



Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

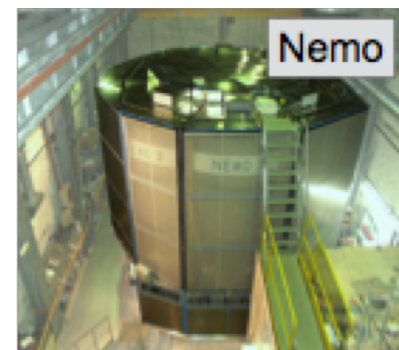
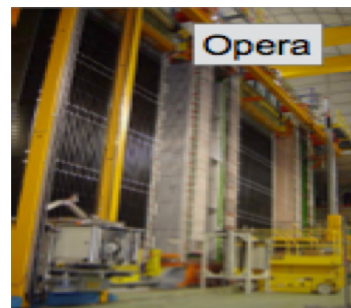
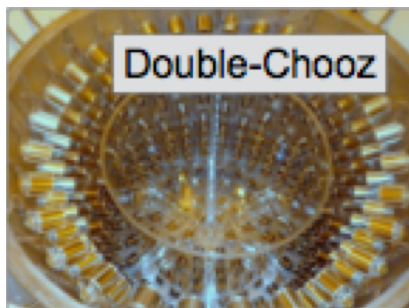
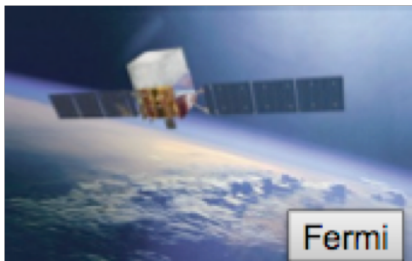
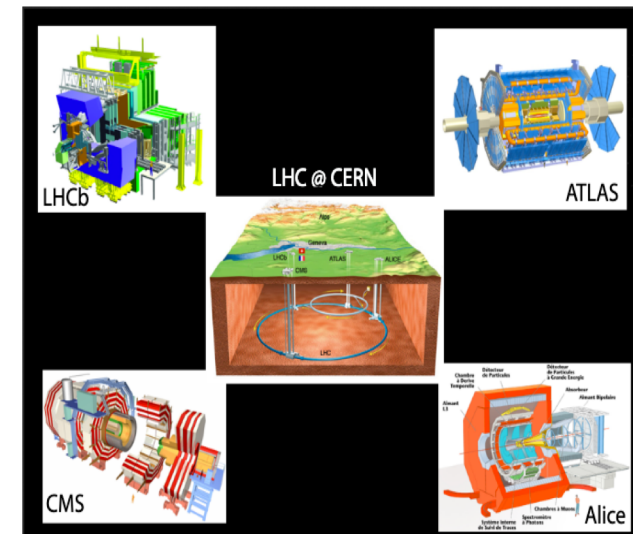
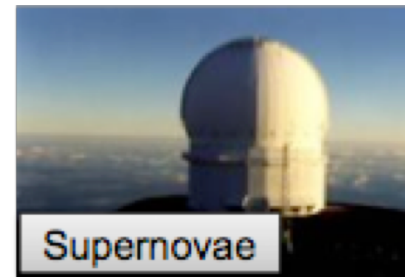
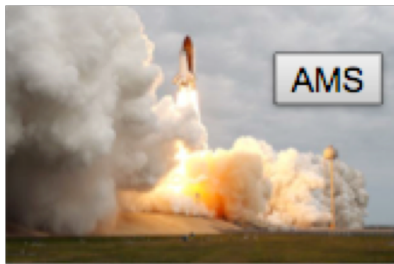
Database Services at CC

6th Feb 2019

Osman AÏDEL



Physics experiments



► RDBMS

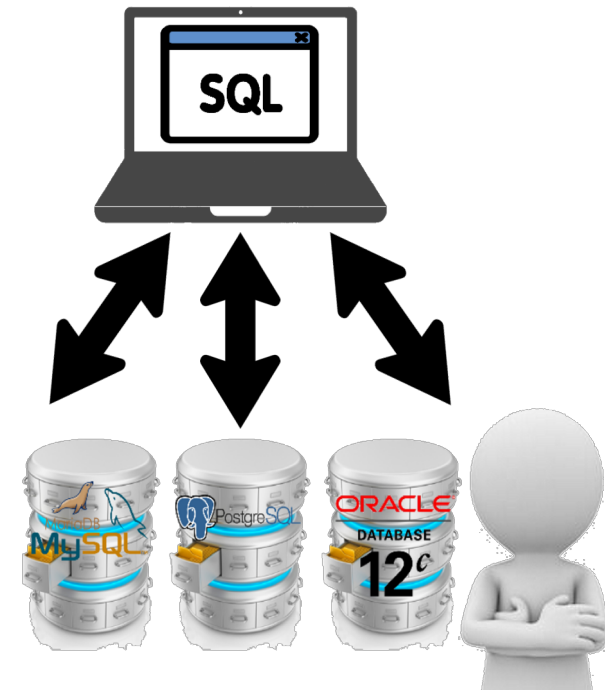
- Mysql (ver. 5.6)
- MariaDB Cluster (ver. 10.2) with Galera (ver. 25.10)
- PostgreSQL (ver. 10.6)
- Oracle (ver. 11.2 -> 12 on progress)

