OSG Technology 2013

Brian Bockelman

OSG AHM 2013

aka, The Talk of Lies

(with all due deference Todd's annual Condor Week talk)

A Year of Technology Transitions

- 2012 was a year of remarkable transitions in software distribution. The bulk of sites moved from Pacman to RPM-based OSG3.
  - We're not done yet, but certainly in "mop-up" mode for the missing critical pieces.
  - The planning, development, and initial releases for OSG3 were done in 2011.
- For OSG Technology, 2013 will be similar to OSG Software's 2012.
  - We will begin a serious rollout of development started in 2012.

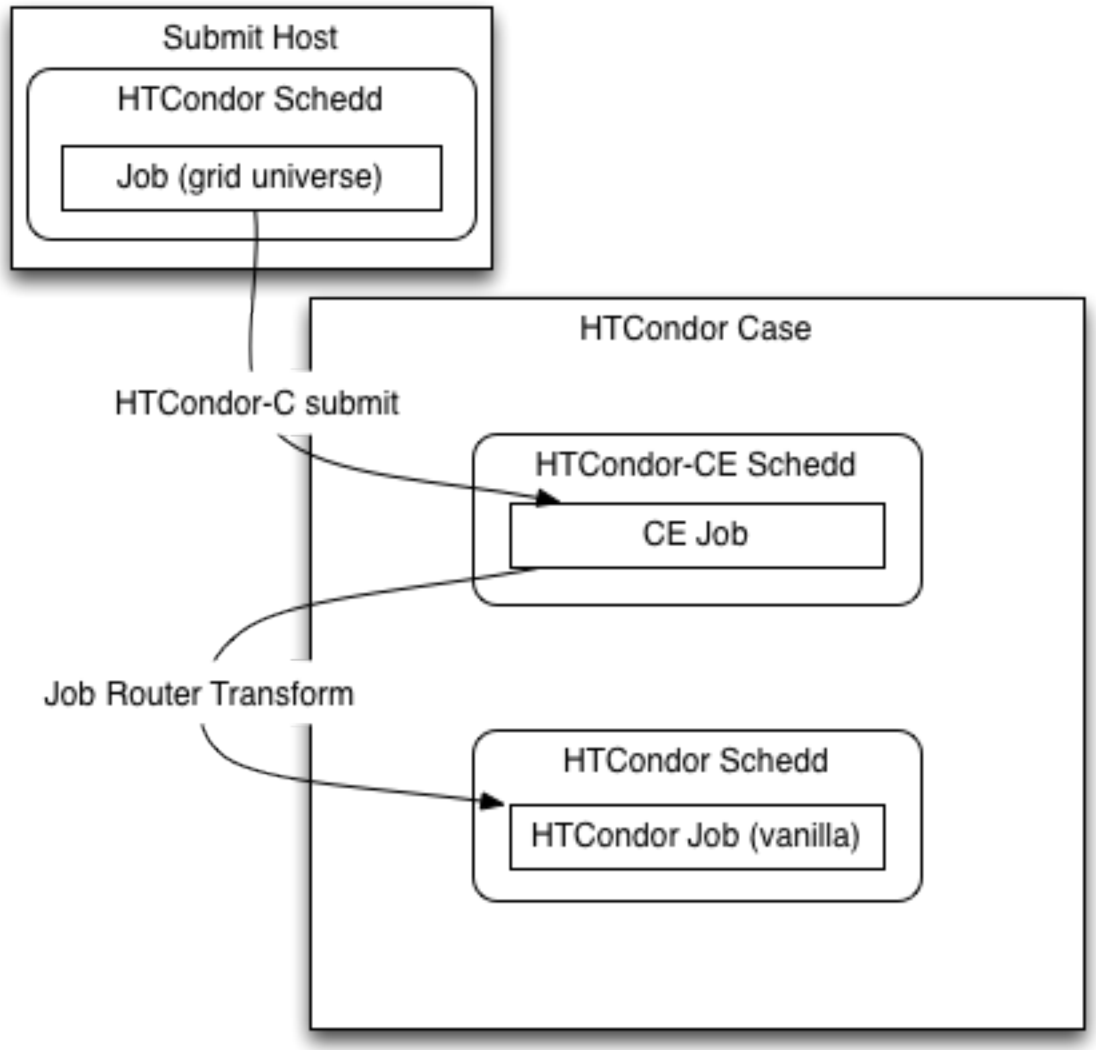
Big Changes

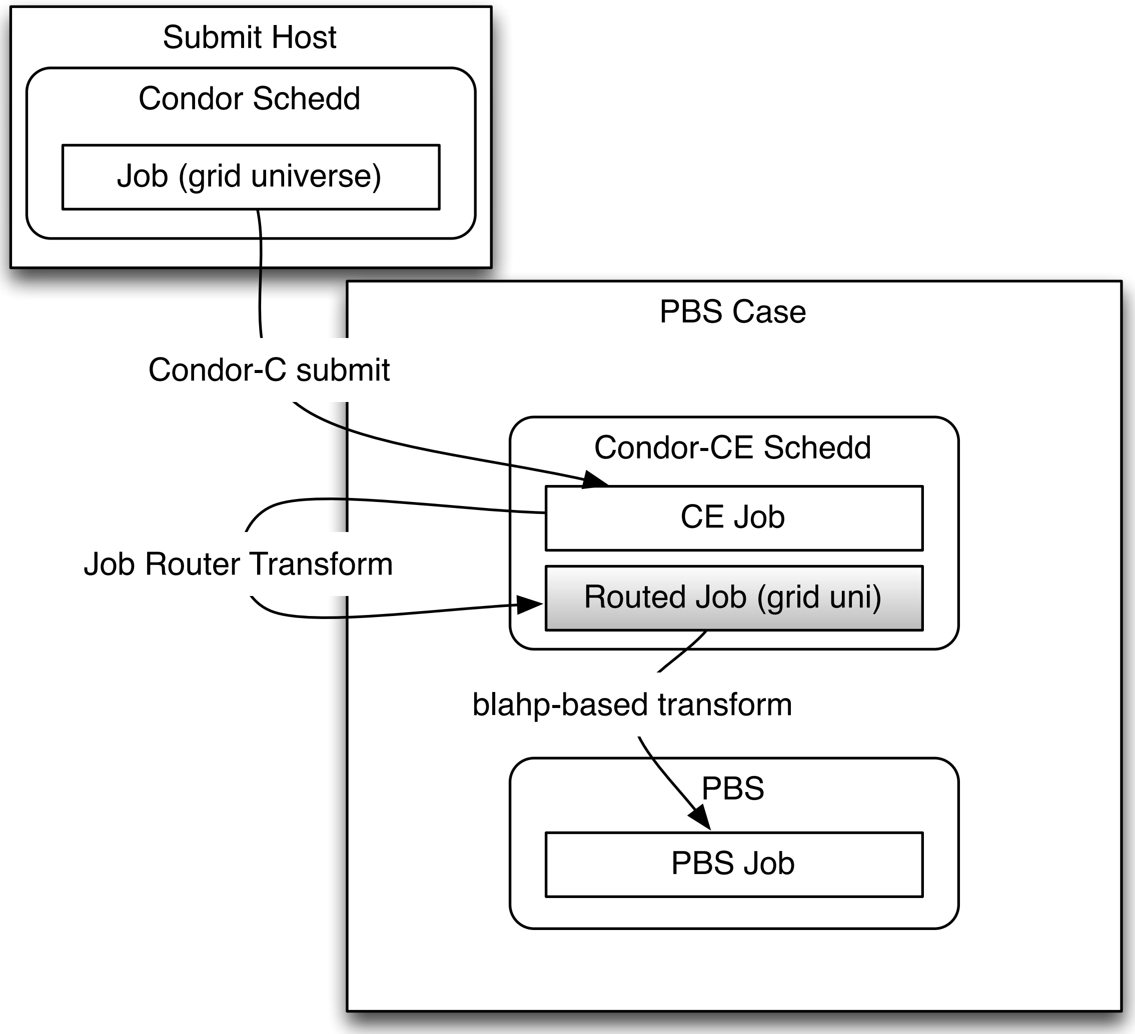
- 2013 should see:
    - New base gatekeeper software (HTCondor-CE, BOSCO).
    - New methods for VO software distribution (OASIS).
    - Less emphasis on SRM for storage management and data movement - especially at non-archive sites.
- Less important changes: information services, glexec improvements.
- Important, but perhaps not in time for 2014: CrossCE.
- There are actually many other activities I won't have time to cover.

