Engagement VO

Matchmaking with ReSS

MPI Jobs on OSG

Engagement VO

- Recruit new users from diverse scientific domains
  - Beyond the physics community
- Demonstrate the power and ease of OSG to new communities
- Today's topics
  - Resource selection with ReSS
  - MPI jobs on OSG

User Environment

- All about usability
  - Make it easy to run simple (but real) jobs
- Engagement VO provides to its users:
  - Submit host
  - Resource selection system
  - Site verification

Why resource selection?

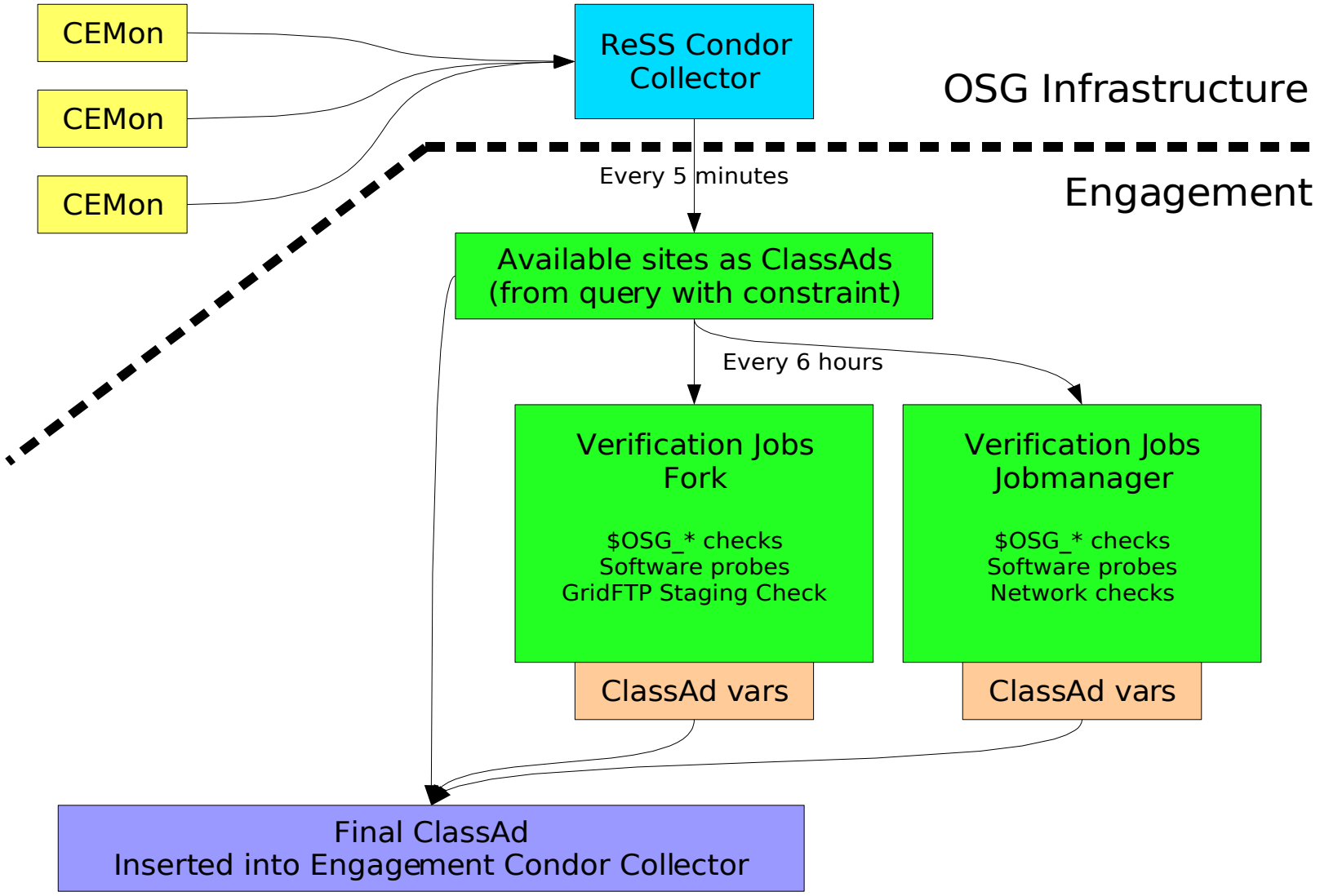
- Engagement users
  - Small labs/experiments
  - Grid?
  - Little or no CS/IT resources
- Example: Kuhlman lab at UNC
  - Protein folding
  - 1 job run (~ 3 days)
  - 4 weeks in wet lab, using results from run
- No knowledge / time / resources / need for a complex job handling system

Engagement VO Jobs

- Mostly very simple jobs
  - No inter job dependencies
  - Simple staging requirements
    - Inputs/outputs staged with job
- Job independence makes it easy to spread a run across many sites
- Great candidates for matchmaking

ReSS

- CEMon
- Condor ClassAds
- User expresses job requirements in Condor submit file
- Condor matches jobs to resources
- Condor handles resubmits
- Use hold/release expressions to resubmit “stuck” jobs
- Post job scripts to workaround exit codes


