HTCondor-CE
Overview and Architecture

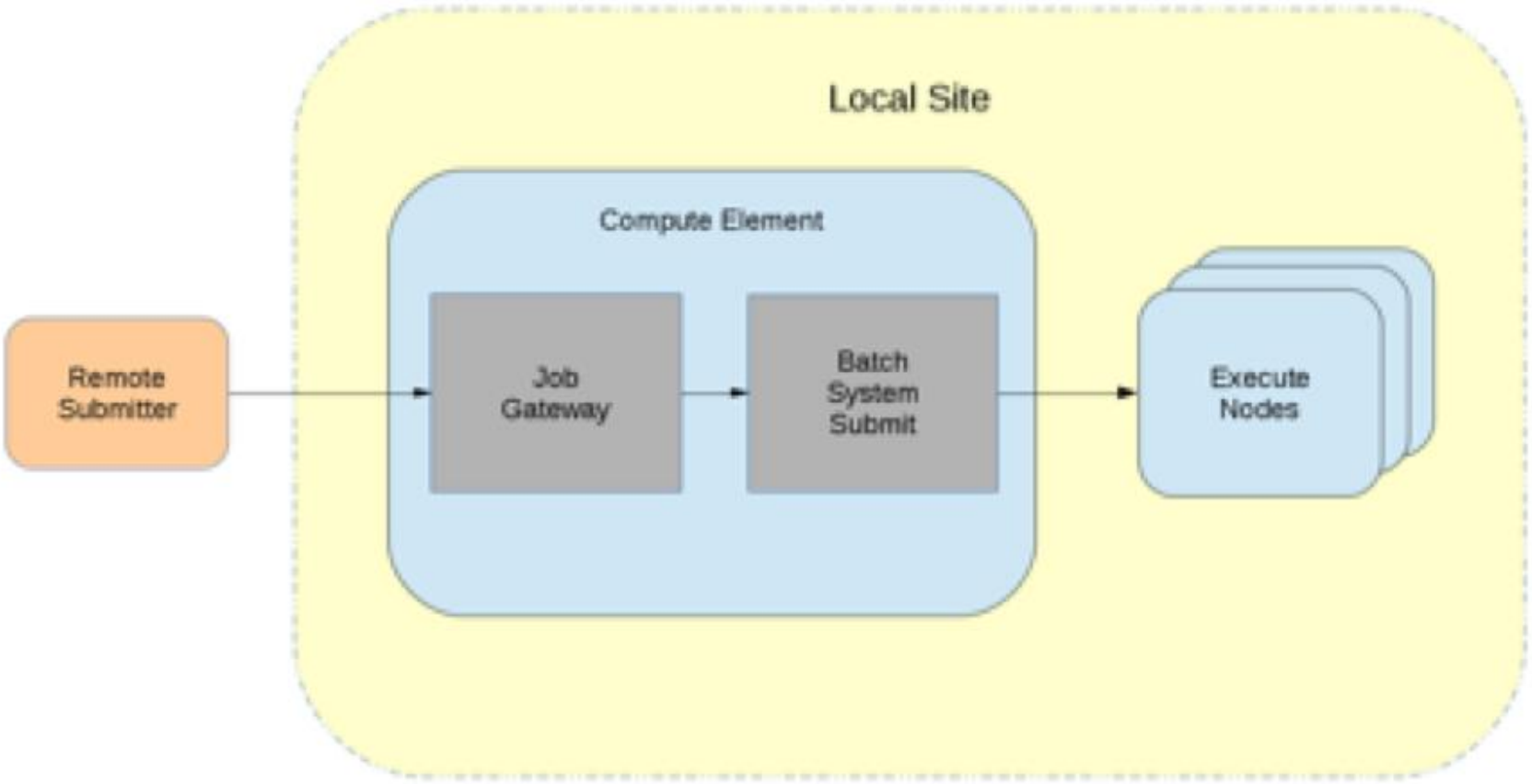
HTCondor-CE

- In 2013, OSG began an evaluation of its choice of CE technology
  - Did we want to keep the same technology? Try a new one?
- *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)

Anatomy of a Compute Element (CE)