► Nosql

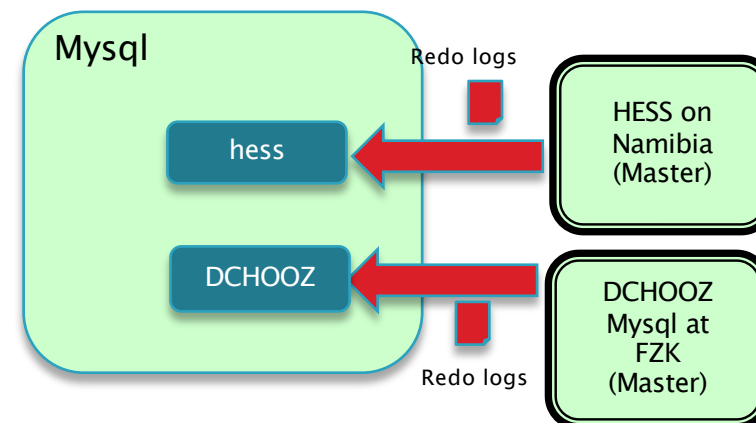
- Mongodb (ver. 3.6)
- ElasticSearch

► Monitoring

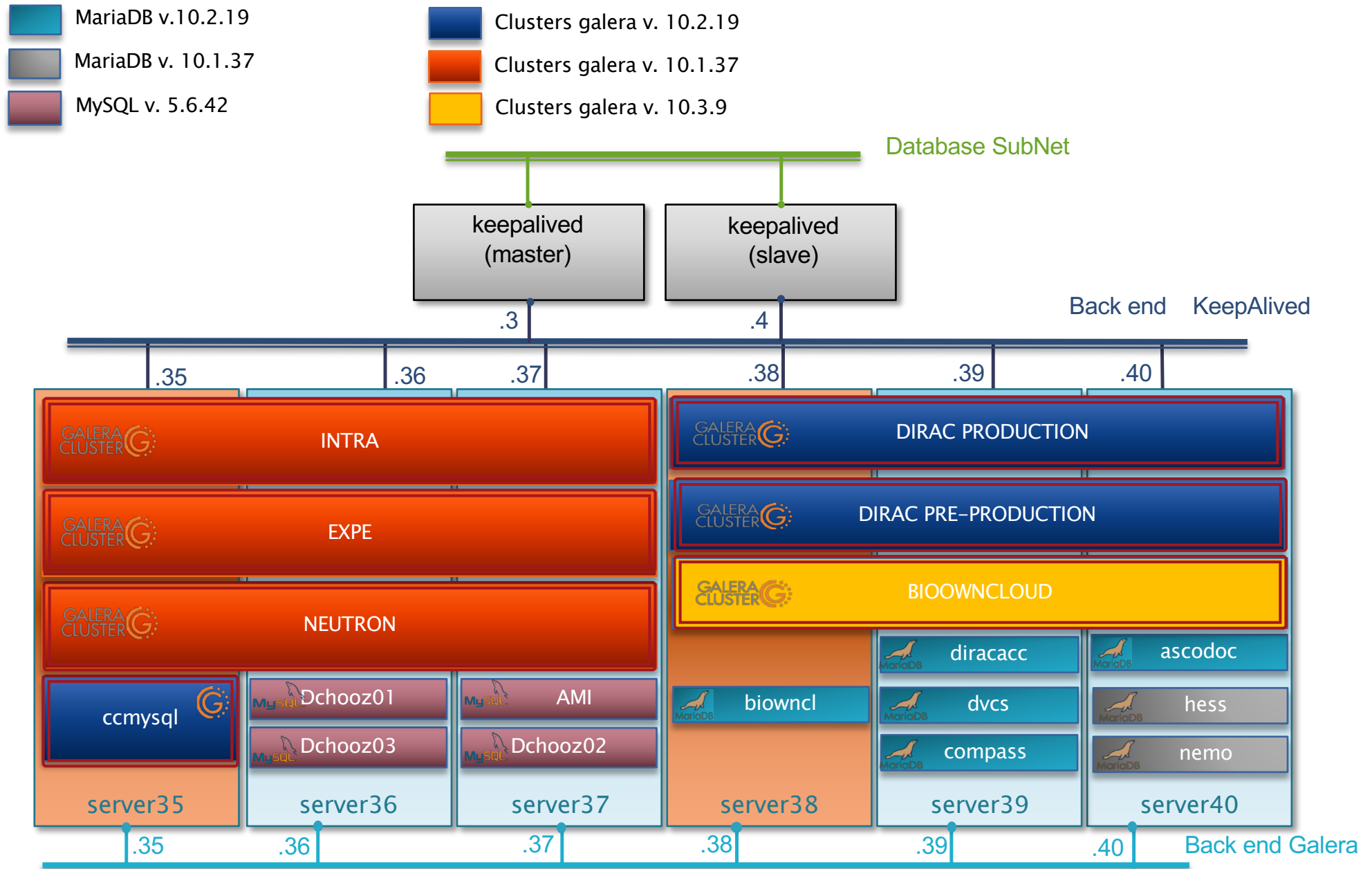
► Spark



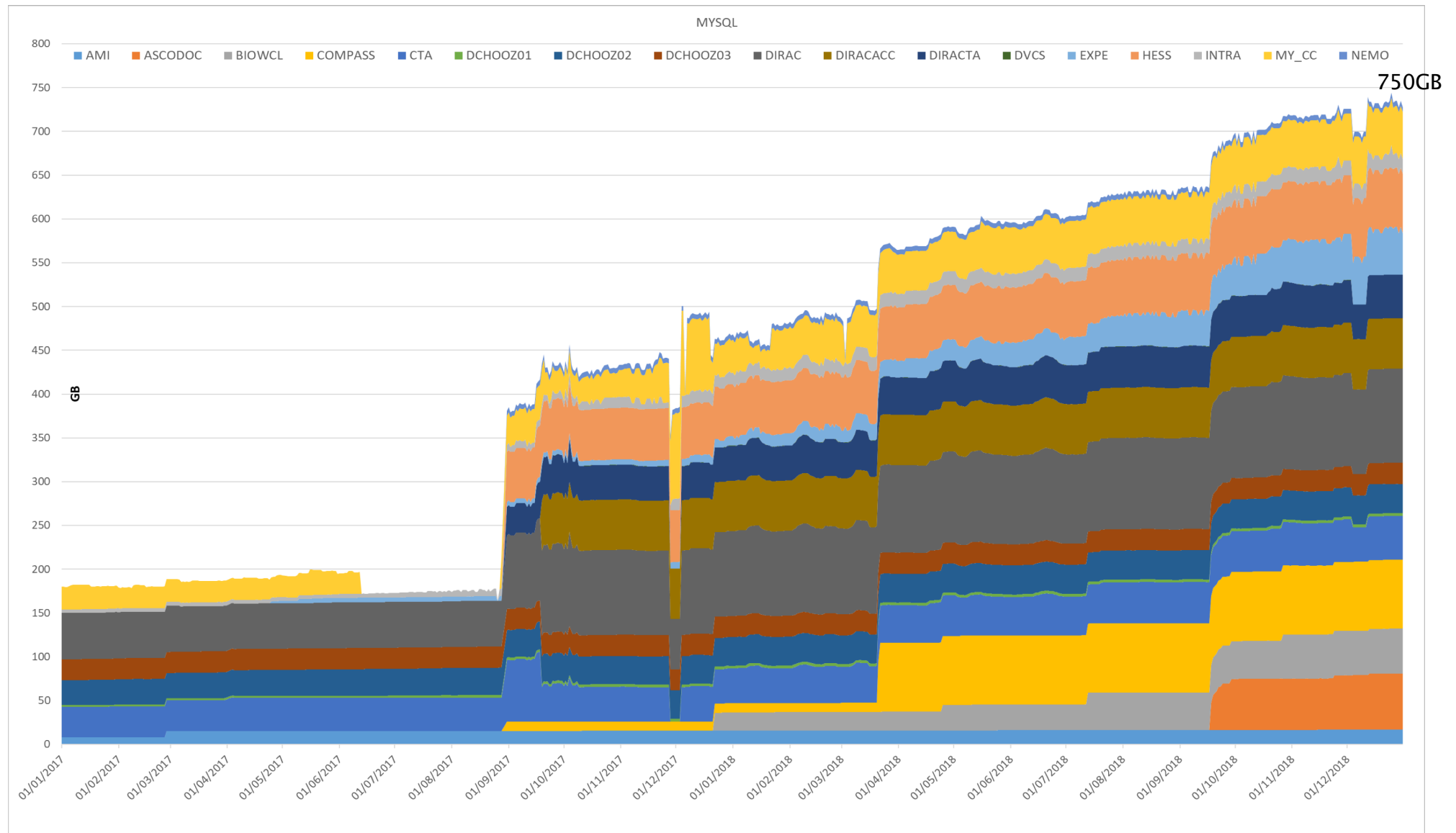
- ▶ Shared platform
 - 6 galera clusters (3 nodes)
 - 12 Mariadb / Mysql standalone instances
 - 610 accounts
- ▶ Each instance is reachable through its own VIP
- ▶ Keepalived is used for security and high availability
 - Kernel patch in rolling mode
 - Major / Minor upgrade in rolling mode
 - Node failures
- ▶ Daily backup :
 - Mysql : SQL Dump plus archived logs
 - Local copy with a retention period of 1 day
 - Tape storage : TSM with a retention period of 1 month
 - TSM : Tivoly Storage System is a backup solution on tape
- ▶ Some specific configurations :
 - One-way Replication (Hess – Dchooz)



