CTA Computing Grid resources and technical aspects of CTA MC simulations

L. Arrabito, J. Bregeon

LUPM CNRS-IN2P3 France

CCF Meeting 20th-23rd June 2016, Barcelona

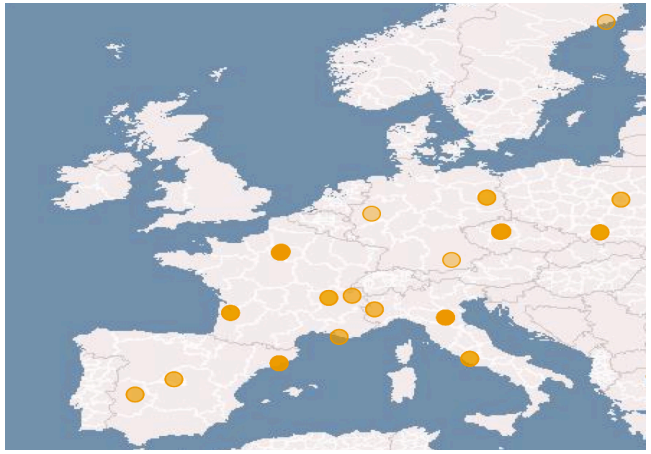
Outlook

- Current status of CTA Computing Grid (CTACG)
  - CTACG resources
  - Computing model and operations
  - DIRAC for CTA
  - Past productions: prod3
- Organization process for official productions
- How are handled specific requests?
- Conclusions and future plans

CTA Computing Grid (CTACG)

CTA Computing Grid (DATA WP)

- Use of EGI grid through the CTA Virtual Organization (since 2008)
- Use of DIRAC to access grid resources (since 2011)

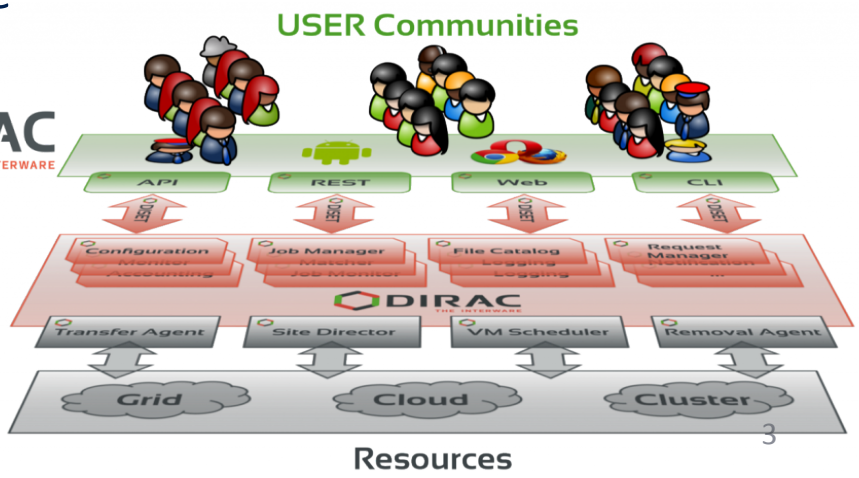


CTA Virtual Organization

- Open to any CTA member having a grid certificate
- Supported by 20 EGI sites in 7 countries + 1 ARC site in Sweden
- Eventual new OSG (US) resources in future

DIRAC for CTA

- Workload and Data Management System
- Dedicated server instance at CC-IN2P3, PIC and DESY
- CTA-DIRAC software extension



CTACG resources

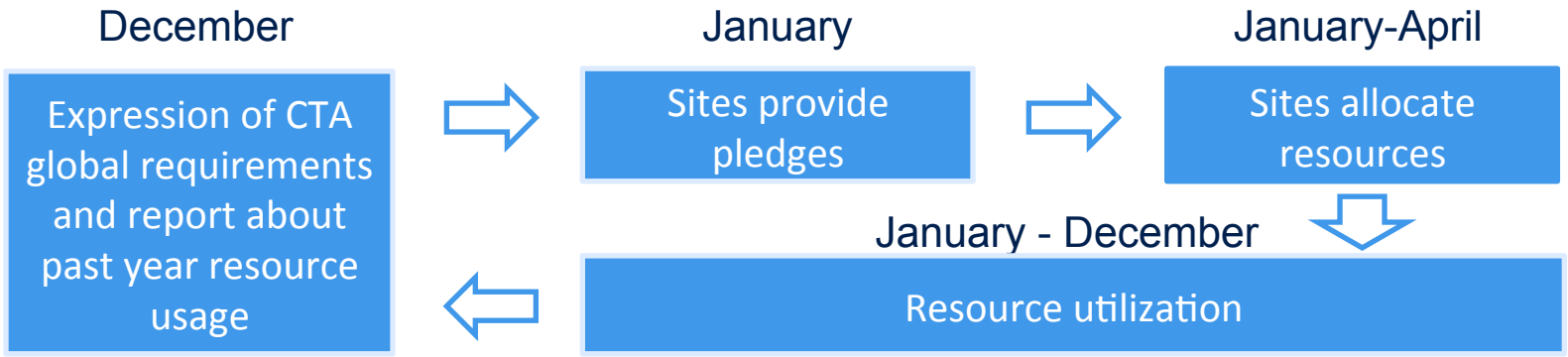
- Reference pages
  - https://forge.in2p3.fr/projects/cta_dirac/wiki/Estimation_of_computing_resources_usage
  - https://forge.in2p3.fr/projects/cta_dirac/wiki/CTA-DIRAC_MC_PROD3_Status
- CPU: 6000 – 8000 cores available on average
- About 2.2 PB (+ 0.7 PB of tape) distributed among 6 main sites
- Additional 50 TB at Frascati and Torino for specific studies
- Disk fills rapidly during MC massive productions -> usually the main bottleneck