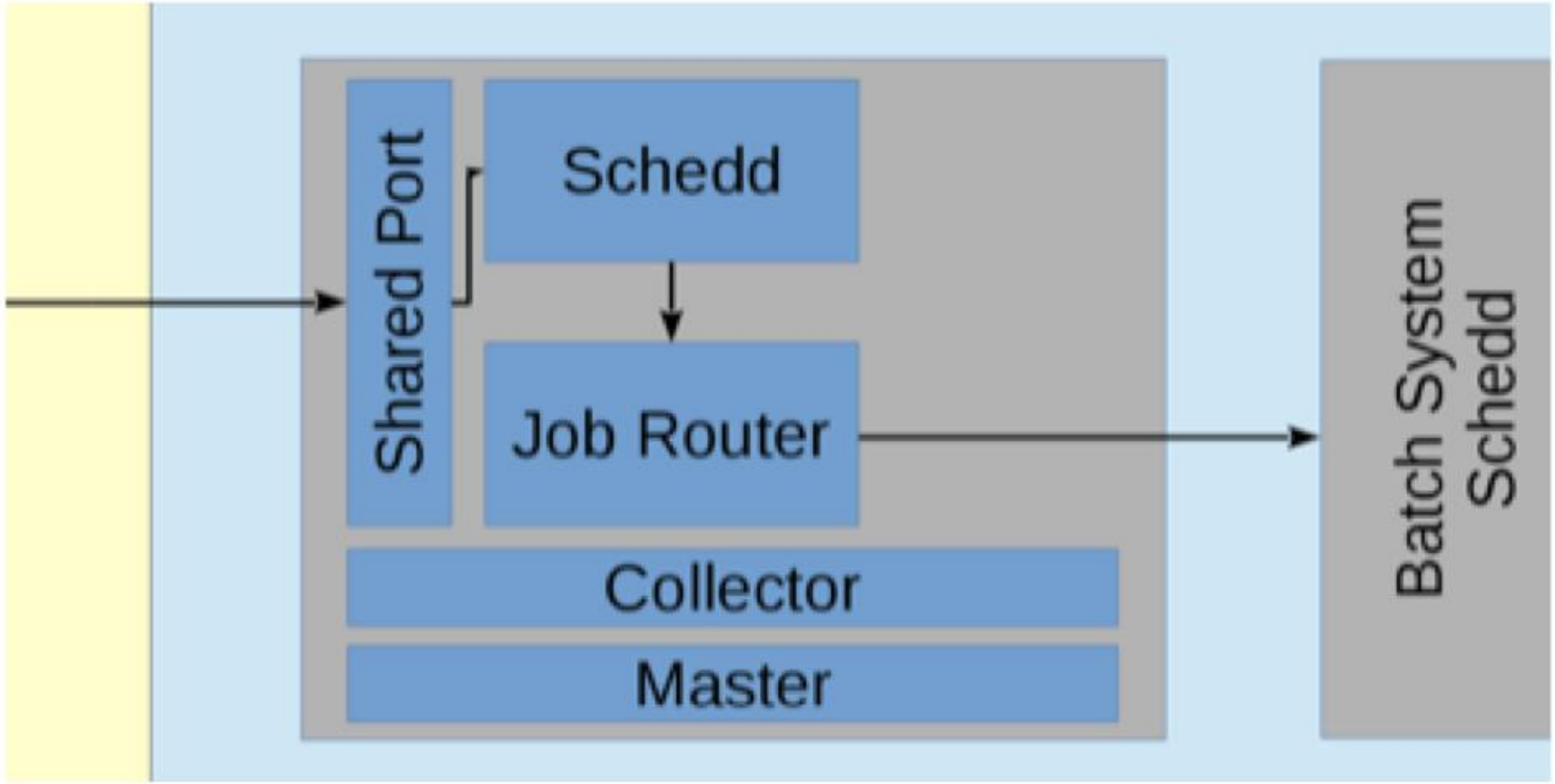
HTCondor-CE

- 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
  - Job Router provides transformation
- Simply need to put things together!

