Calcul LSST
Architecture de l'infrastructure

R. Ansari - Lyon (CC-IN2P3)

6 Février 2013

1an : ~ 10^8 images-CCD → ~ 10^9 files
10 an : ~ 10^9 images-CCD → ~ 10^10 files

LSST data set

- Images : 300 MBytes/s , 15 MBytes/s/raft , 1.5 MBytes/s/ccd
  - 15 TB/night , 1500 exp/night
  - 5000 TB / year ➟ 100 × 50 TB; 50 000 TB / full survey ➟ 1000 × 50 TB
  - Need high level of parallelism in storage / processing
- Catalogs : 100-200 TB for object catalogs, 1000-3000 TB for light curves
- Ancillary / calibration data

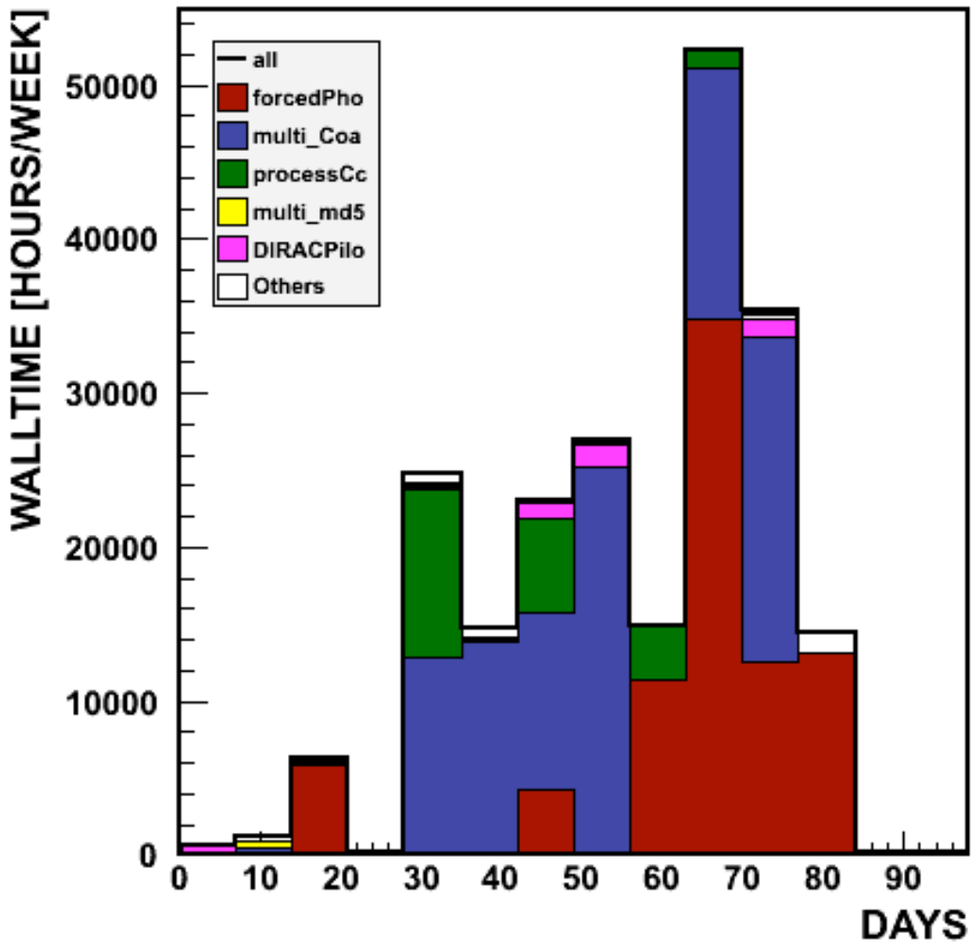
R.Ansari - Déc 2011

DC Summer 2013 (CC-IN2P3)

