Communicating Machine Features to Batch Jobs

Originally to WLCG MB, March 8th 2011

Tony.Cass@cern.ch

Reminder

◆ The HEPiX Virtualisation Working Group proposed the creation of 3 files in `/etc/machinefeatures` with machine information during virtual machine instantiation:

- `hs06` — the HS06 rating for a single core,

- `shutdowntime` — timestamp for when the node is expected to be shutdown, and

- `shutdowncommand` — the full path to a command that can be invoked to arrange early termination of the machine.

◆ Passing similar information to jobs in real machines is considered to be useful.

Proposal for real machines

◆ Physical machines, just as virtual machines, should have in `/etc/machinefeatures` the files

 – `hs06` — the HS06 rating for a single core

 – `shutdowntime` — timestamp for when the node is expected to be rebooted (may be empty)

◆ Additionally, the batch system should provides each job with an environment variable, `$JOBFEATURES`, pointing to a directory with the files

 – `jobstart`   timestamp for the start of job execution

 – `cpu_limit`   the allowed number of cpu-seconds that can be used by the job. The value must be comparable directly against system reported cpu consumption.

 – `wall_limit`   the allowed number of wall clock seconds. The job will be terminated at time `jobstart+wall_limit` whether or not `cpu_limit` has been reached.

With thanks to Steve Traylen and Ulrich Schwickerath

Implementation Status

◆ LSF

– Initial prototype available in SVN repository; initial tests suggest the implementation works OK, no firm date for deployment as yet, but possible before end-March.

» Developed by Ulrich Schwickerath; `$JOBFEATURES` points to `/var/tmp/jobfeatures/<jobid>`

◆ PBS

– NIKHEF asked to deliver an implementation; waiting to see whether or not this request will be accepted.

◆ Other workload management systems

– Which?

– Who?

MB Conclusion

◆ The implementation for LSF may have some CERN specifics. Other sites interested in this implementation are suggested to contact the developer, Ulrich Schwickerath.

◆ NIKHEF clarified that it is not *responsible within WLCG for PBS, but in the context of best effort, a project plan will be submitted to a national project in the Netherlands and if accepted NIKHEF will provide a PBS implementation.*

◆ Compared to the data in the Information System, this proposal makes data dynamically available at run time for the specific machine the job is running on.

◆ Whilst this proposal brings some value, making it compulsory for all batch systems was questioned. Tony Cass clarified that the principle had been agreed in the MB meeting on 28 Sep 2010 and that the goal of this presentation had been to present a deployment plan.

◆ A WLCG discussion would be needed on the distribution mechanism once packages for different