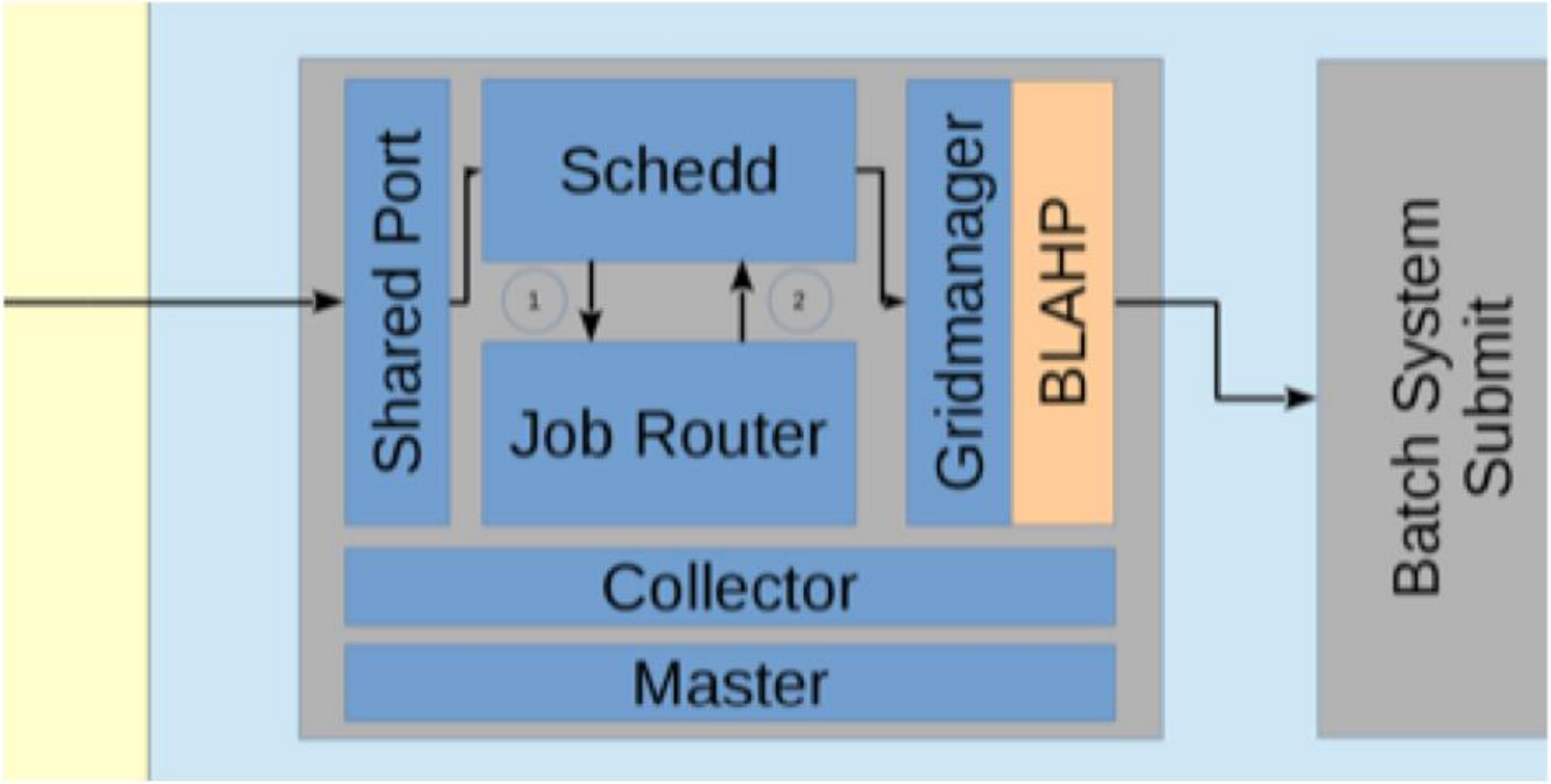
HTCondor-CE

- Special configuration of HTCondor
- Installs small wrappers around Condor CLI
  - **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`)