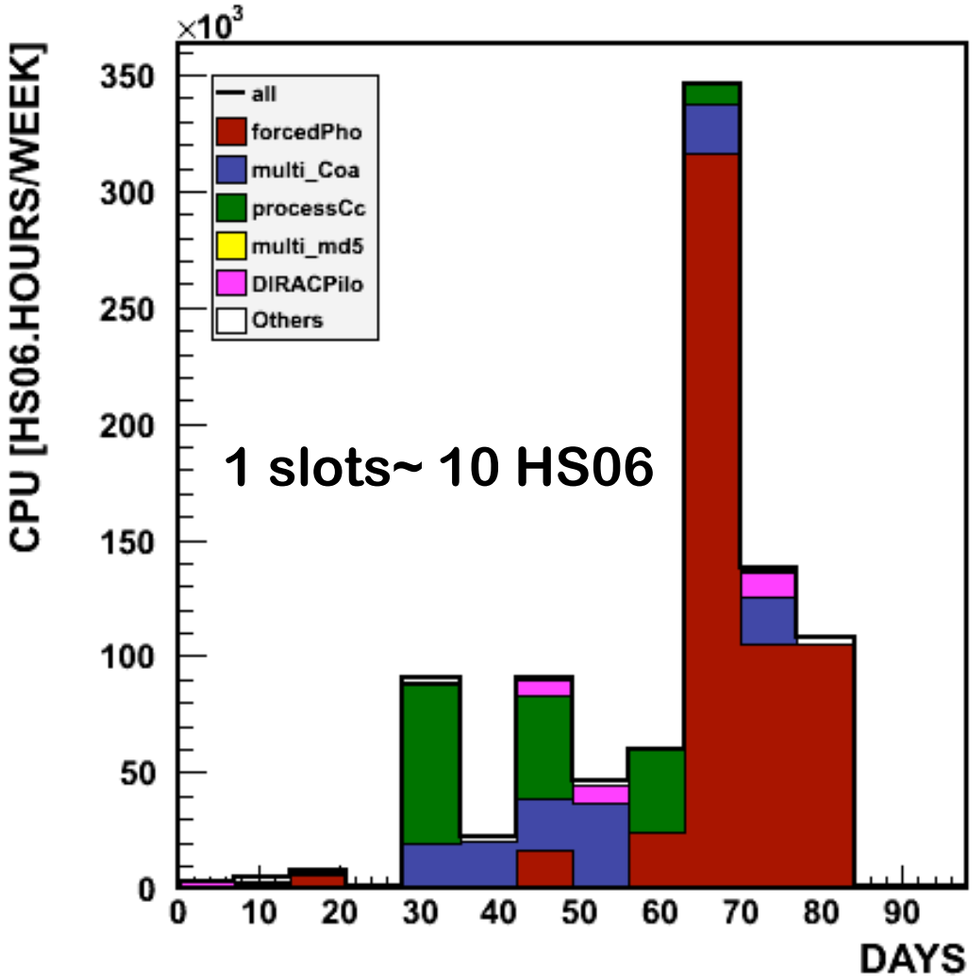
- Stripe 82 : ~1.5 10^6 CCD-images, ~ 7 TB “raw data”
- Comparable to one night of LSST : 4. 10^5 CCD-images, 15 TB (raw data)
- DC2013: 7 TB input , ~100 TB output/processed data , few 10^7 files
- ~40 000 jobs and ~100 000 h CPU (10^6 HS06) over ~ 2 months , ~100 jobs in parallel (100 cores)
- I/O rate : few x 100 TB over ~ 1000 hours → ~30 MB/s
- We need to gain more than a factor 100 in efficiency to perform the first year LSST DRP !

CPU per week

Walltime (Hours)



CPU (HS06.Hours)



➡ co-add and forced-photometry

➡ forced-photometry

Slide by R. Lemrani

Current LSST / DC2013 computing model

Ressource allocation and management (GridEngine)

Distributed File System server (GPFS)

FileServer

Network/Ethernet

Computing nodes

Current LSST / DC2013 computing model

- DC 2013 : 30 MB/s , 100 cores in ǁ
- LSST 1 year : 5 GB/s , 10 000 cores in ǁ
- LSST 10 years : 50 GB/s , 100 000 cores in ǁ

Ressource allocation and management (GridEngine)

Distributed File System server (GPFS)

