Progress Report on the HTCondor-CE

Brian Bockelman

HTCondor Week Europe

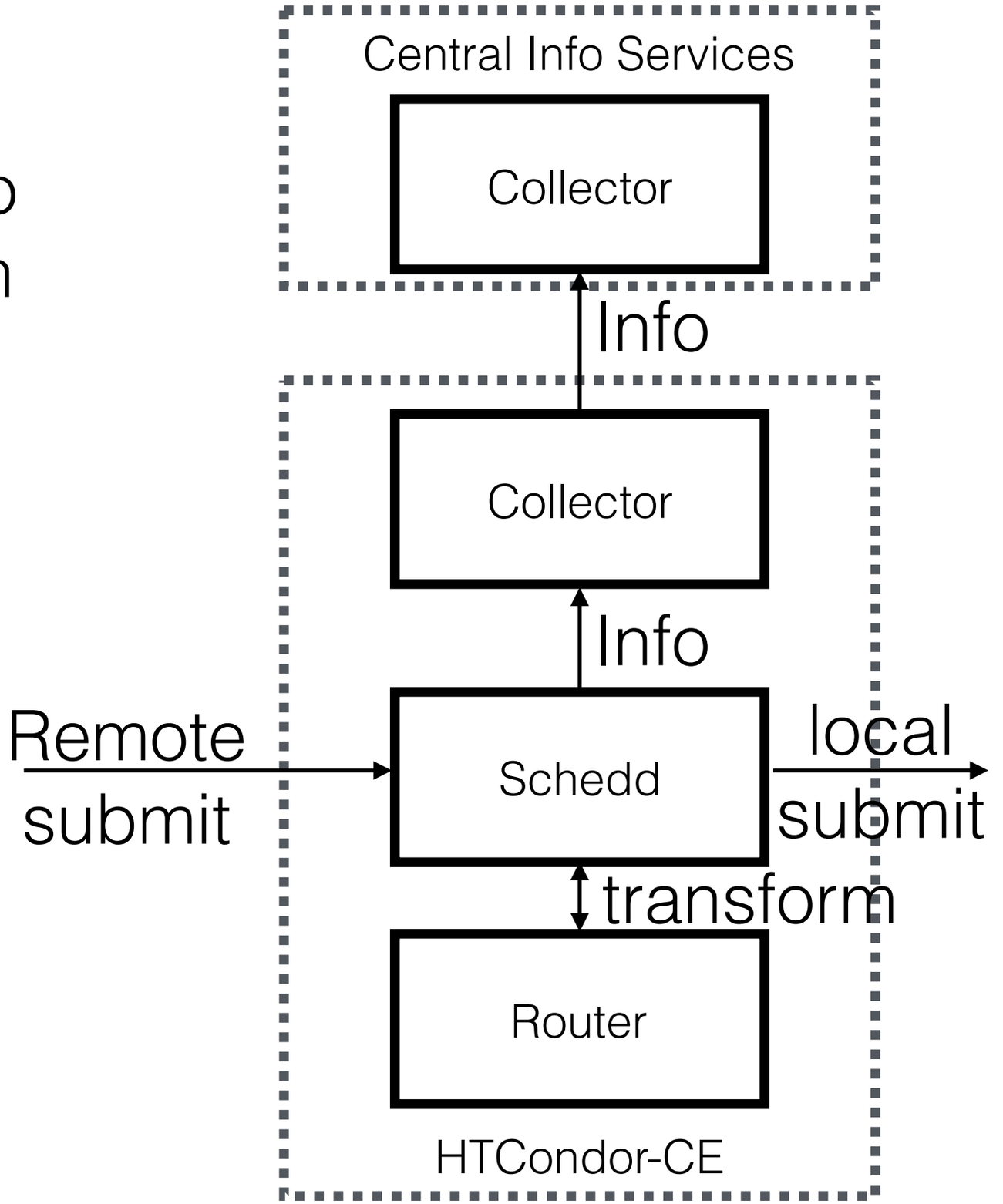
The HTCondor-CE

- What is needed in a CE?
  - Secure authentication.
  - Remote submission.
  - Support for a variety
- The HTCondor-CE started with the observation that **HTCondor already provides this functionality**.

The HTCondor-CE

- The HTCondor-CE aims to be a special configuration of HTCondor that performs the functionality of the site's CE.

