



cherenkov
telescope
array

CTA DATA Management challenges

N.Neyroud (LAPP/IN2P3/CNRS)

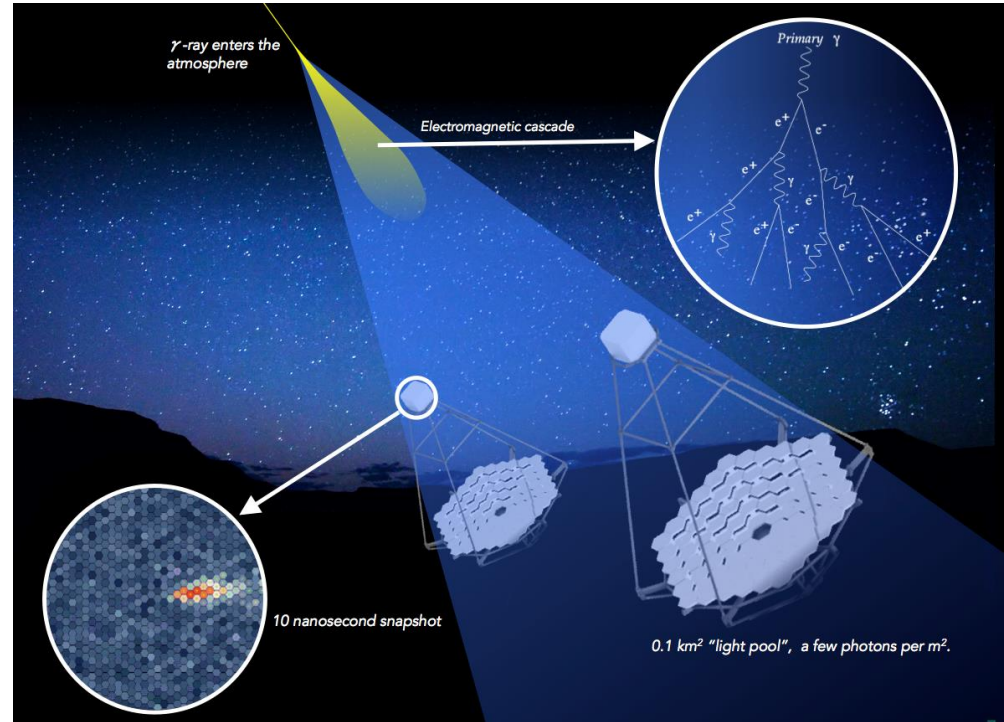


- CTA Overview
- CTA Data volume
- CTA Computing Model
- CTA Data access
- Current status
- Conclusions

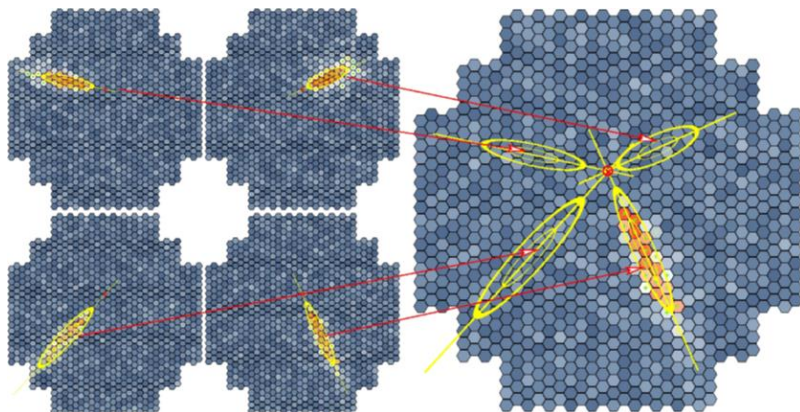
Cherenkov Telescope Array: how it works



- The next generation instrument in VHE gamma-ray astronomy (1350 scientists and engineers in 32 countries)
 - Cosmic ray origins, High Energy astrophysical phenomena, fundamental physics and cosmology