CEMon
CEMon
CEMon
ReSS Condor Collector
OSG Infrastructure
Engagement
Every 5 minutes
Available sites as ClassAds (from query with constraint)
Every 6 hours
Verification Jobs Fork
$OSG_* checks
Software probes
GridFTP Staging Check
Verification Jobs Jobmanager
$OSG_* checks
Software probes
Network checks
ClassAd vars
ClassAd vars
Final ClassAd
Inserted into Engagement Condor Collector

condor_grid_overview

```
ID      Owner   Resource        Status    Time Sta Sub
======= ====== ============== ======== ======== ===
46381   rynge   (DAGMan)                  1:58:54
46382   rynge   GLOW            Running   1:55:43   1
46384   rynge   UWMilwaukee     Pending   1:57:04   1
46387   rynge   Nebraska        Running   1:00:43   1


Site          Jobs   Subm   Pend   Run   Stage   Fail   Rank
============ ===== ===== ===== ===== ===== ===== =====
ASGC_OSG        17      0      0    15      2      0   155
FNAL_GPFARM     14      4      0    10      0      0   720
GLOW            36      6      5    22      3      0   372
Nebraska        17      0      5    12      0      0   288
Purdue-Lear     15      4      0    10      1      0   372
TTU-ANTAEUS     15      2      0    11      2      0   372
Vanderbilt      45      4      4    37      0      0   350
```

Site Rank

- Simple – but works!
- Integer between 0 and ~1000
- Jobs submitting/staging/pending/running provides the baseline
- Last step: ratio between matched jobs, and the max number we want on that site

Condor Submit File

```
globusscheduler = $$(GlueCEInfoContactString)

requirements = (
  (TARGET.GlueCEInfoContactString =!= UNDEFINED) &&
  (TARGET.Rank > 300) &&
  (EngageSoftwareWget == True) &&
  (TARGET.EngageCENetworkOutbound == True))
```

```
# when retrying, remember the last 4 resources tried
match_list_length = 4
Rank              = (TARGET.Rank) -
  ((TARGET.Name =?= LastMatchName0) * 1000) -
  ((TARGET.Name =?= LastMatchName1) * 1000) -
  ((TARGET.Name =?= LastMatchName2) * 1000) -
  ((TARGET.Name =?= LastMatchName3) * 1000)
```

Condor Submit File (cont.)

```
# make sure the job is being retried and rematched
periodic_release = (NumGlobusSubmits < 10)
globusresubmit = (NumSystemHolds >= NumJobMatches)
rematch = True
globus_rematch = True
```

```
# only allow for the job to be queued or running for a while
# then try to move it
#  JobStatus==1 is pending
#  JobStatus==2 is running
periodic_hold = (
  ((JobStatus==1) && ((CurrentTime - EnteredCurrentStatus) >
   (5*60*60))) ||
  ((JobStatus==2) && ((CurrentTime - EnteredCurrentStatus) >
   (24*60*60))) )
```

MPI Jobs on OSG

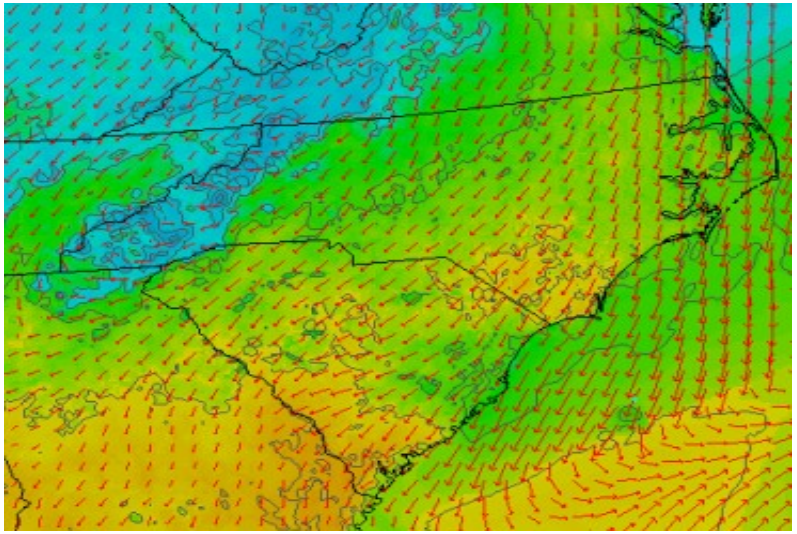
- Why?
  - During our interactions with potential users, one recurring question was whether OSG supports MPI jobs
- If OSG wants to move beyond HEP, MPI will be a requirement

Let's explore!

- Determine the current support for MPI applications
- Document findings
  - white paper
- Start a discussion in the OSG community

Approach

- Selected an MPI simulation
  - WRF (Weather Research Forecast model)
- Selected a set of test sites
  - Purdue x 2, Buffalo, GLOW and NERSC
- Workarounds?
- Time consuming?



Success?

- 3 ways we were able to submit MPI jobs
  - Interactive logins
  - Fork + local scheduler submit
  - (jobtype=mpi)
- It works, but just barely...

Areas of Concern

- MPI Implementations and Interconnects
- MPI Compilers / Interactive Logins
- Job Environment

Advertising Capabilities

- MPI details will need to be advertised
  - Implementation and version
  - Interconnect
  - Necessary RSL variables
  - Compiler location
- You can't just submit an i386 statically linked binary
  - Need to match binary with site MPI
  - Or, build in place

Compilers / Interactive Logins

- Access to MPI compilers?
  - Easiest way to match code to site MPI
  - Will pick up site defaults
- Interactive Logins?
- Necessary?
  - Usability will be compared to local campus resources and TeraGrid

Recommended Action Items

- Read the paper!
  - Feedback appreciated
- Short term
  - Fix job environment
  - Set up and maintain (jobtype=mpi) interfaces
  - Advertise capabilities

Questions?