Site	Available Disk (TB)	Used Disk (TB)	Total Disk (TB)
CYFRONET-LCG2	16	627	643
DESY-ZEUTHEN	7	648	655
IN2P3-CC	105	249	354
GRIF (LPNHE+CEA)	17 (7+10)	182 (112+70)	200 (120+80)
IN2P3-LAPP	29	89	118
INFN-T1	114	172	286
Total	288	1967 (87%)	2255

CTACG resource managament

Organization

- A few meetings per year between CTACG and MC group to estimate the needed resources
- 1 meeting per year between CTACG and site contacts (CTA contact + site admin)
- 2 working documents: 'CTA VO Requirements' and 'CTACG Planning' sheet
- Resource requirements/pledges/provision on an annual basis
- Informal agreements up to now, but used resources accounted as in-kind contributions
- See also https://forge.in2p3.fr/projects/cta_dirac/wiki#CTACG
- If you want to contribute with resources, contact us



Computing Model and Operations

Computing Model

- MC productions are run everywhere (about 20 grid sites)
- Output data are stored (on the fly) at 6 main Storage Elements (SE)
- MC analysis is run at a restricted nb of sites (for efficiency reasons)
- We can freely specify different job scheduling rules for each ‘job type’

Operations (production team)

- Resource management
- DIRAC servers administration
- Development of CTA-DIRAC sw extension (essentially the job interface to configure CTA jobs)
- Launch and follow the productions
- Users support for specific MC productions and analysis

DIRAC for CTA

DIRAC is a framework for the Workload and Data Management on distributed resources

- Used to access CTACG resources, but it can integrate also other types of resources (clouds, local clusters, etc.)
- Based on a Services Oriented Architecture
- CTA has a dedicate server installation (recently upgraded) at IN2P3, PIC, DESY
  - 5 core servers running all Services and Agents
  - 2 MySQL servers
  - 2 web servers
- To access CTACG resources, users just need to install the DIRAC client (don't need any grid middleware)

DIRAC for CTA: main Systems in use

- Workload Management System
  - Job brokering and submission (pilot mechanism)
  - It provides a common interface to hetereogenous resources (CREAM, ARC, clusters)
  - Central management of CTA VO policies
- Data Management System
  - All data operations (download, upload, replication, removal)
  - It includes a Replica and Metadata Catalog (DIRAC File Catalog, DFC)
- Transformation System
  - Used by production team to handle 'repetitive' work (many identical tasks with a varying parameter), i.e. MC productions, MC analysis, data management operations (bulk removal, replication, etc.)

DIRAC File Catalog (I)

- Replica and Metadata catalog
- More than 27 M of replicas registered in a logical namespace
- About 10 meta-data defined to characterize MC datasets
- Simple commands to retrieve list of files, *e.g.*:

cta-prod3-query --site=Paranal --particle=gamma --tel_sim_prog=simtel --array_layout=hex --phiP=180 --thetaP=20 --outputType=Data

- Typical queries return hundreds of thousands of files
- Main useful queries in CTA-DIRAC redmine wiki:
  https://forge.in2p3.fr/projects/cta_dirac/wiki/CTA-DIRAC_MC_PROD3_Status#CTA-DIRAC-MC-PROD3-Status

DIRAC File Catalog (II)

- It supports ‘datasets’, i.e. aliases to given queries
- We have created datasets for the most common queries, *e.g.*:

$ cta-prod3-show-dataset
Available datasets are:
Paranal_electron_North
Paranal_electron_North_20deg_3HB8
…

$ cta-prod3-show-dataset Paranal_gamma_South
Enter eventsPerRun (default 20000): Paranal_gamma_South: ... 1 MetaQuery {'thetaP': 20.0, 'particle': 'gamma', 'array_layout': 'hex', 'tel_sim_prog': 'simtel', 'outputType': 'Data', 'MCCampaign': 'PROD3', 'phiP': 0.0, 'site': 'Paranal'} 2 EventsPerRun 20000 3 TotalNumberOfEvents 0.10e9 4 NumberOfFiles 49112 5 TotalSize 75.5 TB