HTCondor-CE

- Currently, Globus GRAM provides the abstraction, sandbox movement, and remote submission layers for the OSG-CE.
- In the April/May timeframe, we are targeting a new stack based on a HTCondor schedd.
  - Goals is to have HTCondor serve as a complete gatekeeper - only a special configuration, no additional OSG-maintained scripts.

The Big Picture





Remote Submission

- The HTCondor schedd has allowed remote job submission for several years.
  - May not have been completely bug free...
- We have helped validate its scale and performance on high latencies.
- Unlike with Globus GRAM, the state of the pilot job (submitted to the CE, staging in, submitted to the site batch) is easy to track.
  - Just a "condor_q" away!

Abstraction

- Pilot factories tend to want a high-level resource descriptions; there is a desire to minimize the amount of "site knowledge" necessary to describe each pilot.
- Some software is needed as the glue layer between.
  - Right now, we use either Globus GRAM's built in functionality or hack their perl scripts.

JobRouter

- We use the *condor_jobrouter* daemon for transforming the job for the local site.
- This daemon creates a copy of the job and applies a set of admin-prescribed transformations.
    - These can either be done via a classad policy or a script callout.
    - The site customizations will no longer be overwritten by RPM upgrades. Celebrate!
- JobRouter can create the job copy directly in a site schedd, doing the site batch system submission for HTCondor sites.

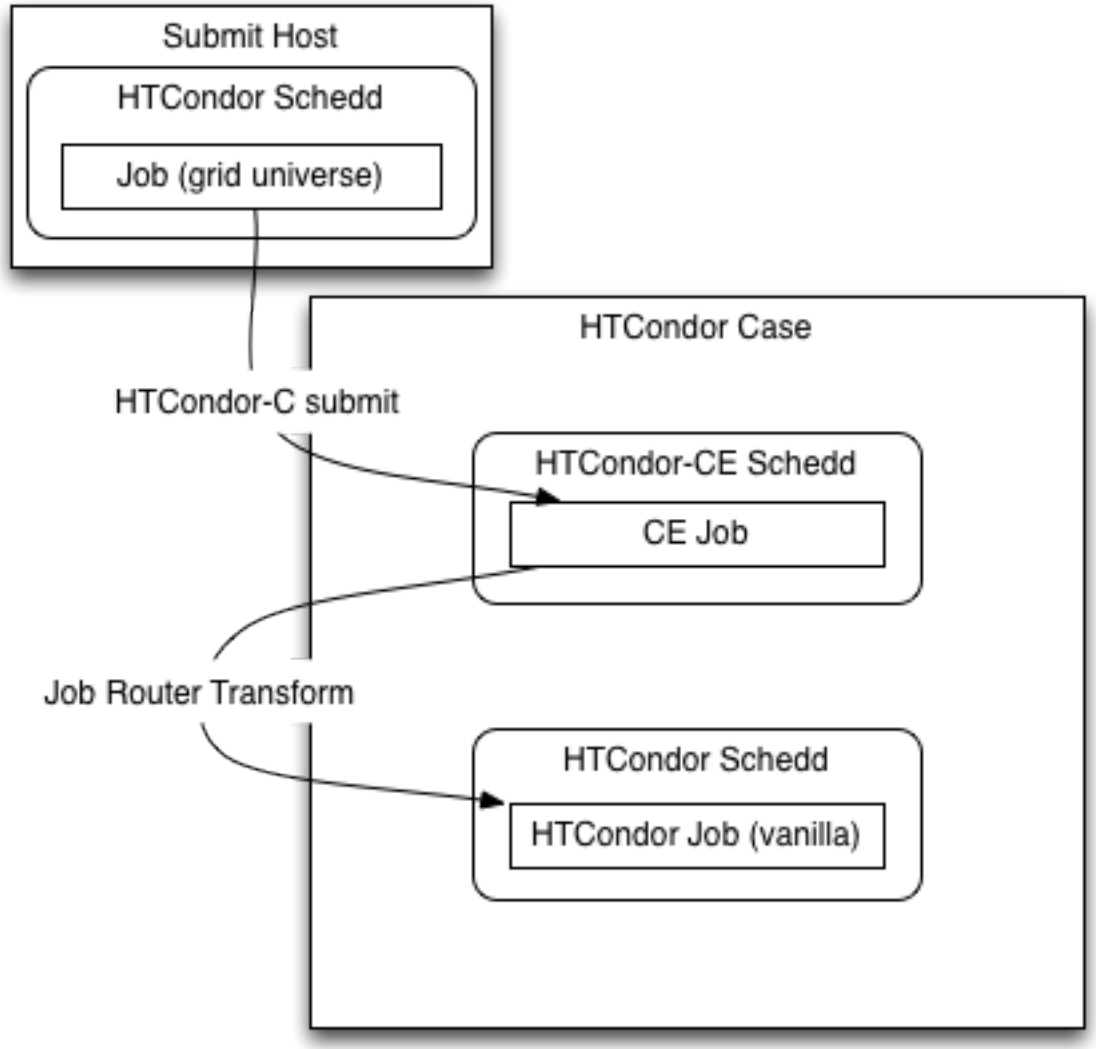
ClassAd Policy

```
JOB_ROUTER_ENTRIES = \
   [ \
     GridResource = "condor localhost localhost"; \
     eval_set_GridResource = strcat("condor ", $
(FULL_HOSTNAME), $(FULL_HOSTNAME)); \
     TargetUniverse = 5; \
     name = "Local_Condor"; \
     Requirements = regexp("^/hcc/", x509UserProxyFirstFQAN); \
     eval_set_AccountingGroup = strcat("hcc.", Owner); \
   ]
```

