JUNO DCI Challenges

Xiaomei Zhang and João Pedro Athayde Marcondes de André

July 3rd, 2020

Outline

- Why do we organize the DCI challenges
- DCI productions
- Future planning

Why do we organize the DCI challenges

- We have deployed the DCI for JUNO $\rightarrow$ it works!
- But that doesn't tell us if it works well enough for JUNO...
- There are 2 main categories for DCI challenges:
  - Stress test the system to see up to what level it works
  - Run usual JUNO software to make sure it works in 'realistic' scenario
- Need to make sure tests done are 'representative' of real usage
  - Contact with physics & simulation and offline groups to get requests for datasets

JUNO DCI simulation productions

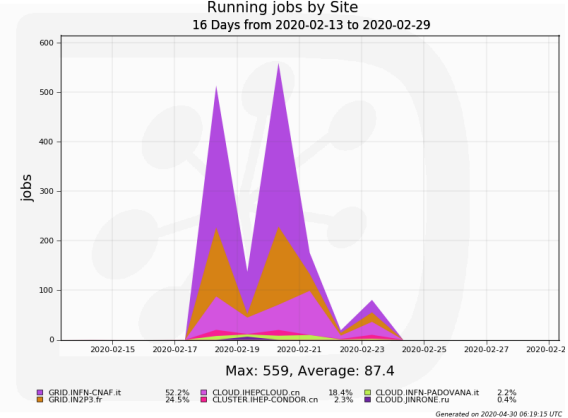
1. ML request: e^- and e^+ samples (see DocDB-5730)
   - Produced around end February, beginning of March 2020
2. ML request: e^-, e^+, γ, α fixed-energy & fixed-position samples
   - For PID testing, not very large
   - Produced around mid to end of May 2020
3. Large scale μ production:
   - Started mid June 2020, still on going
   - Complexity tied to high variance in resources used for different jobs
4. Software validation: ideas to use DCI for that purpose, nothing produced yet
   - The validation would test many different setups, not a lot of events for each case

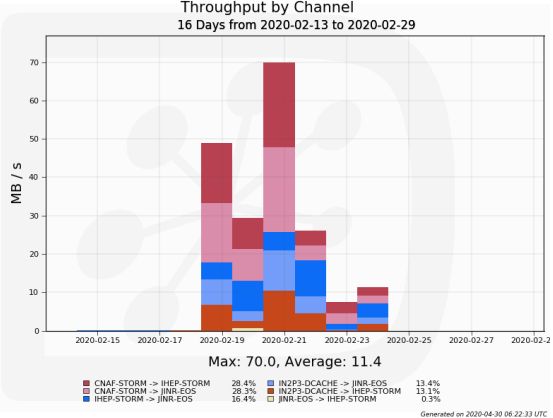
The life-cycle of a production currently

1. A user contacts someone from the DCI to request a production
2. User provides script and specification of what samples are needed
3. Test sample is produced
4. The user making the request validates the test sample
5. Run full production
6. Produced files should be checked by user making request
7. JP (ideally) reports during physics meeting that the sample is available for all

- While it looks ‘easy’ some steps might take multiple days as a lot of this requires interaction by different people with varying availability
- For now, all job submissions are done by Xiaomei

First production: ML request of e^- and e^+ samples

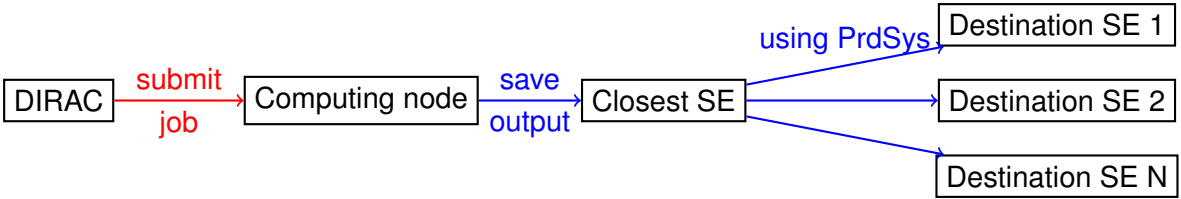




- 5400 jobs submitted, output transferred automatically to JINR & IHEP
- peak speed for transfer > 3 Gb/s
- Share of jobs: INFN – 54%, IN2P3 – 25%, IHEP – 21%, JINR – 0.4%
  - noticed some IN2P3 restricted to 100 CPUs in the beginning, later increased to 200