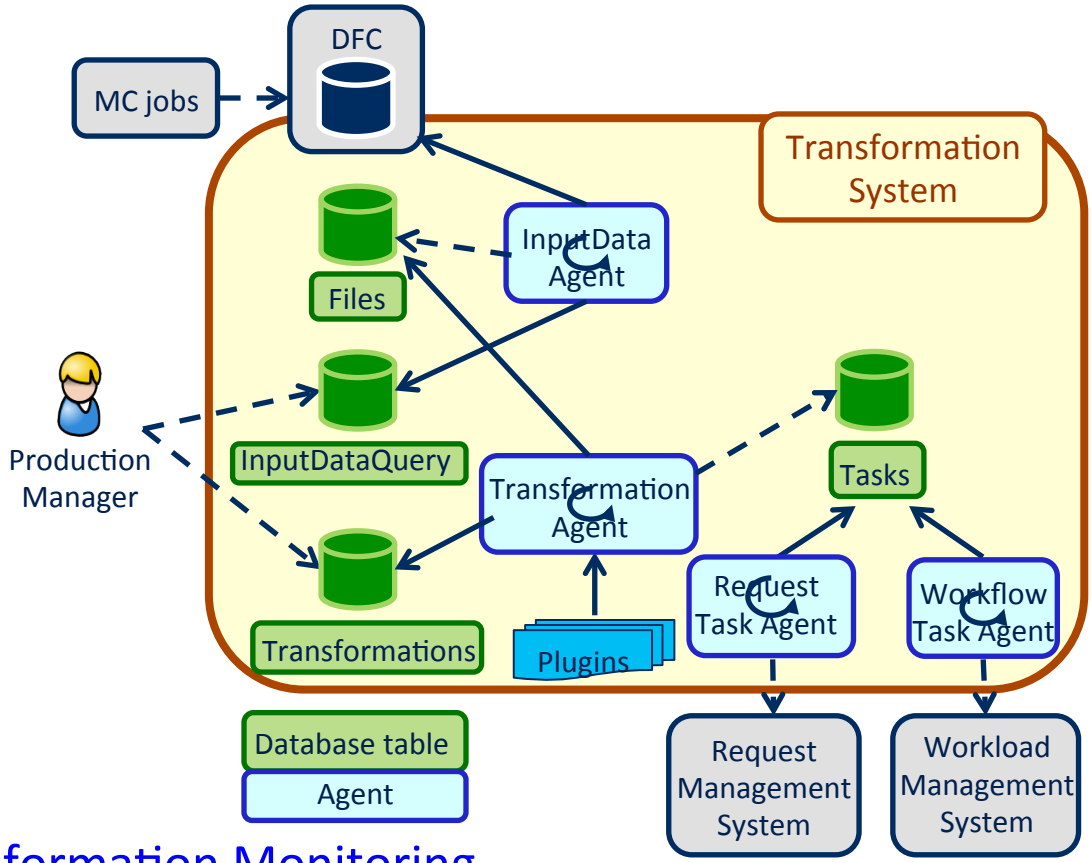
$ cta-prod3-dump-dataset Paranal_gamma_South

- Web interface available for free in DIRAC, but not really used for productions

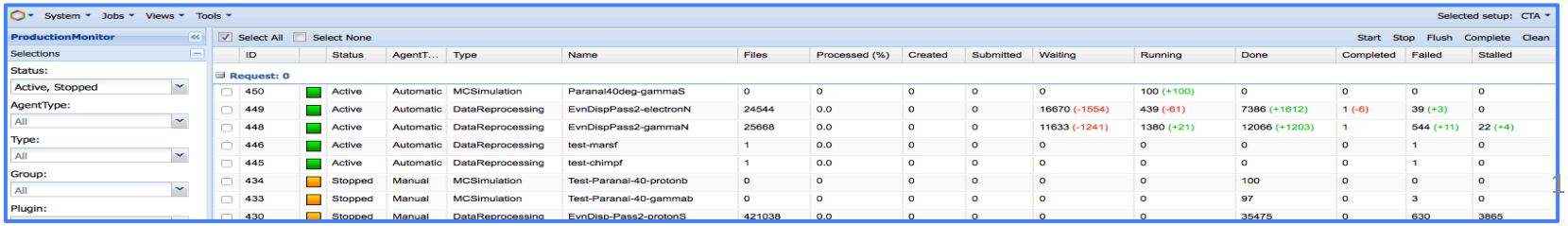
DIRAC for CTA: Transformation System

Transformation System Architecture

- The Production Manager defines the transformations with meta-data conditions and ‘plugins’
- InputData Agent queries the DFC to obtain files to be ‘transformed’
- Plugins group files into tasks according to desired criteria
- Tasks are created and submitted to the Workload or Request Management System



Transformation Monitoring



ID	Status	AgentT...	Type	Name	Files	Processed (%)	Created	Submitted	Waiting	Running	Done	Completed	Failed	Stalled
450	Active	Automatic	MCSimulation	Paranal40deg-gammaS	0	0	0	0	0	100 (+100)	0	0	0	0
449	Active	Automatic	DataReprocessing	EvnDispPass2-electronN	24544	0.0	0	0	16670 (-1554)	439 (-61)	7386 (+1612)	1 (-6)	39 (+3)	0
448	Active	Automatic	DataReprocessing	EvnDispPass2-gammaN	25668	0.0	0	0	11633 (-1241)	1380 (+21)	12066 (+1203)	1	544 (+11)	22 (+4)
446	Active	Automatic	DataReprocessing	test-marsf	1	0.0	0	0	0	0	0	0	1	0
445	Active	Automatic	DataReprocessing	test-chimpf	1	0.0	0	0	0	0	0	0	1	0
434	Stopped	Manual	MCSimulation	Test-Paranal-40-protonb	0	0	0	0	0	0	100	0	0	0
433	Stopped	Manual	MCSimulation	Test-Paranal-40-gammab	0	0	0	0	0	0	97	0	3	0
430	Stopped	Manual	DataReprocessing	EvnDisp-Pass2-protonS	421038	0.0	0	0	0	0	35475	0	630	3865

Past productions: resource usage

Prod2, Prod3 run on CTACG resources and MPIK, DESY clusters:
https://forge.in2p3.fr/projects/cta_dirac/wiki/Estimation_of_computing_resources_usage

