Production System report

Luisa Arrabito

LUPM CNRS/IN2P3, France

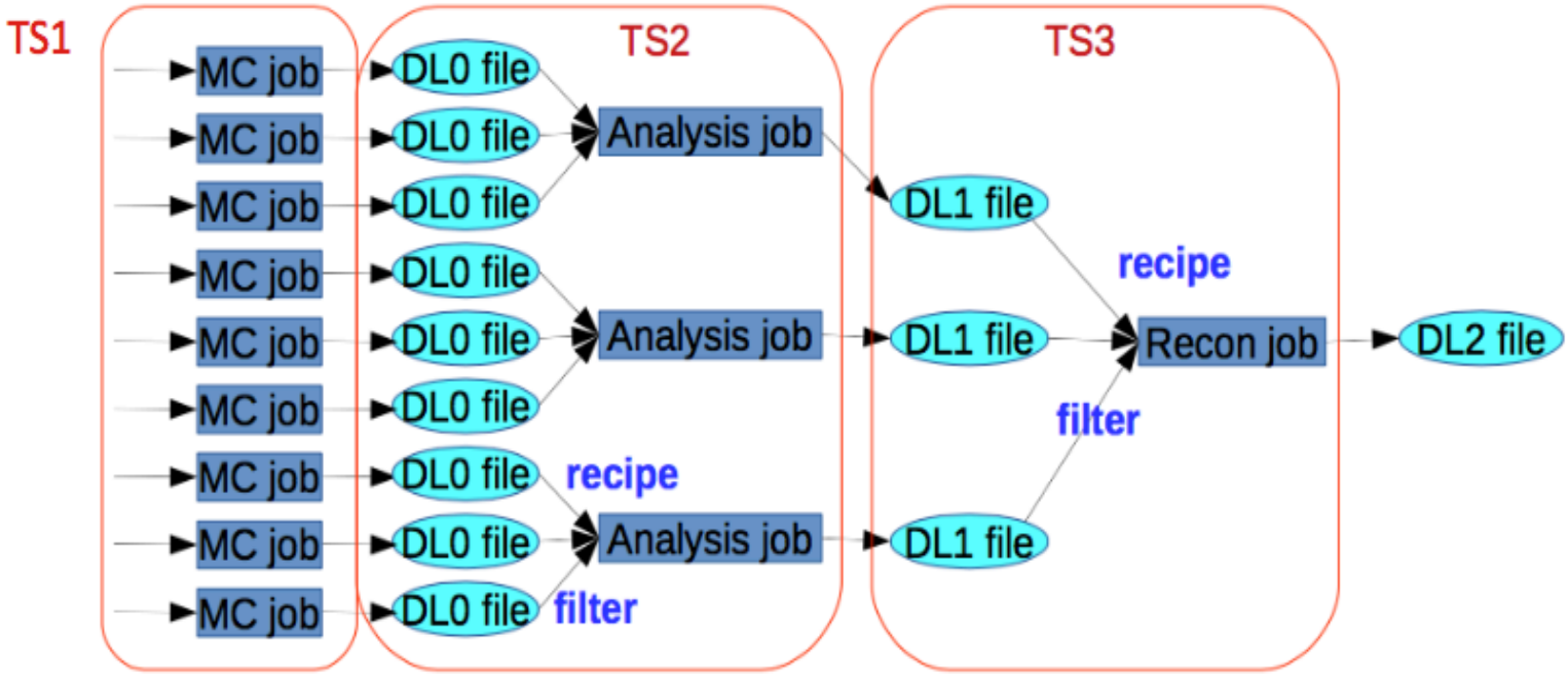
9th DIRAC User Workshop 14th – 17th May 2019, London

Workflows with TS

- Transformations process input data and produce output data which in turn can be specified as input data for yet another Transformation
  - Allow to define chains and graphs of Transformations of arbitrary complexity
  - Transformations creating computing tasks and data management requests can be grouped together in a single workflow