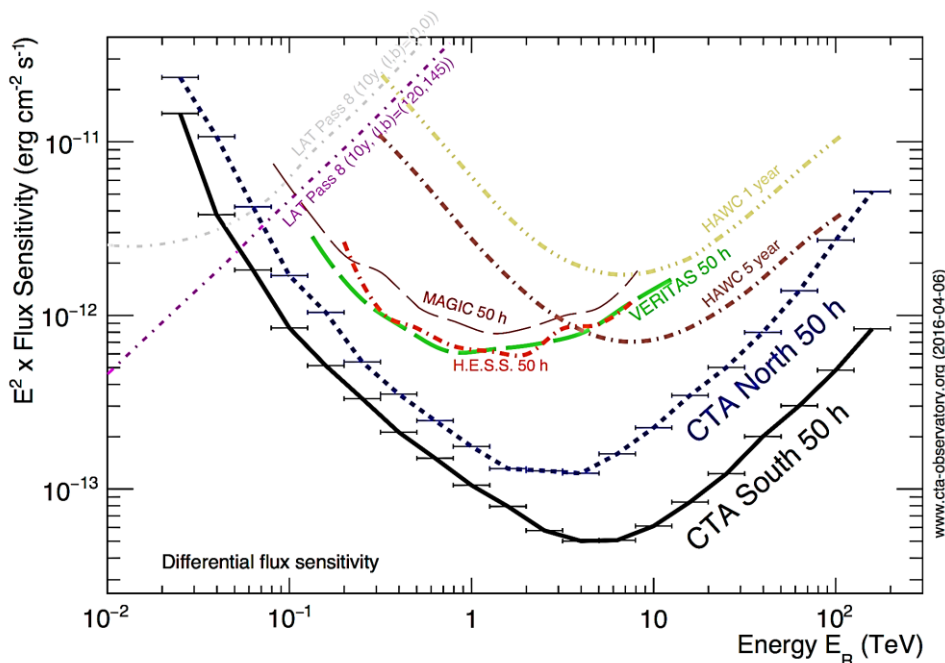
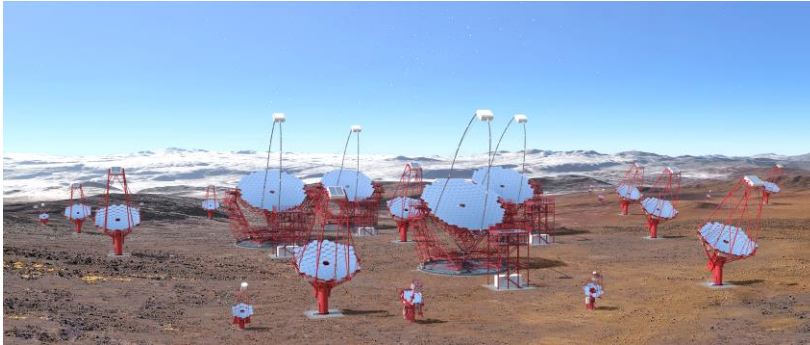
Cherenkov light



Stereoscopic reconstruction

HESS figure (K. Bernlohr credit)

Cherenkov Telescope Array (CTA)



Energy range: 20 GeV – 300 TeV

Two arrays of Cherenkov telescopes

- Northern hemisphere (La Palma, Spain): 4 LSTs, 15 MSTs
- Southern hemisphere (Paranal, Chile): 4 LSTs, 25 MSTs, 70 SSTs

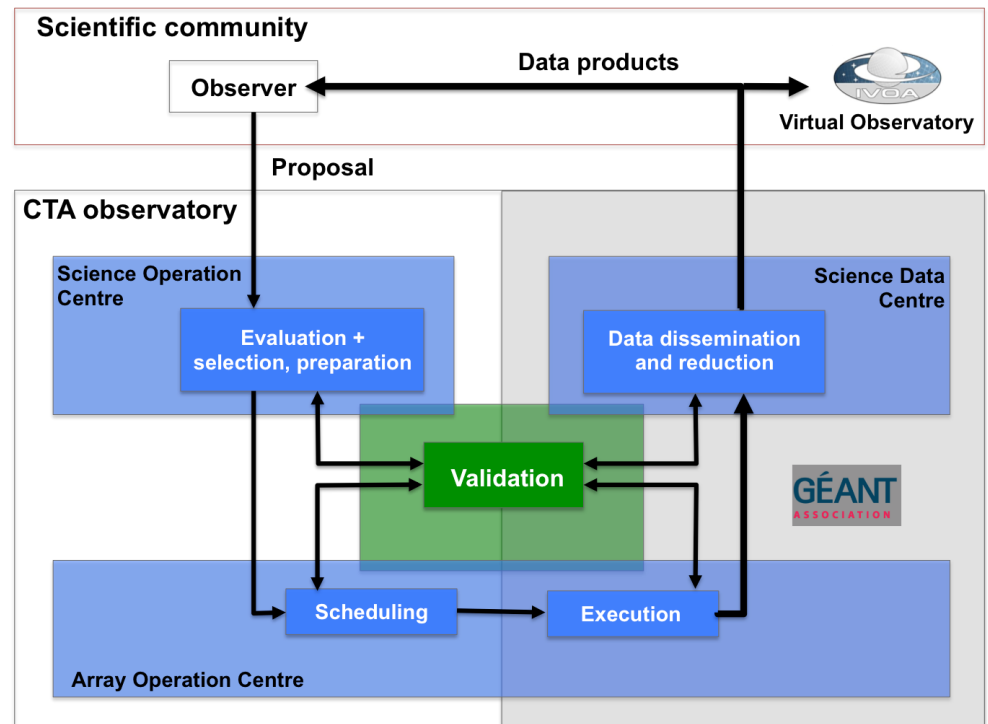
Project schedule

- Construction and deployment: 2017-2024
- Early science: 2018 or 2019
- Science full operations: start in 2022 for ~30 years