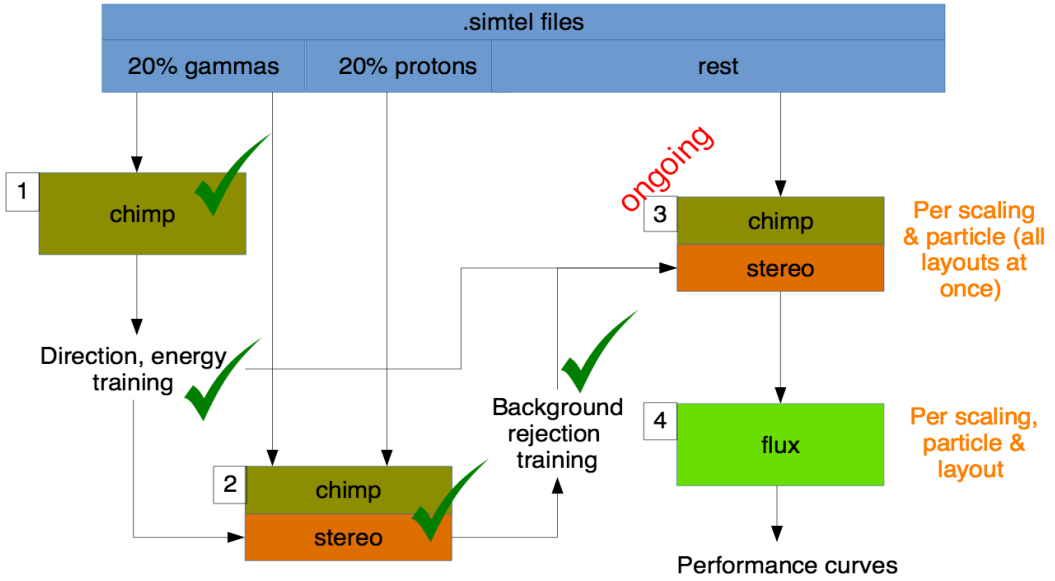
- CTACG
  - ~100 M hours HS06 per year and 2 PB of disk (+ 0.3 PB of tape)
  - Prod3 MC production for Paranal and first stages of Prod3 analysis (MARS and evndisplay)
  - Prod2 MC production
- MPIK
  - ~38 M hours HS06 per year and 950 TB
  - Prod3: MC production for La Palma and Paranal test
  - Prod2: MC production and analysis
- DESY
  - ~30 M hours HS06 per year and 1.2 PB
  - Prod2+Prod3 analysis

Prod3 on CTACG (I)

Prod3 running since August 2015

- Study the different possible layouts of telescope arrays, pointing configurations, hardware configurations, etc.
- 800 telescope positions, 7 telescope types, multiple possible layouts, 5 different scalings
- Run 2 different analysis chains on the simulated data (MARS, evndisplay)
  - Each one processing about 500 TB and 1 M of files for 36 different configurations