Mysql / Mariadb infrastructure

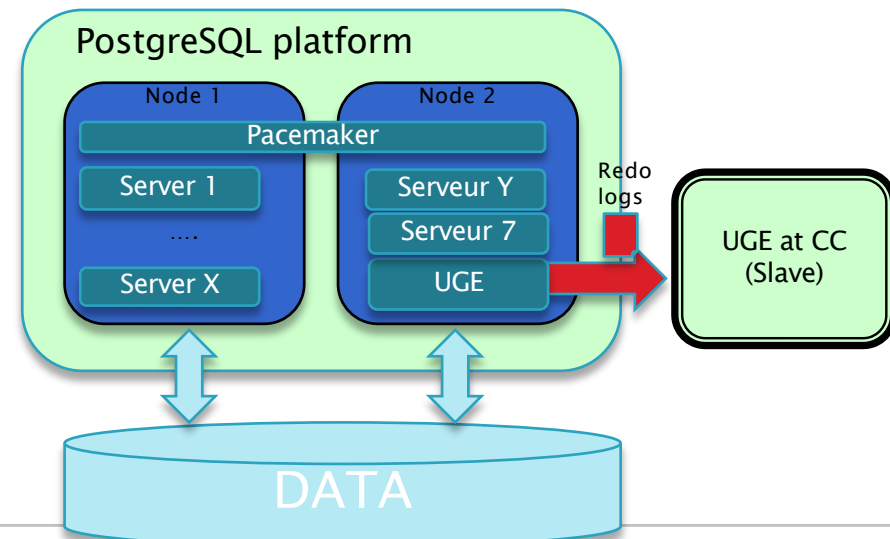


MYSQL/Mariadb



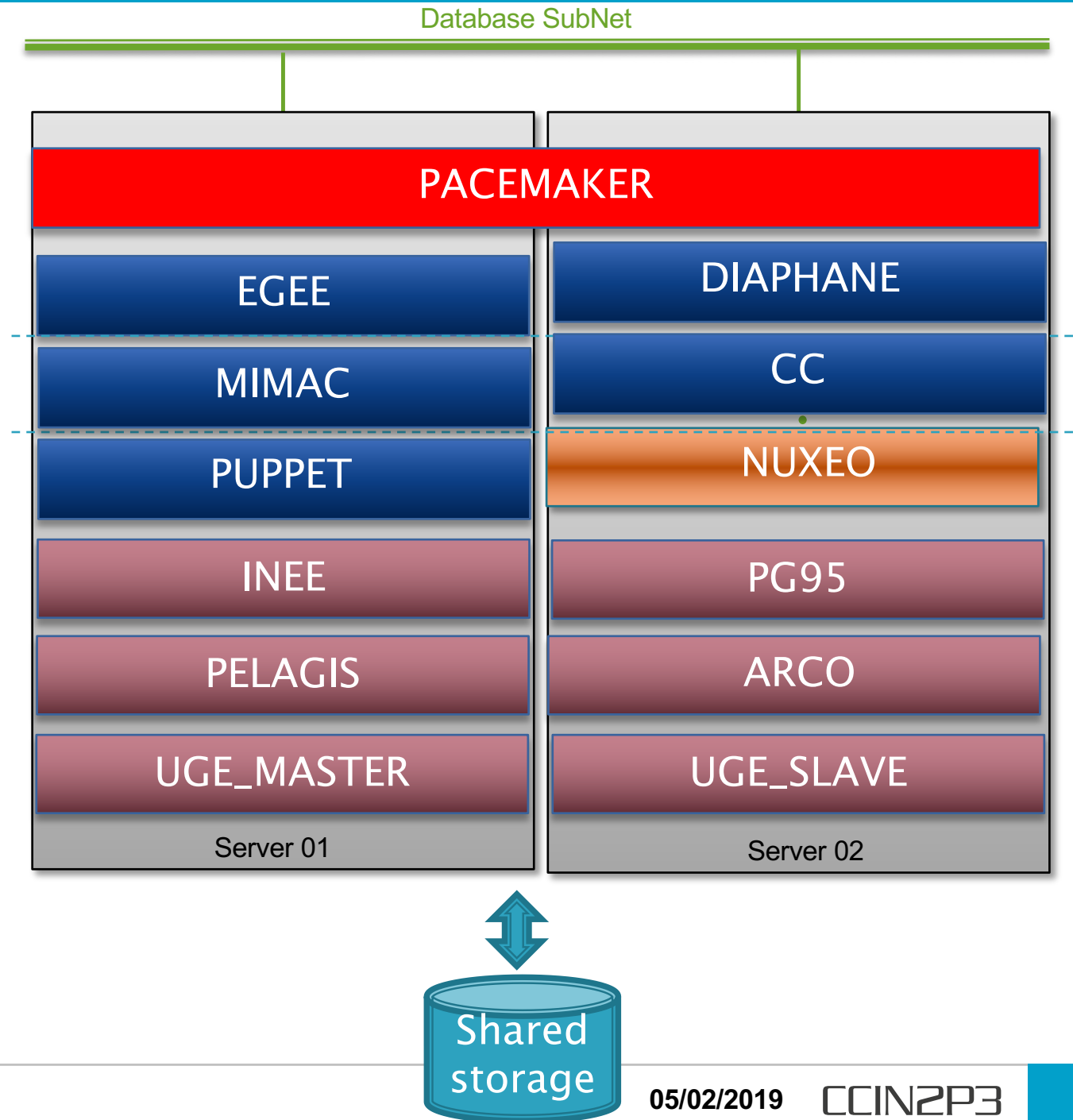
PostgreSQL architecture

- ▶ Shared platform
 - 12 PostgreSQL standalone instances
 - 76 accounts
- ▶ Each instance is reachable through its own VIP
- ▶ Resource manager (Pacemaker)
 - Kernel patch in rolling mode
 - Minor upgrade in rolling mode
 - Node failures
- ▶ Daily Backup :
 - SQL dump / binary copy
 - Local copy with a retention period of 1 day
 - Tape storage : TSM with a retention period of 1 month
- ▶ Some specific configurations :
 - UGE (batch system) slave on read only

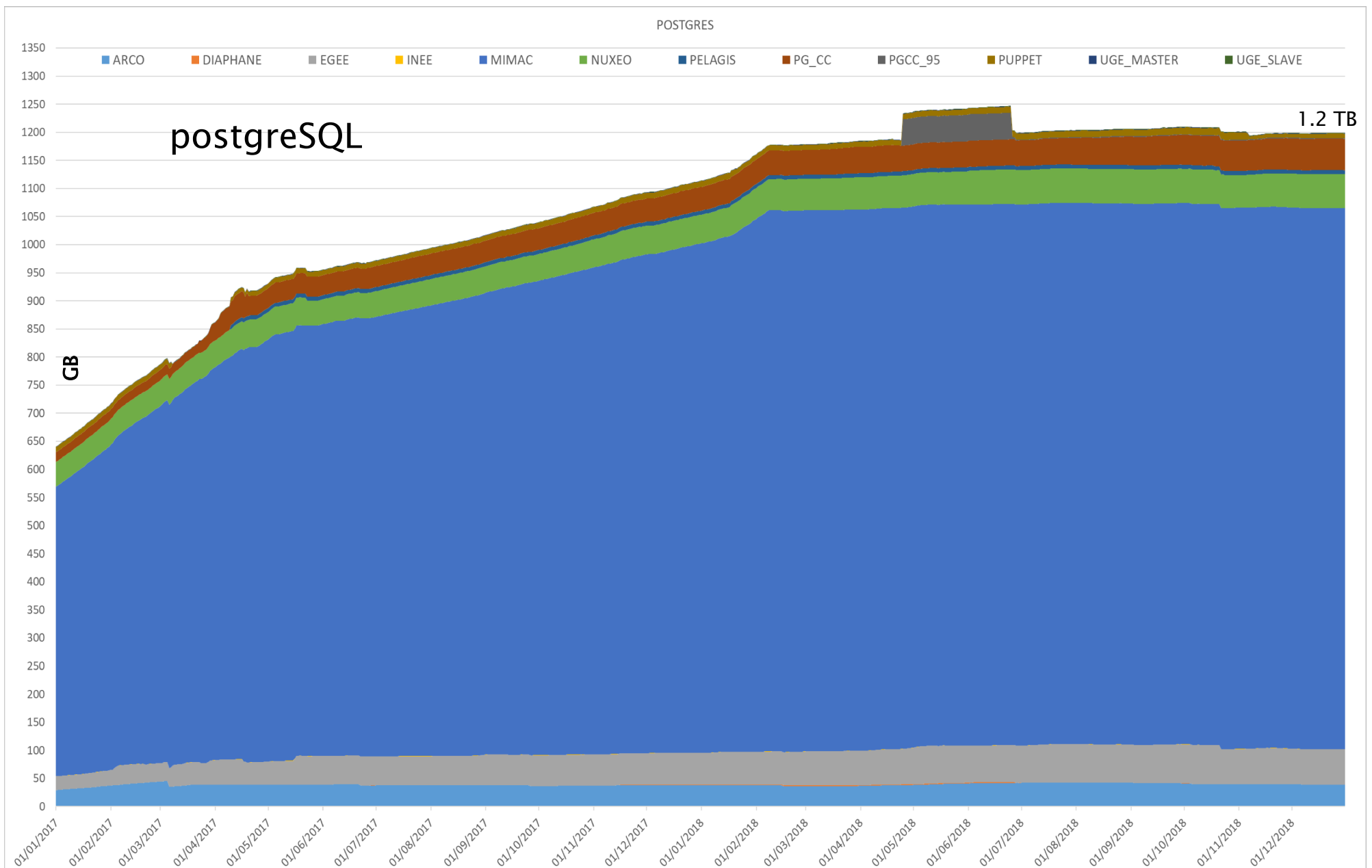


PostgreSQL architecture

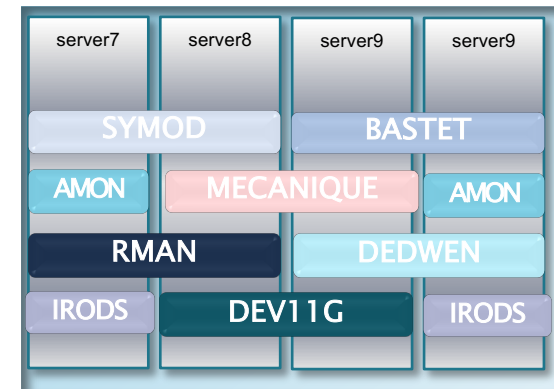
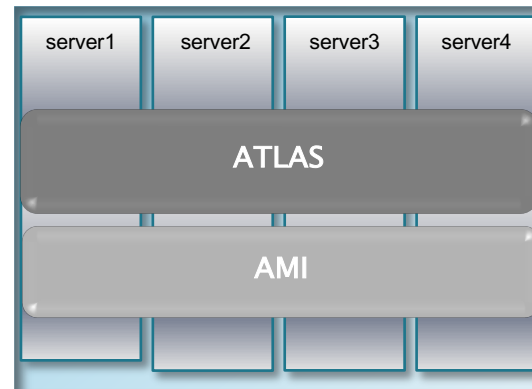
- Postgres v. 10.6
- Postgres v. 9.6.11
- Postgres v. 9.5.7