JobRouter script

```
#!/usr/bin/python

import sys
import classad

route_ad = classad.ClassAd(sys.stdin.readline())
separator_line = sys.stdin.readline()
assert separator_line == "------\n"
ad = classad.parseOld(sys.stdin)

ad["Universe"] = 5
ad["GridResource"] = "condor localhost localhost"
if "x509UserProxyFirstFQAN" in ad and "/cms" in ad.eval("x509UserProxyFirstFQAN"):
    ad["AccountingGroup"] = "cms.%s" % ad.eval("Owner")
else:
    ad["AccountingGroup"] = "other.%s" % ad.eval("Owner")

print ad.printOld(),
```

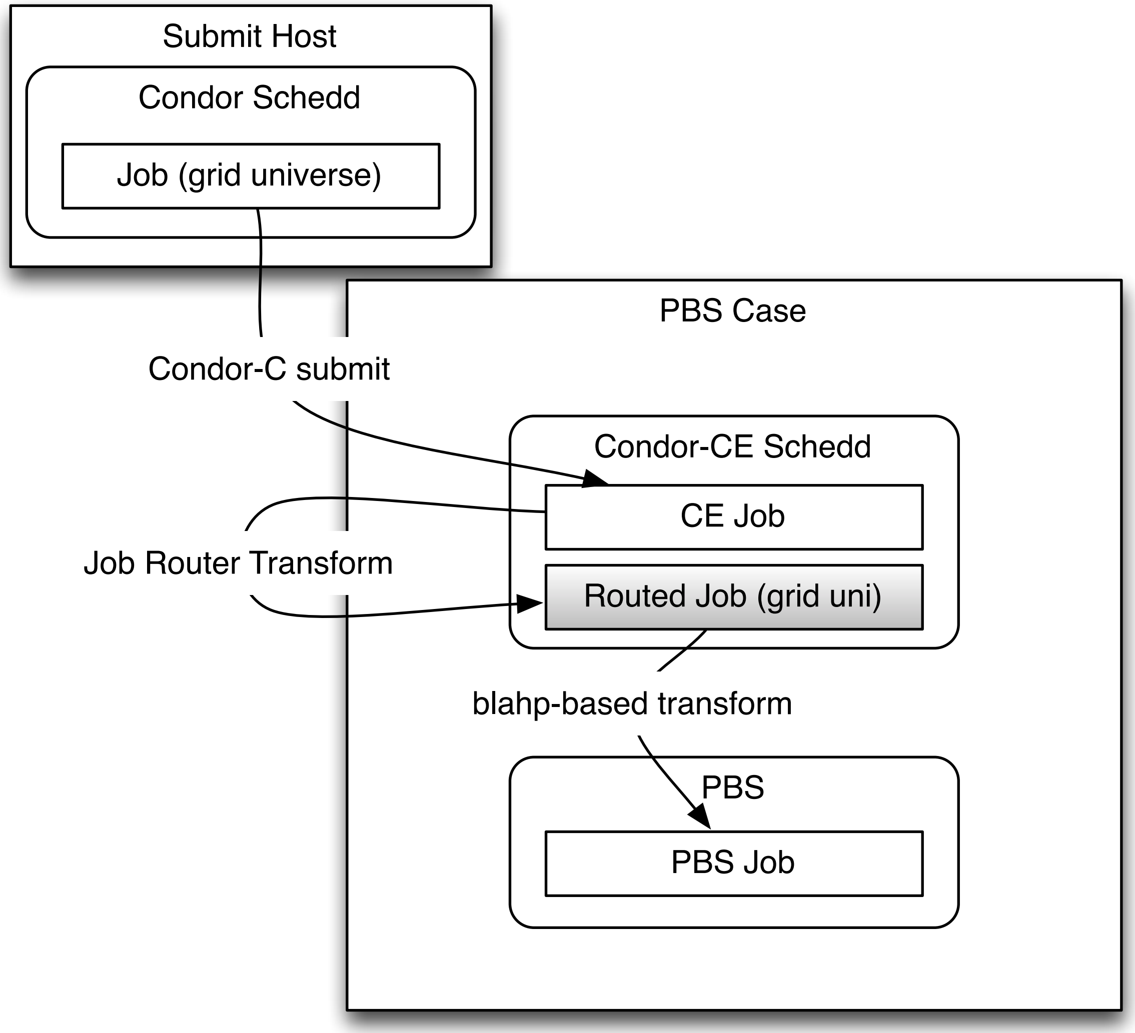
HTCondor Sites, Again



blahp

- An additional step is needed for non-HTCondor sites.
- HTCondor-G has the ability to submit to non-HTCondor sites using blahp.
    - blahp is the executable which then calls, for example, *qstat* / *qsub* / *qdel*.
- blahp has another layer of customization if, for example, you need to tweak *qsub* arguments. Most useful things can be done via the JobRouter transform.

