The HTCondor-CE

Brian Bockelman
HTCondor / Arc-CE Workshop, February 2016
HTCondor-CE

• In summer 2013, OSG began an evaluation of its choice of CE technology.

• Did we want to keep the same technology? Try a new one?

• One of the more intriguing possibilities: Could we construct a CE from a special configuration of HTCondor?

• We’ll get to the technical aspects later, but this was a unique opportunity: no new dependency on an external team.

Out of this work came the HTCondor-CE
What’s in a CE?

• A CE must:
  
  • Expose a **remote API** for resource acquisition.
  
  • Provide authentication and **authorization**.
  
  • Interact with the **resource layer** (batch system).

Note we deal with resource acquisition and provisioning, *not* job submission!
Anatomy of a Compute Element (CE)
HTCondor-CE

• HTCondor already has many of the pieces necessary:
  
  • Remote job submission is possible.
  
  • Extensive authentication and authorization system (including GSI).
  
  • Grid universe integration with blahp (same underlying component as CREAM) allows submission to other batch systems.
  
  • JobRouter provides transformation
  
• Simply need to put things together!
Anatomy of HTCondor-CE: HTCondor Batch System
Anatomy of HTCondor-CE: Non-HTCondor Batch System
HTCondor-CE

• Special configuration of HTCondor.
  
  • (Mostly - historically shipped a few custom ClassAd functions.)
  
• Installs small wrappers around condor CLI (i.e., `condor_ce_status` sets a few config variables and calls `condor_status`).
  
• Runs a complete set of condor daemons:
  
  • Port 9619 (instead of 9618).
  
  • Configs from `/etc/condor-ce` instead of `/etc/condor`.
  
  • Separate `condor_master` process and Linux service (`condor-ce`).
Running Daemons

<table>
<thead>
<tr>
<th>Process</th>
<th>Status</th>
<th>Username</th>
<th>CPU</th>
<th>Memory</th>
<th>Disk</th>
<th>Time</th>
<th>Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>condor</td>
<td>2495</td>
<td>root</td>
<td>0.0</td>
<td>103072</td>
<td>7080</td>
<td>Feb 18</td>
<td>condor_master -pidfile /var/run/condor-ce/condor_master.pid</td>
</tr>
<tr>
<td>root</td>
<td>2518</td>
<td>root</td>
<td>0.1</td>
<td>24524</td>
<td>6100</td>
<td>Feb 18</td>
<td>condor_procd -A /var/lock/condor-ce/procd_pipe -L /var/log/condor-ce/ProcLog -R</td>
</tr>
<tr>
<td>condor</td>
<td>2519</td>
<td>root</td>
<td>0.0</td>
<td>102368</td>
<td>4604</td>
<td>Feb 18</td>
<td>condor_shared_port -f -p 9619</td>
</tr>
<tr>
<td>condor</td>
<td>2521</td>
<td>root</td>
<td>0.8</td>
<td>400144</td>
<td>175800</td>
<td>Feb 18</td>
<td>condor_collector -f -port 9619</td>
</tr>
<tr>
<td>condor</td>
<td>2523</td>
<td>root</td>
<td>0.5</td>
<td>176504</td>
<td>66132</td>
<td>Feb 18</td>
<td>condor_schedd -f</td>
</tr>
<tr>
<td>condor</td>
<td>2524</td>
<td>root</td>
<td>1.4</td>
<td>205100</td>
<td>100888</td>
<td>Feb 18</td>
<td>condor_job_router -f</td>
</tr>
<tr>
<td>condor</td>
<td>2742</td>
<td>root</td>
<td>0.0</td>
<td>97504</td>
<td>7620</td>
<td>Feb 18</td>
<td>condor_master -pidfile /var/run/condor-ce/condor_master.pid</td>
</tr>
<tr>
<td>root</td>
<td>2750</td>
<td>root</td>
<td>0.1</td>
<td>24616</td>
<td>6116</td>
<td>Feb 18</td>
<td>condor_procd -A /var/lock/condor-ce/procd_pipe -L /var/log/condor-ce/ProcLog -R</td>
</tr>
<tr>
<td>condor</td>
<td>2751</td>
<td>root</td>
<td>0.2</td>
<td>200520</td>
<td>101812</td>
<td>Feb 18</td>
<td>condor_schedd -f</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3033878</td>
<td>root</td>
<td>0.0</td>
<td>94604</td>
<td>8152</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821805.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3041926</td>
<td>root</td>
<td>0.0</td>
<td>94604</td>
<td>8184</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821815.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3041927</td>
<td>root</td>
<td>0.0</td>
<td>94604</td>
<td>8196</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821814.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3043312</td>
<td>root</td>
<td>0.0</td>
<td>94604</td>
<td>8184</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821825.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3056870</td>
<td>root</td>
<td>0.0</td>
<td>94604</td>
<td>8184</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821848.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3057151</td>
<td>root</td>
<td>0.0</td>
<td>94584</td>
<td>8184</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821849.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3061095</td>
<td>root</td>
<td>0.0</td>
<td>94604</td>
<td>8176</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821852.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3066118</td>
<td>root</td>
<td>0.0</td>
<td>94600</td>
<td>8176</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821857.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3070732</td>
<td>root</td>
<td>0.0</td>
<td>94600</td>
<td>8132</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821864.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3073572</td>
<td>root</td>
<td>0.0</td>
<td>94600</td>
<td>8144</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821866.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3078308</td>
<td>root</td>
<td>0.0</td>
<td>94600</td>
<td>8136</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821886.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3084233</td>
<td>root</td>
<td>0.0</td>
<td>94600</td>
<td>8180</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821888.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3092091</td>
<td>root</td>
<td>0.0</td>
<td>94600</td>
<td>8172</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821889.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3099541</td>
<td>root</td>
<td>0.0</td>
<td>94604</td>
<td>8176</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821897.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3105248</td>
<td>root</td>
<td>0.0</td>
<td>94600</td>
<td>8140</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821932.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
<tr>
<td>cmsprod</td>
<td>3107777</td>
<td>root</td>
<td>0.0</td>
<td>94604</td>
<td>8148</td>
<td>Feb 26</td>
<td>condor_shadow -f 5821943.0 --schedd=&lt;129.93.239.132:39830?addr[2600-900-6</td>
</tr>
</tbody>
</table>
JobRouter

