cherenkov telescope array

CTA report

Luisa Arrabito, Johan Bregeon

LUPM CNRS-IN2P3, France

on behalf of CTAC and CTAO

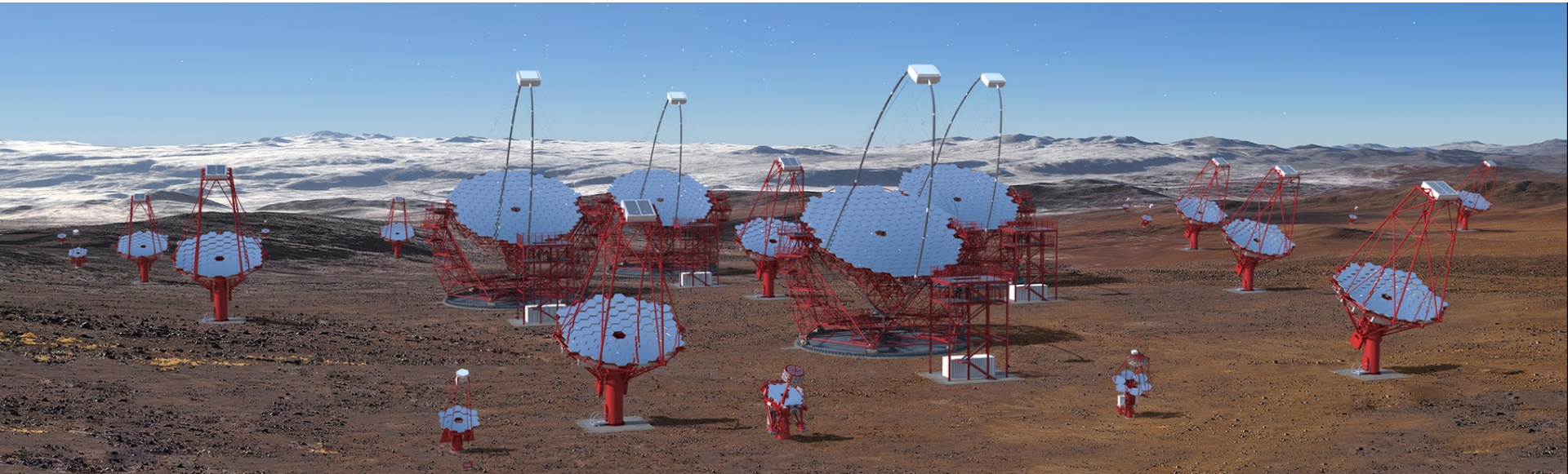
9th DIRAC User Workshop 14th – 17th May 2019, London

Outline

- CTA project
- Current DIRAC usage in CTA
  - Hardware setup
  - DIRAC functionalities in use
  - DIRAC systems extended
  - Externals, new DIRAC extensions, new DIRAC systems
  - DIRAC usage over the past year
- Computing model for future CTA operations
- Conclusions and plans

CTA (Cherenkov Telescope Array)

- Next generation IACT, VHE gamma-rays Observatory
- Worldwide collaboration, 1500 members
- Scientific goals
  - Cosmic ray origins, High Energy astrophysical phenomena, fundamental physics and cosmology
- Two Cherenkov telescope arrays
  - Northern Site (La Palma, Spain): 4 large size, 15 mid-size telescopes
  - Southern site (Paranal, Chile): 4 large size, 25 mid-size, 70 small size telescopes
- Project schedule
  - Construction and deployment: 2020-2025
  - Science operations: from 2022, for ~30 years



CTA and DIRAC

- CTA has data management challanges with large-scale data processing and simulation needs
- Distributed data processing under investigation
- DIRAC is a candidate for the CTA Computing Resource and Workflow Management (CRWMS) in a distributed infrastructure (i.e. DIRAC WMS, TS and PS)
- Integration of DIRAC Data Management capabilities (i.e. DMS) for CTA Bulk Archive under consideration
- DIRAC currently used (since 2011) for large-scale MC simulations in CTA preparatory phase (see next slides)

Current computing model for MC simulations (preparatory phase)

- Use EGI grid resources (CTA Virtual Organization)
  - ~ 15 sites in Europe
  - 6 sites provide in total ~ 6 PB
- MC production jobs run at all sites
  - Output data are stored at 6 SEs (1 distributed replica)
- MC analysis jobs run at sites with good connectivity to SEs
- Users jobs also running in parallel