Comment on file transfers during productions

- In previous slide I said “output transferred automatically to JINR & IHEP”
- What happens in fact is the following:

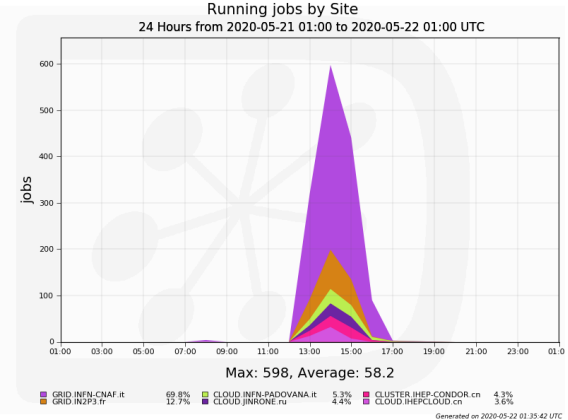


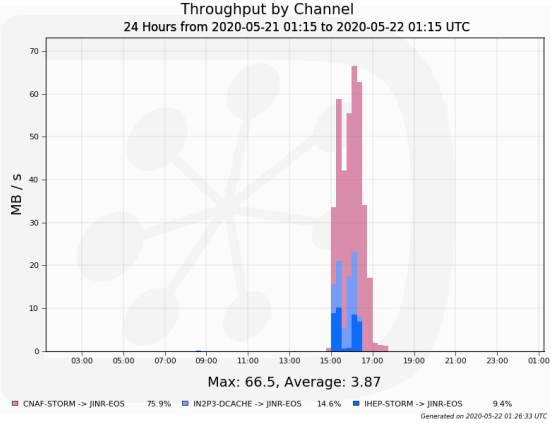
- So outputs will be saved where they were processed + JINR + IHEP, in this case
- Having all outputs in given cluster as it makes ‘local’ processing easier
  - At IHEP, CNAF and JINR user doesn’t need ‘grid’ certificate to access files locally
  - At CC-IN2P3, grid certificate is required, but files can be accessed locally with xrootd

Second production: ML request of samples for PID: the sample

- Simulate 4 types of particles (e^-, e^+, α, γ) with fixed kinetic energy at 3 positions
  - positions: center $(0, 0, 0)$, equator $(17, 0, 0)$ m, top $(0, 0, 17)$ m
  - true energies defined so events have same visible energy: e^- at 2.0000 MeV, e^+ at 1.1086 MeV, α at 11.600 MeV, γ at 2.0781 MeV
  - for each particle at each position: 100k events simulated, 2k events/job
- This is a “small” production
- Output data registered in DFC under
  `/juno/production/yuri/prd003/J19v1r0-Pre3/MultiParticles/`
- JINR copy available from
  `/eos/juno/dirac/juno/production/yuri/prd003/J19v1r0-Pre3/MultiParticles/`

Second production: ML request of samples for PID





- 660 jobs submitted, output transferred automatically to JINR
- speed for transfer of 665 Mb/s (350 GB, 1200 files, transferred)
- Share of jobs: INFN – 75%, IN2P3 – 13%, IHEP – 8%, JINR – 4%

Third production: μ sample

- Simulating μ is particularly complicated due to the inhomogeneous requirements
- Most jobs use 2–4 GB of memory, but some can require a few times more memory
  - it's wasteful to run it first in high memory cores
  - but 'usual' cores are certain to fail sometimes
- Some muons are simulated in minutes, others take days
- All this is tied to where the μ pass through in the detector & their interactions (so we don't know in advance which jobs will fail)