FileServer

Network/Ethernet

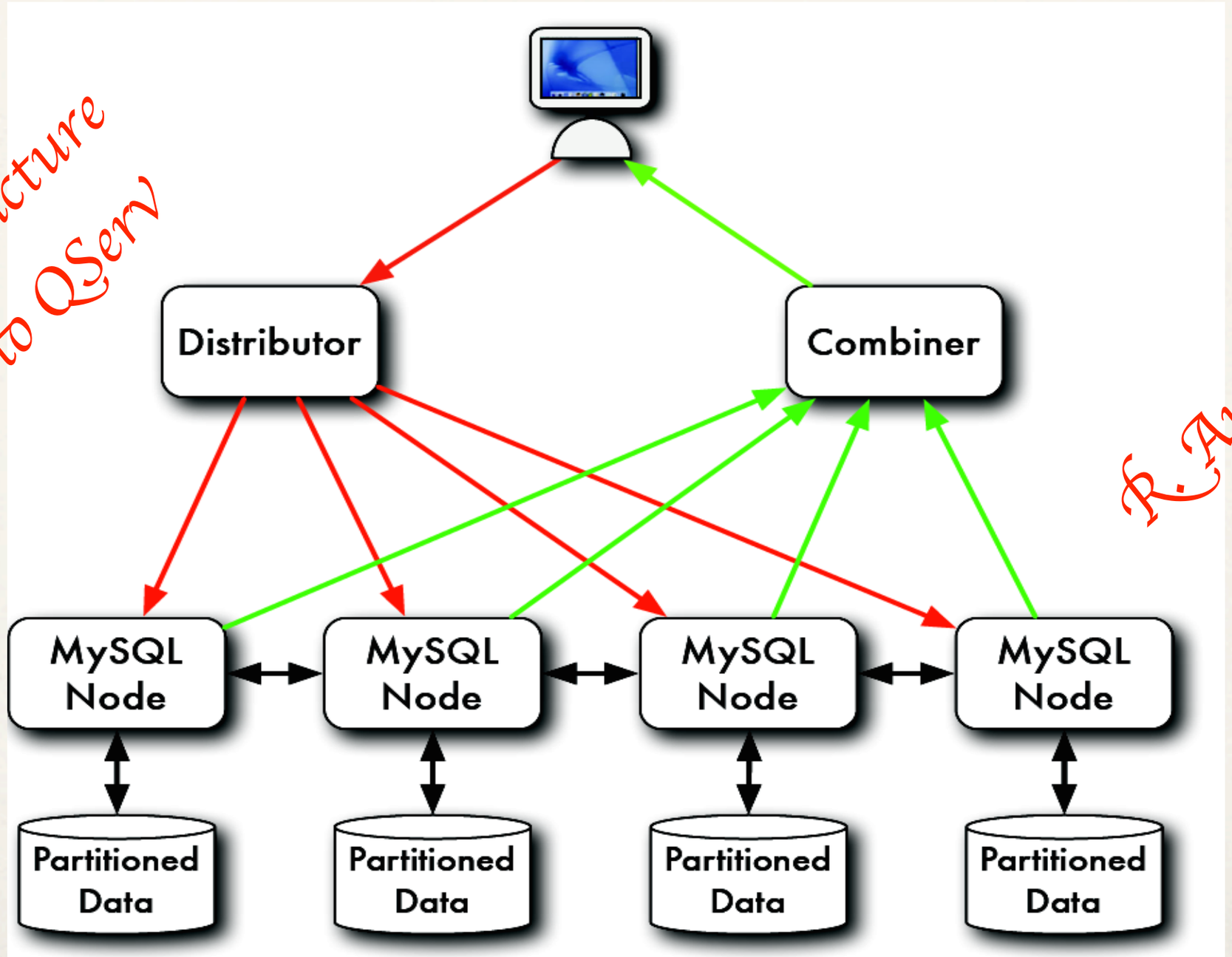
Computing nodes

LSST computing infrastructure

R. Ansari - Déc 2011

- Need an efficient, parallel computing and storage system :
  - ~ [100-200 nodes] x [50-100 TB + CPU's] for the first few years
  - ~ 1000 nodes x [50-100 TB + CPU's] for the full survey
- few MB/s / node data rates should allow several processing runs for DE type complex analysis
- Need a powerful map/reduce (or scatter/gather) software tool (or layer) to enable efficient use of the underlying infrastructure

Parallel processing/storage (MPP)



An infrastructure similar to QServ

R.Ansari - Déc 2011

LSST computing infrastructure architecture (1)

- Use of a distributed computing/storage infrastructure
- Data partitioning according to the sky position (with some overlap)
- LSST files are of the type Write Once, Read Many (WORM)
- A light weight DFS can be used , and the LSST code and tools can transparently access data using C++ classes inheriting from ifstream/ofstream
- The same ressources can be used for the catalogue database (QServ …)
- Possibility to deploy such an infrastructure at CC-IN2P3 ?
- The question ressources (computing and storage) allocation and management ?

LSST computing infrastructure architecture (2)
Possible arrangements:

- A set of SuperNodes , ~64 (first year) to ~256 (final configuration)
- Typical SuperNode: SN-A = [FileServer: 96-128 TB] + [64-128 cores]
- SuperNode-Final config: SN-D = [FileServer: 512 TB] + [512 cores]
- I/O rate will be around ~ 50 - 100 MB/s per SuperNode

LSST computing infrastructure architecture (3)
Possible Deployment plan (& cost)

- SN-0 : [2 × (2 × 8 c + 64-128 GB mem) = 32 cores] + 32-48 TB disk
- 2 × SN-0 would be enough for stripe 82 or CFHTLS DC's (2015)
- SN-1 : [64 cores + 72-96 TB] , 4 × SN-1 (2017)
- SN-A : [128 cores + 128 TB] , 8 × SN-A (2019)
- Pre-survey : 32 × SN-A (2021) at CC-IN2P3 (50 % DRP)
- Survey start (1st,2nd year : 48 × SN-A at CC-IN2P3 (2022)
- Cost : SN-0 → 50 k€ ? , SN-A → 150 k€ ???

LSST computing infrastructure architecture
Other issues

- Reduce storage costs : file compression (per image-CCD basis)
- Avoid long term storage of processed images ?
  - On demand creation of processed images (Apply photometric/ astrometric calibration)
- Reduce computing cost : use of GPU ...