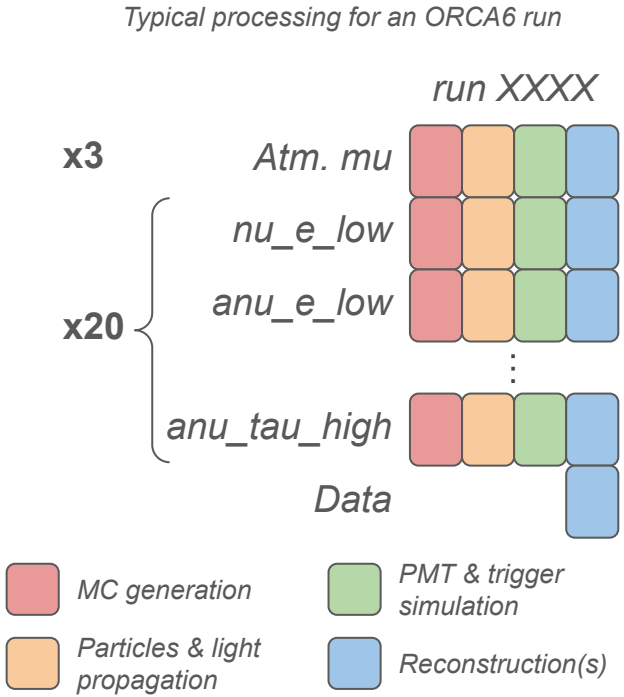
Run-wise data processing: Current status

Valentin Pestel,
KM3NeT France meeting - 08/02/2023

A complex workflow



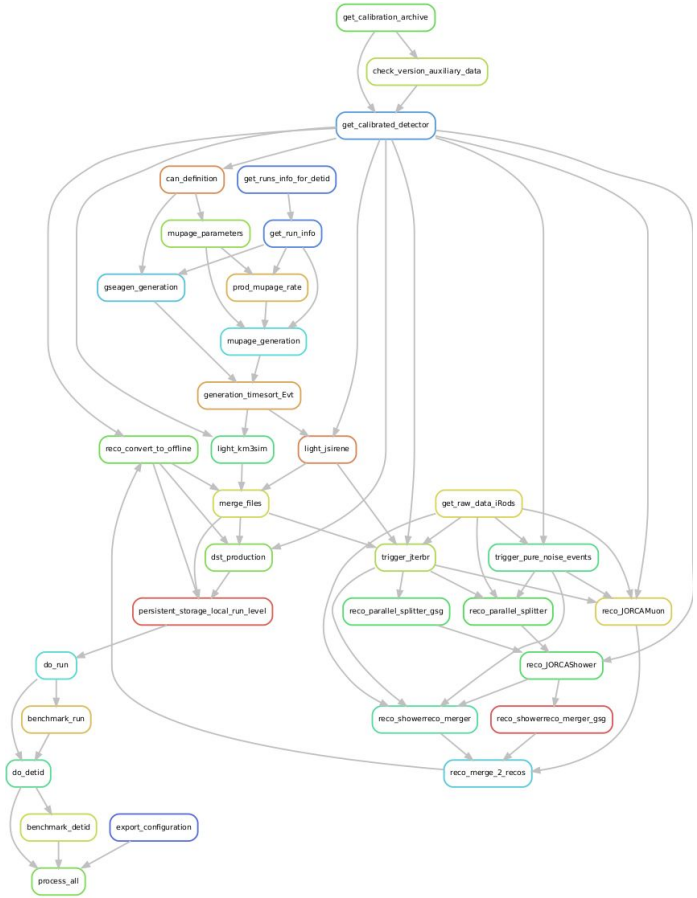
Our data processing is an extensive task:

- 24 contributions, 4 steps per contribution, ~3000 runs -> 288000 “jobs” to perform

Currently a bottleneck for all physics group

- Require too much humans and computing resources
- No automatic monitoring, manual input collection
- For ORCA6, never succeed to provide a “monolithic” production
- Difficult to deploy anywhere except Lyon

New implementation



New workflow: run-wise philosophy with Snakemake

- Allow for a better IO usage
- Relies on containerized softwares
- Automatic inputs collection
  - Solution developed in parallel, based on a git repository

Provide a scalable, easy to deploy workflow:

- Can be run on any cluster of the collaboration
  - CVMFS -> no setup needed
  - Without CVMFS -> half a day of setup, maximum
- Can be used for exploration and development
  - E.g. test production with a modified light generator
- Can be deployed on your laptop/desktop
- In the future, running on the GRID too ?
  - Test on-going with Mieke

Current status

Large scale test on-going, with a dedicated team (french colleagues are welcome):

- Goal: running on multiple cluster an ORCA6 test production
- Currently being run at Lyon, CNAF, Nikhef, Demokritos, Caen
  - Still a bit slow to take off … Somehow considered as a secondary task

It runs only for ORCA6 , but provides:

- Data, atm muons (1:1 livetime), neutrinos and pure-noise MC (1:10 livetime)
- Track and shower reco
- Offline files and DSTs

What is needed to use-it as official collaboration solution:

- Run a full scale processing (on track) -> **help always welcome**
- Test the results validity -> **help needed**
- Make other detectors available -> **help needed to validate that too**
  - Define calibration versioning strategy, implement ARCA reconstruction

156 processing steps per run