HTCondor and the European Grid

Andrew Lahiff

STFC Rutherford Appleton Laboratory

European HTCondor Site Admins Meeting 2014

Outline

- Introduction
- CREAM CE
- ARC CE

Introduction

- Computing element requirements
  - Job submission from LHC VOs
    - AliEn: ALICE
    - HTCondor-G: ATLAS, CMS
    - DIRAC: LHCb
  - EMI WMS job submission
    - Still used by some non-LHC VOs
    - Usage (probably) likely to decrease
  - Information system
    - Information about jobs & worker nodes needs to go into the BDII

CREAM CE

- Does not work out-of-the-box with HTCondor
  - HTCondor not officially supported (anymore)
- However
  - BLAH supports HTCondor, so there is hope…

CREAM CE

- Problems
  - No YAIM function for configuring CE to use HTCondor
  - Scripts to publish dynamic information about state of jobs missing
  - YAIM function for configuring BLAH doesn't support HTCondor
- RAL solution
  - Make use of scripts from very old versions of CREAM which did support HTCondor
    - Updated these for current EMI-3 CREAM CE
    - Needed modernizing, e.g. to support partitionable slots

CREAM CE: queues

- We wanted the following queues
  - Different memory limits (1GB, 2GB, 3GB, 4GB)
  - Multi-core
- Make sure site-info.def contains details about the queues
- Modifications
  - condor_submit.sh (from glite-ce-blahp RPM)
    - Add request_memory, request_cpus to the submit file
  - lrmsinfo-condor
    - Runs condor_q with constraints for each queue

https://twiki.cern.ch/twiki/bin/view/EGEE/QueueSimulationInstructionsForCondor

CREAM CE: accounting

- APEL accounting
  - At RAL
    - Script which modifies blahp.log files

      Appends .hostname to job ids (required for sites with multiple schedds)
    - Script which converts condor history files to PBS-style accounting files
    - PBS APEL parser
  - Now there's a fork of APEL containing HTCondor integration
    - Script which writes some information from condor history into new files
    - HTCondor APEL parser which reads these files

https://github.com/alahiff/cream-for-htcondor

CREAM CE

- RAL has been running 2 CREAM CEs with HTCondor in production for over a year
  - Was used by ALICE, LHCb, non-LHC VOs
    - Now mainly used by a single non-LHC VO
  - Over 3 million jobs run

ARC CE

- NorduGrid product
  - In EMI, UMD
- Features
  - Simpler than CREAM CE
  - Can send APEL accounting data directly to central broker
  - File staging: can download & cache input files; upload output to SE
- Configuration
  - Single config file /etc/arc.conf
    - No YAIM required

ARC CE

- Can the LHC VOs submit to ARC?
  - ATLAS
    - Use HTCondor-G for job submission

      Able to submit to ARC
    - ARC Control Tower for job submission
  - CMS
    - Use HTCondor-G for job submission
  - LHCb
    - Last year added to DIRAC the ability to submit to ARC
  - ALICE
    - Recently regained the capability to submit to ARC
- Submission via EMI WMS
  - Works (uses HTCondor-G)

Science & Technology Facilities Council

ARC CE

- Integration with HTCondor
  - Like CREAM CE, HTCondor scripts had gotten a bit out of date
  - Older versions (< 4.1.0) required lots of patches
    - E.g. assumptions made which were not true with partitionable slots
  - Current version (4.2.0) works out-of-the-box with HTCondor

ARC CE

- 4.1.0
  - Contains many patches provided by RAL for HTCondor backend scripts
- 4.2.0
  - Bug fix for memory limit of multi-core jobs (HTCondor)
- Future release
  - Have submitted patch to enable CE to make use of per-job history files
- Repository
  - At RAL we use the NorduGrid repository. For 4.2.0:

    http://download.nordugrid.org/repos/13.11/redhat/6.4/x86_64/

ARC CE: accounting

- APEL node not required
  - JURA component of ARC sends accounting data directly to APEL central broker
    - Note some sites prefer to have an APEL publisher node
- Scaling factors
  - Unlike Torque, HTCondor doesn't scale CPU & wall time
  - How this is handled at RAL
    - Startd ClassAds contain a scaling factor
    - An ARC auth plugin applies scaling factors to completed jobs (1 line added to /etc/arc.conf)

https://github.com/alahiff/ral-arc-ce-plugins/blob/master/scaling_factors_plugin.py

ARC CE

- Information passed to HTCondor from jobs
  - Max wall time, max CPU time, number of cores, total memory
  - Generates appropriate Periodic_remove for submit file
    - Memory, wall time, CPU time
  - Example snippet from an ATLAS job

```
request_cpus = 1
request_memory = 3000
+JobTimeLimit = 345600
+JobMemoryLimit = 3072000
Periodic_remove = FALSE || RemoteWallClockTime > JobTimeLimit ||
                           ResidentSetSize > JobMemoryLimit
```

Science & Technology Facilities Council

ARC CE: queues

- Can specify HTCondor requirements expression for each queue in /etc/arc.conf
- request_memory can be taken from queue configuration

  (if job doesn't specify memory)
- Cannot setup queues with different time limits

  (without hacking scripts)

ARC CE

- Current issues
  - Currently no proxy renewal for jobs submitted via EMI WMS
    - Affects non-LHC VOs only
    - A workaround exists (*)
  - Job status information is not real-time
    - LHCb have commented on this
    - If ARC WS interface used, situation much better

      Requires RFC proxies

* https://www.gridpp.ac.uk/wiki/Delegating_Proxies_to_ARC_CEs

Science & Technology Facilities Council

Summary

- There are no blocking issues preventing European sites from using HTCondor as a batch system
- Both CREAM and ARC CEs work!
  - CREAM currently requires more effort to setup
- HTCondor-CE?
  - Integration with BDII