- Mostly achieved: code that exists is for wrappers, config generation, or debugging utilities.



Where the Magic Happens

- HTCondor-CE is hosted on GitHub under the OSG organization.
  - https://github.com/opensciencegrid/htcondor-ce
- Currently at version 2.2.0.
- We welcome external contributions: features, bug fixes, new regression tests!
  - Would love feedback on how to make contributing easier.
- Try to maintain a monthly release cadence to get your favorite fix “on the street” quickly!

Battle-Hardened

- HTCondor-CE scales from the smallest sites to the largest.
  - Performs reliably at 10k pilot jobs; up to 20k in testing.
- About 75 CEs in the OSG collector; estimated 100 worldwide.
  - 60% of CEs are on top of a HTCondor batch system; 13% PBS; 13% SLURM; 13% other/unknown.

What's New?

- What have we been up to in the last year?
  - Hosted CE service.
  - Improved integration with AGIS.
  - Automated regression test suite.
  - Variety of code cleanups.

Remote CE Service

- By default, HTCondor-CE submits to a *local* batch system.
  - This utilizes the schedd's "Condor-G" mode, which allows an external system (the site's batch system) manage the job execution.
  - Actual interface with batch system done by `blahp`.
- Prior work (BOSCO) extended HTCondor so it can submit to a *remote* batch system.
  - The `blahp` is managed remotely over a SSH connection.
  - Hence, the **HTCondor-CE can run on host A submitting to host B over SSH**.
- Particularly useful for sites where there's a strong split between the batch system and grid teams.

Hosted CE Service

- Why stop there? HTCondor-CE on site A can submit to a batch system on site B.

- Several small OSG sites have really struggled to find enough effort to run a CE.

- **OSG will run a hosted HTCondor-CE**, connecting to your site via SSH.
  - Idea: a simple SSH connection is the lightest-weight way to “get in” to a site.
  - Caveat: this means you are delegating a lot of scheduling decisions to OSG instead of locally.

- Scaling? Not 100% clear: suggestion is to keep this to <1k pilots per CE.

https://indico.fnal.gov/getFile.py/access?contribId=6&sessionId=12&resId=0&materialId=slides&confId=12973

Information Services

- We tend to think of `condor_collector` as simply holding machine status - default output of `condor_status`. However, it also contains:
  - Submitter and fairshare information.
  - Performance statistics of the various daemons.
  - DNS-like location of each daemon.
- Basically, the collector can be used a generic message board!

Information Service

- For the schedd ad, HTCondor-CE injects information about:
  - Allowed VOs
  - Available resources.
  - How to allocate resources.
  - CE information (site name)
- All ClassAd and matchmaking based!
- The schedd ad is forwarded to a central collector. There, a process serves the information in several formats.
  - In 2017, we added an AGIS-specific JSON.