Grid sites supporting CTA Virtual Organization

Current CTA-DIRAC hardware setup

- Dedicated DIRAC instance 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)
  - 1 hosting all other DBs at PIC
    - Almost lost the whole DB for a crash disk? It took 2 weeks to rebuild
    - Discussing SLA with PIC about DIRAC servers hosting
- 1 server for the Web portal (at CC-IN2P3)
- Installed DIRAC version v6r19p20 –> To be updated soon

DIRAC functionalities in use

- Accounting
- Data Management (DMS)
  - Used for simple operations so far (upload/downalod by jobs and users clients)
  - Used for simple bulk operations (dataset replication to a given site, removal of old datasets) through the TS
  - Limited usage of Tape backend (no pre-staging campaigns)
  - No customized TS plugins for advanced DataManagement yet
  - No FTS usage
- 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)
  - Will be used for MC Simulation, data-processing
- VMDIRAC
  - Scalability tests with Clouds
- WebApp
- Workload Management (WMS)
  - Targeted resources: CREAM CE, ARC CE, HT-Condor CE, PBS cluster

DIRAC systems extended (CTADIRAC)

- Mainly 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
  - Specific development of a job interface for the official CTA software calibration and reconstruction pipeline (good progress made)
- Utilities
  - SoftwareManager -> source software environment mostly from CVMFS
  - Prod3DataManager -> used by production jobs to upload and register their output data and to set user-defined meta-data
- Agent reporting SE usage
  - Querying BDII → To be replaced by RSS
- See our code at https://github.com/cta-observatory/CTADIRAC

New DIRAC systems and externals

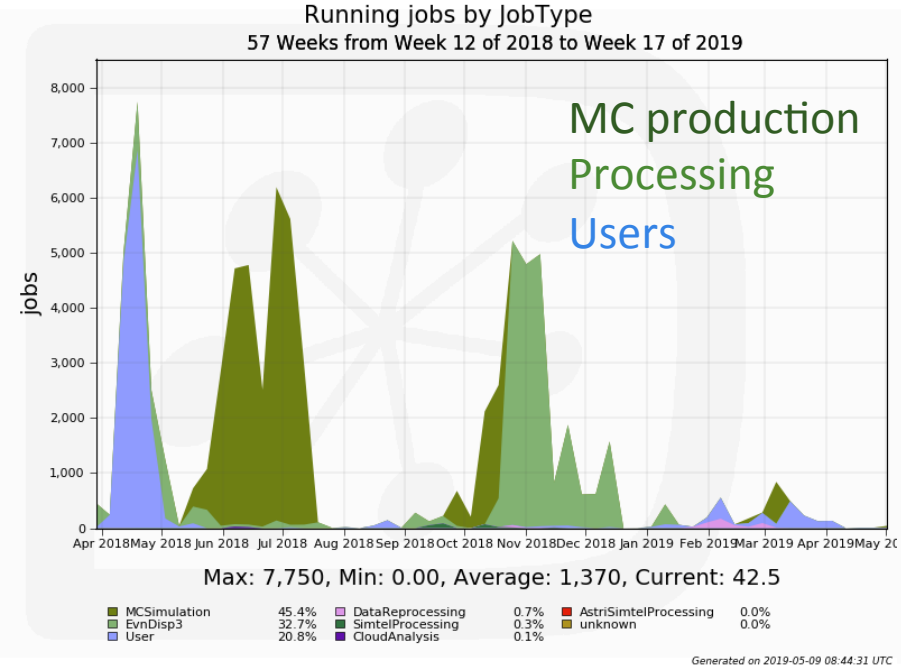
- New DIRAC systems
  - Prototype of **Production System** to easily chain several transformations and handle them in a coherent way (see Monday session)
- Externals
  - CVMFS CTA repository
    - 1 Stratum 0 (at CC-IN2P3) and 2 Stratum 1 (at CC-IN2P3, DESY)
    - Almost all CTA sites configured to access the CTA repository
  - In future: CTA A&A system
    - Supporting single-sign on
  - In future: CTA Bulk Archive System
    - Different solutions being explored to implement the CTA Bulk Archive System, including DIRAC DMS

