Building Campus HTC Sharing Infrastructures

Derek Weitzel
University of Nebraska – Lincoln
(Open Science Grid Hat)
HCC: Campus Grids Motivation

- We have 3 clusters in 2 cities.
- Our largest (4400 cores) is always full
• Workflows may require more power than available on a single cluster.
  − Certainly more than a full cluster can provide.

• Offload single core jobs to idle resources, making room for specialized (MPI) jobs.
HCC Campus Grid Framework Goals

- **Encompass:** The campus grid should reach all clusters on the campus.

- **Transparent execution environment:** There should be an identical user interface for all resources, whether running locally or remotely.

- **Decentralization:** A user should be able to utilize his local resource even if it becomes disconnected from the rest of the campus. An error on a given cluster should only affect that cluster.
HCC Campus Grid Framework Goals

- **Encompass:** The campus grid should reach all clusters on the campus.

- **Transparent execution environment:** There should be an identical user interface for all resources, whether running locally or remotely.

- **Decentralization:** A user should be able to utilize his local resource even if it becomes disconnected from the rest of the campus. An error on a given cluster should only affect that cluster.
Encompass Challenges

• Clusters have different job schedulers: PBS & Condor?

• Each cluster has their own policies
  – User Priorities
  – Allowed users

• We may need to expand outside the Campus
HCC Model for a Campus Grid

- Me, my friends and everyone else
Preferences/Observations

• Prefer not installing Condor on every worker node when PBS is already there.
  – Less intrusive for sysadmins.

• PBS and Condor should coordinate job scheduling.
  – Running Condor jobs look like idle cores to PBS.
  – We don’t want PBS to kill Condor jobs if it doesn’t have to.
Problem: PBS & Condor Coordination

- **Initial:** Condor is running a job.
Problem: PBS & Condor Coordination

- PBS Starts a job – Condor restarts job
Problem: PBS & Condor Coordination

- Real Problem: PBS doesn’t know about Condor
  - Sees nodes as idle.
Campus Grid Goals - Technologies

• Encompassed
  – BLAHP
  – Glideins (See earlier talk by Igor/Jeff)
  – Campus Grid Factory

• Transparent execution environment
  – Condor Flocking
  – Glideins

• Decentralized
  – Campus Grid Factory
  – Condor Flocking
Encompassed – BLAHP

- Written for European Grid Initiative
- Translates Condor job into PBS job
- Distributed with Condor

- With BLAHP: Condor can provide a single interface for all jobs, whether Condor or PBS.
Putting it all Together

Campus Grid Factory

http://sourceforge.net/apps/trac/campusfactory/wiki
Putting it all Together

- Provides on-demand Condor pool for unmodified clients with Flocking.
Putting it all Together

- Creates an on demand condor cluster
  - Condor + Glideins + BLAHP + GlideinWMS + Glue
• Glideins on worker nodes create on-demand overlay cluster
Advantages for the Local Scheduler

- Allows PBS to know and account for outside jobs.
- Can co-schedule with local user priorities.
- PBS can preempt grid jobs for local jobs.
Advantages of the Campus Factory

- User is presented with an uniform Condor interface to resources.
- Can create overlay network on any resource Condor (BLAHP) can submit to PBS, LSF,…
- Uses well established technologies: Condor, BLAHP, Glidein.
Problem with Pilot Job Submission

- Problem with Campus Factory: If it sees idle jobs, it assumes they will run on Glideins.
  - Jobs may require specific software, ram size.
  - Campus Factory will waste cycles submitting idle Glideins.
  - Solutions in past were filters, albeit sophisticated.
Advanced Pilot Scheduling

What if we equated:

Completed Glidein = Offline Node
Advanced Scheduling: OfflineAds

- OfflineAds were put in Condor for power management
  - When nodes were not needed, Condor can turn them off
  - Condor needs to keep track of what nodes it has turned off, and their (maybe special) abilities.

- OfflineAds describe an turned off computer.
Advanced Scheduling: OfflineAds

- Submitted Glidein = Offline Node
  - When a Glidein is no longer needed, turns off.
  - Keep Glidein description in an OfflineAd
  - When a match is detected with the OfflineAd, submit an actual Glidein.
- It is reasonably expected that one can get a similar Glidein when you submit to the local scheduler (BLAHP).
Extending Beyond the Campus

- Nebraska does not have idle resources:

  - Running jobs on Firefly.
  - ~4300 cores
Extending Beyond the Campus - Options

- In order to extend transparent execution goal, need to send Condor outside the campus.

- Options for getting outside the campus
  - Flocking to external Condor clusters
  - Grid workflow manager: GlideinWMS
Extending Beyond the Campus: GlideinWMS

- Expand further with OSG Production Grid
- GlideinWMS
  - Creates a on-demand Condor cluster on grid resources
  - Campus Grid can flock to this on-demand cluster just as it would another local cluster
Campus Grid at Nebraska

- **Prairiefire PBS/Condor** (Like Purdue)
- **Firefly** – Only PBS
- **GlideinWMS interface to OSG**
- **Flock to Purdue**
HCC Campus Grid– 8 Million Hours

Cumulative CPU Hours for the Campus Grid
365 Days from Week 12 of 2010 to Week 12 of 2011

Total: 8,060,645, Average Rate: 0.26 /s
Questions?

- Me, my friends and everyone else