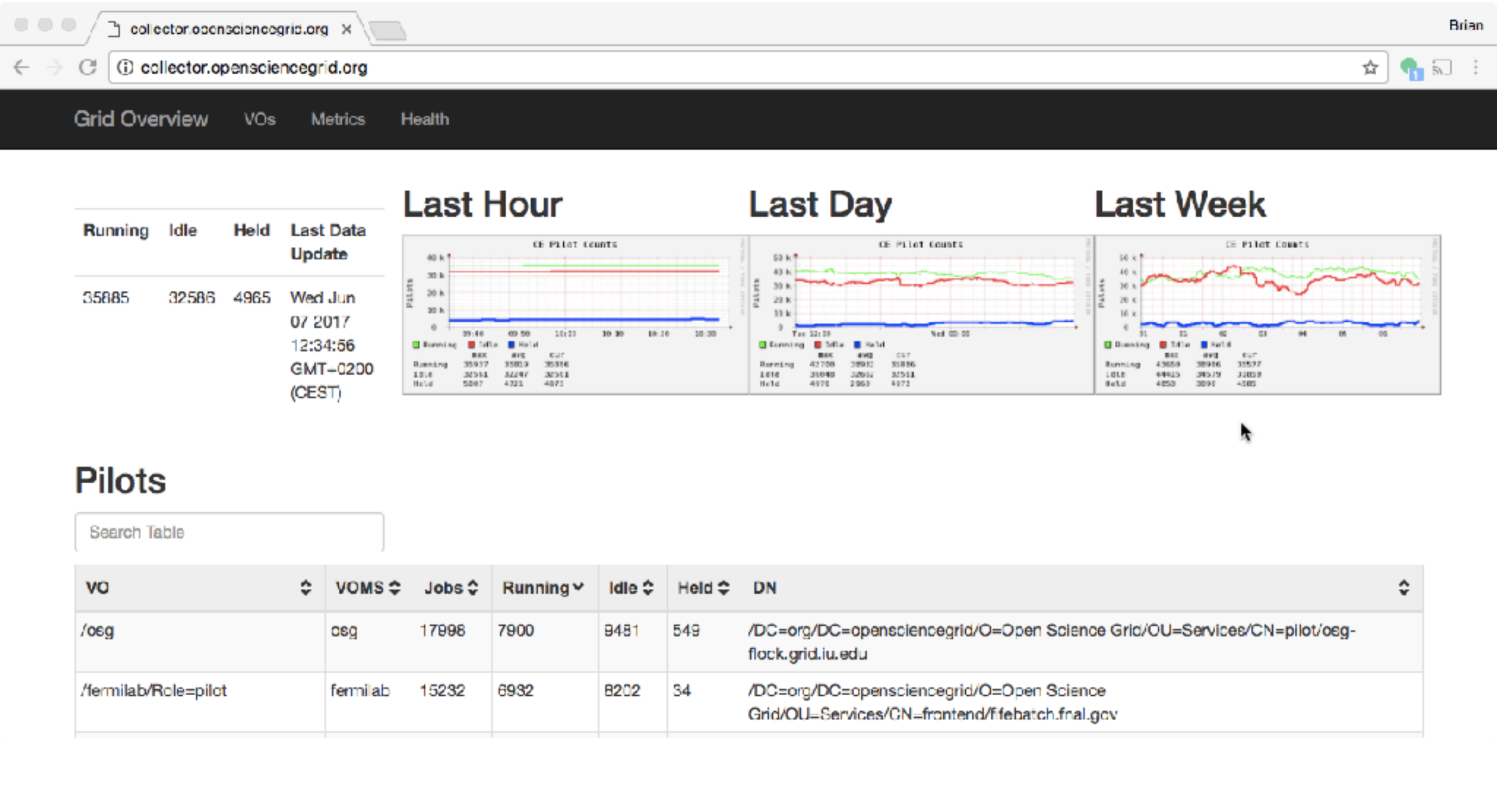
bbockelm — bbockelm@hcc-briantest7:~ — ssh hcc-briantest7.u...

```
[bbockelm@hcc-briantest7 ~]$ condor_ce_status -schedd -pool collector.openscienc
egrid.org
Name                 Machine              RunningJobs   IdleJobs   HeldJobs

CE01.CMSAF.MIT.EDU   CE01.CMSAF.MIT.EDU           264        799         14
CE02.CMSAF.MIT.EDU   CE02.CMSAF.MIT.EDU           126         16          0
CE03.CMSAF.MIT.EDU   CE03.CMSAF.MIT.EDU           147         29          0
atlas-ce.bu.edu      atlas-ce.bu.edu              654        380          0
bonner06.rice.edu    bonner06.rice.edu             11          2          4
ce.grid.unesp.br     ce.grid.unesp.br               0          0          0
ce01.brazos.tamu.edu ce01.brazos.tamu.edu           2          2        354
ce1.accre.vanderbilt.e ce1.accre.vanderbilt.e     586        152        321
cecc7test.hep.wisc.edu cecc7test.hep.wisc.edu       0          0          0
cit-gatekeeper.ultrali cit-gatekeeper.ultrali     632        219          7
```

