Investigating HEP workloads on the Dutch HPC

R. Aaij, M. Kalamker-Stam, C. Schrijvers

Cartesius and Future HPC

- Dutch HPC
- Bullx with extra sequana island
- 1.84 Pflop/s
- Infiniband interconnect
- Lustre shared filesystem
- Most nodes: 24 cores, 64G RAM
- So far not used for HEP HTC work
- Goal of the project: test a few typical HEP workloads

Pilot Scenarios and Requirements

Jobs

- LHCb-II simulation jobs: CPU-heavy little I/O
- LHCb NTuple-making job: Some CPU, heavy input, some output
- In addition: stage data for testing
- So far purely functional testing, no large-scale running

Requirements

- CVMFS on interactive and worker nodes
- Data
- Support contact

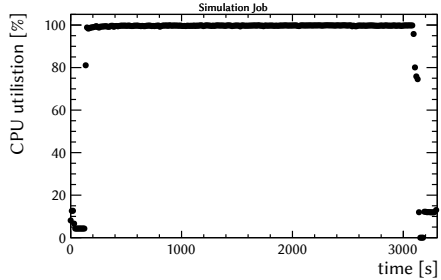
Technical Considerations

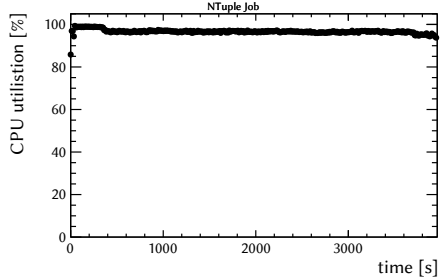
- Worker nodes have no outside access
- Low memory/core: (2.7 GiB) compared to standard grid nodes
- A job is a node, not a core

- LHCb simulation not multi-threaded
- Cannot fully load node with N “grid” jobs
- Multiprocessing capabilities exist: GaudiMP
- GaudiMP not generally used in production: see talk by F. Stagni for another example
- GaudiMP distributes events to workers and collects output
- Histograms and counters also taken care of

Running Jobs

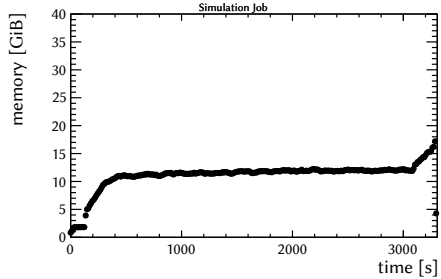
- Started by running jobs by hand
- Quite a few GaudiMP issues to debug
- NTuples not handled by GaudiMP
  $\rightarrow$ custom solution
- Once MP worked, memory was OK
- Can fully load a node
- “Standard” issues also encountered
  i.e. broken CVMFS failed all jobs

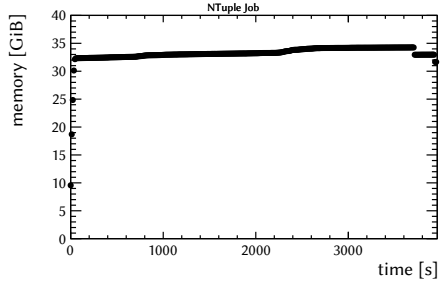




Running Jobs

- Started by running jobs by hand
- Quite a few GaudiMP issues to debug
- NTuples not handled by GaudiMP
  $\rightarrow$ custom solution
- Once MP worked, memory was OK
- Can fully load a node
- “Standard” issues also encountered
  i.e. broken CVMFS failed all jobs





Summary

Funcional Test

- LHCb jobs run on Cartesius
- Support effort needed to install and maintain CVMFS
- No outside connectivity, so data had to be staged to shared filesystem
- Can fully load nodes
- Memory usage not an issue with multi-processing

Going Forward

- Tested only O(few) jobs running simultaneously
- Further investigation required for running at scale
- Investigate automated staging (input and output) and job management
- Scale-out of SURFsara grid cluster to Cartesius also being tested