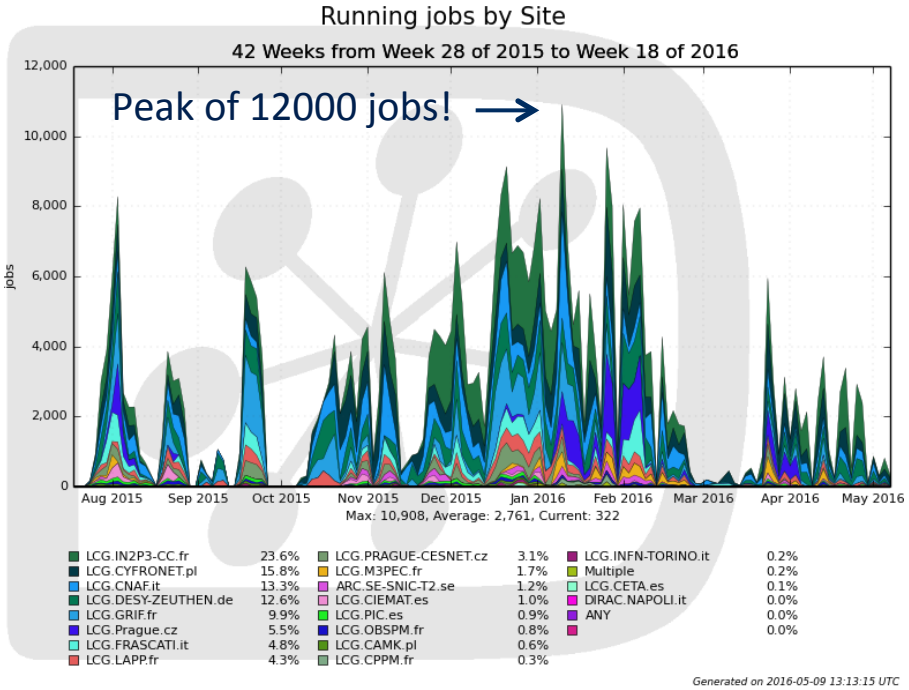
Example of analysis with MARS

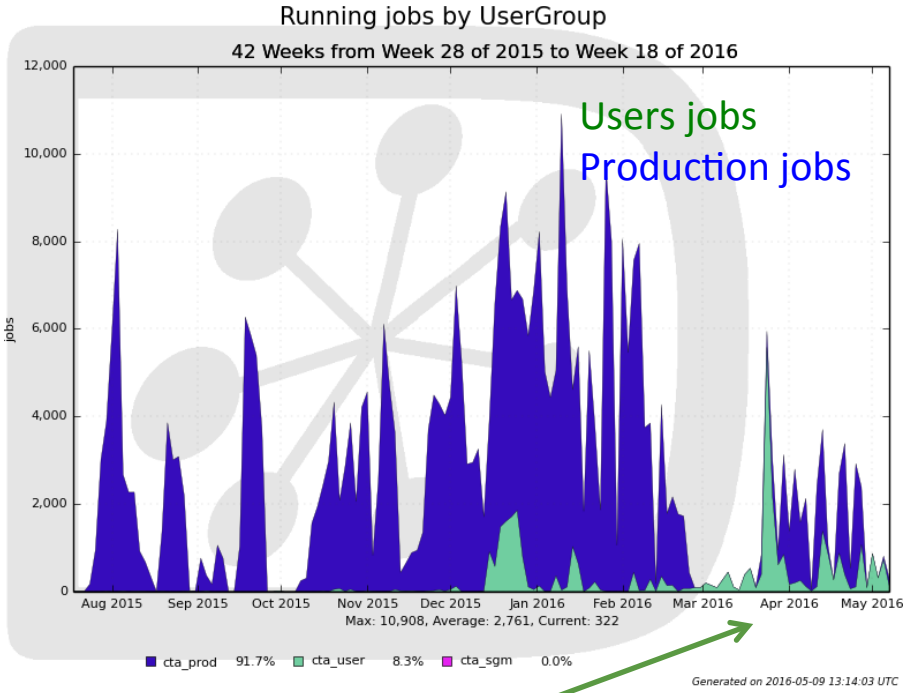


picture by Abelardo Moralejo

- Very complex workflow
- Using many sub-samples
- Several steps performed on the grid by the production team while others locally by analysis teams
- The whole chain far from being automatised

Prod3 on CTACG (II)

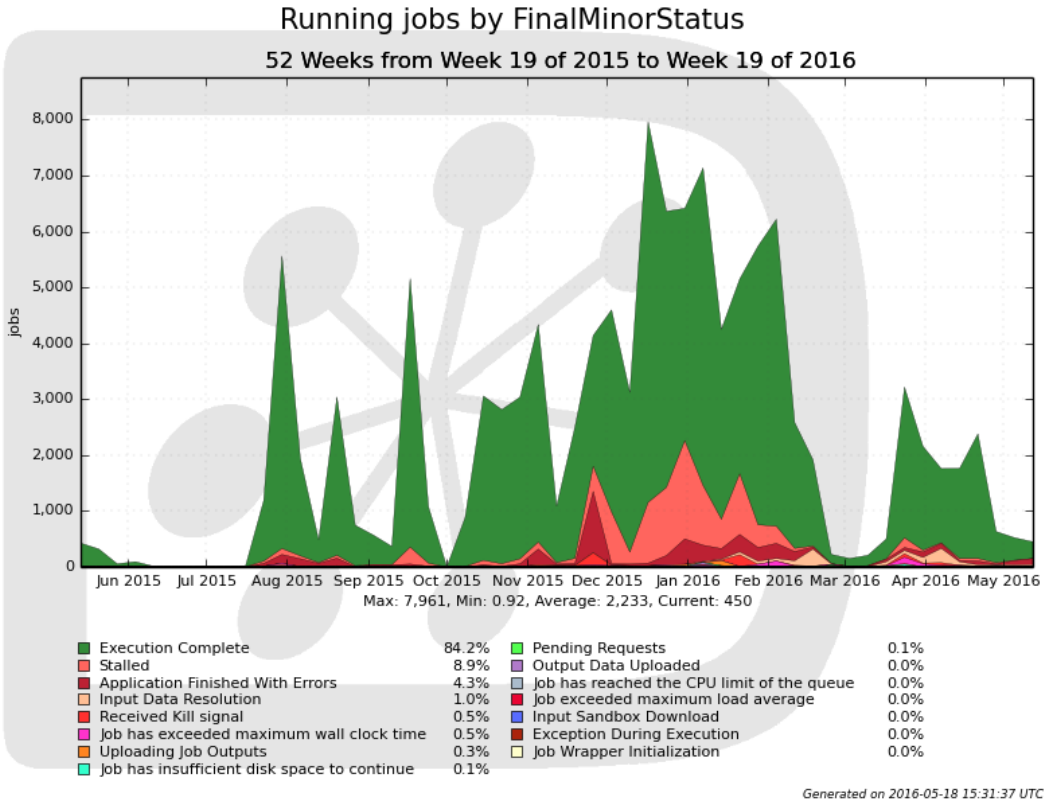




Specific productions run by users
(atmospheric studies, SST mini array, etc.)

