Liverpool HTC-related Developments

Ste Jones
(Liverpool University, GridPP)
HepSysMan, RAL, May 2019

Contents

- Things that have not changed
- Things that we've been given
- Things that we've bought
- Clusters that we use now
- Storage that we use now
- New baselines we've installed
- New software development we've done

Network

- WNs 1G uplink to TopOfRack switches (internal)
- TopOfRack switches uplink to core network at 20G (2x10G, internal)
- WNs use NAT for connections off-site
- Main storage uplinks to Core Network at 20G (10G internal+10G external)
- Core network uplinks to University core network at 20G (2x10G) with redundant failover
- Internal research network physically and logically isolated from external network. Runs with Jumbo Frames.
- Also use an isolated management network for IPMI, hardware control, monitoring.
- IPv6 enabled on perfsonar test systems so far.

Server room

Bought this year

- We replaced one RAIDed DPM storage server (~ 40TB) with a new ZFS DPM storage server (~ 220 TB), balance 180TB, giving total DPM storage capacity of around 1.6 PB.
- We also bought Gold 6132 CPU systems, each with 56 hyperthreaded cores, 128 GB RAM. 5 units, giving 280 slots, or ~ HS06 of 3656.
- Total Site HS06 is 28482
- Total Site Cpus is 2691

Clusters

- VAC is off at present.
- We had trouble due to kernel lockups on early C7 kernels. The problem is solved now (C76+), but I haven't remade the systems since code has changed quite a bit, e.g. new features like vac pipes that I have to get my head around.
- The site is all CentOS7.
- All the service nodes (bar storage) are virtual (KVM)

Clusters - HTCondor-CE

- HTCondor-CE + HTCondor
  htcondor-ce-3.2.0-1.el7
  condor-8.6.12-1.el7
- totHs06 – 15859.82
- totCpus – 1507

Clusters - ARC-CE

- ARC-CE + HTCondor
  nordugrid-arc-5.4.1-1.el7
  condor-8.6.3-1.el7
- totHs06 - 12622.08
- totCpus – 1184
-

Storage

- 1.6 PB
- 1 x headnode, DPM, not DOME yet
- 17 x storage nodes in total
- 15 x raid6 of all kinds: 3ware, areca, megaraid, adaptec.
- 2 x ZFS – these are the newest
- John seems very happy with ZFS.

New baselines we've installed

- I spent the winter f nding and writing up a stable baseline for HTCondor-CE + HTCondor.
- Some parts were missing; APEL, BDII (glue1).
- Don't need glue1... sort of.
- APEL: No support. Worked with WLCG Accounting Group and HTCondor-CE team to make a portable implementation based on APEL client software.
- Fed that back to HTCondor devs; RPMs "available" (not released).

New baselines we've installed

- **HTCondor-CE+HTCondor baseline doc**
  - https://www.gridpp.ac.uk/wiki/Example_Build_of_an_HTCondor-CE_Cluster
- **Has two modes to install.**
- **One is manual, with no conf ig control system. Package list, conf ig as tarball.**
- **The second method involves CERN puppet module.**
- **Previous talk here (search for jones):**
  - https://indico.cern.ch/event/780766/timetable/

Software development we've done

- **Added APEL support to HTCondor-CE**
  - https://twiki.cern.ch/twiki/bin/view/LCG/HtCondorCeAccounting
- **This is the "manual" install version. Doc changes needed when RPMs come out.**
- **At present, it only works when using HTCondor batch system backend.**
- **Github commits (subject to change):**
  - https://github.com/opensciencegrid/htcondor-ce/tree/master/contrib/apelscripts
- **Will need volunteers using other batch systems to help us write (or test existing) schemes to ingest data from other batch systems such as SGE, PBS, LSF, SLURM, …**

Software development we've done

- NOTE: The APEL client software obtains data from two sources.
- It gets data from the CE and then data from the batch system.
- The CE data is in a standard format called BLAH, hence there is only one BLAH parser in the APEL client software.
- Each type of batch system needs its own parser (or "adapter" … TBD).

Parsers or adapters

- We have only tested with HTCondor .
- I used the existing HTCondor batch system APEL client software parser (originally for CREAM.)
- And I wrote an adapter using condor_ce_history formatting language (printf-style) to get the data in the correct format for the parser to ingest.
- So HTCondor-CE + HTCondor is basically solved, see links above.

Parsers or adapters

- **Some of the other batch systems supported by HTCondor-CE (SGE, PBS, LSF, SLURM) may or may not have existing parsers in APEL client software (for use with older CEs), but I haven't tried them out.**
- **If they don't already have existing parsers in APEL client software, it will be necessary to either**
  - Write a whole new parser or
  - Write an “adapter” to convert batch logs to the format of an existing parser.

Developing parsers or adapters

- **To write a new parser, it is necessary to clone the APEL client software git repo, add the new parser into the software suite, create a pull request to ingest your changes into the main APEL client software tree. The maintainer of that material is Adrian Coveney (RAL) and this is the repo:**
  - https://github.com/apel/apel

Developing parsers or adapters

- **To write an adapter, use the same process, but in the HTCondor-CE repo.**
- **Clone the HTCondor-CE repo, add a new adapter in the contrib/apel (or /apelscripts) directory, create a pull request to ingest your changes into the main HTCondor-CE repo**
- **The maintainer of that material is Brian Lin (HTCondor dev) and this is the repo:**
  - https://github.com/opensciencegrid/htcondor-ce

Developing parsers or adapters

- **If you manage to make a parser (for APEL client software) or an adapter (for HTCondor-CE contrib), then the maintainers will create RPMs containing the changes to the benef t of all subsequent users.**
-

Budget

- Less (Fewer?) FTE in future.
- Not enough to continue software dev work, planning, baseline work or other ancillary tasks, WLCG task force work, vomssnooper, VOMS RPMs, documentation, talks and communications, user and sysadmin advice and/or assistance.
- Leaving only the routine sysadmin work.
- So it goes. In future, we'll have todo more “roll your own.”

Questions?

Thanks,

Ste