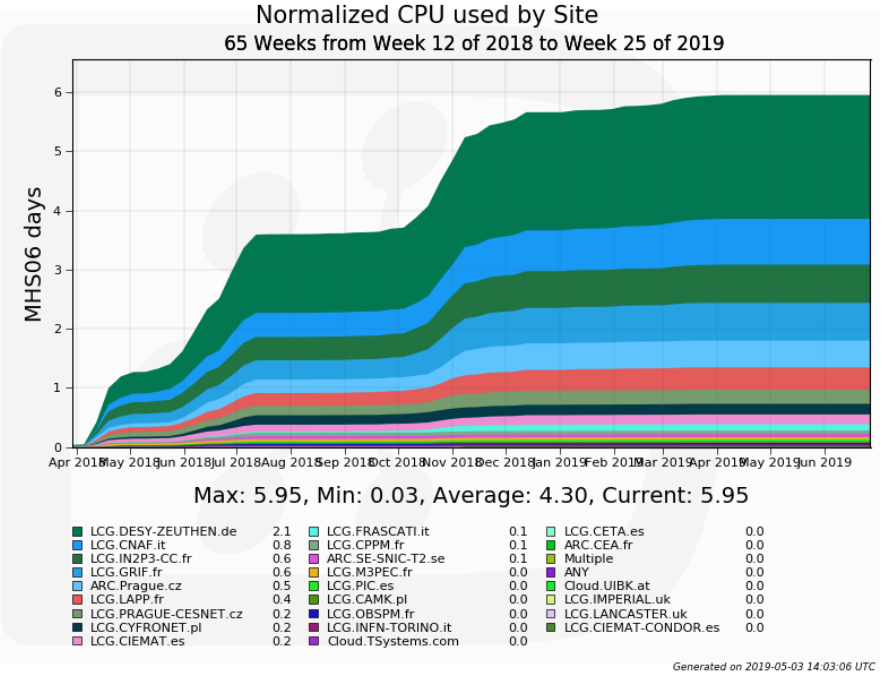
DIRAC usage since last year

- Extended simulations with the final array layouts for both North and South sites
  - Simulations with High Night Sky background (for moonlight observations)
- Special set of simulations with different configurations and models for the small telescopes (3 telescope structures and 4 cameras)
  - Compare performances of different prototypes
  - Support CTA Observatory in the harmonization process (for the small-size telescopes)
- Data processing
  - Performed with the current analysis chain (*EventDisplay,* ROOT based)
  - Developments for the *ctapipe* python framework chain
- Still heavily relying directly on the Transformation System
  - Using meta-filters through datasets to chain production and analysis
  - No real-life test of the Production System yet

DIRAC usage since last year

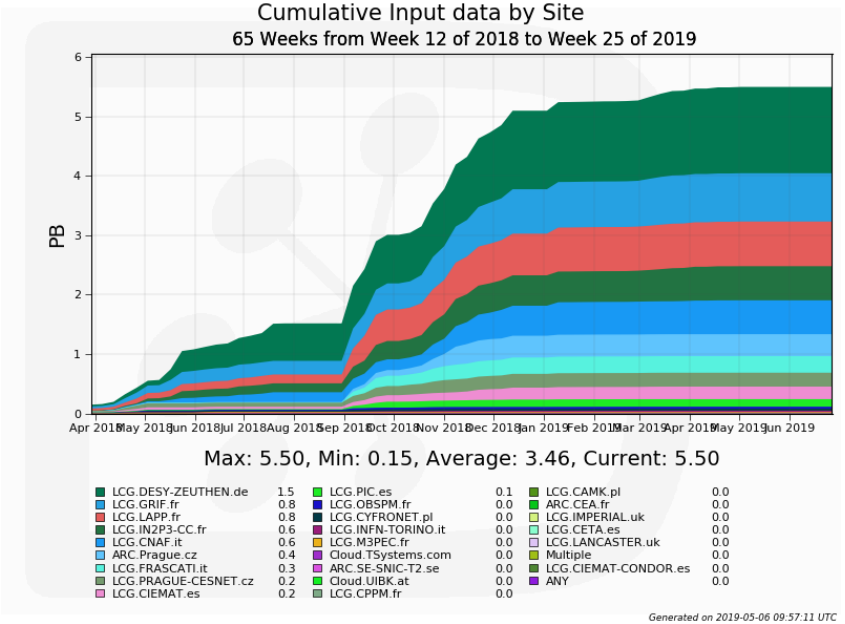
- MC production and analysis running in parallel
- 144 M HS06 hours
- 1.9 M executed jobs