CTA (Cherenkov Telescope Array) Observatory



- Consortium
 - To build the instrument
 - Specific rights during the operations
- Observatory
 - Legal entity that will operate the instrument during 30 years
 - Announcement of Opportunity for observation proposal collection
 - Data will be public after predefined proprietary period
 - Virtual Observatory compatibility



Data acquisition (South site example)



Telescope	Data rate	Data rate (Central Trigger)
LST	110Gb/s	40Gb/s
MST	450Gb/s	150Gb/s
SST	60Gb/s	30Gb/s
Total	610Gb/s	220Gb/s

Central Trigger

Full waveform signal from photodetectors (Total 1314h):

307 PB/year

Pixel integration

3% full waveform signal, remaining signal integrated:

21 PB/year

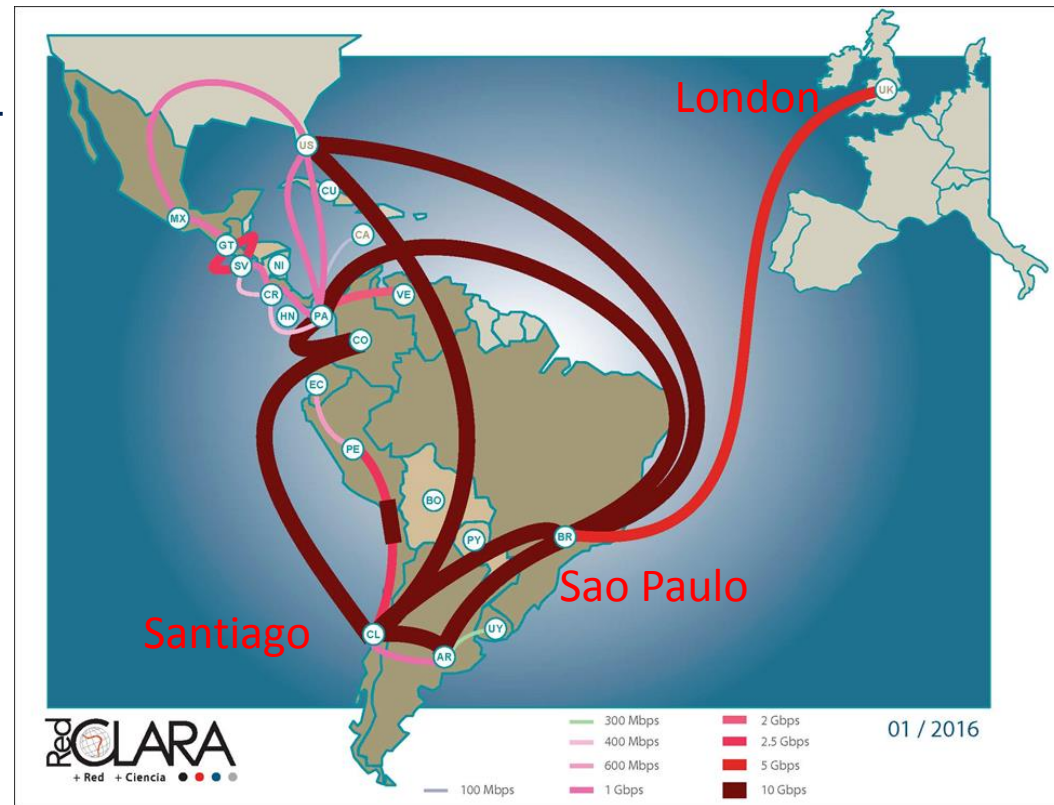
Telescope	Data rate (sampled pixels)	Data rate (Integrated)	Total
LST	2.2Gb/s	8.6Gb/s	11Gb/s
MST	4.5Gb/s	15.5Gb/s	20Gb/s
SST	1Gb/s	4.1Gb/s	5.1Gb/s
Total			36 Gb/s

- **Resulting data rate** (Including 20% of technical data, calibration...)
 - CTA south: 5.4 GB/s
 - CTA North: 3.2 GB/s
 - Average observation time per year limited to 15% of time => 1300 hours/year
 - Distribution over year is maximized in winter for South hemisphere and summer for North one (maximum 12 hours per day)
- ⇒ 40PB/year (Max 370TB/day)**

Connectivity to Paranal (Chile)



- Last-mile connection to Antofagasta (15€/meter – 130 km)
- Antofagasta to Santiago (One 10Gbps wavelength from REUNA (NREN) shared between REUNA, ESO, ALMA, RedClara)
- Santiago to Sao Paulo (Brazil) by RedCLARA
- To Europe (London): 5Gbps shared



