CTA report

Luisa Arrabito, Johan Bregeon

LUPM CNRS-IN2P3 France

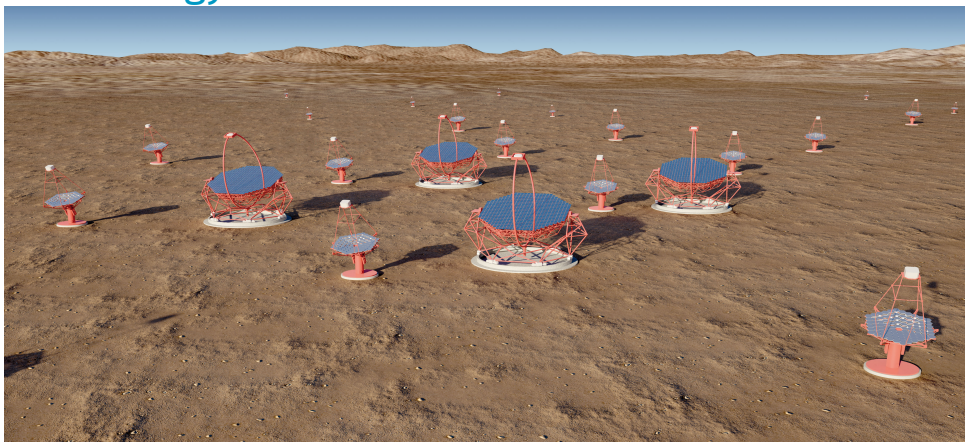
8th DIRAC User Workshop 22nd – 25th May 2018, Lyon

Outline

- CTA project
- DIRAC for CTA
  - Hardware setup
    - Galera cluster
  - DIRAC functionalities in use
    - Transformation system
    - Resources integration (cloud, HPC, GPU)
  - DIRAC systems extended
  - Externals, new DIRAC extensions, new DIRAC systems
    - Production System
  - DIRAC usage
- Conclusions and plans

CTA project

- The next generation instrument in VHE gamma-ray astronomy (1200 scientists in 32 countries)
  - Cosmic ray origins, High Energy astrophysical phenomena, fundamental physics and cosmology



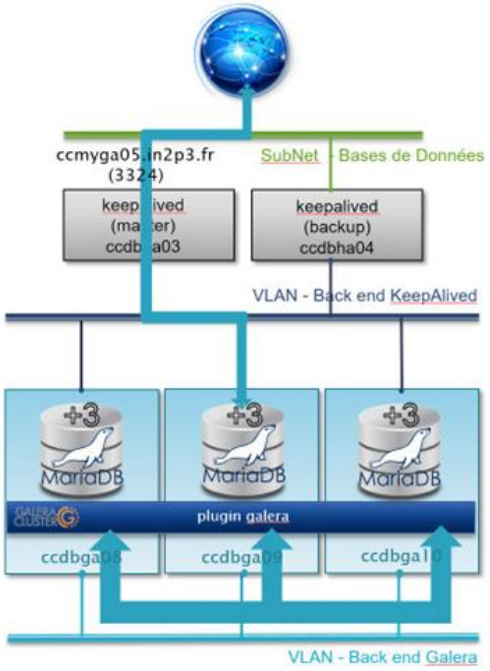
- Two arrays of Cherenkov telescopes
  - Northern hemisphere (La Palma, Spain): 4 LSTs, 15 MSTs
  - Southern hemisphere (Paranal, Chile): 4 LSTs, 25 MSTs, 70 SSTs
- Project schedule
  - Construction and deployment: 2020-2025
  - Science operations: start in 2022 for ~30 years

CTA-DIRAC hardware setup

- DIRAC instance dedicated to CTA distributed at 3 sites (CC-IN2P3, PIC, DESY)
- 5 core servers
  - 1 running WMS services (32 cores, 32 GB RAM)
  - 1 running WMS agents and executors (32 cores, 32 GB RAM)
  - 1 running TS and RMS (16 cores, 8GB RAM)
  - 1 running DMS + 1 DIRAC SE (16 cores, 8GB RAM, 2 TB of disk for the SE)
  - 1 running duplicated DMS, TS, RMS services (8 cores, 32 GB RAM)
- 2 MySQL servers
  - 1 hosting FileCatalogDB, TransformationDB, ReqDB (dedicated server at CC-IN2P3 recently migrated to *MariaDB/Galera cluster*)
    - After a few initial hiccups, excellent service quality so far in 2018
  - 1 hosting all other DBs at PIC (to be renewed this year)
- 1 server for the Web portal (at CC-IN2P3)
- Installed DIRAC version v6r19p20