bbockelm — bbockelm@hcc-briantest7:~ — ssh hcc-briantest7.unl.edu —

```
[bbockelm@hcc-briantest7 ~]$ condor_ce_info_status
Name                    CPUs   Memory  MaxWallTime AllowedVOs

T2_US_Nebraska Dell SC     4     8053         1440 cms, belle, cigi, des, fermilab, g
GridUNESP                  8    16384          720 cdf, gridunesp, mis, star, cms, dz
OrangeGrid                 1     4000         1440
gpgrid.fnal.gov            8    12010         1440 osg, minos, nova, cdf, minerva, ma
OU_OSCER_ATLAS_2650       20    32768         1440 atlas, dosar
OU_OSCER_ATLAS_2670       24    65536         1440 atlas, dosar
GridUNESP_CENTRAL          8    16384          720 gridunesp, cdf, fermilab, accelera
IceCube                   32    64000        43200 glow, icecube, osg
gpgrid.fnal.gov            8    12010         1440 osg, minos, nova, cdf, minerva, ma
GLOW CE                    8    16030         1440
GLOW g12                   8    16384         1440
GLOW g14                   8    16384         1440
GLOW g18                  16    48305         1440
GLOW g19                  24    48305         1440
GLOW g20                  16    32049         1440
GLOW g22                  24    64335         1440
GLOW g23                  24    64335         1440
GLOW g24                  24    64335         1440
GLOW g26                  32    96000         1440
GLOW g27                  40   128000         1440
GLOW g28                  40   128000         1440
GLOW g29                  40   128000         1440
GLOW g30                  40   128000         1440
```

Information Service



Whole Node Jobs

- WLCG tends to divide its resources into:
  - 1-core / 2GB RAM slots
  - 8-core / 16GB RAM slots
- This does not always map neatly onto existing hardware, leaving either CPU or memory underutilized by the batch system.
- HTCondor-CE 2.2.0 supports the concept of a “whole node pilot”.
  - Set `WantWholeNode=True` in job.
  - Currently in testing at Nebraska, where we have some very strange hardware (56 HT-cores / 256GB RAM).
- Potential to massive simplify sites — *if* it meets the local needs.

Interfacing with OpenStack

- The job router is an extraordinarily powerful transformation mechanism.
    - Whole node jobs are a great example of this.
- The transforms themselves can be an arbitrary script.
- For example, one could transform a pilot job into a VM universe job.
- Or launch a VM on OpenStack corresponding to the incoming pilot.
- See: http://research.cs.wisc.edu/htcondor/HTCondorWeek2017/presentations/ThuCaballero_OpenStack.pdf

Auditing Payload Jobs

- What job is the pilot running? Who "owns" the pilot job?
  - Devilishly difficult question to answer!
  - One approach is to require **each payload to have an X509** certificate and have the **pilot invoke a setuid binary** to record the certificate.
- Alternate approach: when pilot launches a payload it asserts the user's identity.
  - Uses `condor_advertise` (or C library) to push this information to the `condor_collector` on the CE.

Auditing Payload Jobs

- `condor_ce_status` shows a snapshot of payload jobs currently running (nicely formatted).

- New (HTCondor 8.6.3), the `condor_collector` can invoke a python function for each new ad.

- Expected (HTCondor-CE 2.2.1), we will log all payload ownership information to disk.

bbockelm — root@red:~ — ssh hcc-briantest7.unl.edu — 95

```
[root@red ~]# condor_ce_status | head -n 24
Worker Node          State     Payload ID      User       Scheduler

cmsprod-660286.0-red Unclaimed
cmsprod-660286.0-red Claimed   7665986.0       cms287     crab3@vocms0196.cer
cmsprod-660286.0-red Claimed   7665935.0       cms287     crab3@vocms0196.cer
cmsprod-660286.0-red Claimed   7665928.0       cms287     crab3@vocms0196.cer
cmsprod-660286.0-red Claimed   7665929.0       cms287     crab3@vocms0196.cer
cmsprod-660286.0-red Claimed   7665990.0       cms287     crab3@vocms0196.cer
cmsprod-660286.0-red Claimed   6877117.0       cms1104    crab3@vocms0197.cer
cmsprod-660286.0-red Claimed   10545027.0      cms287     crab3@vocms0121.cer
cmsprod-660286.0-red Claimed   9560711.0       cms1868    crab3@vocms0107.cer
cmsprod-660286.0-red Claimed   7665842.0       cms287     crab3@vocms0196.cer
cmsprod-660286.0-red Claimed   7665783.0       cms287     crab3@vocms0196.cer
cmsprod-660286.0-red Claimed   7665798.0       cms287     crab3@vocms0196.cer
cmsprod-660286.0-red Claimed   9556936.0       cms287     crab3@vocms0107.cer
cmsprod-660286.0-red Claimed   7665872.0       cms287     crab3@vocms0196.cer
cmsprod-660286.0-red Claimed   9560748.0       cms1868    crab3@vocms0107.cer
cmsprod-660286.0-red Claimed   9556950.0       cms287     crab3@vocms0107.cer
cmsprod-660286.0-red Claimed   7665810.0       cms287     crab3@vocms0196.cer
cmsprod-660286.0-red Unclaimed
cmsprod-660416.0-red Unclaimed
cmsprod-660416.0-red Claimed   6902320.0       cms1046    crab3@vocms0197.cer
cmsprod-660416.0-red Claimed   7665825.0       cms287     crab3@vocms0196.cer
cmsprod-660416.0-red Claimed   10548537.0      cms336     crab3@vocms0121.cer
[root@red ~]#
```