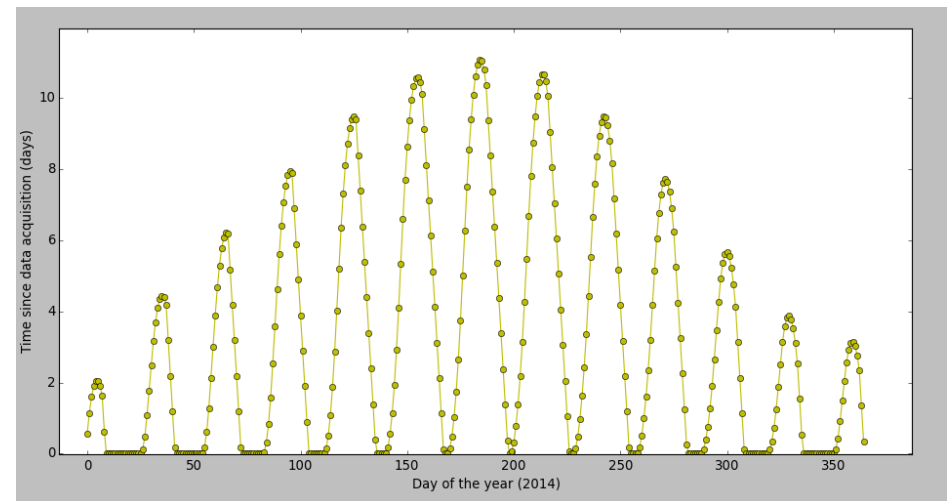
CTA data volume



- Constraints:
 - Limited bandwidth of 1Gb/s
 - Transfer data in less than 10 days

**=> Data volume reduction factor of 10 challenge:
Compression, Event selection,...**

- Resulting annual raw-data volume
 - **4PB/year**



Daily data transfer duration/ Day of the year

On-site Real Time Analysis (level-A)



- Alert/Data Quality management
- On data stream
- High-performance computing techniques
 - Speed up common algorithms (up to 900x!)
 - Reduce storage size with better compression
 - GPU/ARM computing



DATA MODEL

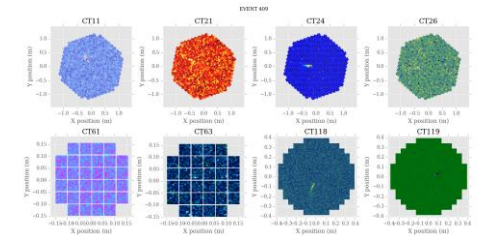
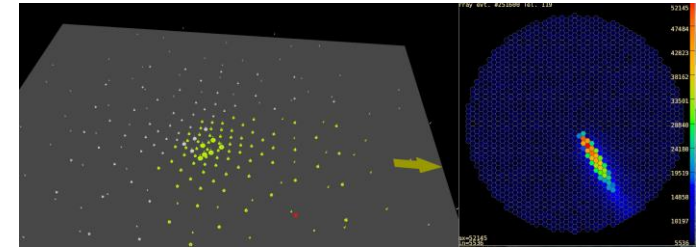


Data Level	Short Name	Description	Data reduction factor
Level 0	RAW	Data from DAQ written to disk.	1
Level 1	CALIBRATED	Physical quantities measured in the camera: photons, arrival times etc. (Preliminary image shape parameters could be also included within)	1
Level 2	RECONSTRUCTED	Reconstructed shower parameters such as energy, direction, and particle ID. Several increasingly sophisticated sub-levels are envisaged.	10^{-1}
Level 3	REDUCED	Sets of selected (e.g. gamma-candidate) events.	10^{-2}
Level 4	SCIENCE	High-level binned data products like spectra, skymaps, or lightcurves.	10^{-3}
Level 5	OBSERVATORY	Legacy observatory data, such as CTA survey sky maps or the CTA source catalog.	10^{-5} - 10^{-3}

CTA processing pipeline



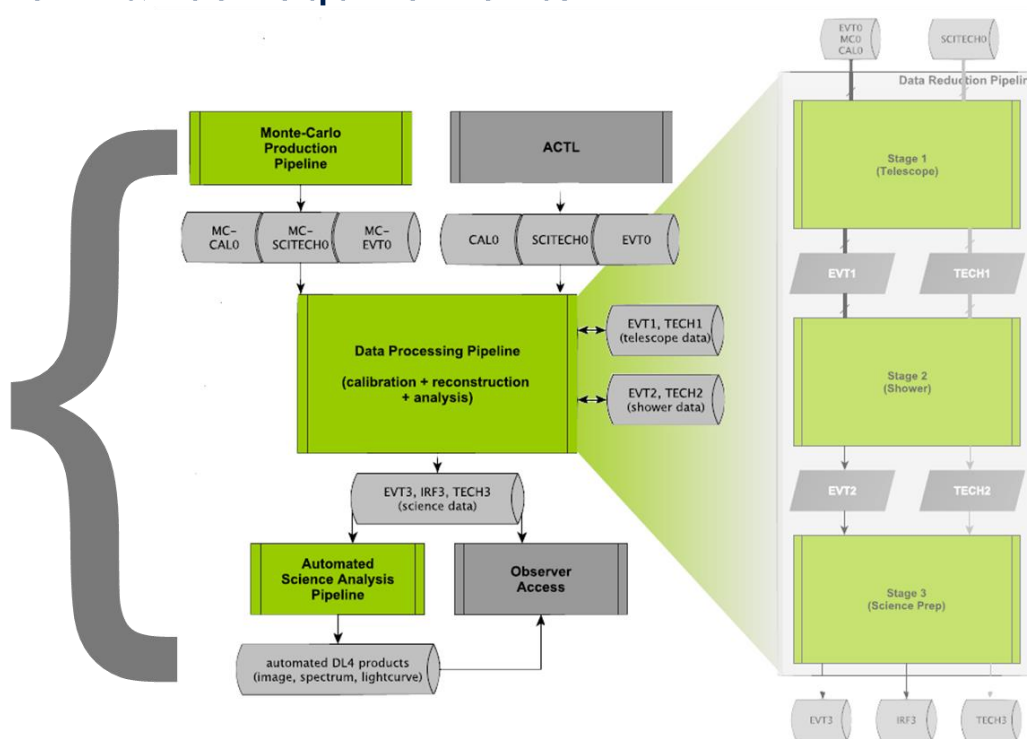
- Traditional batch system
- Will run on distributed computing environment
- Much tighter I/O and performance requirements