PBS/LSF/SGE/SLURM Sites



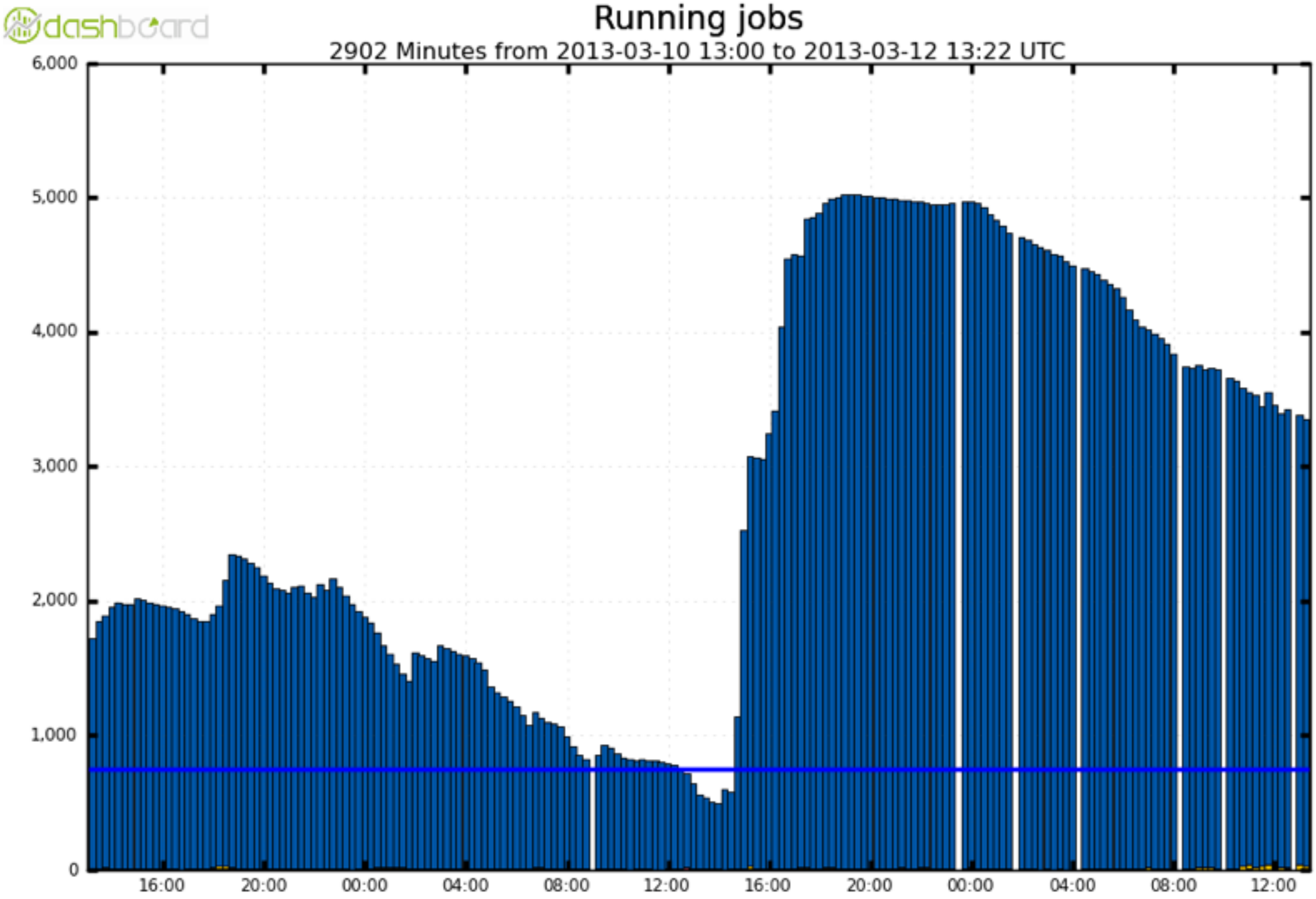
What about BOSCO?

- BOSCO and HTCondor-CE both share common components - HTCondor-G and blahp.
  - If it works for BOSCO, it works for HTCondor-CE.
- BOSCO is, in itself, a gateway service. Instead of HTCondor-C or GRAM for remote submission, BOSCO uses SSH.
  - Refer to Tuesday's presentations for more information.

BOSCO and CMS

- CMS has been using BOSCO to run jobs at the Gordon XSEDE resource at SDSC.
  - The glideinWMS factory uses Condor-G to submit via a SSH login. BOSCO scripts were used to stage blahp to the remote side.
    - So, we are launching real CMS pilots via BOSCO.
  - (I hope) this has been useful for BOSCO - found 4-5 very instructive bugs.
- I believe BOSCO could be repurposed like this again in the future. I see great value here for making the "simplest site possible".

BOSCO in Action



Other Tidbits

- HTCondor-CE integrates with the expected services: Gratia, GIP, LCMAPS, glideinWMS, AutoPy Factory.
- Improvements in HTCondor (sandbox and transfer management) also benefit the CE.
    - Features X, Y, and Z (whatever they may be!) in future HTCondor versions will also benefit the CE.

Project Status

- Except for improved security audit logs, the HTCondor-CE is feature-complete. Any site should be able to test out the osg-development version.
  - Documentation exists (https://www.opensciencegrid.org/bin/view/Documentation/InstallHTCondorCE), but needs more work.
  - blahp continues to get battle-hardened as it is part of CREAM and BOSCO.
- HTCondor-CE has passed scale tests (10k running jobs) for a HTCondor-based site. We are still working on validating other batch systems (PBS will be next).
- HTCondor-CE will require HTCondor 8.0; this sets the deployment calendar (production release no earlier than May, likely June).
- GRAM and HTCondor-CE will run side-by-side for a very long time.