• The JobRouter is responsible for taking a job, creating a copy, and changing the copy according to a set of rules.

  • When running in HTCondor, the copy is inserted directly into the site batch schedd.

  • Each chain of rules is called a “route” and is defined by a ClassAd.

• Once the copy has been created, attribute changes and state changes are propagated between the source and destination jobs.

• JobRouter directly accesses the schedd’s transaction log: most efficient way of mirroring jobs!
Example HTCondor Job Route

Cameron has an HTCondor pool and she wants CMS jobs submitted to her CE to be forwarded to her pool and requesting x86_64 Linux machines and setting the attribute “foo” on her routed job to “bar”. All other jobs should be submitted to the pool without any changes.

```
JOB_ROUTER_ENTRIES = [ 
    name = "condor_pool_cms"; 
    TargetUniverse = 5; 
    Requirements = target.x509UserProxyVOName =?= "cms"; 
    set_requirements = (Arch == "X86_64") && (TARGET.OpSys == "LINUX"); 
    set_foo = "bar"; 
]

[ 

    name = "condor_pool_other"; 
    TargetUniverse = 5; 
    Requirements = target.x509UserProxyVOName !== "cms"; 
]
```

Example PBS Job Route

Cameron has a PBS pool and she wants CMS jobs submitted to her CE to be forwarded to her pool. All other jobs should be submitted to her pool without any changes

```python
JOB_ROUTER_ENTRIES = [ \
    name = "pbs_pool_cms"; \ 
    TargetUniverse = 9; \ 
    GridResource = "batch pbs"; \ 
    Requirements = target.x509UserProxyVOName =?= "cms"; \ 
] \ 
[ \ 
    name = "pbs_pool_other"; \ 
    TargetUniverse = 9; \ 
    GridResource = "batch pbs"; \ 
    Requirements = target.x509UserProxyVOName =!= "cms"; \ 
]
```

Where does a pilot go?

Pilot Factory

Factory process (non-condor) \( \rightarrow \text{(CEDAR)} \rightarrow \text{condor_submit} \rightarrow \text{Schedd} \rightarrow \text{Gridmanager/GAHP} \)

HTCondor-CE

- CE Schedd
- JobRouter
- Batch Schedd
Where does a pilot go?
Where does a pilot go?
Where does a pilot go?

![Diagram of HTCondor-CE architecture with components like Pilot Factory, Schedd, Gridmanager/GAHP, CE Schedd, JobRouter, and Batch Schedd.]

- Pilot Factory
  - Factory process (non-condor)
  - Schedd
  - Gridmanager/GAHP

- HTCondor-CE
  - CE Schedd
  - JobRouter
  - Batch Schedd
Client Tools

- `condor_ce_trace`: Test each step of job submission individually; determine where failures may occur.

- `condor_ce_run`: Run a single job against a remote host (either local or through batch; great for debugging!).

- `condor_ce_ping`: Test authorization for various actions (read, write, administer).
condor_ce_trace

[bbockelm@hcc-briantest ~]$ condor_ce_trace red.unl.edu
Testing HTCondor-CE collector connectivity.

2016-02-28 11:07:05 Failed to ping <2600:9006:1101:5054:ff:fe76:711a:9619>; authorization check exited with code 1. Re-run the command with '-d' for more verbose output.