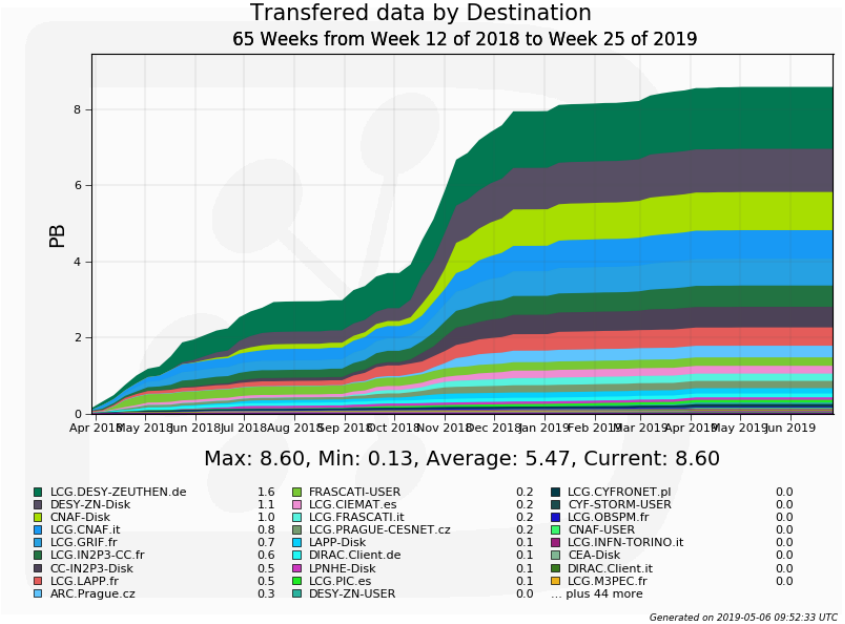




DIRAC usage since last year

- 8.6 PB of transfered data
- 5.5 PB of processed data
- Total: 4.3 PB currently on disk/tape
- Total: 30.5 M replicas in DFC





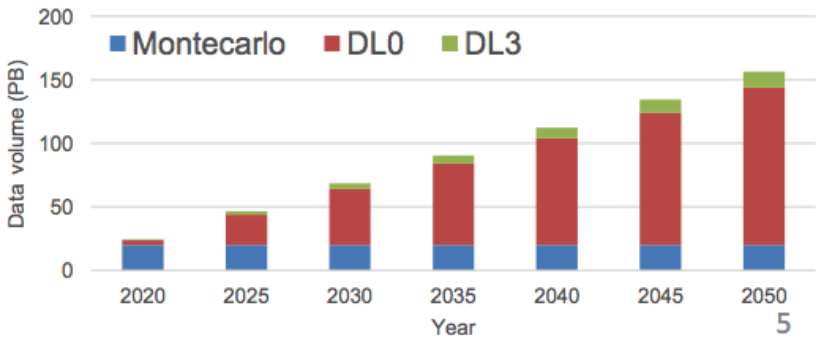
From Preparatory to Operation phase

- Current computing model for MC simulations (Preparatory phase)
  - Distributed model with ~ 20 grid sites for processing
  - Data distributed at 6 grid sites (only 1 distributed replica)
  - No automated data replication, removal, pre-staging campaigns etc.
  - Relatively simple meta-data model implemented (using DFC)
  - Job failures are tolerated -> Not a big deal for MC jobs
- In-Operation Computing Model (next slides)
  - Designed to cope with availability, performance and cost constraints (cf. TDR v2.0 March 2016)
  - More strict requirements on data-processing and archive (fast data processing and re-processing, long-term data preservation, security, etc.)