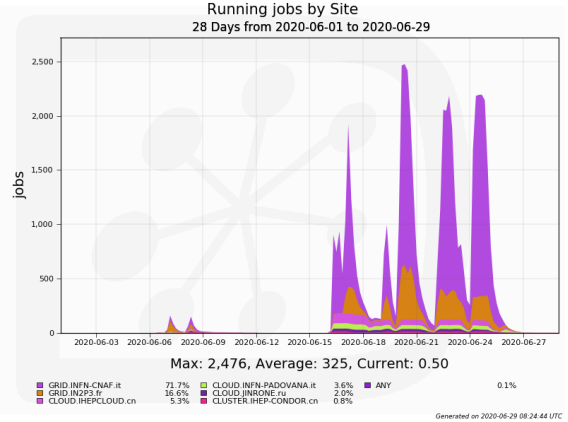
$\Rightarrow$ each job has to simulate only a few events

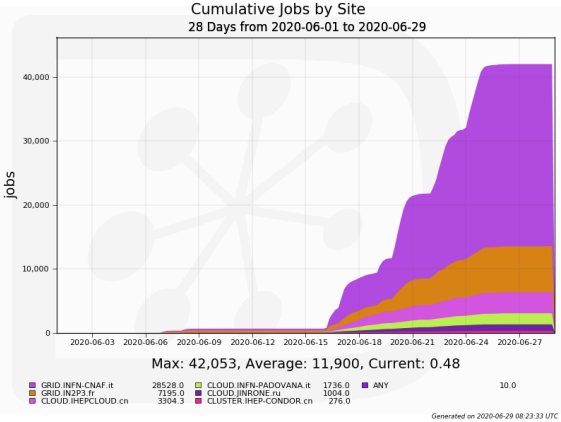
- Many small files (file sizes range from 16K to 430M)
- Lots of short jobs and some very long jobs
- Expect a fraction of the jobs will fail in first pass, in any case

Third production: μ sample – procedure

- Modified procedure for μ simulation (this time):
  1. Xiaomei submits 10k jobs
  2. JP identifies failed/missing (typically 200–500 jobs) and re-submits in regular cores (at CC-IN2P3)
  3. JP identifies which ones still failed (typically 50–70 jobs) and JP & Jilei re-submit them to CNAF and IHEP as high memory jobs
     - ⋆ about 80 high memory cores available at IHEP, not dedicated to JUNO
     - ⋆ @CNAF ask for more memory with multi-core jobs (even if using single core)
     - ⋆ could potentially also use CC-IN2P3/JINR for this, not set up yet though
  4. If there are still failed jobs, repeat previous step
  5. Once finished, add re-processed files to FC
- As stated, we haven't quite finished this process yet... Currently at step 3
- Once the queues start freeing up a bit, Xiaomei will start again from 1, while the other steps are run in parallel.
- Goal is to reach 200k jobs (ie, 1 M μ events $\rightarrow \sim 1/2$ day of μ)

Third production: μ sample – number of jobs





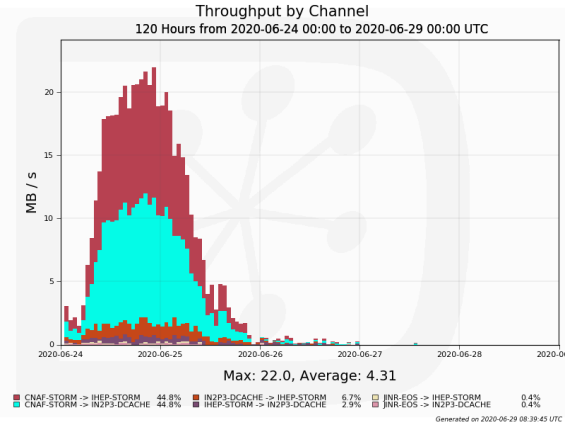
- Each big peak corresponds to 10k jobs submitted, 2 small peaks are test runs
- In total $\sim$ 40k jobs submitted already
- Share of jobs: INFN – 75%, IN2P3 – 17%, IHEP – 6%, JINR – 2%