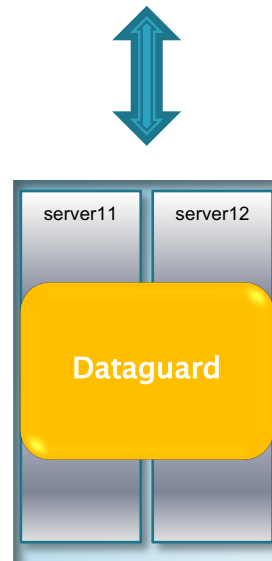
Database volumetry



Oracle infrastructure

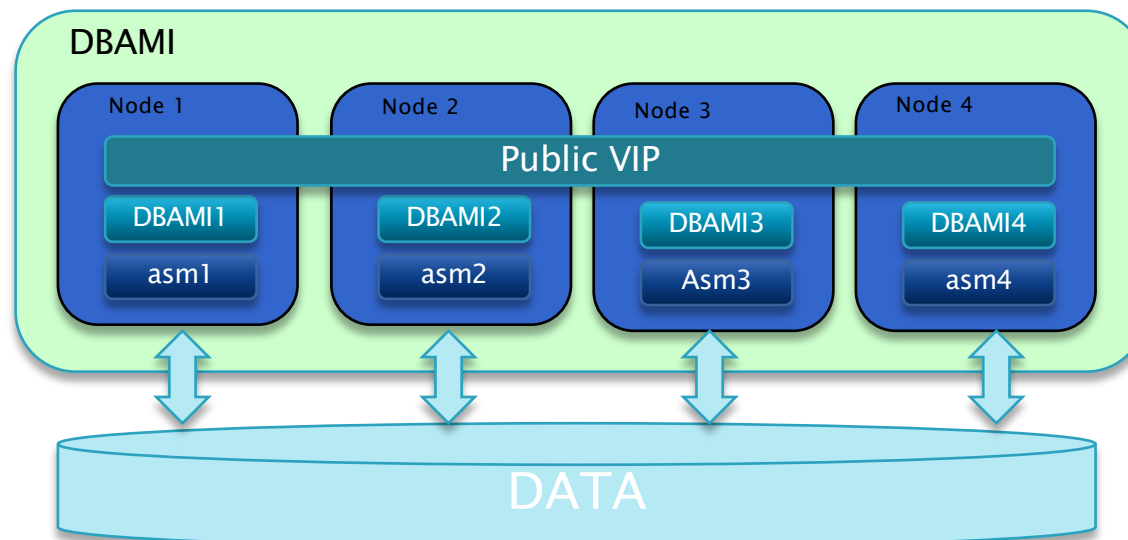


494 accounts



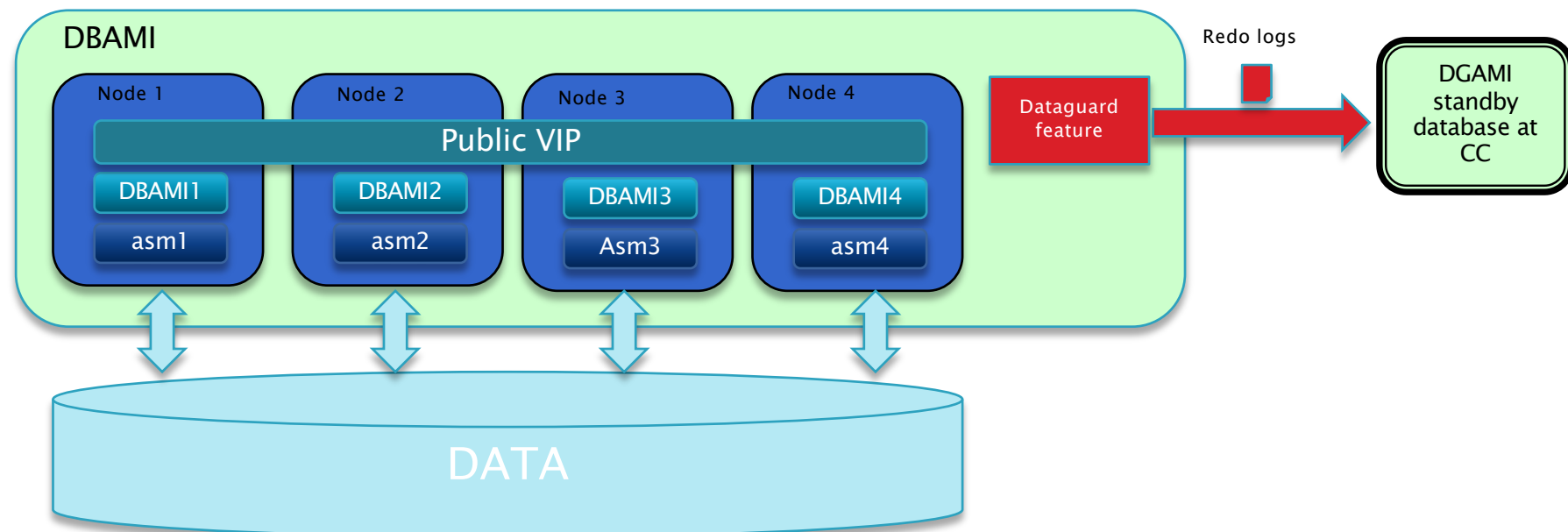
► Real Application Cluster database

- N actives nodes simultenaously
- Clients only access to one public VIP
- Node failure does not impact the availability of the database
- Data is shared via Automatic Storage Management (ASM)
- ASM makes transparent the physical devices
 - Renewing of storage does not require downtime
 - Data may be replicated at logical level



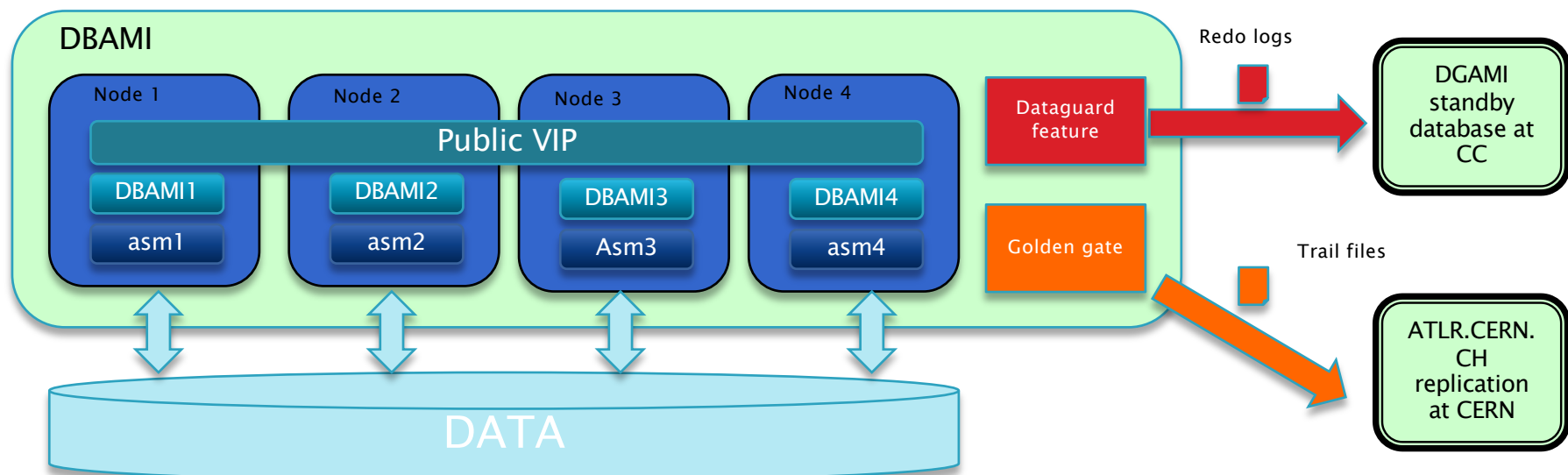
► Dataguard

- Oracle's standby database solution for disaster recovery
- Exact copy of the database production
- All clients can point to the standby db without any change at the application level
- Switchover very fast (some minutes)
- Major upgrade only requires some minutes of unavailability
- Severe crash on production might involve a small lost (few transactions)