Level A: on-site
realtime

Level B: on-site
delayed

Level C: off-site
advanced

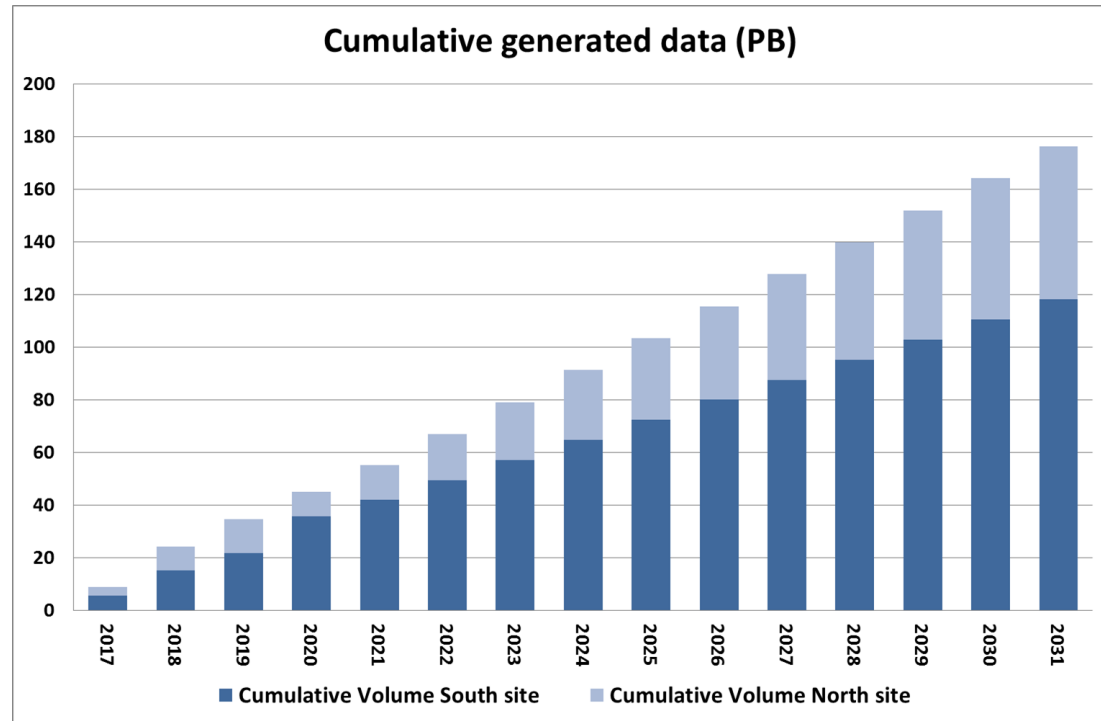


- Instrument Response Function to be able to reconstruct and analyze data
 - Standard Monte-Carlo analysis
 - Deployment period specific Monte-Carlo data
 - Dedicated to on-site analysis (Level A and Level B)
 - Per-period Monte-Carlo analysis
 - Based on technical data (Atmosphere, ...)
- Computing & storage resources to be estimated

CTA data volume



- Data to be archived for 30 years of operation + 10 years
 - One reprocessing per year (2 versions kept)
 - Resulting new data per year:
 - Raw data: 4PB/y
 - Processed data: 4PB/y
 - Monte-Carlo data: 20PB
- => 12 PB/year