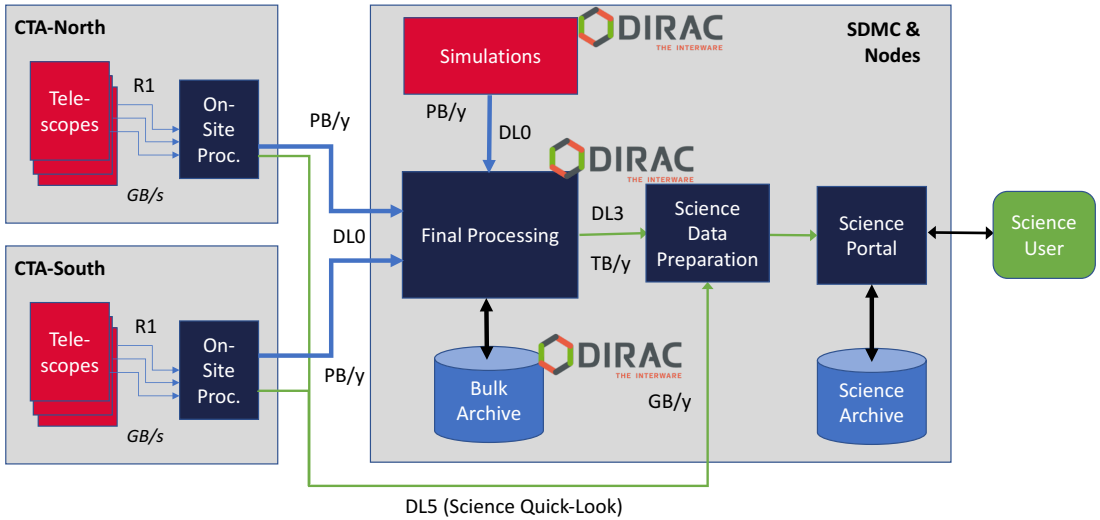
Data Levels and Storage Requirements

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}

- Strong data reduction along the processing steps
  - From PB/y (DL0) to GB/y (high-level science data, DL3-DL5)
- Total storage requirements
  - +6 PB/y on Disk
  - +21 PB/y on Tape



Data Flow Overview



- On-site computing
  - Near real-time processing
  - Next-day data processing for quick-look and science alerts
  - On-site buffer and data transfer

→ DIRAC for on-site processing + data transfer?

- Off-site computing
  - Simulations and final processing (DL0 -> DL3)
  - Bulk Archive (DL0-DL3)
  - Science data preparation (DL3 -> DL5)
  - Science Archive (DL3-DL5)
  - Open access through Science Portal

→ Main potential DIRAC usage

→ Eventual DIRAC usage?