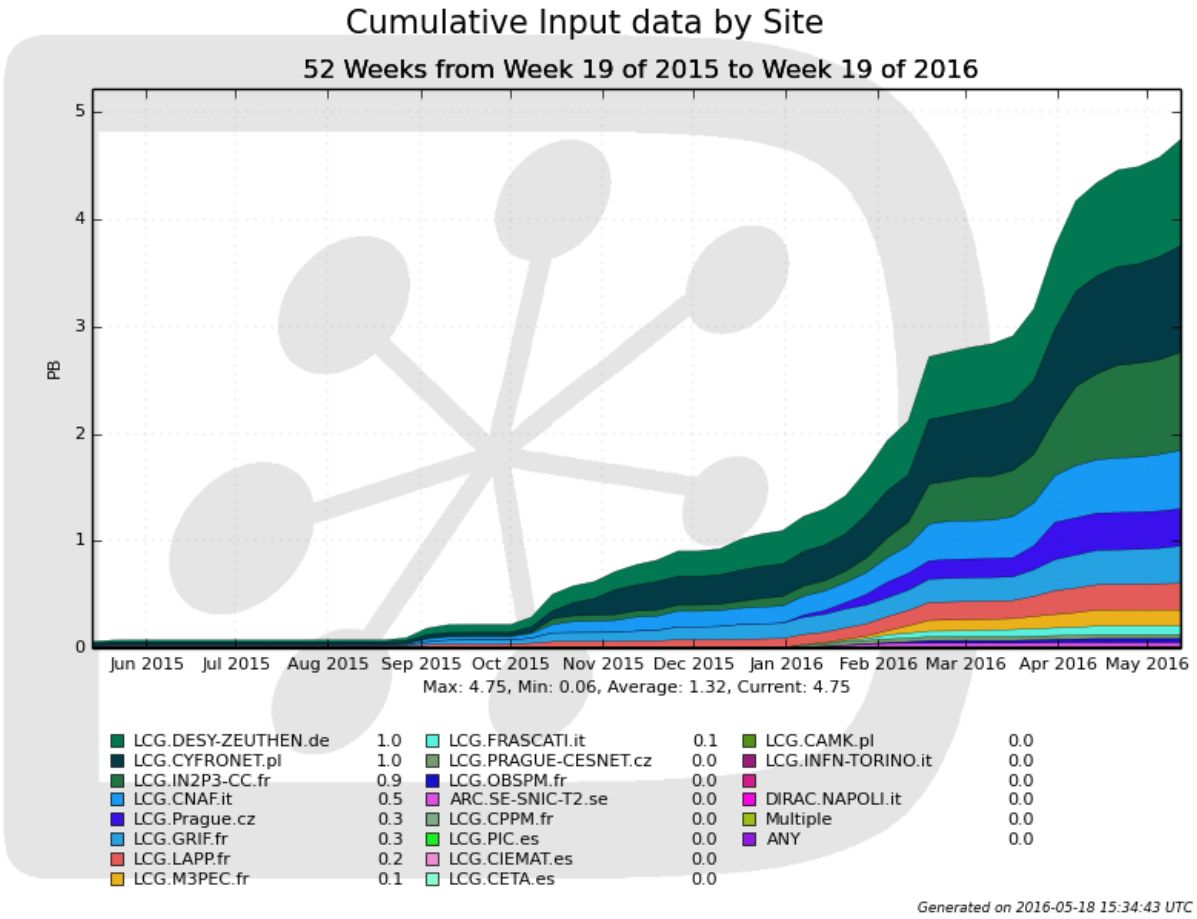
Prod3 on CTACG (III)

- Job success rate: 84%, main reasons for failures
  - Stalled jobs due to high memory consumption
  - Application errors



Prod3 on CTACG (IV)

- About 4.7 PB of processed data



Organization process for official productions (I)

- Discuss plans details
  - Priorities of simulation productions
  - Simulation scheme
  - Needed statistics
  - Resource estimations (CPU, storage, memory req.)
  - Sw to use and configuration parameters
- Where?
  - MC calls
  - redmine MC forum, *e.g.*:
    https://forge.in2p3.fr/boards/201/topics/1536?r=1582

Organization process for official productions (II)

- Run productions sharing tasks
  - MC group
    - provides sw for production and analysis
    - runs productions and analysis on local clusters and on the grid
    - produces final results
  - Production team (DATA WP)
    - ports MC sw into DIRAC workflows
    - runs massive productions on the grid and low-level stages of analysis (using the Transformation System)
  - Use CTA-DIRAC redmine issues for the follow-up: https://forge.in2p3.fr/projects/cta_dirac/issues
- Really a team work!

How are handled specific requests?

Example of atmospheric simulation studies

- Interested people should estimate the needed resources (CPU and storage) and contact the production team
  - Usually CPU is not a problem
- If the requirements are low wrt to official productions (especially for storage, i.e. a few TB)
  - Users run the production by themselves submitting parametric jobs (usually enough) -> See next slide to get started with DIRAC
  - Power users could use the more advanced Transformation System if needed
  - Users jobs have highest priorities wrt to production jobs
  - Production team provides support (best effort basis) for job debugging, porting sw into DIRAC workflows, etc.
- If the requirements are high
  - These requests should be discussed within the MC group
- In any case -> use CTA-DIRAC redmine for your requests

Run your production: getting started with DIRAC

- General documentation on CTA-DIRAC wiki:
  https://forge.in2p3.fr/projects/cta_dirac/wiki/CTA-DIRAC_Users_Guide
- Pre-requisites:
  - Request a grid certificate
  - Register to CTA VO
- Install the DIRAC client. Note that:
  - Access to grid SE relies on lcg-bindings -> recommended platform is SL6
  - Other platforms (other Linux and Mac OS) can access grid SE using an alternative DIRAC service but with lower performances
  - Consider using a VM
- Learn DIRAC basics (job and data management) from the wiki
- Ready-to-use examples are available for prod3 simulation and 2 analysis sw (read_cta, evndisplay)
  - https://forge.in2p3.fr/projects/cta_dirac/wiki/CTA-DIRAC_Prod3_Users_Guide
  - Simple python scripts using DIRAC API
- Adapt the scripts for your use case

Conclusions

- CTACG resources are open to any member of CTA
- The entry point to access resources is DIRAC
- With the current CTA-DIRAC prototype, we manage the MC production, analysis and data archive
- In future:
  - CTA computing model baseline is a distributed model using 4 data-centers
  - DIRAC will be used to manage the different steps of the level 1 data processing
  - To do this, it will interface the final CTA Archive (under development)
- If you have any simulation request, just contact us
- Simulation priorities are discussed within MC group and production team