Computing Model: Data preservation on disk/tapes



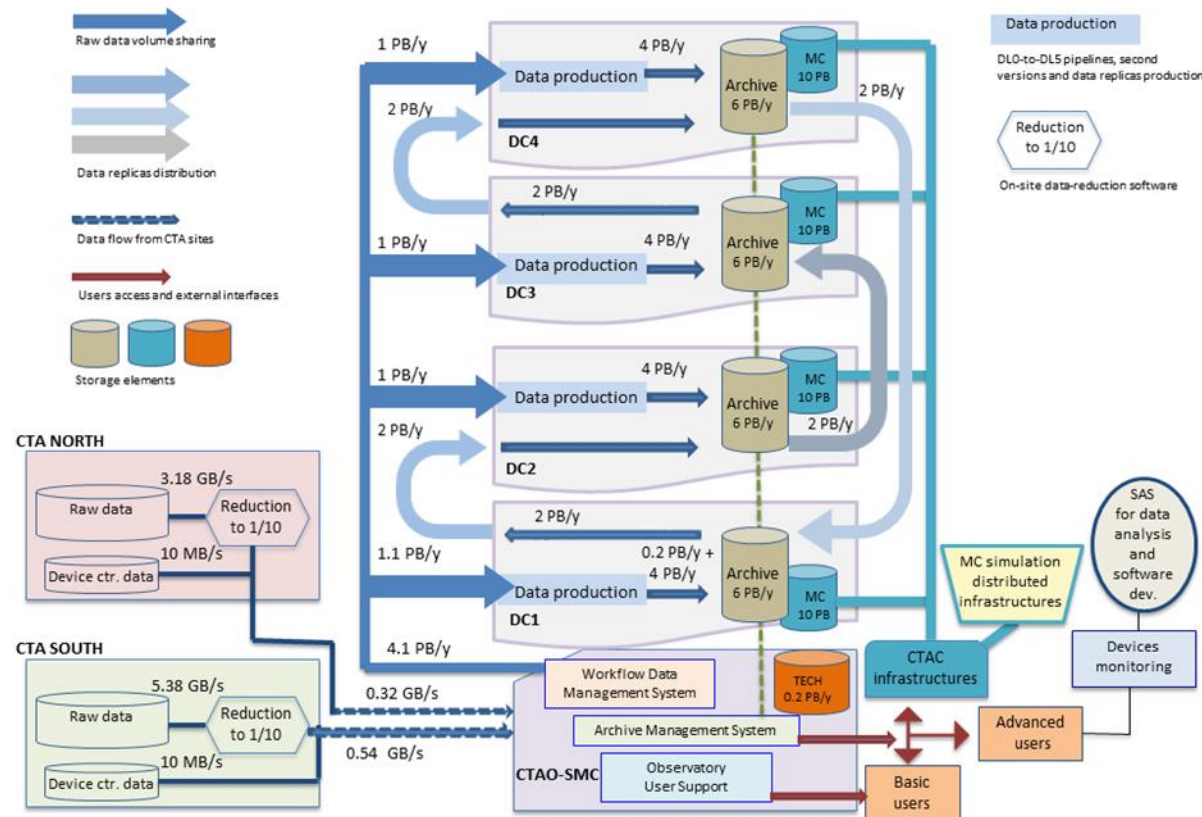
- Total required storage
 - **27 PB/year** (including reduced data and all replicas)
 - 6PB on Disk, 21PB on Tapes

Data occupying storage - Planned Model (Data)					
	Event Size	Data Access	Disk replicas of each version	Number of versions	Number of Tape Copy
Raw (DL0)	14000 to 41000 bytes depending on camera	Write once, low read rate	10% kept on disk (Tape cache is 10% of the global data volume)	1	1+1 (backup)
Calibrated (DL1)	20% to 100% of RAW prior to any data volume reduction		0	0	0
Reconstructed (DL2)	10% of RAW camera data prior to any data volume reduction	New version per year. Low read rate.	100% kept on disk for the last version, 0% for the old version	2	1+1 (backup for last version only)
Reduced (DL3)	1% of DL2	High read rate	1 (100%)	2	1+1 (backup)
Science (DL4)	0.1% of DL2	High read rate	1 (100%)	2	1+1 (backup)
Observatory (DL5)	0.1% of DL2	High read rate	1 (100%)	2	1+1 (backup)
MC Data	100% of Observation data (Min 5PB, Max 20PB)	Read/Write	100% during commissioning phase (3 years), 1PB afterwards	1	1+1 (backup)