Anatomy of HTCondor-CE: HTCondor Batch System



Anatomy of HTCondor-CE: Non-HTCondor Batch System



Running Daemons

bbockelm — root@red-gw1:~ — ssh hcc-briantest — 150×25

```
condor      2495  0.0  0.0 103072  7080 ?        Ss   Feb18   0:25 condor_master -pidfile /var/run/condor-ce/condor_master.pid
root        2518  0.1  0.0  24524  6100 ?        S    Feb18  15:46  \_ condor_procd -A /var/lock/condor-ce/procd_pipe -L /var/log/condor-ce/ProcLog -R
condor      2519  0.0  0.0 102368  4604 ?        Ss   Feb18   9:16  \_ condor_shared_port -f -p 9619
condor      2521  0.8  1.0 400144 175800 ?       Ss   Feb18 114:32  \_ condor_collector -f -port 9619
condor      2523  0.5  0.4 176504 66132 ?        Ss   Feb18  80:29  \_ condor_schedd -f
condor      2524  1.4  0.6 205100 100888 ?       Ss   Feb18 192:16  \_ condor_job_router -f
condor      2742  0.0  0.0  97504  7620 ?        Ss   Feb18   0:27 condor_master -pidfile /var/run/condor/condor_master.pid
root        2750  0.1  0.0  24616  6116 ?        S    Feb18  16:29  \_ condor_procd -A /var/run/condor/procd_pipe -L /var/log/condor/ProcLog -R 100000
condor      2751  0.2  0.6 200520 101812 ?       Ss   Feb18  33:56  \_ condor_schedd -f
cmsprod  3033878  0.0  0.0  94604  8152 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821805.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3041926  0.0  0.0  94604  8184 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821815.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3041927  0.0  0.0  94604  8196 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821814.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3043312  0.0  0.0  94604  8184 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821825.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3056870  0.0  0.0  94604  8184 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821848.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3057151  0.0  0.0  94584  8184 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821849.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3061095  0.0  0.0  94604  8176 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821852.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3066118  0.0  0.0  94600  8176 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821857.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3070732  0.0  0.0  94600  8132 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821864.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3073572  0.0  0.0  94604  8144 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821866.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3078308  0.0  0.0  94600  8136 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821886.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3084233  0.0  0.0  94600  8180 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821888.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3092091  0.0  0.0  94600  8172 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821889.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3099541  0.0  0.0  94604  8176 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821897.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3105248  0.0  0.0  94600  8140 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821932.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
cmsprod  3107777  0.0  0.0  94604  8148 ?        SN   Feb26   0:01  |   \_ condor_shadow -f 5821943.0 --schedd=<129.93.239.132:39830?addrs=[2600-900-6
```

Job Router

- Responsible for taking a job and creating a copy modified according to a set of rules
  - Each chain of rules is called a “route” and is defined by a ClassAd
- Attribute changes and state changes are propagated between the source and destination jobs
- Job Router 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.

Example HTCondor Job Route

```
JOB_ROUTER_ENTRIES @=jre
[
    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";
]
@jre
```