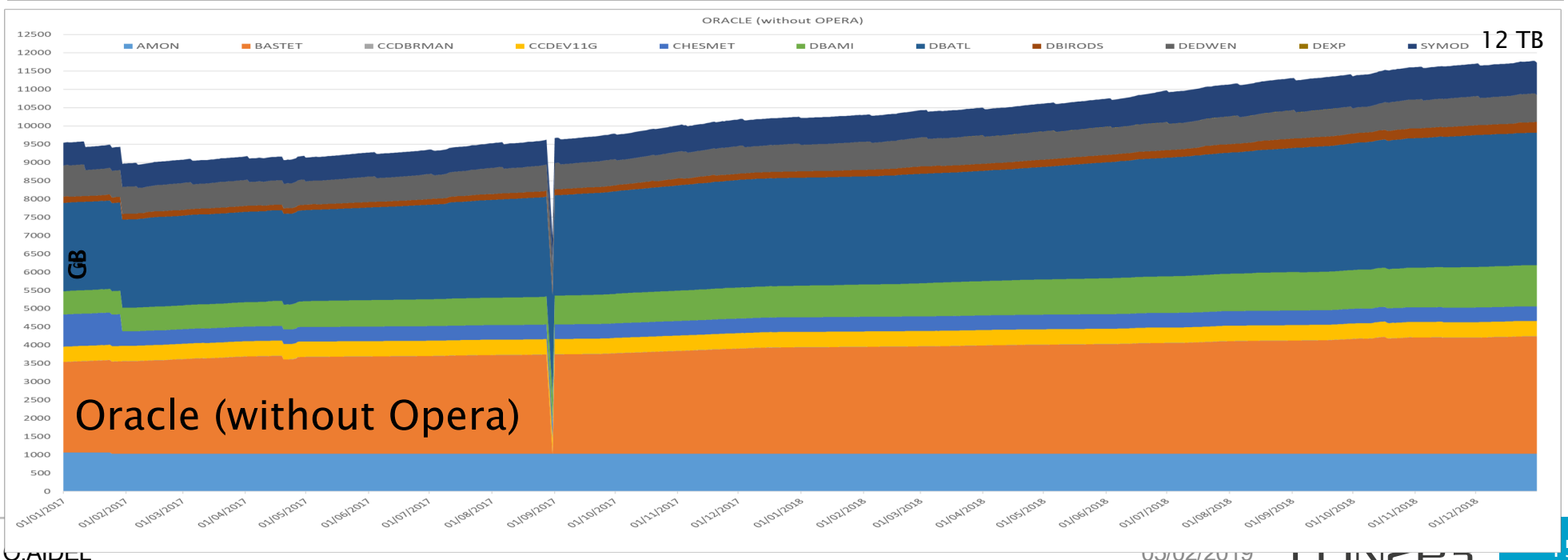
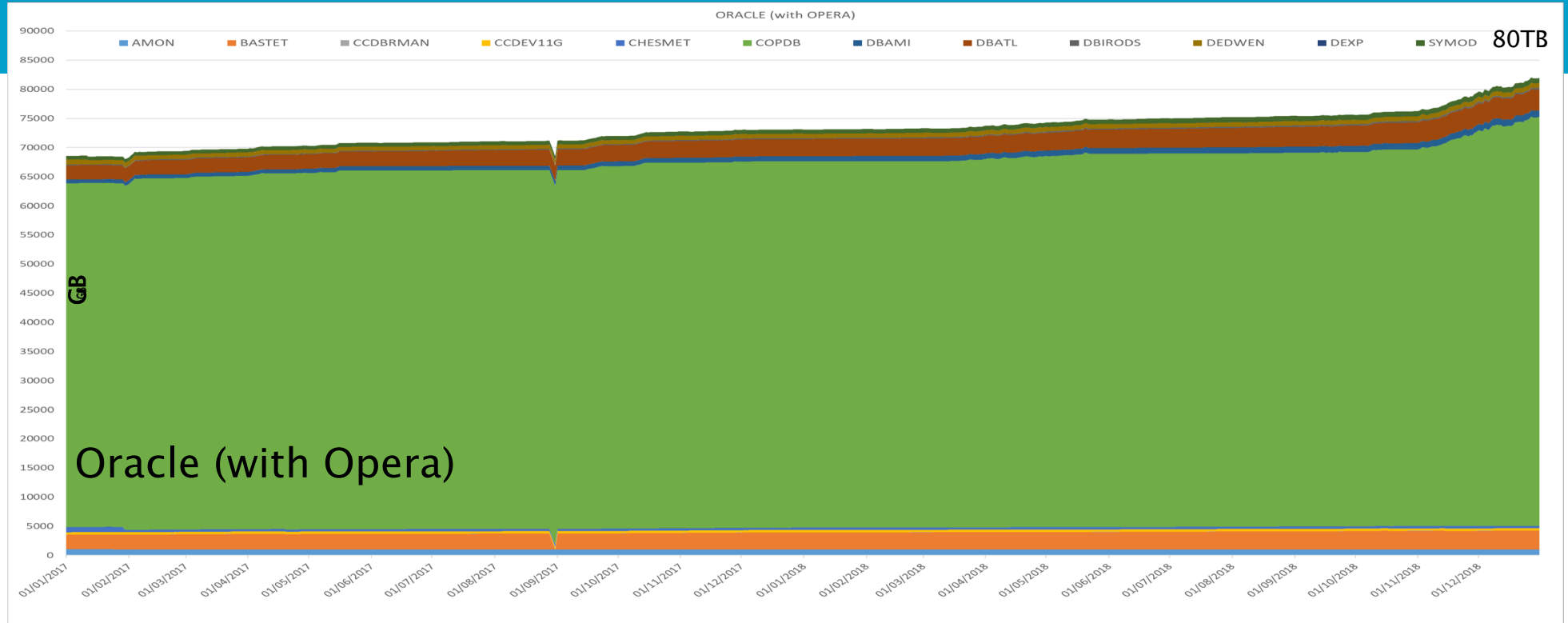
► Golden Gate (streams)

- Oracle replication solution for heterogeneous databases
- Used for the Atlas metadata catalog (AMI)
- Replication at user account / table level
- DBAMI : one-way replication from CC-IN2P3 to CERN



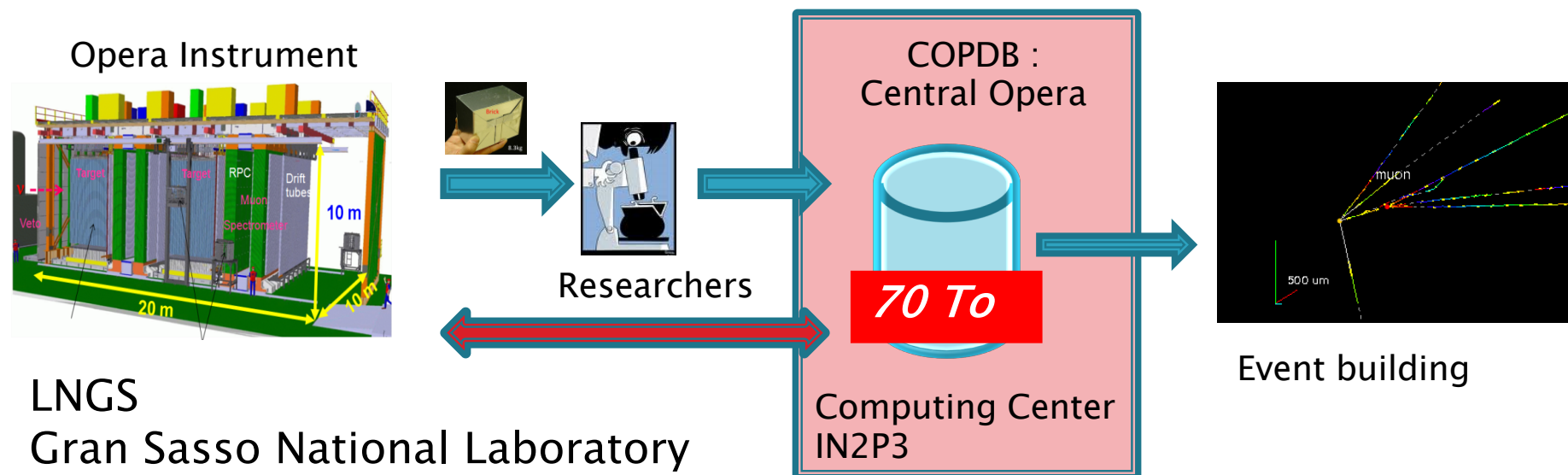
► Backup

- Recovery MANager
- RMAN Catalog
- Weekly full backup and incremental backup every day to TSM
- Retention depends on databases

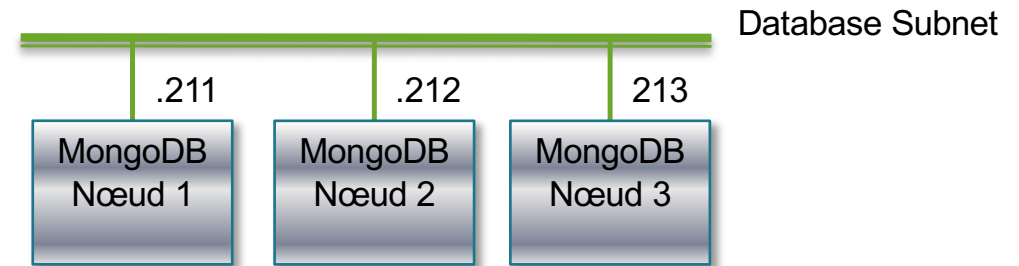


▶ OPERA

- Oscillation Project with Emulsion tRacking Apparatus
- 34 institutes spread out 12 countries
- Decommissioning at CC on progress