On-site Computing and intercontinental link

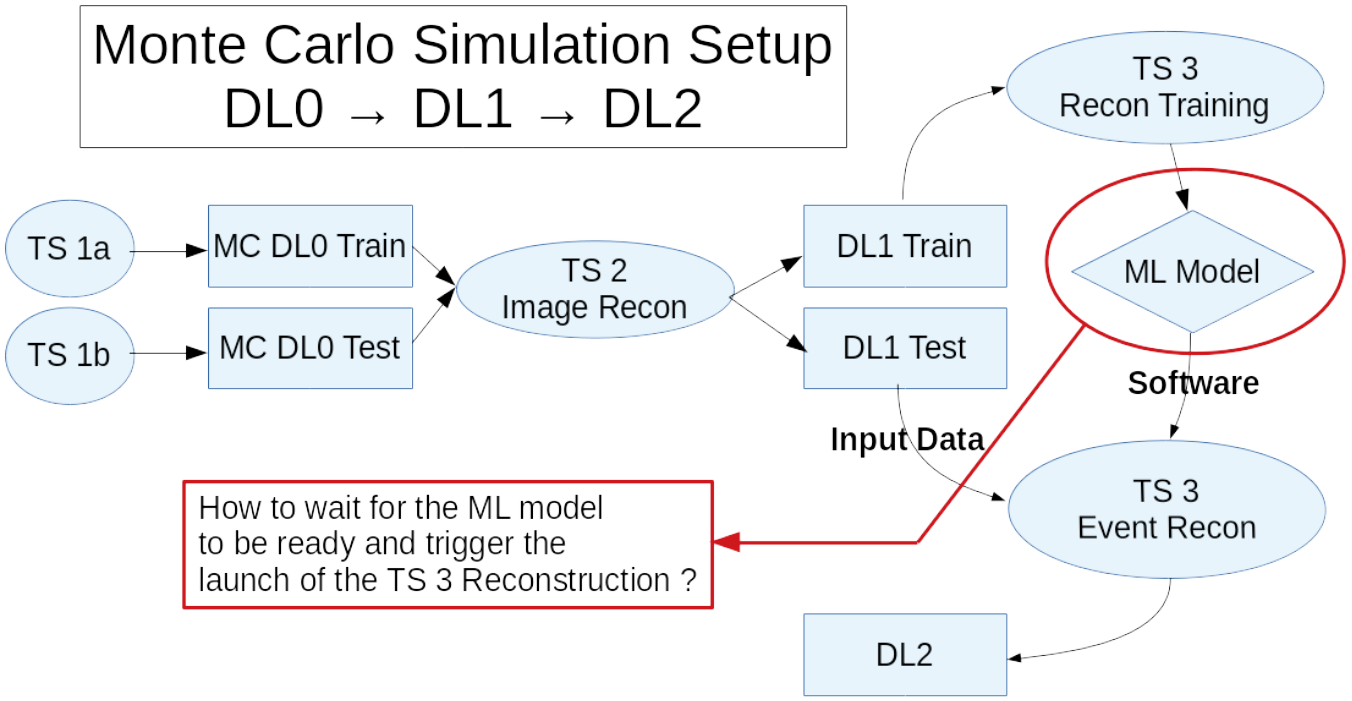
- « Relatively » small capacity data centers planned on-site (Chile, La Palma – Spain)
- Network link quality/stability obviously not as good as between grid Tier-1
  - Requirement and current baseline is 1 Gbps for both sites, higher bandwidth under investigation
- On-site processing
  - Only small-scale processing (high availability required)
  - DIRAC as CRWMS is an option, but simple batch processing could do the job too
- On-site buffering and transfer to Europe
  - No archiving on-site, only buffering of data for limited period
  - 'Ingest in archive (file registry+metadata)' may be used for data transfer from on-site to off-site
- Requirement to have a unique workload/workflow management interface or on-site and off-site processing
- Questions
  - Which is the best DIRAC installation to guarantee efficient on-site processing?
  - Single DIRAC instance in Europe with some services eventually duplicated on-site?
  - Need a dedicated separate DIRAC instance on-site?
  - Alternatively, special handling of on-site processing and then only transfer data to Europe
    - Can we use DIRAC as an interface from on-site buffer to off-site archive (i.e. registry of file, metadata, transfer to Europe)?
    - How reliable is the data transfer mechanism used by DIRAC (*e.g.* limited bandwidth, instability of link, long transfer times)?

Off-site Computing

- Large-scale processing (DL0->DL3)
  - Baseline is to use a distributed e-infrastructure with 4 Data Centers (Tier1-like)
  - Main potential customer for DIRAC
- Requirements
  - Raw data to be processed within 1 month
  - 1 full re-processing per year
  - Average CPU: 1200 to 9000 cores/year after 15 years of operation (constant growth)
  - CPU peak value: 1200 to 9000 cores/year after 15 years of operations (more rapid growth)
- CTA aims for fully automatic processing, including automatic recovery model
  - Are there optimized or best practices workflows that we (the scientific community / large-scale processing facilities) should follow? Is this documented somewhere?

Example of CTA workflow (Split/Merge & Train/Test)

- How do we fully automatize this kind of workflow?
  - Problem is that the look-up table/ML model/BDT are « part » of the software, how do we know that these are ready and that the next step can be run?



CTA Bulk Archive

- Bulk Archive
  - No public access, only CTAO staff
  - Archive bulk raw-data (DL0-DL3) and associated metadata information over the lifetime of CTA
  - Preserve DL0 simulation data for at least 3 years after production
  - Located at least 2 sites with 300 km distance
  - Handle increasing data volume of ~10 PB/yr
  - Allow fast (re-)processing of data (annual reprocessing of all data within 1 month)
  - No data loss over the full lifetime of CTA
- Option 1
  - An Archive that interfaces to DIRAC WMS, TS (as CRWMS)
- Option 2
  - Archive is built on DIRAC DMS

CTA Bulk Archive requirements 1/2

- Bulk Archive will follow OAIS standard (or reference model or reference architecture)
- OAIS is an ISO standard, defined by Committee for Space Data Systems