[bbockelm@hcc-briantest ~]$ condor_ce_trace tusker-gw1.unl.edu
Testing HTCondor-CE collector connectivity.
- Successful ping of collector on <129.93.227.123:9619>.

Testing HTCondor-CE schedd connectivity.

[ ]
- Machine = "tusker-gw1.unl.edu";
- CondorPlatform = "$CondorPlatform: X86_64-CentOS_6.6 $";
- Name = "tusker-gw1.unl.edu";
- MyType = "Scheduler";
- MyAddress = "<129.93.227.123:9619?noUDP&sock=5472_8b22_23>";
- CondorVersion = "$CondorVersion: 8.3.5 Apr 06 2015 $"

Submitting job to schedd <129.93.227.123:9619?noUDP&sock=5472_8b22_23>
- Successful submission; cluster ID 3071635
Resulting job ad:
[ ]
- BufferSize = 524288;
- NiceUser = false;
- CoreSize = -1;
- CumulativeSlotTime = 0;
- OnExitHold = false;
- RequestCpus = 1;
condor_ce_ping

```
[bbockelm@hcc-briantest ~]$ condor_ce_ping -pool tusker-gw1.unl.edu -name tusker-gw1.unl.edu -table ALL

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Authentication</th>
<th>Encryption</th>
<th>Integrity</th>
<th>Decision</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLOW</td>
<td>GSI</td>
<td>none</td>
<td>MD5</td>
<td>ALLOW</td>
<td><a href="mailto:uscmsPool018@users.opensciencegrid.org">uscmsPool018@users.opensciencegrid.org</a></td>
</tr>
<tr>
<td>READ</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>ALLOW</td>
<td>unauthenticated@unmapped</td>
</tr>
<tr>
<td>WRITE</td>
<td>GSI</td>
<td>none</td>
<td>MD5</td>
<td>ALLOW</td>
<td><a href="mailto:uscmsPool018@users.opensciencegrid.org">uscmsPool018@users.opensciencegrid.org</a></td>
</tr>
<tr>
<td>NEGOTIATOR</td>
<td>GSI</td>
<td>none</td>
<td>MD5</td>
<td>DENY</td>
<td><a href="mailto:uscmsPool018@users.opensciencegrid.org">uscmsPool018@users.opensciencegrid.org</a></td>
</tr>
<tr>
<td>ADMINISTRATOR</td>
<td>GSI</td>
<td>none</td>
<td>MD5</td>
<td>DENY</td>
<td><a href="mailto:uscmsPool018@users.opensciencegrid.org">uscmsPool018@users.opensciencegrid.org</a></td>
</tr>
<tr>
<td>OWNER</td>
<td>GSI</td>
<td>none</td>
<td>MD5</td>
<td>DENY</td>
<td><a href="mailto:uscmsPool018@users.opensciencegrid.org">uscmsPool018@users.opensciencegrid.org</a></td>
</tr>
<tr>
<td>CONFIG</td>
<td>GSI</td>
<td>none</td>
<td>MD5</td>
<td>DENY</td>
<td><a href="mailto:uscmsPool018@users.opensciencegrid.org">uscmsPool018@users.opensciencegrid.org</a></td>
</tr>
<tr>
<td>DAEMON</td>
<td>GSI</td>
<td>none</td>
<td>MD5</td>
<td>DENY</td>
<td><a href="mailto:uscmsPool018@users.opensciencegrid.org">uscmsPool018@users.opensciencegrid.org</a></td>
</tr>
<tr>
<td>ADVERTISE_STARTD</td>
<td>GSI</td>
<td>none</td>
<td>MD5</td>
<td>DENY</td>
<td><a href="mailto:uscmsPool018@users.opensciencegrid.org">uscmsPool018@users.opensciencegrid.org</a></td>
</tr>
<tr>
<td>ADVERTISE_SCHEDD</td>
<td>GSI</td>
<td>none</td>
<td>MD5</td>
<td>DENY</td>
<td><a href="mailto:uscmsPool018@users.opensciencegrid.org">uscmsPool018@users.opensciencegrid.org</a></td>
</tr>
<tr>
<td>ADVERTISE_MASTER</td>
<td>GSI</td>
<td>none</td>
<td>MD5</td>
<td>DENY</td>
<td><a href="mailto:uscmsPool018@users.opensciencegrid.org">uscmsPool018@users.opensciencegrid.org</a></td>
</tr>
</tbody>
</table>
```
```
Deployment in OSG

- Latest version is 2.0.0.
- 60 CEs registered with the central collector.
  - 60% are on the latest version.
  - 70% are HTCondor
Interaction Examples
```
Worker Node | State | Payload ID | User      | Scheduler                                                    |
------------|-------|------------|-----------|--------------------------------------------------------------|
red-c0001.unl.edu | Unclaimed |          |           |                                                              |
red-c0001.unl.edu | Unclaimed |          |           |                                                              |
red-c0001.unl.edu | Unclaimed |          |           |                                                              |
red-c0001.unl.edu | Unclaimed |          |           |                                                              |
red-c0001.unl.edu | Unclaimed |          |           |                                                              |
red-c0001.unl.edu | Unclaimed |          |           |                                                              |
red-c0001.unl.edu | Unclaimed |          |           |                                                              |
red-c0001.unl.edu | Unclaimed |          |           |                                                              |
red-c0001.unl.edu | Unclaimed |          |           |                                                              |
red-c0001.unl.edu | Unclaimed |          |           |                                                              |
red-c0001.unl.edu | Unclaimed |          |           |                                                              |
scheduled for 18730904.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18730904.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18730904.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 21597039.0 | yxS | 4@xd-login.opengrid.org |                                                              |
scheduled for 18739992.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18712583.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18742089.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18735888.6205 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18730904.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18730904.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18730904.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18730904.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18730904.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18730904.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18730904.0 | zcx | login01 osgconnect.net |                                                              |
scheduled for 18730904.0 | zcx | login01 osgconnect.net |                                                              |
```

condor_ce_status -schedd
Job Query

