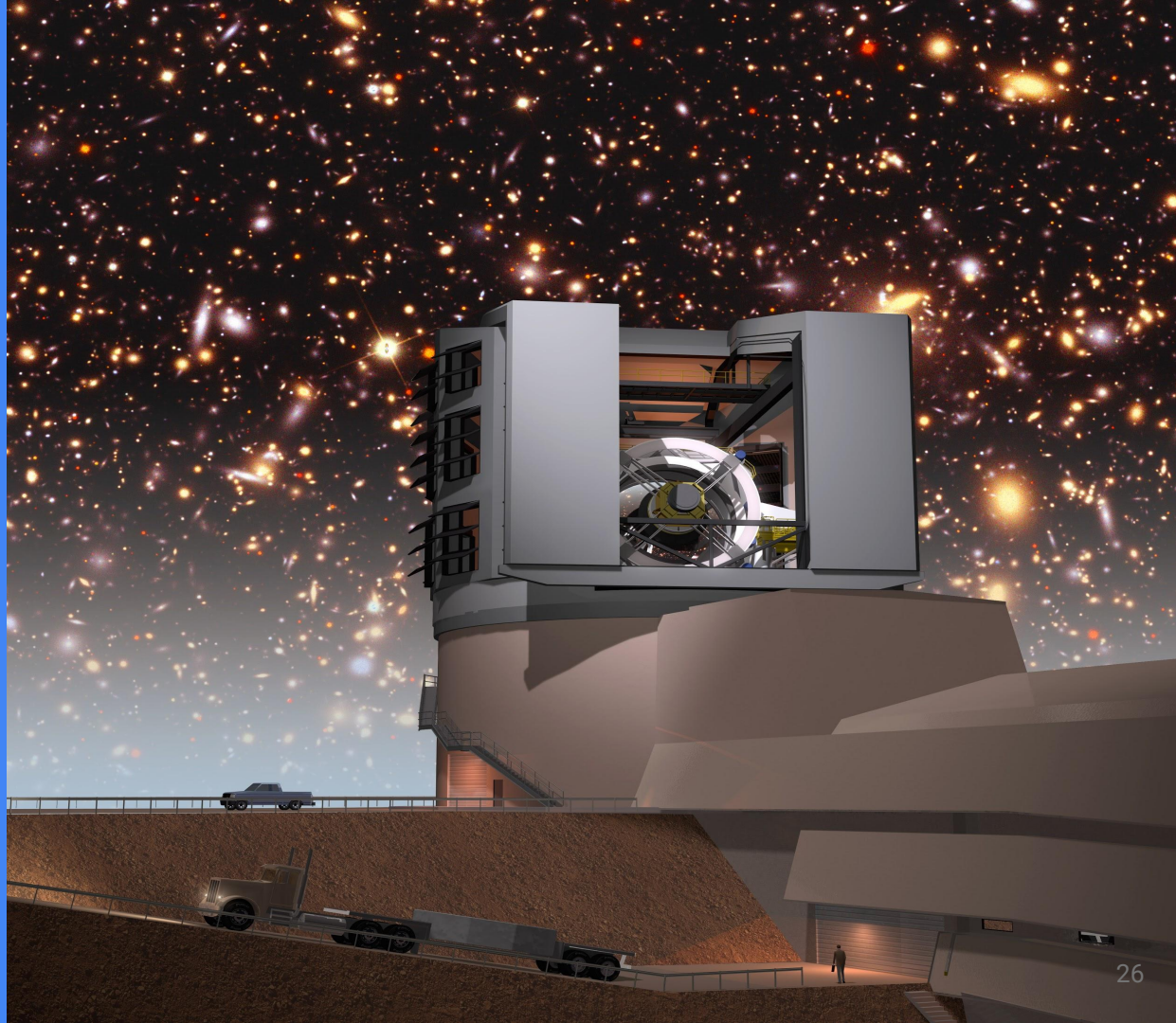
- ★ Want to learn more?
  - http://ls.st/4gh (Database Design doc)
  - http://ls.st/6ym (User Manual)
- ★ Are you an adventurous super early adopter? You can try it now
  - http://ls.st/89y (Qserv Documentation)

Thanks!

Contact:

Fabrice JAMMES
LPC
Clermont-Ferrand

fabrice.jammes@in2p3.fr



Implementation Details