Migration to MariaDB/Galera

- Migration done by CC-IN2P3 team mid-september 2017
- From 1 MySQL server to a 3 nodes MariaDB/Galera cluster (shared with FG-DIRAC instance)
  - It concerns 2 DBs: TransformationDB, ReqDB (all others are at PIC)
- Goals
  - Ensure functionality evolution of the service on long term
  - High availability and load balancing
  - Upgrade to more powerful servers
- Migration tested on a dedicated platform
  - Supposed to be transparent for us



Migration to MariaDB/Galera

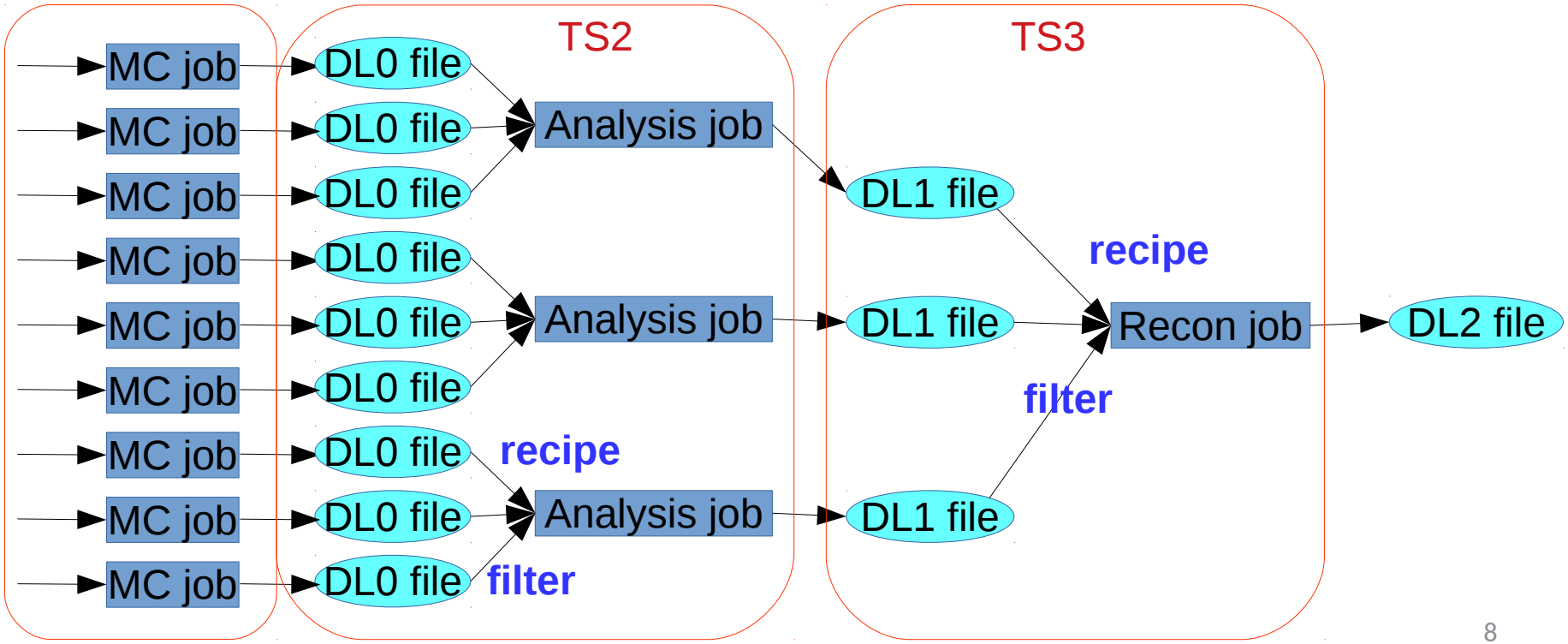
- An issue due to the new configuration caused several days of CTA-DIRAC downtime
  - Conflicts during insertions on different nodes causing the freezing of the whole cluster
- Another non-blocking issue also discovered
  - Auto_increment property passed from +1 to +3 (to avoid conflicts during insertions on a multi-node cluster)
    - Property used for instance for TransformationID, JobID, etc.
    - Not really a problem for DIRAC, but not that 'users friendly'
      - E.g. User sees TransID: 1563, 1566, 1569, …
- Agreed with CC-IN2P3 team to reconfigure the cluster architecture
  - Come back to a single node mode
  - Use 2 separate nodes for CTA and FG-DIRAC DBs
  - All DBs replicated to the third node, used for backup on a separate server
  - High availability still preserved

