What’s new in HTCondor?
What’s coming?

HTCondor Week 2016
Madison, WI -- May 18, 2016

Todd Tannenbaum
Center for High Throughput Computing
Department of Computer Sciences
University of Wisconsin-Madison
Release Timeline

› Stable Series
  • HTCondor v8.4.x - introduced Aug 2015
    (Currently at v8.4.6)

› Development Series
  • HTCondor v8.5.5 frozen, in beta test, release to web later this month.

› HTCondor v8.6.0 expected summer 2016.

<table>
<thead>
<tr>
<th></th>
<th>All Time</th>
<th>12 Month</th>
<th>30 Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commits:</td>
<td>39067</td>
<td>2349</td>
<td>141</td>
</tr>
<tr>
<td>Contributors:</td>
<td>152</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td>Files Modified:</td>
<td>11588</td>
<td>1665</td>
<td>169</td>
</tr>
<tr>
<td>Lines Added:</td>
<td>12352208</td>
<td>444401</td>
<td>29395</td>
</tr>
<tr>
<td>Lines Removed</td>
<td>6810332</td>
<td>187595</td>
<td>7835</td>
</tr>
</tbody>
</table>

Source: https://www.openhub.net/p/condorproject
Some enhancements in HTCondor v8.4

› Scalability and stability
  • Goal: 200k slots in one pool, 10 schedds managing 400k jobs
  • Resolved developer tickets: 240 bug fix issues (v8.2.x tickets), 234 enhancement issues (v8.3 tickets)

› Docker Job Universe

› Tool improvements, esp condor_submit

› IPv6 mixed mode

› Encrypted Job Execute Directory

› Periodic application-layer checkpoint support in Vanilla Universe

› Submit requirements

› New packaging
Scalability
Enhancement
Examples
Condor_shadow resources

- Reduce memory footprint of Shadow
- Eliminate need for authentication step to schedd, startd (on execute host)

<table>
<thead>
<tr>
<th>Version</th>
<th>Nominal (KB)</th>
<th>With GSI Authentication (KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>v7.8.7</td>
<td>860</td>
<td>1860</td>
</tr>
<tr>
<td>v8.4.0</td>
<td>386</td>
<td></td>
</tr>
</tbody>
</table>
Authentication Speedups

- FS (file system) and GSI authentication are now performed asynchronously
  - So now a Condor daemon can perform many authentications in parallel
  - CMS pool went from 200 execute nodes (glideins) per collector to 2000
- Can cache mapping of GSI certificate name to user name
  - Mapping can be heavyweight, esp if HTCondor has to contact an external service (LCMAPS…)
  - Knob name is GSS_ASSIST_GRIDMAP_CACHE_EXPIRATION
Faster assignment of resources from central manager to schedd

- Negotiator can ask the schedd for more than one resource request per network round trip.

```plaintext
NEGOTIATORRESOURCE_REQUEST_LIST_SIZE = 20
```
Impact of multiple resource requests
Negotiation times for 1000 slot pool

# of job autoclusters

- 8.2.8 LAN
- 8.3.5 LAN 20reqs
- 8.3.5 LAN 100reqs
- 8.2.8 WAN
- 8.3.5 WAN 20reqs
- 8.3.5 WAN 100reqs
Eliminate CCB service pauses
Query Responsiveness

› Improvement: Collector will not fork for queries to small tables
  • Load Collector with 100k machine ads
  • Before change: \(~4.5\) queries/second
  • After change: \(~24.4\) queries/second

› Improvement: Schedd condor\_q quantum adjusted (to 100ms)
  • Load schedd with 100k jobs ads, 40Hz job throughput
  • Before change: \(~135\) seconds per condor\_q
  • After change: \(~22\) seconds per condor\_q
Container Support
(Black Box Applications)

› HTCondor cgroup support now manages swap space in addition to CPU, Memory

› New job universe to support Docker Containers

  • Please talk to us if you have interest in using Docker with HTCondor!
Docker Universe Job
Is still a job

- Docker containers have the job-nature
  - `condor_submit`
  - `condor_rm`
  - `condor_hold`
  - Write entries to the job event log(s)
  - `condor_dagman` works with them
  - `Policy expressions` work.
  - `Matchmaking` works
  - `User prio / job prio / group quotas` all work
  - `Stdin, stdout, stderr` work
  - `Etc. etc. etc.*`
Many condor_submit improvements