(Pause for Demo)

OASIS

- OSG has offered a poor experience to VOs for application software. We ask sites to create an NFS-mounted directory, $OSG_APP, export it to users, and give guidelines with respect to total size.
  - We give sites no mechanisms for managing this directory.
  - It is up to the VOs to figure out how to install software at every site and deal with inconsistencies.
    - Not everyone can afford a few FTE-years for software management!
- **We can do better**.

Introducing OASIS

- OASIS - the OSG Application Software Installation Service.
  - An OSG-provided mechanism for VOs to install software at sites and for sites to manage the VO areas.
- Currently, the implementation is based on CVMFS.
  - See prior talks for CVMFS details. To sites, it is a FUSE-based, read-only file system which distributes data via a series of HTTP caches.

OASIS 1.0 - Features and Limitations

- A VO can enroll in the OASIS service via OIM and manage the list of VO software admins via the web.
- VO software admins can use gsissh to login to *oasis-login.opensciencegrid.org*; there, they can install their software.
- Once done, run a tool to create a synchronization between the login server and the repository.
  - Only one VO can run a synchronization at a time.
- Sites can then mount *oasis.opensciencegrid.org* repository on their worker noes as $OSG_APP.

Project Status

- GOC is tuning the service, based on 2 months of beta testing:
  - Rolling out additional replica servers.
  - Security improvements based on an internal audit.
  - Write SLA documents.
- I expect the first release to coincide with an April production update.
  - You can access it now -

```
[bbockelm@hcc-briantest ~]$ cat /cvmfs/oasis.opensciencegrid.org/hcc/hello_world
Hello world!.
```

Post-OASIS-1.0

- We'd like to see better support for multiple repos:
  - Unnatural to submit from one node and install from another - user must leave the "local environment".
  - Perhaps the VO's host site wants to run the repository server?
- Distribute OSG worker node client via OASIS.
- Improved monitoring at the GOC.
- Require a HTTP proxy as a part of an OSG-CE.
- Integrate parrot/CVMFS with the OSG glideinWMS distribution.
- Upgrade to CVMFS 2.1 (?)

Changes in Data Management

- The changes in data management are more philosophical than technical.
- There is less emphasis on SRM-based management, more emphasis on using HTTP caches, GridFTP, or a *protocol-agnostic* layer.
- For smaller data volumes, more emphasis on implicit data movement.

SRM at non-archival sites

- At non-archival sites, SRM provides:
  - *Load balancing* for transfers - can be done natively with GridFTP, HTTP, or Xrootd.
  - *Metadata queries* like rm/ls/mkdir - can be done natively with GridFTP, HTTP, or Xrootd
  - *Storage management* - unique to SRM. Most SRM functionality not used via grid although some aspects ('du' of pieces of namespace) are used. Quite a few local sites find SRM useful for local management.
- SRM may be the biggest fish in the OSG sea, but it is not the only one! We have alternates .

Making Life Easier

- WLCG VOs mostly take care of their own, so the most important to focus on new VOs:
    - Provide a protocol-agnostic layer using file transfer plugins. Users will switch between protocols by changing the letters before "://".
    - Provide additional HTTP proxies to improve the experience of running cache-friendly workflows.
        - Improve circuit breakers for failures.

Project Status

- SRM is not going anywhere.
  - But I think we are seeing a sea-change in approaches.
  - For example, it is possible to run a CMS site without SRM. CMS has found storage management is less important than transfer management and remote IO.
- In summer, I believe we will provide an HTTP proxies as a required part of the OSG job runtime environment.
- Summer-to-fall, we will integrate file transfer plugins with our glideinWMS distribution.
- The "holy grail" of opportunistic storage appears to still be out of reach for 2013. We will be moving forward in small, evolutionary steps.

Conclusion

- While the OSG has existed for about 6 years, we are in the midst of a large-scale changes in the underlying technology.
    - Some of this is driven by opportunities in external calendars (LHC long shutdown allows us the opportunity to be disruptive).
    - Some is driven by new emphases in the OSG, such as better serving new communities (BOSCO, OASIS).
    - Some is driven by long evolutionary processes (changes in data management and information services) that are culminating.