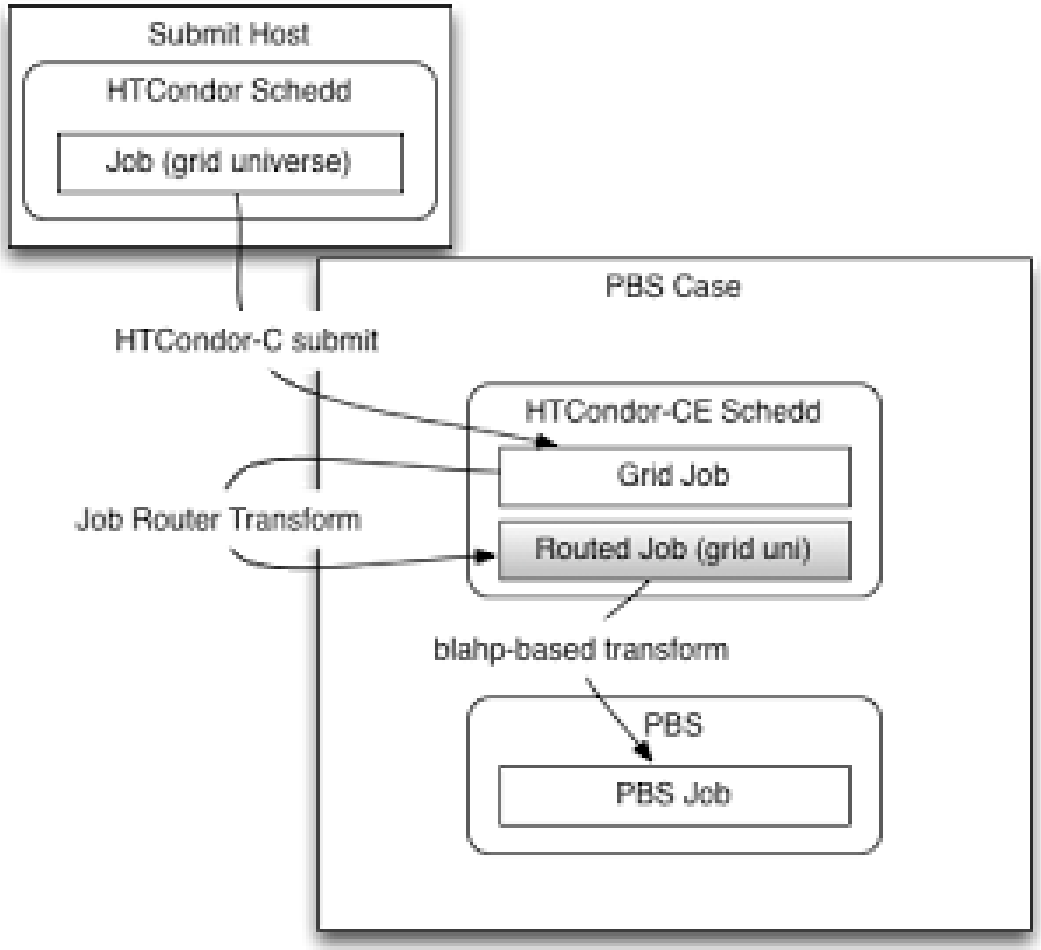
Example PBS Job Route

Cameron has a PBS pool and she wants CMS jobs submitted to her CE to be forwarded to her pool under the “cms” queue. All other jobs should be submitted to her pool without any changes.

Example PBS Job Route

```
JOB_ROUTER_ENTRIES @=jre
[
    name = "pbs_pool_cms";
    Requirements = target.x509UserProxyVOName =?= "cms";
    TargetUniverse = 9;
    GridResource = "batch pbs";
    set_BatchQueue = “cms”;
]
[
    name = "pbs_pool_other";
    Requirements = target.x509UserProxyVOName =!= "cms";
    TargetUniverse = 9;
    GridResource = "batch pbs";
]
@jre
```

Submitting to the CE



Example Submit File

```
universe = grid
grid_resource = condr condorce.example.com \
  condorce.example.com:9619
use_x509userproxy = true
executable = myjob.sh
output = myjob.out
...
queue
```

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, administrator)

condor_ce_trace

bbockelm — bbockelm@hcc-briantest:~ — ssh hcc-briantest -v — 188×35

```
[bbockelm@hcc-briantest ~]$ condor_ce_trace red.unl.edu
Testing HTCondor-CE collector connectivity.
- Failed ping of collector on <2600:900:6:1101:5054:ff:fe76:711a:9619>.

*****************************************************************************
2016-02-28 11:07:05 Failed to ping <2600:900:6:1101:5054: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.
- Successful ping of schedd on <129.93.227.123:9619?noUDP&sock=5472_8b22_23>.


    [
        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 — bbockelm@hcc-briantest:~ — ssh hcc-briantest -v — 107×24

```
[bbockelm@hcc-briantest ~]$ condor_ce_ping -pool tusker-gw1.unl.edu -name tusker-gw1.unl.edu -table ALL
         Instruction Authentication Encryption Integrity Decision Identity
               ALLOW            GSI       none       MD5    ALLOW uscmsPool018@users.opensciencegrid.org
                READ           none       none      none    ALLOW unauthenticated@unmapped
               WRITE            GSI       none       MD5    ALLOW uscmsPool018@users.opensciencegrid.org
           NEGOTIATOR            GSI       none       MD5     DENY uscmsPool018@users.opensciencegrid.org
        ADMINISTRATOR            GSI       none       MD5     DENY uscmsPool018@users.opensciencegrid.org
                OWNER            GSI       none       MD5     DENY uscmsPool018@users.opensciencegrid.org
               CONFIG            GSI       none       MD5     DENY uscmsPool018@users.opensciencegrid.org
               DAEMON            GSI       none       MD5     DENY uscmsPool018@users.opensciencegrid.org
     ADVERTISE_STARTD            GSI       none       MD5     DENY uscmsPool018@users.opensciencegrid.org
     ADVERTISE_SCHEDD            GSI       none       MD5     DENY uscmsPool018@users.opensciencegrid.org
     ADVERTISE_MASTER            GSI       none       MD5     DENY uscmsPool018@users.opensciencegrid.org
[bbockelm@hcc-briantest ~]$
[bbockelm@hcc-briantest ~]$
[bbockelm@hcc-briantest ~]$
[bbockelm@hcc-briantest ~]$
[bbockelm@hcc-briantest ~]$
[bbockelm@hcc-briantest ~]$
[bbockelm@hcc-briantest ~]$
[bbockelm@hcc-briantest ~]$
[bbockelm@hcc-briantest ~]$
[bbockelm@hcc-briantest ~]$
[bbockelm@hcc-briantest ~]$
```

Interaction Examples

condor_ce_status

bbockelm — root@red-gw1:~ — ssh hcc-briantest — 187×31