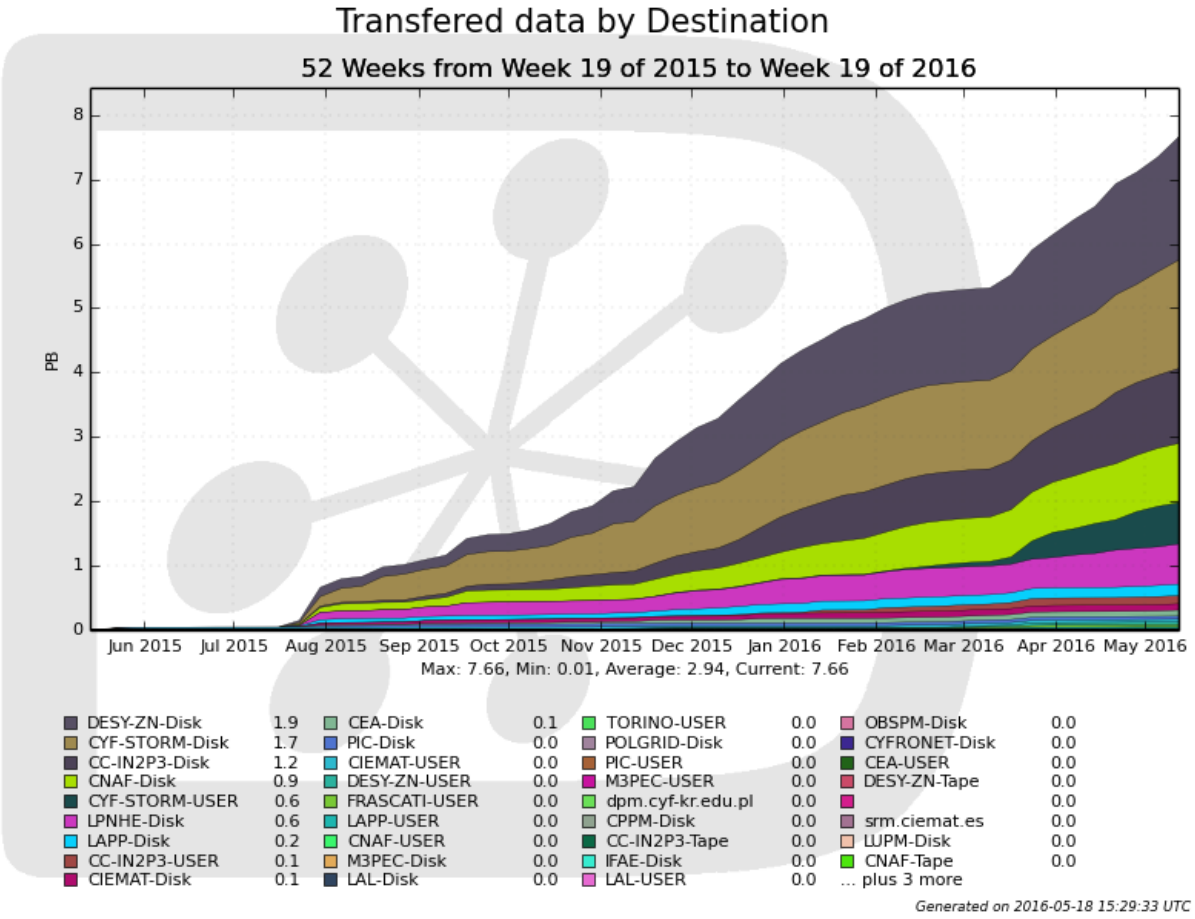
Future plans

- The next production to be started soon is the simulation of the HB9 array layout (approved baseline layout for CTA South) -> it will require about 500 TB
- Development of a production system, based on the Transformation System to automatise the whole processing chain (on an input data-driven model)
- Use CVMFS as software repository for grid jobs