Authorization Overhaul

- OSG is in the midst of an authorization overhaul in 2017:
    - Retiring support for our central authorization service, GUMS.
    - Focusing on mapfiles / ban-files for VOMS FQANs and DNs.
    - *Finally* dropping support for `edg-mkgridmap`.
- Achievement: Due to the use of LCMAPS, the only change was to remove an OSG-specific LCMAPS config.

Quality of Life Fixes

- **Less code than last year!** Back to pure python.
    - Dropped custom C++ ClassAd functions. Upstreamed functionality to HTCondor.
        - Now assume condor-python bindings are installed.
- External contributor fixed up GLUE2 bindings.
- Purge the old Globus `xcount` semantics - switch to `RequestCpus`.
- Small UI improvements to the builtin CEView.
- Explicit SLURM support.

Where do we need to go?

- Better distribution.
- Improve accounting.
- Plan for closed-source Globus Toolkit.
- Alternate authentication libraries.

Better Distribution

- Right now, the HTCondor-CE RPM is distributed as part of the OSG Software stack.
    - Source tarballs are available on the relevant GitHub page.
- Additional channels would be beneficial:
    - **Docker**: The current community standard for distributing services. Minimizes effort needed to try out HTCondor-CE!
    - **Debian**: Important for reaching the world beyond the WLCG!
    - **EPEL**: Base community repo for RHEL.
- More non-OSG distribution channels would help decouple OSG-specific components from HTCondor-CE.
- Looking for collaborators!

Improved Accounting

- Each pilot job is recorded in the CE's history log.
  - For OSG accounting, a batch-system-specific probe correlates the CE information (pilot DN) with batch information (i.e., CPU / memory usage).
  - Thus, **OSG maintains one probe per batch system** :(.
    - With minor modification, HTCondor can add the required attributes to the CE ad. Already there for SLURM and HTCondor!
- Goal: **One accounting probe** to rule them all!
- Goal: **Integrate probe into the CE** itself. Zero-config for OSG sites (or run a single probe centrally for all of OSG).
- Stretch goal: Need **similar integration with APEL**.

Globus Toolkit -> Closed Source

- The grid community is still digesting last week's announcement that Globus Toolkit is going closed source.

- HTCondor-CE currently uses GT for *authentication* (GSI) and the authorization callout plugin layer.

- Preliminary strategy:

    - Simplify use of GSI: Look into **dropping support for legacy-** and GT3-style proxies. OpenSSL itself will authenticate RFC proxies: the GT libraries are much thinner.

    - Directly invoke LCMAPS instead of using abstraction layer.

- Do not currently plan on writing a new library or pulling relevant code into HTCondor proper: will rely on OSG's support guarantee.

Next-Gen AAI

- Globus Toolkit going closed source is a *harsh reminder of the need to modernize our authorization and authentication infrastructure (AAI).* Ongoing overhaul is not enough!
- HTCondor's authentication handshake includes authentication protocol negotiation: provides a mechanism to cleanly transition to a new authentication mechanism.
- No clear proposal yet, but components may be:
  - **Bearer tokens** a-la JSON Web Tokens or macarons.
  - Macaron-style **attenuation** or GSI-style delegation.
  - **Capability-based** (a-la VOMS FQAN) instead of identity-based.
- Other groups - dCache, INDIGO - are headed in a similar direction.
- Community previously rallied around an interoperable XACML profile. **Is there a similar opportunity here?**

Final Thoughts

- The HTCondor-CE “core” has become relatively stable.
    - Focus on the last year has been hardening and incremental features.
        - How can we make it easier to monitor?
        - Enjoy the fact the overall package is smaller! Less code = (hopefully) better support.
- Looking to explore expanding the footprint of the base install.