```
[<root@red-gw1 ~]> condor_ce_q

-- Schedd: red-gw1.unl.edu : <129.93.239.132:28466>
ID  OWNER  SUBMITTED  RUN_TIME   ST  PRI  SIZE  CMD
1505510.0  fermilab  3/27 17:20  +0:00:03  H  0  0:0  whoami
1506580.0  fermilab  3/27 21:28  +0:00:03  H  0  0:0  whoami
1518799.0  fermilab  3/31 15:08  +0:00:03  H  0  0:0  whoami
1802269.0  fermilab  6/2 10:12  +0:00:04  H  0  0:0  whoami
1802270.0  fermilab  6/2 10:15  +0:00:04  H  0  0:0  whoami
1923583.0  fermilab  6/24 13:16  +0:00:04  H  0  0:0  whoami
1923788.0  fermilab  6/24 14:27  +0:00:04  H  0  0:0  whoami
2670540.0  glow  12/11 05:40  +0:06:51  C  0  195.3  glidein_startup.sh
2677852.0  glow  12/10 03:59  +0:06:21  C  0  195.3  glidein_startup.sh
2738000.0  glow  12/30 18:39  +0:04:47  C  0  9.8  glidein_startup.sh
2738113.0  glow  12/30 19:17  +0:05:26  C  0  14.6  glidein_startup.sh
2738114.0  glow  12/30 19:17  +0:05:26  C  0  14.6  glidein_startup.sh
2738115.0  glow  12/30 19:17  +0:05:26  C  0  14.6  glidein_startup.sh
2738145.0  glow  12/30 19:25  +0:02:40  C  0  12.2  glidein_startup.sh
2741874.0  glow  12/31 23:13  +0:08:23  C  0  17.1  glidein_startup.sh
2741880.0  glow  12/31 23:16  +0:08:22  C  0  14.6  glidein_startup.sh
2744310.0  glow  1/1 17:57  +0:08:02  C  0  14.6  glidein_startup.sh
2753580.0  glow  1/3 06:10  +0:08:24  C  0  14.6  glidein_startup.sh
2758819.0  glow  1/3 21:37  +0:04:24  C  0  14.6  glidein_startup.sh
2758843.0  glow  1/3 21:42  +0:04:21  C  0  14.6  glidein_startup.sh
2758845.0  glow  1/3 21:42  +0:04:20  C  0  17.1  glidein_startup.sh
2759289.0  glow  1/3 23:46  +0:02:22  C  0  293.0  glidein_startup.sh
2759291.0  glow  1/3 23:46  +0:02:24  C  0  293.0  glidein_startup.sh
```
Why Consider this CE?

• If you are using HTCondor for batch:
  • One less software provider - same thing all the way down the stack.
  • HTCondor has an extensive feature set - easy to take advantage of it (i.e., Docker universe).

• Regardless, a few advantages:
  • Can scale well (up to at least 16k; maybe higher).
  • Declarative ClassAd-based language.

• But disadvantages exist:
  • Non-HTCondor backends are finicky outside PBS and SLURM.
  • Declarative ClassAd-based language.
Conclusions

• We believe the HTCondor-CE is a drastically different approach to the classic CE.

• It brings quite a few concepts forward from the underlying HTCondor system.

• It has special advantages for HTCondor sites, especially in terms of support and existing knowledge.

• Explicitly supported by the OSG, but we’ve seen a few external sites pop up recently.

• We hope to work with a wider community!