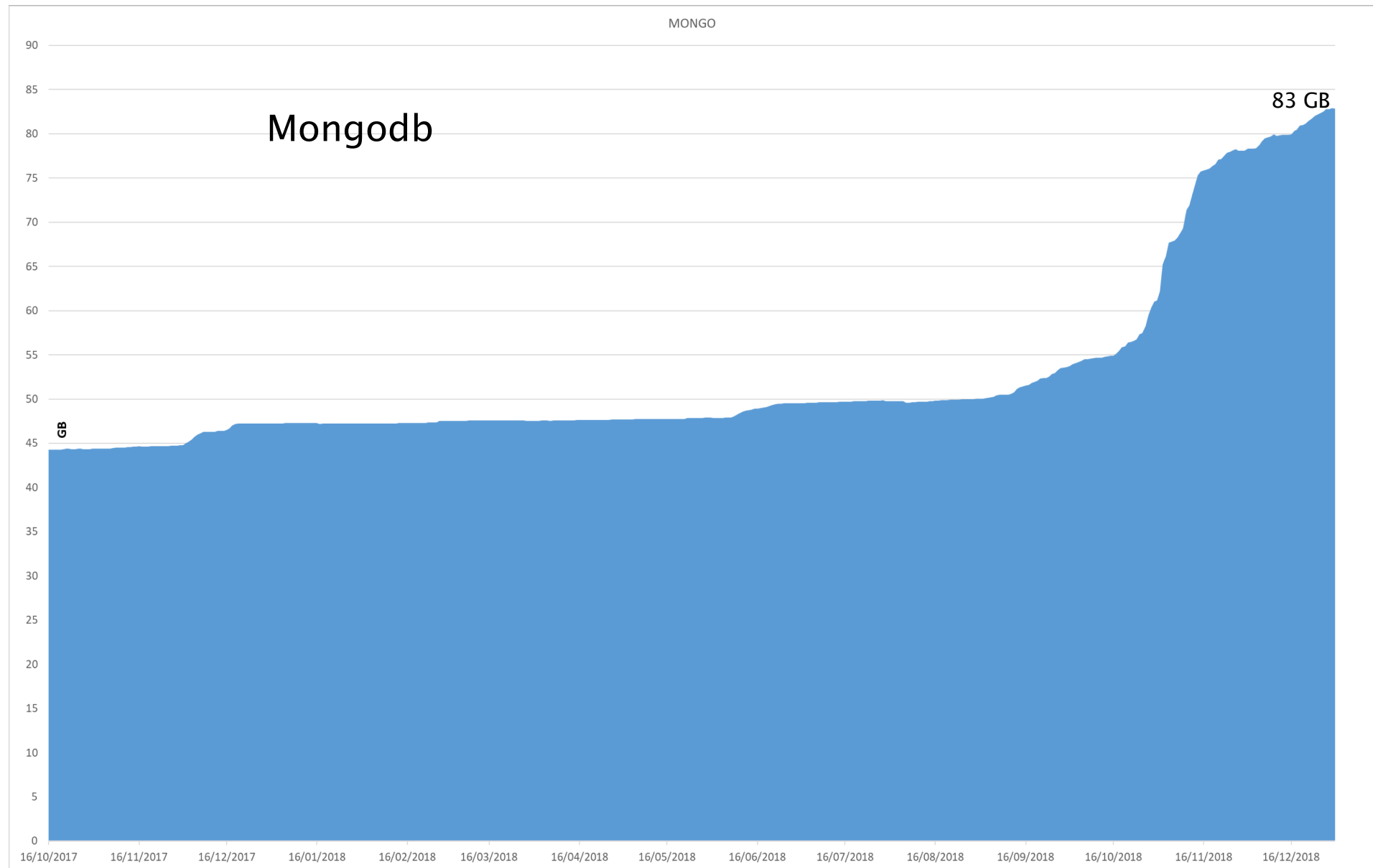


- ▶ 3 Nodes in replication
 - Kernel patch in rolling mode
 - Minor upgrade in rolling mode
 - Node failures

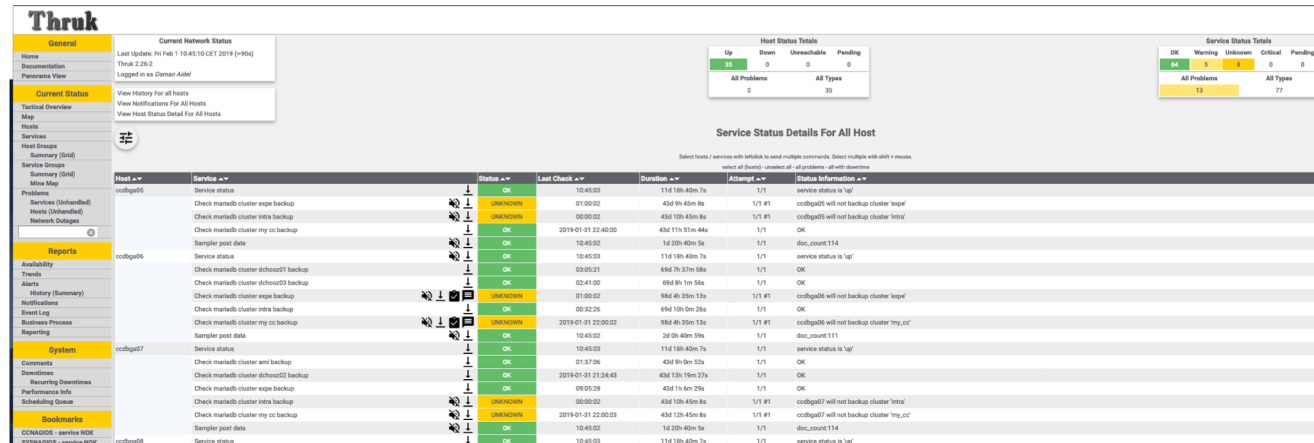


- ▶ 3 nodes in replication
 - One active node and 2 passives
 - MongoDB driver manages connections in case of failure
 - Kernel patch in rolling mode
 - Major and Minor upgrade in rolling mode
 - Node failures
- ▶ Usage (12 accounts)
 - INEE (vigiechiro)
 - Openstack (ceilometer)
 - UGE (Batch system)
 - ...
- ▶ Daily backup to TSM