You submit your jobs with *that* script??!?  You’re braver than I thought!
More ways to Queue 'foreach'

Queue <N> <var> in (<item-list>)
Queue <N> <var> matching (<glob-list>)
Queue <N> <vars> from <filename>
Queue <N> <vars> from <script> |

› Iterate <items>, creating <N> jobs for each item
› In/from/matching keywords control how we get <items>
› There's more. See the manual for details.
Example: Queue matching files

Executable = foo.exe
Arguments = -inputdata $(Item)
Queue 1 Item matching (*.dat, m*)

› Produces a job for each file that matches *.dat or m* (or both)
› $(Item) holds each filename in turn
Condor_q new arguments

- **-dag <dagman-job-id>**
  - Show all jobs in the dag

- **-limit <num>**
  - Show at most <num> records

- **-totals**
  - Show only totals

- **-autocluster -long**
  - Group and count jobs that have same requirements
  - …perfect for provisioning systems
IPv6 Support

› New in 8.4 is support for “mixed mode,” using IPv4 and IPv6 simultaneously.
› A mixed-mode pool’s central manager and submit nodes must each be reachable on both IPv4 and IPv6.
› Execute nodes and (other) tool-hosting machines may be IPv4, IPv6, or both.
› ENABLE_IPV4 = TRUE
  ENABLE_IPV6 = TRUE
Jobs can request (or admins can require) that their scratch directory be encrypted in realtime:
  • /tmp and /var/tmp output also encrypted
  • Put `encrypt_execute_directory=True` in job submit file (or `condor_config`)

Only the `condor_starter` and job processes can see the cleartext:
  • Even a root ssh login / cron job will not see the cleartext
  • Batch, interactive, and `condor_ssh_to_job` works
Periodic Application-Level Checkpointing in the Vanilla Universe

- Experimental feature!
- If requested, HTCondor periodically sends the job its checkpoint signal and waits for the application to exit.
- If it exits with code 0, HTCondor considers the checkpoint successful and does file transfer, and re-executes the application.
- Otherwise, the job is requeued.
Submit Requirements

› Allow administrator to decide which jobs enter the queue via a SUBMIT_REQUIREMENTS constraint
› Rejection (error) message may be customized
HTCondor RPM Packaging

› More Standard Packaging
  • Matches OSG and Fedora package layout
  • Built with rpmbuild
  • Source RPM is released
    • Can rebuild directly from the source RPM
    • Build requirements are enforced by rpmbuild
  • Partitioned into several binary RPMs
    • Pick and choose what you need
<table>
<thead>
<tr>
<th>RPM</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>condor</td>
<td>Base package</td>
</tr>
<tr>
<td>condor-all</td>
<td>Includes all the packages in a typical installation</td>
</tr>
<tr>
<td>condor-bosco</td>
<td>BOSCO – Manage jobs on remote clusters via ssh</td>
</tr>
<tr>
<td>condor-classads</td>
<td>HTCondor classified advertisement library</td>
</tr>
<tr>
<td>condor-classads-devel</td>
<td>Development support for classads</td>
</tr>
<tr>
<td>condor-debuginfo</td>
<td>Symbols for libraries and binaries</td>
</tr>
<tr>
<td>condor-externals</td>
<td>External programs and scripts</td>
</tr>
<tr>
<td>condor-externals-libs</td>
<td>External libraries</td>
</tr>
<tr>
<td>condor-kbddd</td>
<td>HTCondor Keyboard Daemon</td>
</tr>
<tr>
<td>condor-procd</td>
<td>HTCondor Process Tracking Daemon</td>
</tr>
<tr>
<td>condor-python</td>
<td>Python Bindings for HTCondor</td>
</tr>
<tr>
<td>condor-static-shadow</td>
<td>Static Shadow (Use 32-bit shadow</td>
</tr>
<tr>
<td>condor-std-universe</td>
<td>Standard Universe Support</td>
</tr>
<tr>
<td>condor-vm-gahp</td>
<td>VM Universe Support</td>
</tr>
</tbody>
</table>
HTCondor Debian Packaging

More Standard Packaging
• Matches debian package layout
• Built with pbuilder
• Source package is released

<table>
<thead>
<tr>
<th>deb</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>condor</td>
<td>Base Package</td>
</tr>
<tr>
<td>condor-dbgs</td>
<td>Symbols for libraries and programs</td>
</tr>
<tr>
<td>condor-dev</td>
<td>Development files for HTCondor</td>
</tr>
<tr>
<td>condor-doc</td>
<td>HTCondor documentation</td>
</tr>
<tr>
<td>libclassad-dev</td>
<td>Development files for Classads</td>
</tr>
<tr>
<td>libclassad7</td>
<td>Classad runtime libraries</td>
</tr>
</tbody>
</table>
What to do with all these statistics?