```
[root@red-gw1 ~]# condor_ce_status
Worker Node          State     Payload ID      User       Scheduler                        Job Runtime BatchID    BatchUser  Jobs   Pilot Age
red-c0801.unl.edu    Unclaimed                                                             0+00:00:03  5823965.0  glow          5   0+00:56:39
red-c0801.unl.edu    Unclaimed                                                             0+00:29:05  5823965.0  glow          5   0+00:55:41
red-c0801.unl.edu    Unclaimed                                                             0+00:29:07  5823965.0  glow          5   0+00:55:42
red-c0801.unl.edu    Unclaimed                                                             0+00:09:04  5823965.0  glow          5   0+00:55:43
red-c0801.unl.edu    Unclaimed                                                             0+00:34:04  5823965.0  glow          5   0+00:55:44
red-c0801.unl.edu    Unclaimed                                                             0+00:55:45  5823965.0  glow          5   0+00:55:45
red-c0801.unl.edu    Unclaimed                                                             0+00:55:46  5823965.0  glow          5   0+00:55:46
red-c0801.unl.edu    Unclaimed                                                             0+00:55:39  5823965.0  glow          5   0+00:55:39
red-c0803.unl.edu    Unclaimed                                                             0+09:15:38  5818978.0  osg          36   0+09:16:04
red-c0803.unl.edu    Claimed   18730904.0      zcx        login01.osgconnect.net           0+04:01:14  5818978.0  osg          36   0+09:16:04
red-c0803.unl.edu    Claimed   18710260.0      zcx        login01.osgconnect.net           0+04:01:14  5818978.0  osg          36   0+09:16:04
red-c0803.unl.edu    Claimed   18726288.0      fbdescamps login01.osgconnect.net           0+00:42:15  5818978.0  osg          36   0+09:16:04
red-c0803.unl.edu    Claimed   21597039.0      yx5        Q4@xd-login.opensciencegrid.org  0+00:12:41  5818978.0  osg          36   0+09:16:04
red-c0803.unl.edu    Claimed   18739992.0      fbdescamps login01.osgconnect.net           0+02:42:52  5818978.0  osg          36   0+09:16:04
red-c0803.unl.edu    Claimed   18712503.0      zcx        login01.osgconnect.net           0+05:25:24  5818978.0  osg          36   0+09:16:04
red-c0803.unl.edu    Claimed   18742089.0      fbdescamps login01.osgconnect.net           0+01:24:06  5818978.0  osg          36   0+09:16:04
red-c0803.unl.edu    Claimed   18735888.6205   intoy      login01.osgconnect.net           0+03:45:53  5818978.0  osg          36   0+09:16:04
red-c0805.unl.edu    Unclaimed                                                             1+17:41:06  5822017.0  cmsprod     113   1+17:41:26
red-c0805.unl.edu    Claimed   7286.0          cmst1      vocms0311.cern.ch                0+08:17:49  5822017.0  cmsprod     113   1+17:41:26
red-c0805.unl.edu    Claimed   133368.61       cmsdataops cmssrv219.fnal.gov               0+06:10:47  5822017.0  cmsprod     113   1+17:41:26
red-c0805.unl.edu    Claimed   133336.50       cmsdataops cmssrv219.fnal.gov               0+04:23:40  5822017.0  cmsprod     113   1+17:41:26
red-c0805.unl.edu    Claimed   403904.7        cmsdataops cmsgwms-submit1.fnal.gov         0+07:17:03  5822017.0  cmsprod     113   1+17:41:26
red-c0807.unl.edu    Unclaimed                                                             1+18:50:51  5821966.0  cmsprod      77   1+18:51:14
red-c0807.unl.edu    Claimed   133729.66       cmsdataops cmssrv219.fnal.gov               0+05:37:57  5821966.0  cmsprod      77   1+18:51:14
red-c0807.unl.edu    Claimed   7222.4          cmst1      vocms0311.cern.ch                0+08:52:02  5821966.0  cmsprod      77   1+18:51:14
red-c0809.unl.edu    Unclaimed                                                             0+09:16:46  5818960.0  osg          51   0+09:17:10
red-c0809.unl.edu    Claimed   18741865.6      pkilgo     login01.osgconnect.net           0+01:43:19  5818960.0  osg          51   0+09:17:10
red-c0809.unl.edu    Claimed   18741545.0      zcx        login01.osgconnect.net           0+00:22:57  5818960.0  osg          51   0+09:17:10
red-c0809.unl.edu    Claimed   40764350.0      donkri     Q2@xd-login.opensciencegrid.org  0+00:06:18  5818960.0  osg          51   0+09:17:10
```

Job Query

bbockelm — root@red-gw1:~ — ssh hcc-briantest — 104×28