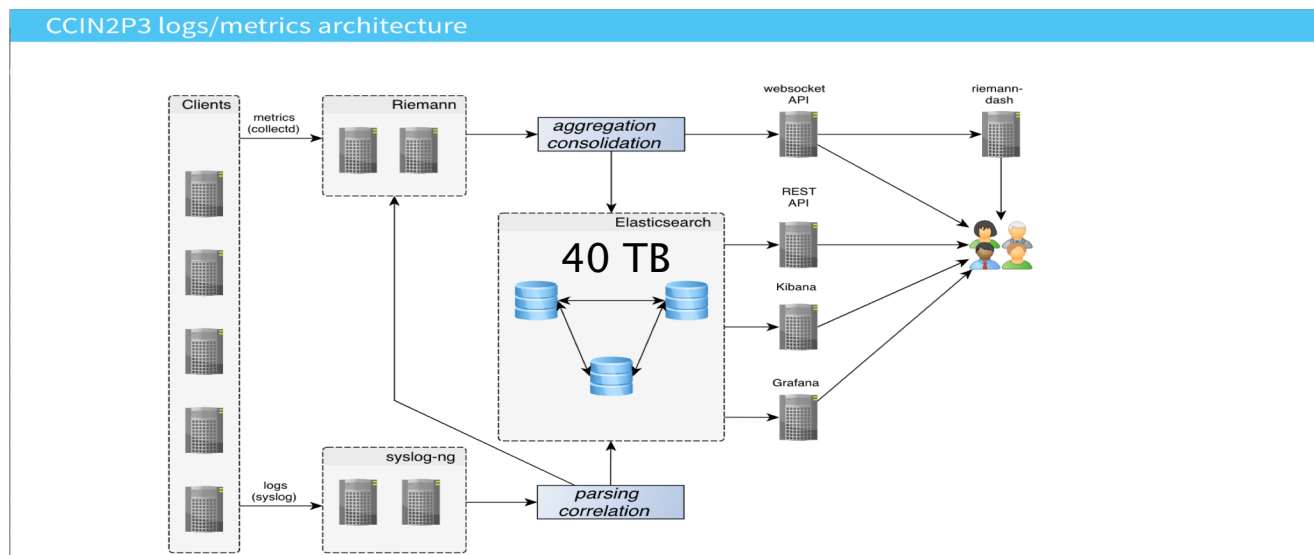
Database volumetry



► Nagios

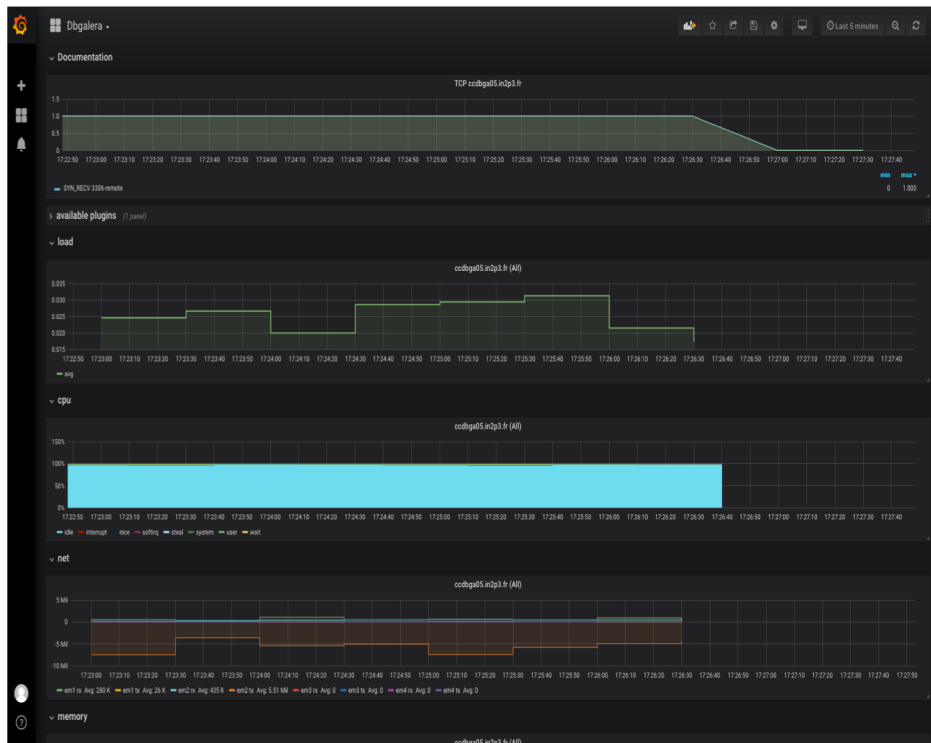


► Coloss

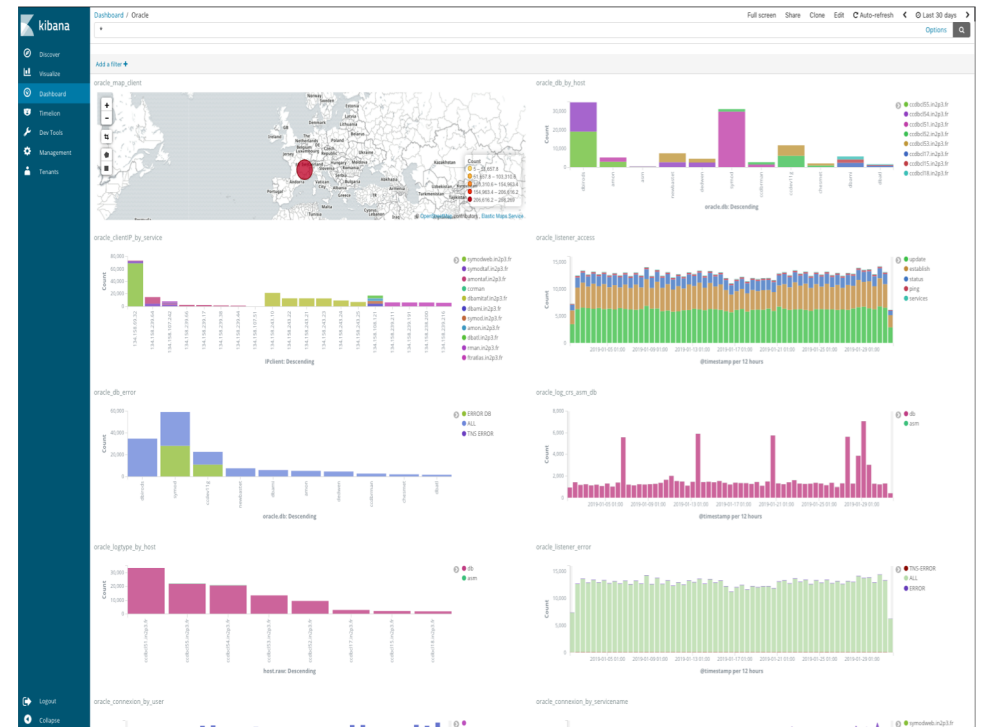


Monitoring

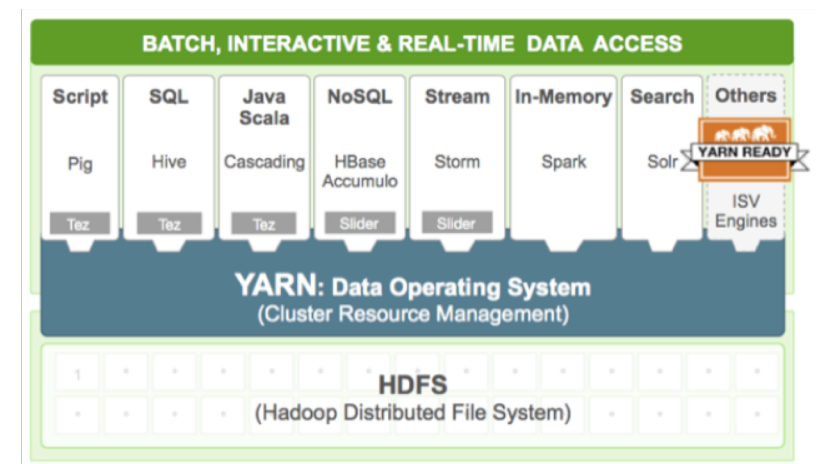
Grafana



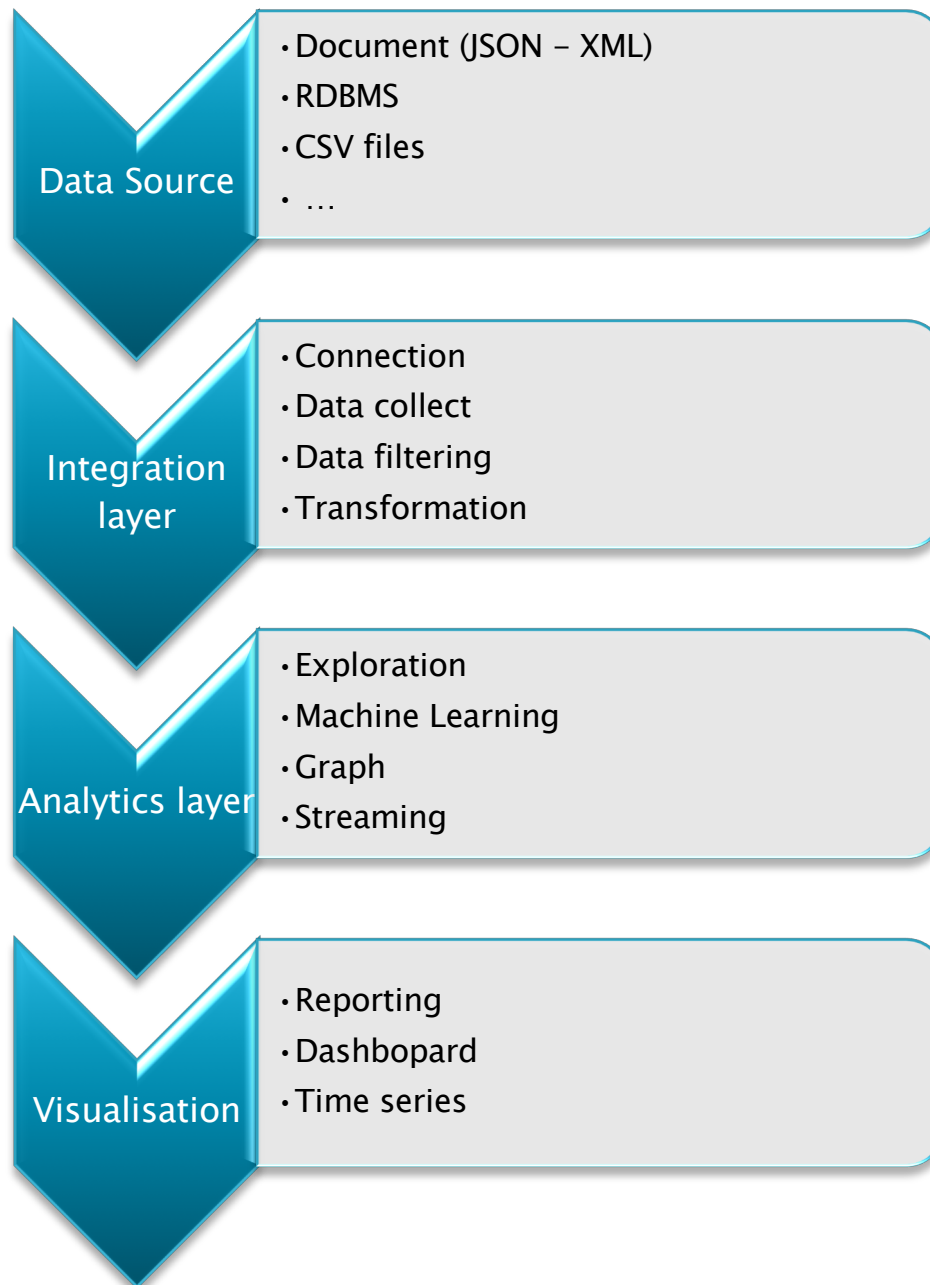
Kibana interface



- ▶ The rise of new framework for BIG DATA / DATA SCIENCE.
- ▶ Spark
 - Large scale data processing
 - High availability
 - Best suited for batch applications that apply the same operation to all elements of a dataset
- ▶ Spark features
 - Mlib (Machine Learning lib)
 - Streaming
 - Spark SQL
 - GraphX