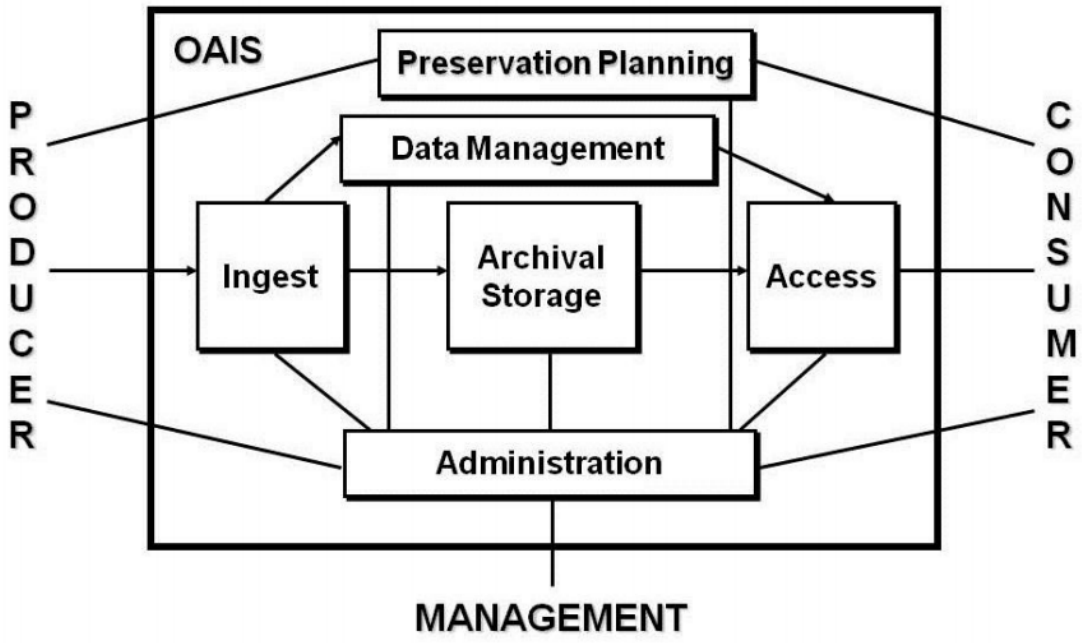


Diagram taken from the OAIS Reference Model by the Digital Preservation Coalition

CTA Bulk Archive requirements 2/2

- Storage-related requirements
  - Management, operation of and access to Bulk Archive only by CTA Observatory staff (no public access)
  - Access rights management supporting specific roles for CTA Observatory staff (Archive Manager, Data Processing Manager)
  - Unique identifiers for all data products that are independent of the storage location or number of copies
  - Versioning of data products
  - Placement, Replacement, Duplication, Migration of data products and metadata
  - Multiple locations and different storage media
  - Bulk Archive validation and preservation of archive organization
- Ingest/Access-related requirements
  - Passive and active ingest
  - Metadata extraction and browsing (metadata could be in a separate DB for faster queries)
  - Update of metadata and regeneration of metadata from data products
  - Confirmation of the availability of requested data products and estimation of the retrieval time within 1s of the search request on average

Conclusions

- CTA will start with Early Science operations in 2022, and will ramp up to produce several PB/year for about 30 years. During that time Bulk Data archive will be supported

- CTA investigates a distributed computing model
  - Baseline with 4 Data Centers (not necesserly using grid middleware)
  - DIRAC is proposed for the CRWMS (i.e. WMS, TS and the Production System)
  - DIRAC DMS to be evaluated for the CTA Bulk Archive

- DIRAC is currently used for MC production and analysis
  - Using all the main DIRAC functionalities (WMS, DMS, DFC, TS, RMS)

- Many improvements in CTA-DIRAC since last workshop
  - TS meta-filters in production with datasets
  - Prototype of Production System available in v7r0
  - Job API extension rewrite