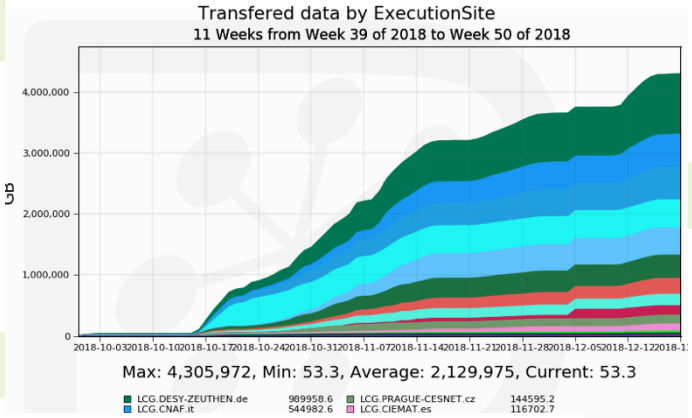
Workflow: CTA example

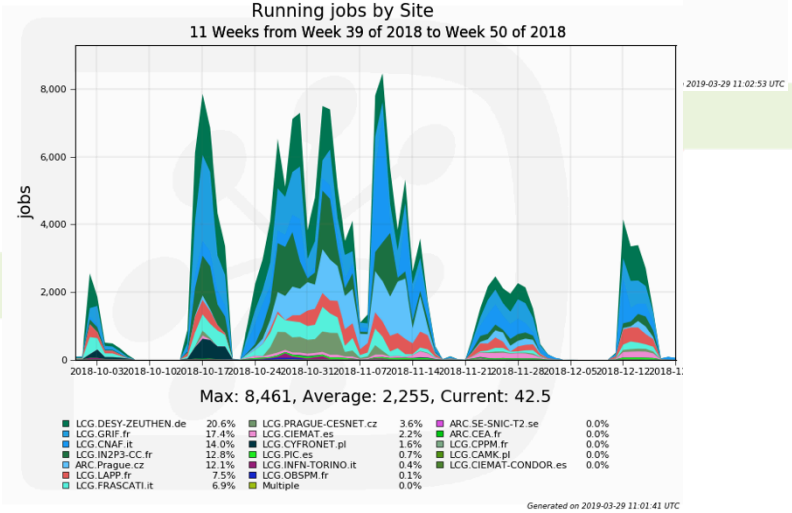
- CTA MC Production workflow (simplified)



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**





Motivations for the Production System

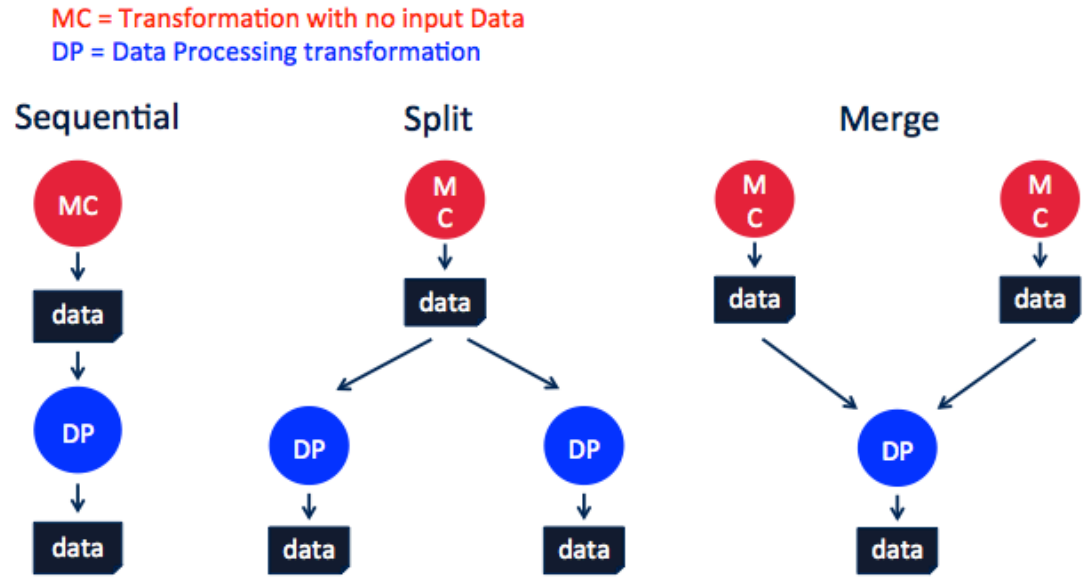
- TS automatizes a single step of workflow execution
  - Need to monitor tens of transformations at once
  - Manually defining each transformation (job description, input data filter, ...) is error prone
- A higher level System is needed to automatize the execution of full workflows
- LHCb, ILC, Belle II developed specific Production Systems on the top of the TS
  - Found many commonalities
  - A common general Production System (PS) can benefit to several communities

Production System

- Enhancement of the transformation definition to characterize its inputs and outputs through meta-queries
- A production is a set of transformations with their associations (‘links’)
- It is specified through a description consisting of several ‘production steps’
- Each production step corresponds to a transformation with the eventual specification of a ‘linked’ transformation
- Two transformations are linked if their Input and Output meta-queries intersect