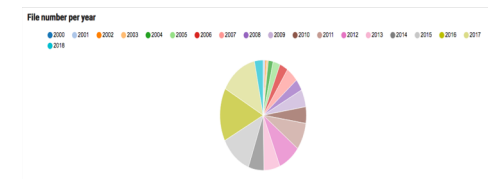
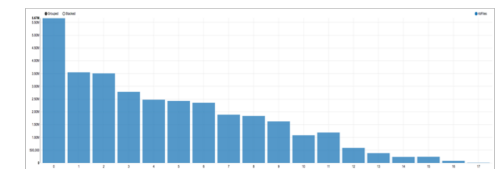
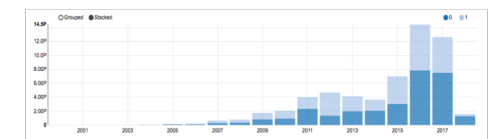
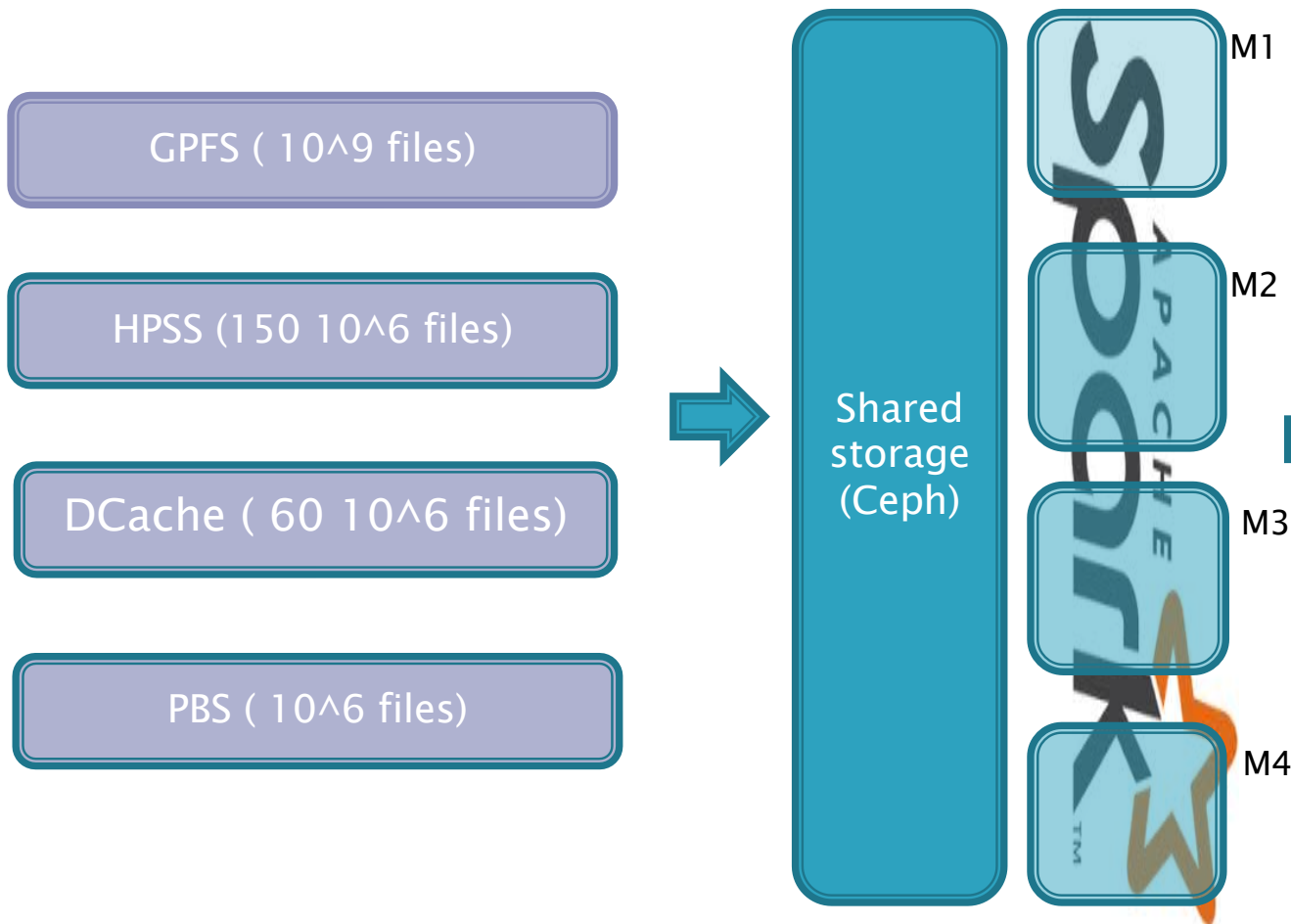


SPARK : data processing



► Storage metrics

Storage system



- ▶ The main idea is to provide an unified view on storage systems for users, experiment manager , CC support, storage experts and direction.
- ▶ User view
 - How storage is used at the experiment or user level ?
 - Number of files, used space, file size
 - How files are distributed through CC storage systems
 - ...
- ▶ Experiment Manager
 - How much space is allocated / used by the experiment ?
 - Which user consumes too much storage ?
 - ...
- ▶ CC Support view
 - Users leaving an experiment do not cleanup their space. The support in collaboration with the experiment manager need to identify user files and what to do with it. Which files is involved ? Are they relevant ?
 - Optimize the storage use by locating files no more accessed for one month / year in order to spill them down to a cheap storage (ex. tape).
- ▶ Expert view
 - For some storage system, it may be interesting to know the number of links or empty files because the number of inodes is not unlimited.
 - Small and Large files have an impact on system performance and in resource consumption. Indeed, a system configured with large block are not adapted for small files, a small file could consume more space it requires. It's why ,it is important for an expert to know the distribution of file size n order to adapt the best block size in order to minimize the wasted space.
- ▶ Direction view
 - How much space is allocated to each experiment ?
 - How files are spread through CC storage systems ?
 - Do we respect experiment agreements

► Storage metrics

Storage system

GPFS (10^9 files)

HPSS (150 10⁶ files)

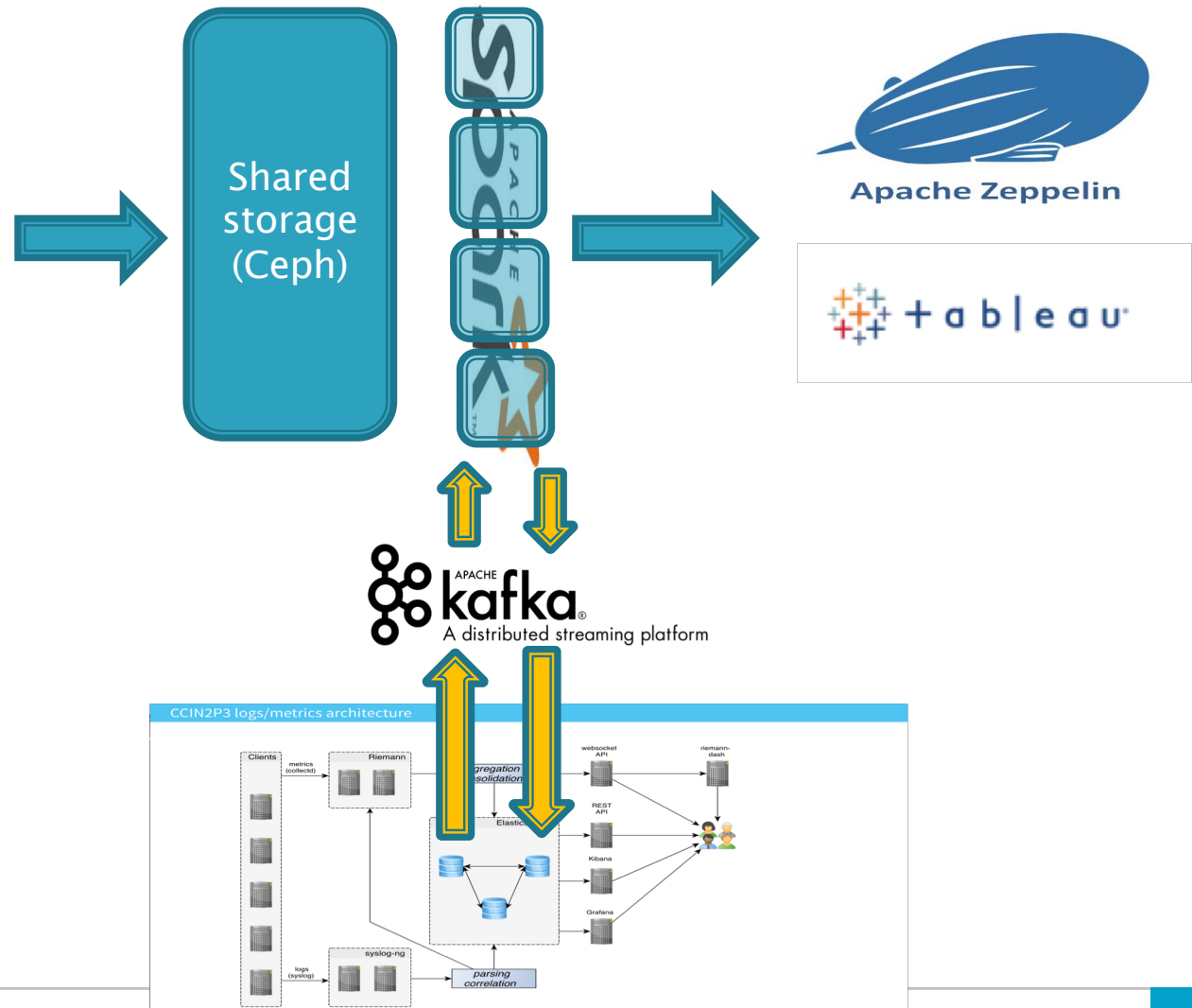
DCache (60 10⁶ files)

PBS (10^6 files)

Batch accounting

Job logs

■ ■ ■



QUESTIONS?

