ALICE workflow efficiency considerations

2016-10-09

Maarten Litmaath

v1.0

Workflow

- Users submit jobs to the central task queue
- A VOBOX for each (real or virtual) site submits pilots (aka job agents) to the site's resources
  - Via CE(s) or directly to the batch system
  - When there are tasks waiting that match the site
- Pilots contact the central task queue and get matching jobs assigned
  - A pilot exits when there are no matching tasks

Matchmaking

- Matchmaking is based on input data locality
  - Most of a task's input data need to be close to the site
  - This requirement is relaxed for the tails of job collections
    - To allow them to finish within a reasonable time
- ~15% remote access can be tolerated today
  - And avoid swamping *networks* or the remote *disk servers*
  - Also needed for fail-over when a local replica is inaccessible
- Analysis has the most heavy I/O and hence highest locality demands
- Reco only runs at T0 and T1 that have the data
- MC simulation can run anywhere

Requirements

- For a typical 10-HEPSpec reference CPU
- Analysis
  - 20 Mbit/sec/core for 80% CPU efficiency
  - RTT penalty example: 150 msec (e.g. CERN-JINR) implies -20% efficiency
- MC simulation
  - 0.3 Mbit/sec/core
  - input: ~few MB/job
  - output: ~350 MB/job
  - average job duration: 6 hours for 300 p+p events or 3 Pb+Pb events

Improving analysis efficiency

- Users are advised to run big analyses as parts of analysis *trains*
  - Input data are read once and made available to all *wagons* of a train, each doing its own analysis
  - Train jobs have higher priority than individual analysis jobs
- The fraction of trains vs. individual analysis has strongly increased and then stabilized in the last years
  - 2015 Jan-Dec avg: 9k train jobs, 3.9k individual jobs, 61k total
  - 2016 Jan-*Sep* avg: 8k train jobs, 3.9k individual jobs, 73k total
    - The small decrease is due to the delayed reconstruction of the 2015 heavy-ion run, now mostly done
    - Since the start of autumn: 10k train jobs on average!
- Users are also advised to run analysis over compact AOD files instead of the 10x bigger, sparsely useful ESD files
  - Try to reduce unnecessary reading of large amounts of unused data
  - AOD analysis generally has higher priority than ESD analysis