CTA Proposed Distributed Computing Model



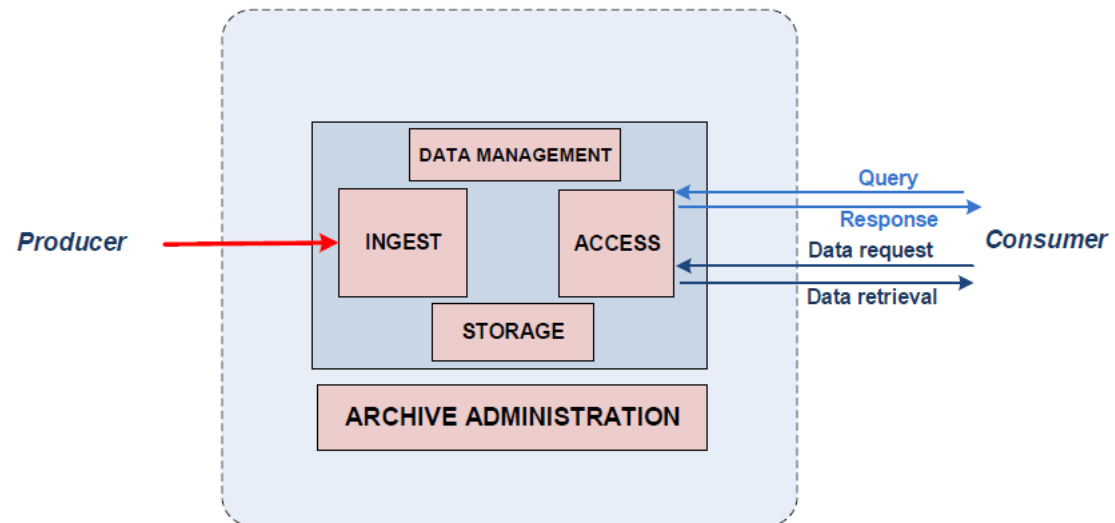
- Distributed model over a few centers, e.g. 4 data centers
 - Trade-off between economy of scale and sustainability
 - Flexibility
 - The underlying technology is not decided yet
- One Science Data Management Centre
 - Archive and Workload management
 - Observatory User Support
 - Network Service Providers management



Related model for Archive & Challenges



- « Open Archival Information System » (OAIS) ISO standard
 - Distributed Archive/File Catalog/Metadata
 - On-site subset/off-site?
 - Able to manage replicas, tapes, disks
 - From low-level to high level




CTA Proposed Computing Model & Challenges



- Workflow management of jobs close to data
- But how to distribute data between sites?
 - Per day, Sub-Array, Proposal,
- Tape management
 - To minimize tape access delays,
 - To be able to easily remove/rewrite processed data,
- How to minimize RAM memory requirements for data reconstruction jobs?

Observatory Data Access



- Open data formats and data discovery 
- Archive must ensure data access in line with the data access rights policy
 - Proprietary period
 - Complex calculation of start time –“For segmented, long duration of observations and surveys the proprietary period for each target/field begins....”
 - Default duration could be extended on request by the CTA Observatory Director General
 - Guest observers access rights
 - Principal Investigator (PI) of an approved observation proposal is able to share his access rights with co-PIs
 - From raw data (on request) to High-Level data (List of events)

- Different categories of Users:
 - CTA Observatory users
 - Two Array sites users (operators, experts, etc....)
 - Science Management Center users
 - HeadQuarter users
 - CTA Consortium users:
 - Observers (Guest Observers, Archive users)
 - Software developers,...
 - Academic users (Observers)
 - Anyone else (Observers)
 - Connected through Observatory Science Gateway (Single-Sign On) or directly to a CTA Observatory application
 - For data access and applications access/rights management
- => A centralized/coherent A&A system**

Central A&A system challenges



- Authentication
 - How to link several authentication methods to point to a single account for a user ?
 - How to manage a user that moved to another institution ?
 - How to deal with users that are outside of eduGAIN identity federation ?
 - How to deal with observatory users on sites that could be disconnected from central system ?
- Authorization
 - How to guarantee data access rights during the proprietary period and manage this complex time limit?
 - How to deal with a group of users associated to the same data access rights defined by the PI himself?
 - How to manage change of default proprietary period duration by CTA Observatory Director General?

Current status: pre-construction phase



- We are still currently in a pre-construction phase
- Developments in progress and some prototypes under evaluation for:

- Archive (Cloud)



- Pipelines (Grid, Cloud)

- Simulation & Analysis Use Case for HNSciCloud
- Simulation and Analysis with Workload management Sytem



- High-level Analysis (science tools)

- Authentication & Authorization



Current computing model for Monte-Carlo simulations



- Use EGI grid resources (CTA Virtual Organization)
 - ~ 20 sites in Europe
 - 6 main sites provide in total 2.8 PB
- Monte-Carlo production jobs run at all sites
- Output data are stored at 6 Storage Elements
- Monte-Carlo analysis jobs run at specific sites
- Users analysis also running in parallel

Grid sites supporting CTA Virtual Organization



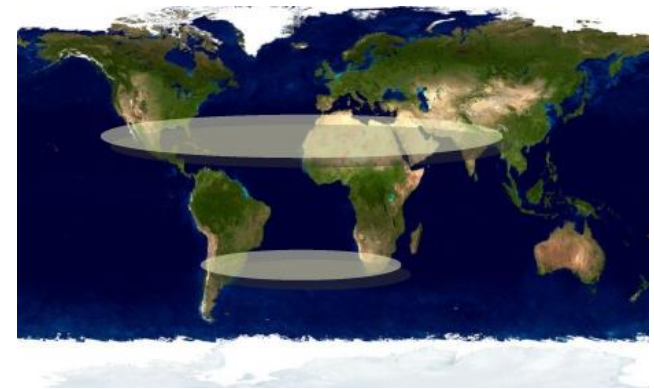
Monte Carlo computing needs for CTA preparation