› Aggregate and send them to Ganglia!
  • `condor_gangliad` introduced in v8.2

› In addition to (or instead of) sending to Ganglia, aggregate and make available in JSON format over HTTP
  • `condor_gangliad` rename to `condor_metricd`

› View some basic historical usage out-of-the-box by pointing web browser at central manager (modern CondorView)…

› Or upload to influxdb, graphite for Grafana
Enabled by default and/or easier to configure

- **Enabled by default**: shared port, cgroups, IPv6
  - Have both IPv4 and v6? Prefer IPv4 for now
- **Configured by default**: Kernel tuning
- **Easier to configure**: Enforce slot sizes
  - `use policy: preempt_if_cpus_exceeded`
  - `use policy: hold_if_cpus_exceeded`
  - `use policy: preempt_if_memory_exceeded`
  - `use policy: hold_if_memory_exceeded`
New condor_q default output

- Only show jobs owned by the user
- Batched output (-batch, -nobatch)
- Proposed new default output of condor_q will show summary of current users jobs.

```
-- Submitter: adam      Schedd: submit-3.chtc.wisc.edu
OWNER      IDLE RUNNING   HELD  SUBMITTED  DESCRIPTION  JOBIDs
adam          -       1      -  3/22 07:20 DAG: 221546  230864.0
              -       -      1  3/23 08:57 AtlasAnlysis 263203.0
              -       1      -  3/27 09:37 matlab.exe   307333.0
133      21      -  3/27 11:46 DAG: 311986  312342.0 ... 313304.0

In the last 20 minutes:
  0 Job(s) were Completed
  5 Job(s) were Started                              312690.0 ... 312695.0
  1 Job(s) were Held                        263203.0
  263203.0  5/11 07:22 Error from slot1@eee.chtc.wisc.edu: out of disk
```
New condor_q default output

› Only show jobs owned by the user
   • disable with `--allusers`

› Batched output (`--batch`, `--nobatch`)

› Proposed new default output of condor_q will show summary of current user's jobs.

```plaintext
-- Submitter: adam      Schedd: submit-3.chtc.wisc.edu
OWNER      IDLE RUNNING HELD SUBMITTED  DESCRIPTION  JOBIDs
adam          -       1      -  3/22 07:20 DAG: 221546  230864.0
              -      1 -  3/23 08:57 AtlasAnalysis  263203.0
              -       1 -  3/27 09:37 matlab.exe    307333.0
          133      21      -  3/27 11:46 DAG: 311986  312342.0 ... 313304.0

In the last 20 minutes:
  0 Job(s) were Completed
  5 Job(s) were Started
  1 Job(s) were Held

263203.0  5/11 07:22 Error from slot1@eee.chtc.wisc.edu: out of disk
```
New condor_status default output

- Only show one line of output per machine
- Can try now in v8.5.4+ with "-compact" option
- The "-compact" option will become the new default once we are happy with it

<table>
<thead>
<tr>
<th>Machine</th>
<th>Platform</th>
<th>Slots</th>
<th>Cpus</th>
<th>Gpus</th>
<th>TotalGb</th>
<th>FreCpu</th>
<th>FreeGb</th>
<th>CpuLoad</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>gpu-1</td>
<td>x64/SL6</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>15.57</td>
<td>0</td>
<td>0.44</td>
<td>1.90</td>
<td>Cb</td>
</tr>
<tr>
<td>gpu-2</td>
<td>x64/SL6</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>15.57</td>
<td>0</td>
<td>0.57</td>
<td>1.87</td>
<td>Cb</td>
</tr>
<tr>
<td>gpu-3</td>
<td>x64/SL6</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>47.13</td>
<td>0</td>
<td>16.13</td>
<td>0.85</td>
<td>Cb</td>
</tr>
<tr>
<td>matlab-build</td>
<td>x64/SL6</td>
<td>1</td>
<td>12</td>
<td></td>
<td>23.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mem1</td>
<td>x64/SL6</td>
<td>32</td>
<td>80</td>
<td></td>
<td>1009.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HTCondor and Kerberos