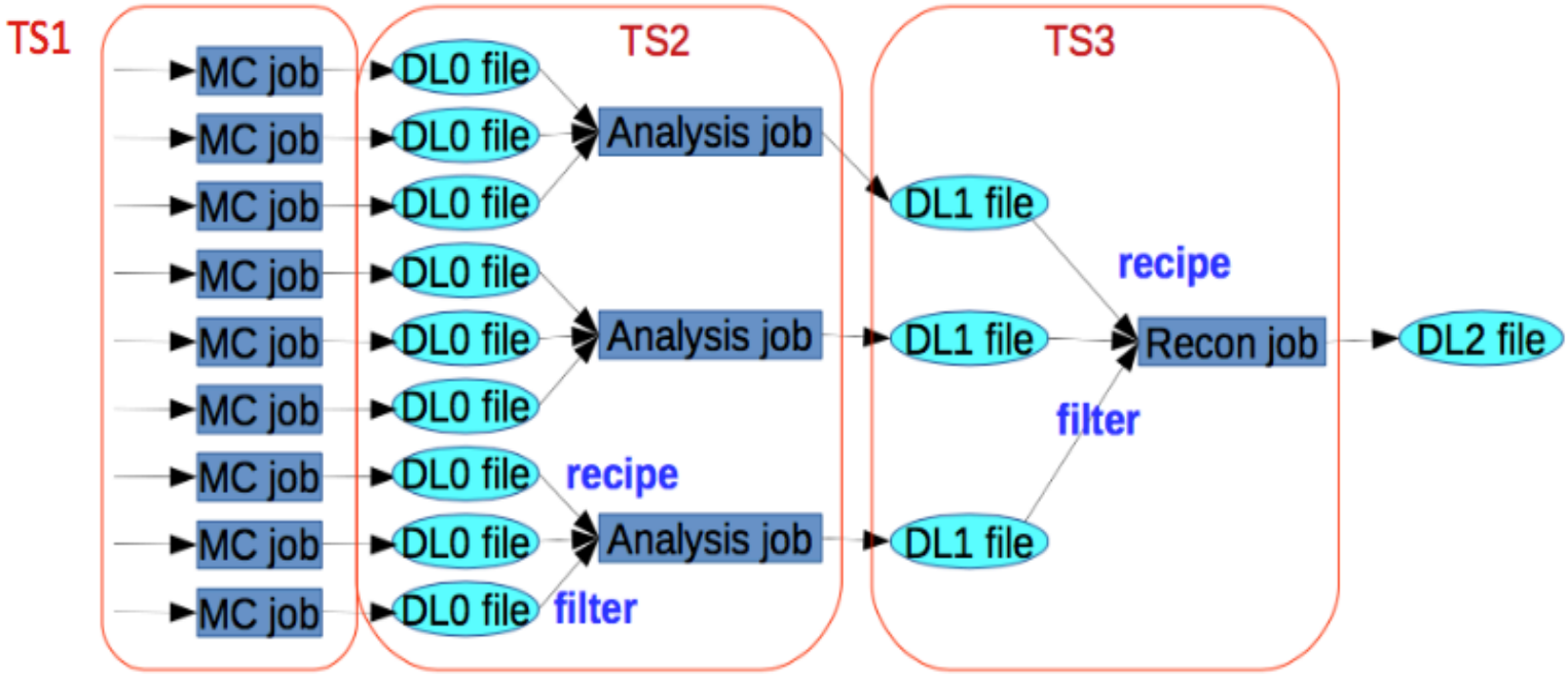
Plans

- Further develop the Production System on the top of the Transformation System
  - Based on users feedback
  - Aim to fully automatize complex CTA workflows
- Refine the main Job API to wrap/interface the CTA pipeline framework
- Resource integration as the need arises, *e.g.* HPC, GPU
- Fully evaluate DIRAC DMS for CTA Bulk Archive
- Work on interfaces with external systems such as A&A
- Go toward a high-available CTA-DIRAC installation in view of CTA operations
  - Enhance service redundancy, DB backup and replication
  - Establish operation procedures and documentation

Backup

Workflow example

– CTA MC Production workflow (simplified)



Production example

- CTA recent production (oct-dec 2018)
  - Goal: evaluate performances of different camera+telescope configurations (for small telescopes only)
  - Total jobs = 563 000
  - Total disk = 592 TB distributed in 3 SEs
  - 1.3 M of replicas in 64 ‘datasets’

- Workflow
  - Air shower simulation
    -> 360 TB of ‘corsika’ data
  - Telescope simulation processing corsika data for 5 different telescope+camera configurations
    -> 230 TB of ‘simtel’ data
  - Processing of ‘simtel’ data for event reconstruction
    -> 0.6 TB
  - **Realized with 68 transformations**