DIRAC functionalities in use

- Accounting
- Data Management (DMS)
- DIRAC File Catalog (DFC)
  - Extensively used as replica and meta-data catalog
  - Using datasets for official productions
- Request Management (RMS)
  - For replication/removal (through TS)
  - For job failover
- Transformation (TS)
  - For MC Simulation, data-processing
  - Data Management: bulk replication/removal
- *Production system prototype (PS)*
  - For MC Simulation, data-processing
- VMDIRAC
- COMDIRAC
- WebApp
- Workload Management (WMS)
  - Targeted resources: CREAM CE, ARC CE, PBS cluster

Data driven Workflow Management

- Transformation System
  - Automated Tasks, workhorse for MC production and analysis
  - A **Transformation** is an input *data filter + a recipe* to create jobs
  - Fully data-driven: jobs are created as soon as data with required properties are registered into the file catalog



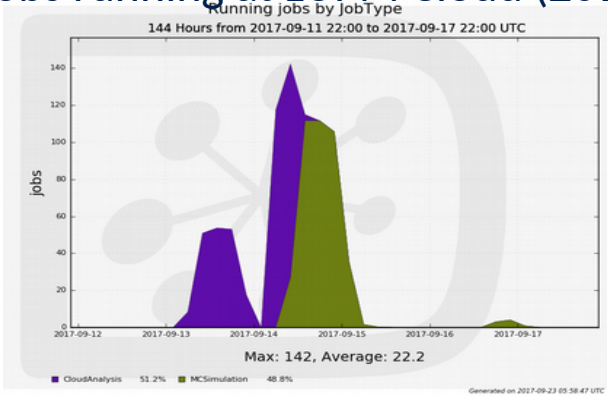
Data driven Workflow Management

- Production System
  - **Production**: set of **Linked Transformations**
    - *E.g.* 2 transformations t1, t2 are 'linked' if they verify
      - InputQuery2 intersects OutputQuery1
        - Files produced by t1 are potentially input of t2
      - TransMetadata1 verifies InputQuery2
        - At least part of the files produced by t1 are potentially input of t2
  - **Prototype implementation done early 2018**
    - Improved the current **transformation** definition to characterise the inputs/outputs of a **transformation** through *meta-queries*
    - Introduced new **transformation** attributes
      - Input/Output Query, Transformation Metadata
    - To be used in production for "La Palma 60 deg" simulation and analysis
  - Proposed as new general *System (Monday session)*

Resources integration

- CTA main resources are on the EGI grid
  - Also tackle a few farms directly through their batch system
  - Willing to integrate other kind of resources
- **Cloud** resources integration – *mostly done*
  - Using VMDIRAC module for transparent integration
    - Clouds are just seen as additional sites
  - Test Clouds
    - Commercial companies in the context of the HNSciCloud project
    - Academic clouds (LUPM, CC-IN2P3)
  - Functional tests successful (~50 VMs) in 2017
  - **Scalability** test (~1000 VMs) planned in 2018

Jobs running at LUPM Cloud (2017)



Resources integration

- **HPC center** resources integration – *started*
  - Using the Dirac **gateway** service
    - Access to HPC centers is usually very restricted (in and out)
  - Test platform
    - Bura SuperComputer in Croatia (Univ. of Rijeka)
- **GPU** resources integration – *coming soon*
  - Using Dirac jobs **tags** to request specific resources
    - Matching jobs to resources is the challenge
  - Test platform
    - CC-IN2P3 in Lyon has 10 Dell C4130 with 4 GPUs and 16 CPU cores per machine
      - Should be accessible through a CREAM CE

DIRAC systems extended

- Job API extensions
  - Simple extension of the Job API to configure and run CTA applications
  - Evolved to using a few job base classes, and use inheritance for specific productions
  - Need specific developments to wrap the official CTA software calibration and reconstruction pipeline (just started)
- Various specialized extensions
  - *Prod3SoftwareManager*
  - *Prod3DataManager*
- Agent reporting SE usage
  - Querying BDII