- HTCondor currently allows you to authenticate users and daemons using Kerberos.
- However, it does NOT currently provide any mechanism to provide a Kerberos credential for the actual job to use on the execute slot.
HTCondor and Kerberos/AFS

- So we are adding support to launch jobs with Kerberos tickets / AFS tokens

- Details
  - HTCondor 8.5.X to allows an opaque security credential to be obtained by `condor_submit` and stored securely alongside the queued job (in the `condor_credd` daemon)
  - This credential is then moved with the job to the execute machine
  - Before the job begins executing, the `condor_starter` invokes a call-out to do optional transformations on the credential
Grid Universe

- Reliable, durable submission of a job to a remote scheduler
- Popular way to send pilot jobs
- Supports many “back end” types:
  - HTCondor
  - PBS
  - LSF
  - Grid Engine
  - Google Compute Engine
  - Amazon EC2
  - OpenStack
  - Deltacloud
  - Cream
  - NorduGrid ARC
  - BOINC
  - Globus: GT2, GT5
  - UNICORE
Improved Scalability of Amazon EC2 grid jobs

Number of jobs running on Spot instances in Amazon AWS

[Graph showing the number of jobs running on Spot instances over time, with a peak around 100,000 jobs at 15:00, followed by a decline towards 0 at 18:00.]
Elastically grow your pool into the Cloud: `condor_annex`

- Leverage efficient AWS APIs such as Auto Scaling Groups and Spot Fleets
  - Implement a “lease” so charges cease if lease expires
- Secure mechanism for cloud instances to join the HTCondor pool at home institution

```
condor_annex --set-size 2000
   --lease 24 --project "144PRJ22"
```
Grid Universe support for SLURM, OpenStack, Cobalt

› Speak native SLURM protocol
  • No need to install PBS compatibility package

› Speak OpenStack’s NOVA protocol

› Speak to Cobalt Scheduler
  • Argonne Leadership Computing Facilities

Jaime: Grid Jedi
Transformation of job ad upon submit

› Allow admin to have the schedd add/edit job attributes upon submission
  (use case: insert trusted group attributes based upon owner)

› In v8.5.1+ can also set attributes as immutable by the user
  › Prevent user from editing protected attributes with condor_qedit or chirp
Docker Universe Enhancements

- Docker jobs get usage updates (i.e. network usage) reported in job classad
- Admin can add additional volumes
  - That all docker universe jobs get
  - Why?
    - CVMFS
    - Large shared data
  - Details

Potential Future Docker Universe Features?

› Advertise images already cached on machine?
› Support for condor_ssh_to_job?
› Package and release HTCondor into Docker Hub?
› Network support beyond NAT?
› Run containers as root??!?!?
› Automatic checkpoint and restart of containers! (via CRIU)
SELinux and systemd

- SELinux
  - (On by default in RHEL 7)

- Systemd Integration
  - Port Reservation - Systemd will reserve 9618 for HTCondor
  - Watchdog - If masters stops responding, systemd will restart it
  - Status messages - display via systemctl status
  - Logging - Daemon log messages can go to systemd-journald
Draining jobs from execute nodes

- Add ability to backfill with pre-emptable jobs while draining
  - Specifically, ability to specify a new startd START expression when entering drain state
- Add ability to shutdown when fully drained
  - Alternative to condor_off -peaceful
- Investigating ability to upgrade HTCondor on execute nodes without restarting jobs
DAGMan Improvements

- Splice Pin connections
  - Allows more flexible parent/child relationships between nodes within splices
  - Parsed when DAGMan starts up
- INCLUDE directive
- Set ClassAd attributes in DAG
- Set Batch Name
Seeking ideas to help users and admins learn

› Move HOWTO recipes on wiki to stackoverflow?
› Sub-reddit instead of email list?
› YouTube videos?
Smarter and Faster Schedd

- User accounting information moved into ads in the Collector
  - Enable schedd to move claims across users
- Non-blocking authentication, smarter updates to the collector, faster ClassAd processing
- *Late materialization of jobs in the schedd* to enable submission of very large sets of jobs
  - More jobs materialized once number of idle jobs drops below a threshold (like DAGMan throttling)
Thank You!

P.S. Interested in working on HTCondor full time? Talk to me! We are hiring! htcodnor-jobs@cs.wisc.edu