Multicore

Alessandra Forti

HEPSYSMAN

13th January 2014

Multicore problem

- Getting multicore on the WNs in WLCG is “easy” enough because the experiments want all the cores on 1 node there is no need to setup MPIs.
- The problem is running them without wasting resources.
- The main factor in wasting resources is the cores draining for multicore jobs.
  - Enabling a multicore queue without changing absolutely anything on the batch system is not advisable even if experiments may come up with some strategies to reduce the impact of draining (see CMS).

Multicore problem in WLCG

- Draining problems affects everyone
  - Dedicated/non dedicated
    - It's also an MPI problem but MPI can get away more easily by allocating CPUs on different nodes
  - In Europe in particular we support multi-LHC and smaller Vos
    - Even within an experiment there will be single core and multi-core jobs
  - In UK our funding depends on CPU efficiency
    - Mishandled multicore can really do some damage.

Atlas

- Has currently no strategy to schedule different payloads in the same pilot.
  - They believe configuring for multi-core is a site problem to solve.
- All they do is requesting 1 node and 8 cores in the JDL
- This is passed to the blah submission scripts and translated to batch system requirements
- Xmas multi-core production was initiated because at the time the resources were empty.
  - Atlas has the most at stake because they have the biggest events and they cannot run production without multicore after LS1.
- AthenaMP works

CMS

- CMS has the opposite approach. They have worked on a mechanism to schedule different workloads even mixed in the same pilot.
  - It was aimed at not asking sites to change anything in their batch system.
- It has drawbacks
  - It requires very long pilot lifetimes
  - It still doesn't solve the draining problem, although if the pilot works for a week it is reduced.
    - They think they can solve the draining problem using machine job features files for pilot batch-system communication..
  - If a payload in such a pilot causes problems all the payloads are affected.
  - Might work at T1s but definitely not at T2s
- Their executable doesn't work yet, plans for October.

Sites

- Some sites have multicore enabled.
  - Static dedicated resources
  - Dynamic scheduling
    - Aim of everyone working on this should be this.
- Some batch system seem to deal with it better than others
  - Worst are Torque+maui and LSF
  - Others Htcondor and SGE use a technique called partitioning.
    - More in Andrews talk (?)
  - Slurm has also some mechanism do ease multi-core scheduling
- Most MC sites currently working in Atlas have dedicated resources
  - Some have dynamic scheduling but not clear how many resources they are wasting.

Manchester experience

- Enabled multi-core with “dynamic” scheduling
  - Without changing anything in the batch system
    - without allocating resources
- Up to 200 empty slots until atlprod 8 core is at the top of the queue.
  - Lower higher priority jobs blocked
- Slots kept empty maybe for hours get filled by incoming higher priority jobs.
  - Reduced the slots that can be kept empty using maui (MAXIPROC, MAXIJOB)
    - Multicore has stopped running
      - Might also be an atlas problem.

Manchester experience

- Node allocation policy optimized for load balance is counter productive for multicore
  - Batch system frees CPU on all the nodes until it finds one with 8
  - BatchHold jobs. ie assigned to a node but not running because the CPUs have been assigned in the mean time.
    - Not clear if this was caused by a change in maui configuration though.
- Atlas tests tracked here
  - https://twiki.cern.ch/twiki/bin/viewauth/AtlasComputing/AtlasMulticore

WLCG TF

- To put all these parties together to work the WLCG TF was created
  - Will be the forum to put all together
  - https://twiki.cern.ch/twiki/bin/view/LCG/DeployMultiCore
- Already 30+ people subscribed
- Approved in mid-December
- We haven't started any activities
  - Collecting all fragments of information first
- Will send first email this week