See our code at https://github.com/cta-observatory/CTADIRAC

Externals and new DIRAC systems

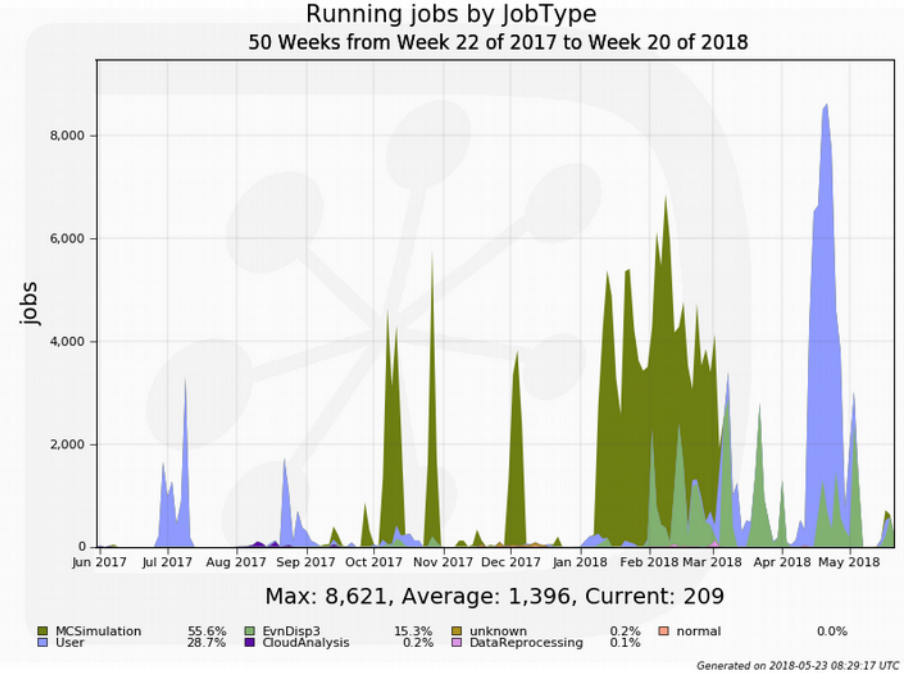
- Externals
  - CVMFS CTA repository
    - 1 Stratum 0 (at CC-IN2P3) and 2 Stratum 1 (at CC-IN2P3, DESY)
    - Almost all CTA sites configured to access the CTA repository
  - In future (2018): CTA Archive system
    - *e.g.* OneData (INDIGO / eXtreme DataCloud)
    - Might eventually/partially replace our use of the DFC
  - In future: CTA A&A system
- New DIRAC systems
  - Prototype of **Production System** to easily chain several transformations and handle them in a coherent way