Times, they are a changin'

- A few thoughts on technology changes:
    - We believe these are huge improvements compared to the prior technologies.
    - Change is messy. There will be bugs. There will be mistakes. There will be something for everyone to love and new things for everyone to hate.
        - Please be patient and forgiving.
    - Technologies may change, but the core principles of OSG do not.
        - "Production" is still critical to the OSG Production Grid. The transitions I discuss will take multiple years - likely even longer than the RPM rollout - before things are ready for every site.

Questions!?

Backup Slides

- (In case if the demo fails)

Simplest Jobs

Test Tool:

```
[bbockelm@brian-test ~]$ condor_ce_run -r red.unl.edu:9619 echo "hello world"
hello world
```

Test HTCondor-G Job

```
universe = grid
grid_resource = condor red.unl.edu red.unl.edu:9619

executable = test.sh
output = test_g.out
error = test_g.err
log = test_g.log

ShouldTransferFiles = YES
WhenToTransferOutput = ON_EXIT

use_x509userproxy = true

queue
```

CE Processes

```
USER       PID %CPU %MEM    VSZ   RSS TTY      STAT START   TIME COMMAND
condor    8556  0.0  0.0  90628  5256 ?        Ss   Jan31   0:49 condor_master -pidfile /var/
root      8559  0.1  0.0  47056 27464 ?        S    Jan31  73:06  \_ condor_procd -A /var/loc
condor    8560  0.0  0.0  90212  5696 ?        Ss   Jan31   8:49  \_ condor_shared_port -f -p
condor    8566  0.0  0.0  91512  6992 ?        Ss   Jan31   9:12  \_ condor_collector -f -por
condor    8567  0.2  0.7 324552 230260 ?       Ss   Jan31 140:28  \_ condor_schedd -f
condor    8570  0.5  3.0 1079916 991740 ?      Ss   Jan31 329:15  \_ condor_job_router -f
```

CE Queue

```
[root@red ~]# condor_ce_q


-- Submitter: red.unl.edu : <129.93.239.129:9620?sock=8556_0571_4> : red.unl.edu
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
112044.0   uscmsPool018    3/12 21:35   0+00:00:02 C  0   0.0  echo
112046.0   uscmsPool018    3/12 21:45   0+00:00:00 I  0   0.0  echo

2 jobs; 1 completed, 0 removed, 1 idle, 0 running, 0 held, 0 suspended
```

CE Queue

```
[root@red ~]# condor_ce_q


-- Submitter: red.unl.edu : <129.93.239.129:9620?sock=8556_0571_4> : red.unl.edu
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
112122.0   uscmsPool018    3/13 08:17   0+00:00:00 H  0   0.0  test.sh
112123.0   uscmsPool018    3/13 08:17   0+00:00:00 I  0   0.0  test.sh
...
112130.0   uscmsPool018    3/13 08:17   0+00:00:00 I  0   0.0  test.sh
112135.0   uscmsPool018    3/13 08:17   0+00:00:00 H  0   0.0  test.sh
112136.0   uscmsPool018    3/13 08:17   0+00:00:00 I  0   0.0  test.sh
112137.0   uscmsPool018    3/13 08:17   0+00:00:00 I  0   0.0  test.sh
112138.0   uscmsPool018    3/13 08:17   0+00:00:00 H  0   0.0  test.sh
112139.0   uscmsPool018    3/13 08:17   0+00:00:00 H  0   0.0  test.sh
112140.0   uscmsPool018    3/13 08:17   0+00:00:00 I  0   0.0  test.sh
112141.0   uscmsPool018    3/13 08:17   0+00:00:00 I  0   0.0  test.sh
112142.0   uscmsPool018    3/13 08:17   0+00:00:00 H  0   0.0  test.sh
112143.0   uscmsPool018    3/13 08:17   0+00:00:00 I  0   0.0  test.sh
112144.0   uscmsPool018    3/13 08:17   0+00:00:00 I  0   0.0  test.sh
112145.0   uscmsPool018    3/13 08:17   0+00:00:00 H  0   0.0  test.sh
112146.0   uscmsPool018    3/13 08:17   0+00:00:00 I  0   0.0  test.sh

105 jobs; 1 completed, 0 removed, 82 idle, 0 running, 22 held, 0 suspended
```