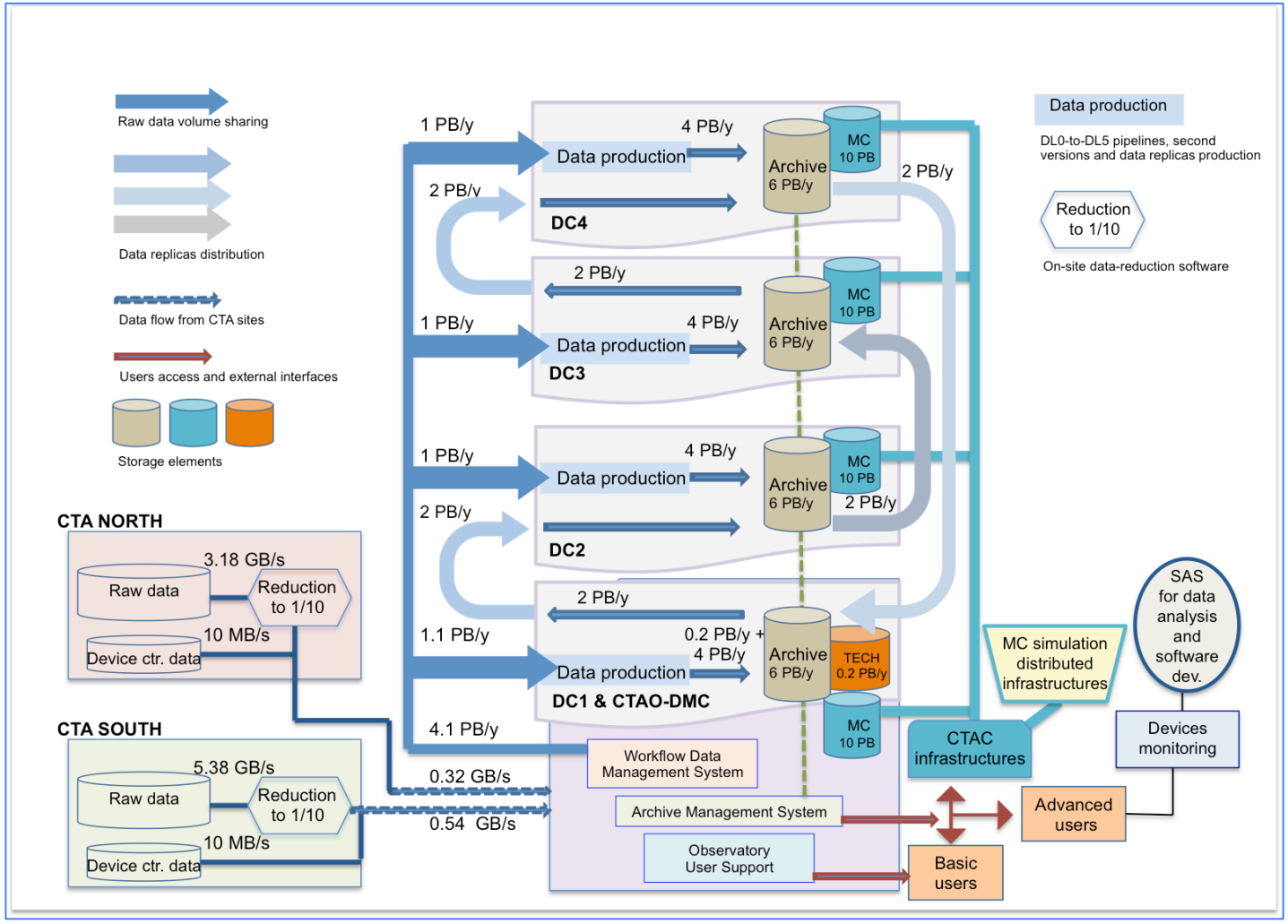
Backup

Prod3 on CTACG (IV)

- About 7.7 PB of transferred data



CTA computing: data flow



CTA data volume

Raw-data rate

- CTA South: 5.4 GB/s
- CTA North: 3.2 GB/s

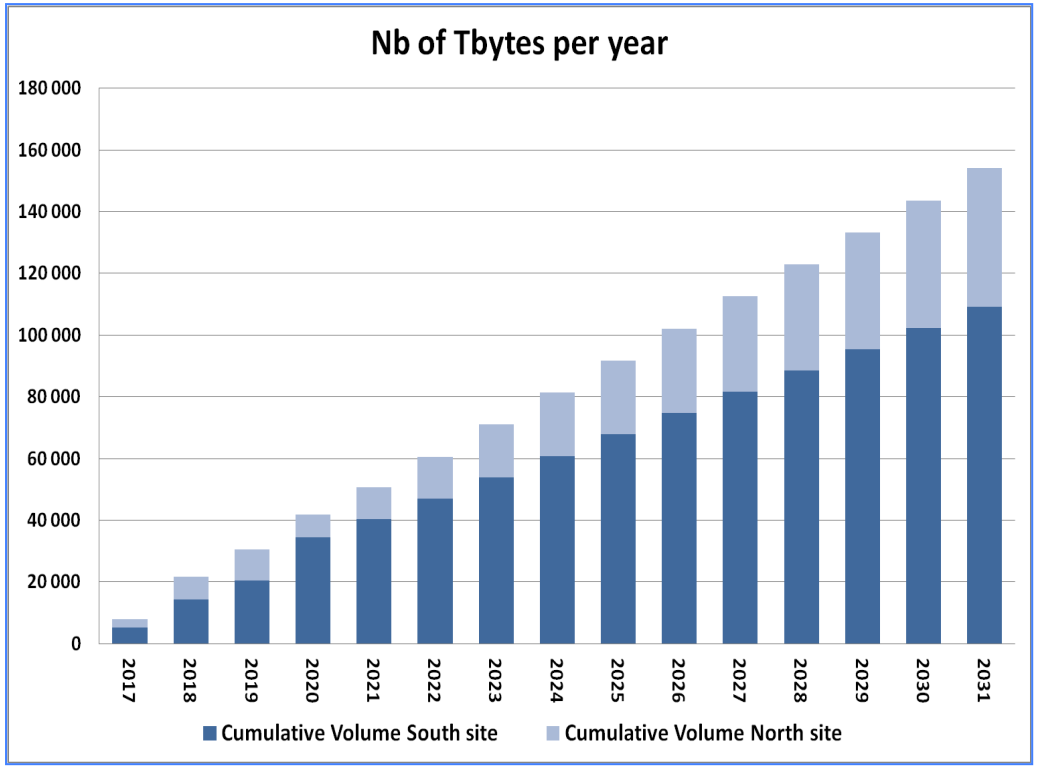
1314 hours of observation per year

Raw-data volume

- ~40 PB/year
- ~4 PB/year after reduction

Total volume

- ~27 PB/year including calibrations, reduced data and all copies



DIRAC hardware setup

- DIRAC instance dedicated to CTA: upgraded in 2016

- 5 Core servers
  - 1 running DMS + 1 DIRAC SE: 16 cores, 8GB RAM, 2 TB of disk for the SE (at IN2P3)
  - 1 running TS and RMS: 16 cores, 8 GB RAM (at IN2P3)
  - 1 running WMS: 32 cores, 32 GB RAM (at PIC)
  - 1 running Accounting, Framework, etc.: 32 cores, 32 GB RAM (at PIC)
  - 1 server installed at DESY last week (thanks to A. Haupt): running redundant services
- 2 MySQL servers
  - 1 hosting FileCatalogDB, TransformationDB, ReqDB (dedicated server with 600 GB at IN2P3 MySQL cluster)
  - 1 hosting all other DBs (old server)
- Web servers
  - 1 hosting the new web portal
  - 1 hosting the old web portal still in use