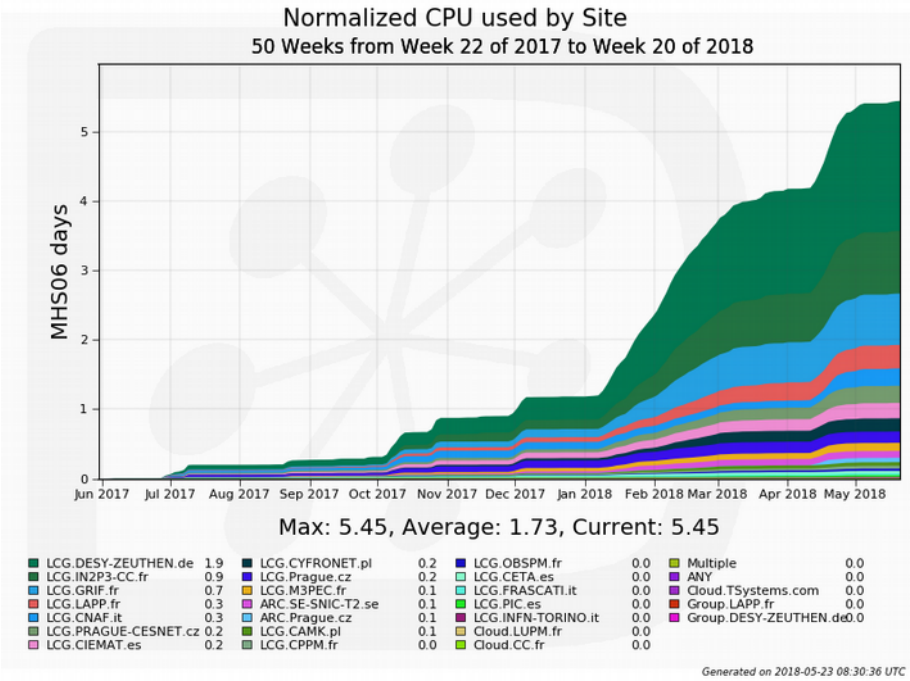
DIRAC usage since last year

- Extended simulations with the final array layouts for both North and South sites
  - 2 azimuth pointing directions and 3 zenith angles
  - New configurations (high night sky background)
  - 130 M HS06 hours and ~ 1 PB stored on disk
- Data processing
  - Reference analysis chain (*EventDisplay*, ROOT based)
  - New analysis chain for Machine Learning study (SciKitLearn)
- Heavily relying on the Transformation System
  - Using meta-filters through datasets to chain production and analysis
  - Still many manual operations :
    - Fold as much as possible into the Production system

DIRAC usage since last year

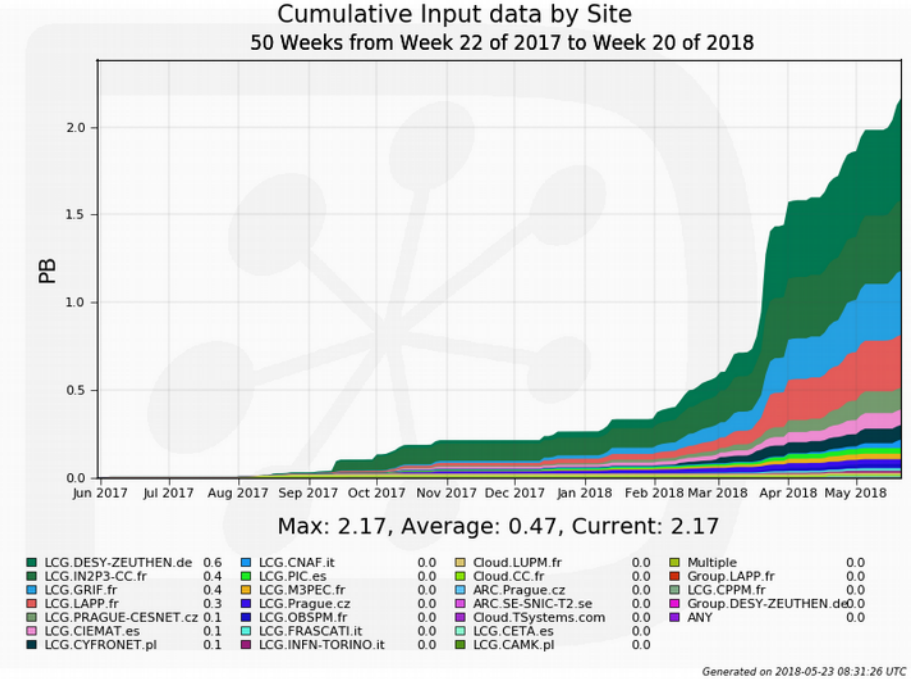
- MC production and analysis running in parallel
- 130 M HS06 hours
- 1.9 M executed jobs