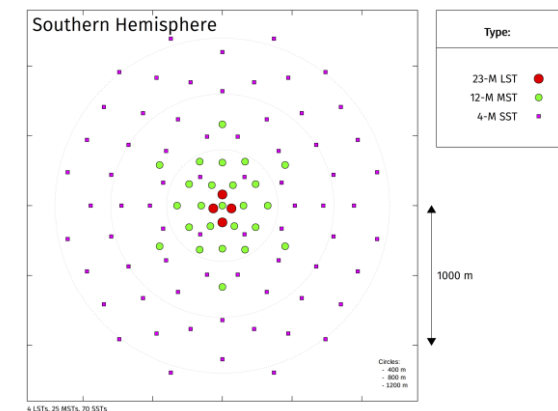
Two main MC campaigns during past 4 years

- Site selection
 - Characterize 8 site candidates to host CTA telescopes
 - 5 B events simulated for each candidate
- Telescope array layout optimization
 - Find the optimal layout for a given number of telescopes
 - Many telescope positions, with alternative telescope/cameras
 - 3 different analysis chains run in parallel on the simulated data

Site candidates

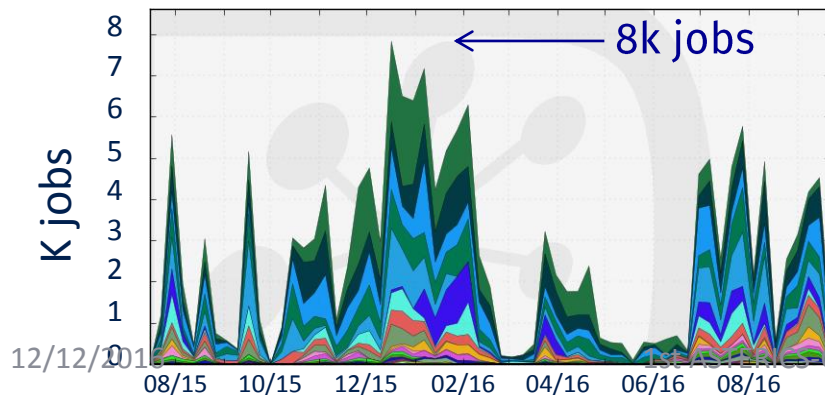


Baseline telescope layout for CTA South

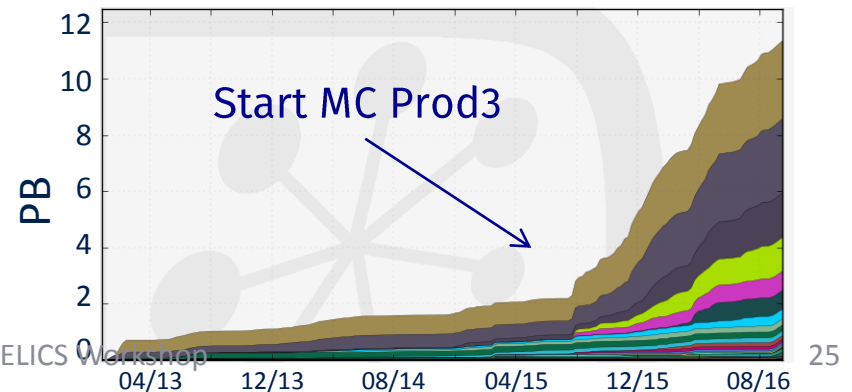


- Workload Management System prototype based on DIRAC
 - 5 DIRAC servers and a CTADIRAC software extension
- Already in use since 4 years to handle massive MC simulations on the grid
 - 360 M HS06 CPU hours (10% for users analysis)
 - 11 PB transferred data (2 PB currently on disk/tape)
 - 25 M files registered in the DIRAC File catalog

Running jobs by site
during last MC campaign



Transferred data by
destination



Challenges conclusion



- Be able to reduce data flow from 600 Gb/s to 4 Gb/s with limited loss of information:
 - Pixel integration
 - Event selection (Algorithms optimization, GPU/ARM,...)
 - Data compression
- Be able to transfer, process, archive and retrieve 4PB/year (37TB/day) of raw data at reasonable cost for the coming 30+10 years
 - File Transfer, Workload management systems, Databases, High-Throughput, File system, Tape, Disk, redundancy, Grid/Cloud,
 - Algorithms: RAM usage, HPC,
- Data discovery
 - Open Data formats in connection with Virtual Observatory
- Be able to guarantee data access rights and proprietary period to Principal Investigators and his/her team
 - Central A&A system & Archive/storage data access management

Acknowledgement

- H2020-Astronomy ESFRI and Research Infrastructure Cluster (Grant Agreement number: 653477).