```
[[root@red-gw1 ~]#  condor_ce_q

-- Schedd: red-gw1.unl.edu : <129.93.239.132:28464>
 ID        OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
1505510.0   fermilab         3/27 17:20   0+00:00:03 H  0   0.0  whoami
1506580.0   fermilab         3/27 21:28   0+00:00:03 H  0   0.0  whoami
1518799.0   fermilab         3/31 15:08   0+00:00:03 H  0   0.0  whoami
1802269.0   fermilab         6/2  10:12   0+00:00:04 H  0   0.0  whoami
1802270.0   fermilab         6/2  10:15   0+00:00:04 H  0   0.0  whoami
1923583.0   fermilab         6/24 13:16   0+00:00:04 H  0   0.0  whoami
1923788.0   fermilab         6/24 14:27   0+00:00:04 H  0   122.1 whoami
2670540.0   glow            12/11 05:40   0+06:51:44 C  0   195.3 glidein_startup.sh
2677852.0   glow            12/12 03:59   0+00:21:17 C  0   195.3 glidein_startup.sh
2738000.0   glow            12/30 18:39   0+00:47:40 C  0   9.8  glidein_startup.sh
2738113.0   glow            12/30 19:17   0+00:15:26 C  0   14.6 glidein_startup.sh
2738114.0   glow            12/30 19:17   0+00:15:26 C  0   14.6 glidein_startup.sh
2738115.0   glow            12/30 19:17   0+00:15:28 C  0   12.2 glidein_startup.sh
2738145.0   glow            12/30 19:25   0+00:20:40 C  0   12.2 glidein_startup.sh
2741874.0   glow            12/31 23:13   0+00:23:46 C  0   17.1 glidein_startup.sh
2741880.0   glow            12/31 23:16   0+00:22:10 C  0   14.6 glidein_startup.sh
2744310.0   glow             1/1  17:57   0+00:22:01 C  0   14.6 glidein_startup.sh
2753580.0   glow             1/3  06:10   0+00:22:49 C  0   14.6 glidein_startup.sh
2758819.0   glow             1/3  21:37   0+00:24:41 C  0   14.6 glidein_startup.sh
2758843.0   glow             1/3  21:42   0+00:21:44 C  0   14.6 glidein_startup.sh
2758845.0   glow             1/3  21:42   0+00:20:45 C  0   17.1 glidein_startup.sh
2759289.0   glow             1/3  23:46   0+00:22:44 C  0   293.0 glidein_startup.sh
2759291.0   glow             1/3  23:46   0+00:24:26 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)

Why Consider this CE?

- 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
- Now available apart from the OSG software stack
  - `htcondor-ce` RPM package
- More information available here:
  - https://opensciencegrid.org/docs/compute-element/htcondor-ce-overview/