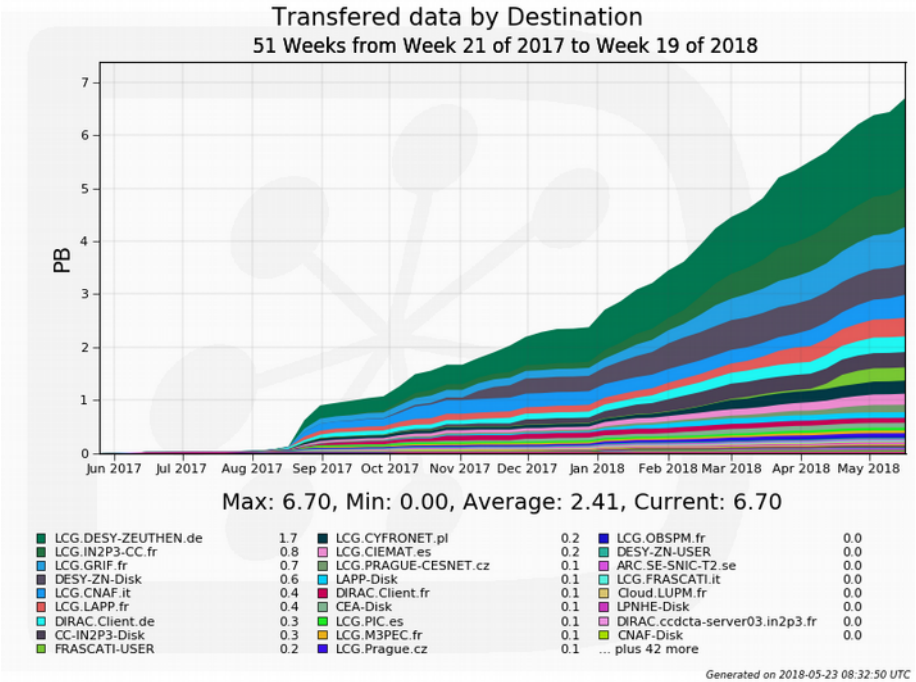




DIRAC usage since last year

- 6.7 PB of transfered data
- 2.2 PB of processed data
- Total: 3.2 PB currently on disk/tape
- Total: 28 M replicas in DFC





Conclusions

- CTA will start operations around 2024 and will be producing several PB/year (for more than 10 years)
- CTA will rely on a distributed computing model
  - Baseline with 4 Data Centers (not necesserly using grid middleware)
  - DIRAC will be used for the WMS and the Production System
- DIRAC is currently used for MC production and analysis
  - Using all the main DIRAC functionalities (WMS, DMS, TS, RMS)
- Many improvements since last workshop
  - DB server upgrade (Galera cluster)
  - TS meta-filters in production with datasets
  - Prototype Production System
  - Job API extension rewrite

Plans

- Further develop the Production System on the top of the Transformation System
- Resources integration : cloud scalability test, HPC, GPU
- Development of the CTA Job API to wrap/interface the CTA pipeline framework (currently under development)
- Work on interfaces with external systems CTA Archive and A&A

➔ Biggest challenge ahead will be to be able to process real data *efficiently*

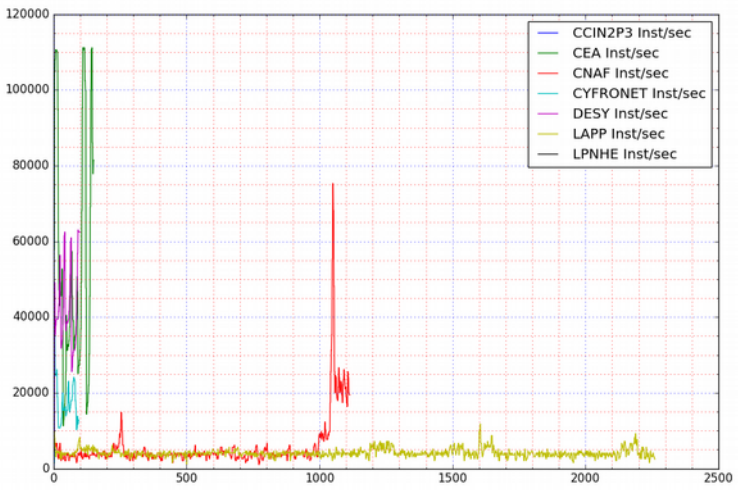
- ➔ And it's actually a very different use case with respect to what we are doing now.

Backup

VMDIRAC tests with LUPM Cloud

- Stalled jobs when accessing Input Data at particular SEs
  - Identified two SEs with low throughput: CNAF and LAPP

SE/Site	Cloud.LUPM	Cloud.IPHC	LCG.CC-IN2P3 (wn/ui)
CC-IN2P3-Disk	84 MB/sec		
CEA-Disk	22 MB/sec		
LPNHE-Disk	72 MB/sec		
LAPP-Disk	5.3 MB/sec	18.2 MB/sec	
DESY-ZN-Disk	33 MB/sec	12 MB/sec	
CYF-STORM-Disk	31 MB/sec		
CNAF-Disk	3.8 MB/sec	37 MB/sec	100 MB/sec /3.5 MB/sec