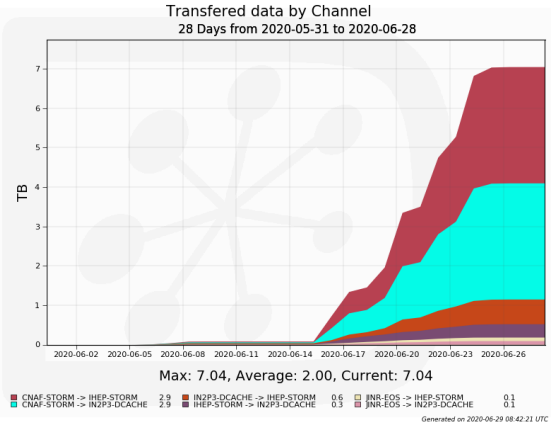
Third production: μ sample – general comments

- In early production: issues with CC-IN2P3 due to maintenance
- In cloud sites had problems with memory
  - increased memory/core to 4GB

  $\Rightarrow$ failure rate decreased
- Extra 1200 new JINR cores not yet on DCI for this production
- CC-IN2P3 mentioned that they having always jobs on the DIRAC queue would help, rather than having peaks of job submission, with the DIRAC queue getting empty in between

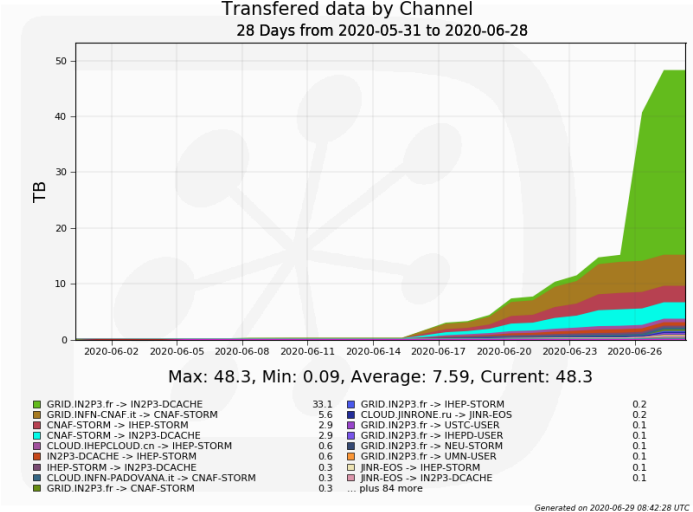
Third production: μ sample – transfer rates





- Files transferred automatically to IHEP & IN2P3
- $\sim$7 TB transfered between sides over FTS3
  - speed can reach 22 MB/s

Third production: μ sample – transfer rates



- 48.3 TB of data moved (including transfer to closest SE and between sites)

Future plan

- Need to organize more Challenges with added resources
  - have extra resources for a week for the Challenge – test system at higher throughput
  - first test at $\times 2$, then $\times 3 - 4$, then $\times 6 - 8$ (if possible)
    - ⋆ usually it's easier to have a quick increment in CPU throughput than in available disk,
    - ⋆ but as long as enough space is available that should be OK
  - make sure to test different kinds of jobs (not just simulation)
- Will inform data centers a week ahead of time so they can prepare
  - doodle will be sent out also to decide on best dates
- Ideally, coordinate with physics/offline to produce useful samples when possible

Summary

- DCI system is working
  - could already participate in 3 productions useful for physics
  - μ production is challenging in any situation, processed 200k μ in a bit more than 1 week
    - not counting time for fixing issues with processing which are done manually at this point
  - files automatically sent to their 'final destinations'
- Still lessons learned:
  - Need better handling for high-memory productions
  - Default memory on cloud sites is small, need to set it properly
- Constant communication between sites was essential to fix issues as they appeared!
- Next: get closer to JUNO real production use case and planned stress testing when resource allows
  - Current challenges were simulation based – high CPU usage, but smaller impact on SE
    - data transfer challenges also needed
  - Need to test scalability of system towards JUNO expected needs
    - agreement to have challenges with more (temporary) resources in future
    - CCs will be informed about them a bit ahead of time