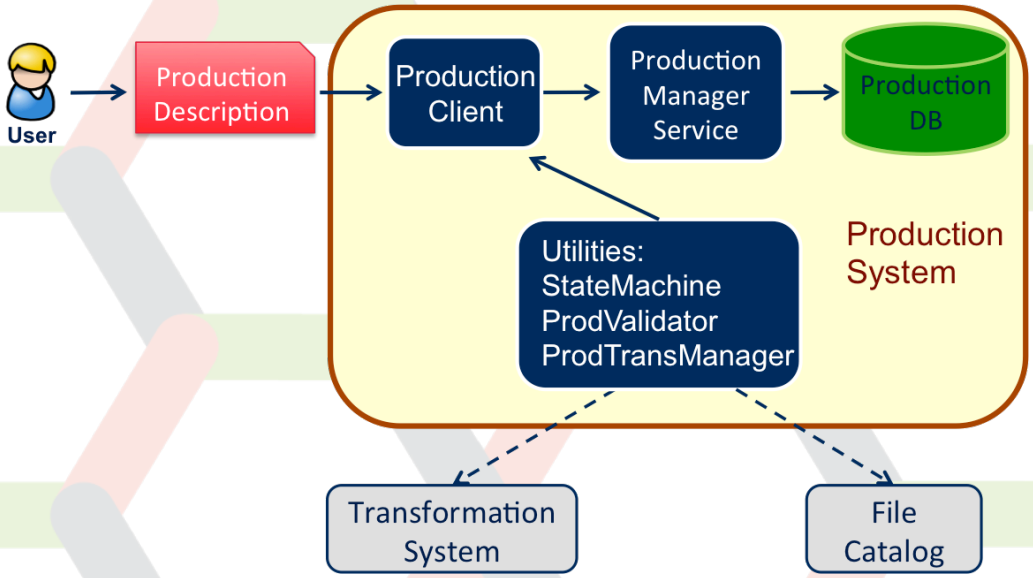
Production System

- Available in the DIRAC v7r0 release
- Automatic transformation instantiation based on the production definition
- Fully data-driven
- Tested for simple workflow schemes



PS Architecture

- User provides a production description
  - All the transformations of the workflows and their meta-data queries
- The ProdValidator utility checks the description validity
  - Veryfing links between transformations
- If valid, the production is stored into the DB
- The user activates the production
  - the ProdTransManager utility creates the associated transformations



Conclusions

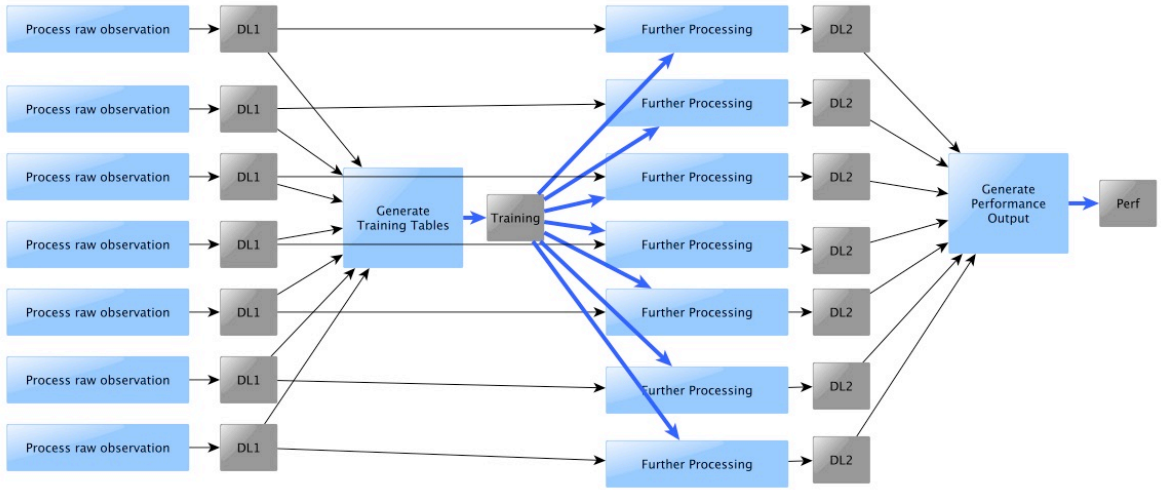
- The first version of the Production System available in v7r0
- Documentation
  - https://github.com/DIRACGrid/DIRAC/wiki/Production-System
- Not yet tested in real life productions
  - Ready to be tried out by the different communities
  - Will be tested in future CTA productions
- Currently only API and CLI interfaces are available
  - In future a dedicated web monitor has to be developed
- The ‘link’ logic to associate transformations is very simple
  - To be improved based on the usage experience
- Further improvements will come after users experience

DIRAC
THE INTERWARE

Backup

Workflow example

- Workflow example from CTA



- First process data per “observation” (or even per telescope)
- Then merge those data before another future step begins, using the merged data as input
- There can be multiple splits and joins
- At each “merge” step, there can be 1000s of files processed