Testing VMDIRAC

- Started testing VMDIRAC in the context of HNSciCloud EU project
  - Goal: run CTA MC production and analysis workflows on commercial and private clouds
  - First basic tests with simple jobs to access commercial cloud resources (only one provider tested, Openstack based)
- Successful tests BUT only after some bug fixes
  - libcloud
    - Bugs found in method to list floating IPs and to check provider Service name -> Maybe due to an old version -> To be confirmed
  - VMDIRAC
    - Essentially missing connection details to the cloud endpoint
      - PR waiting to be merged : https://github.com/DIRACGrid/VMDIRAC/pull/103
- Not succeeded to make WebApp VMDIRAC working
- Prepared VMDIRAC basic documentation

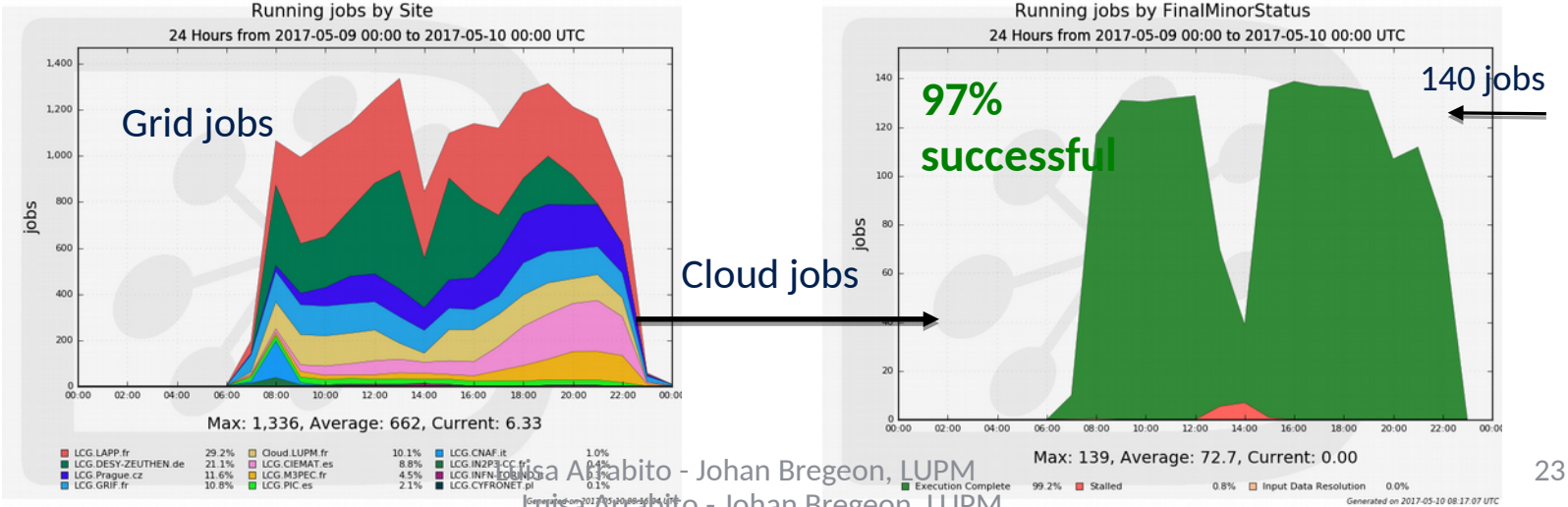
Testing VMDIRAC for LUPM Cloud

- Tests with LUPM Openstack cloud (and recently IPHC cloud)
  - Part of France Grilles Federated Clouds
  - Not large resources available -> Tested with max 150 jobs
- Using customized bootstrap scripts
  - Configure CVMFS to access CTA repository (thanks to A. Haupt)
- Using sligthly customized images
  - Starting from standard CentOS 6
  - Install a library needed to interact with srm SE: (ltd libs)
  - Install a few libraries needed for CTA (*e.g.* gfortran)
- Tested with real CTA production jobs
  - MC production and analysis (corsika, sim_telarray, evndisplay)
  - All types of jobs run well (running only a few jobs as proof of concept)

VMDIRAC tests with LUPM Cloud

- More advanced tests with jobs accessing 'distant' Input Data (at 6 SEs: CC-IN2P3, CNAF, DESY, LAPP, GRIF, CYFRONET)
  - For small Input Data (2 GB) -> Jobs run successfully
  - For larger Input Data (3x2GB)
    - Stalled jobs when accessing data at particular SEs
    - Not a general problem of VM connectivity
    - While CTA grid sites are all well connected to the 6 SEs

Analysis campaign (2 GB input data)



Luisa Arrabito - Johan Bregeon, LUPM