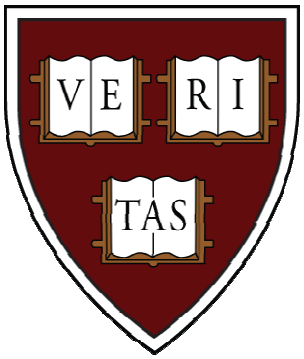


FAS
INFORMATION TECHNOLOGY

Cluster Management and ATLAS at Harvard University

John Brunelle
Harvard FAS IT Research Computing

Faculty of Arts and Sciences

- The Faculty of Arts and Sciences is:
  - Harvard College
  - Graduate School of Arts and Sciences
  - School of Engineering and Applied Sciences
  - Division of Continuing Education
- Over 1000 faculty

FAS INFORMATION TECHNOLOGY

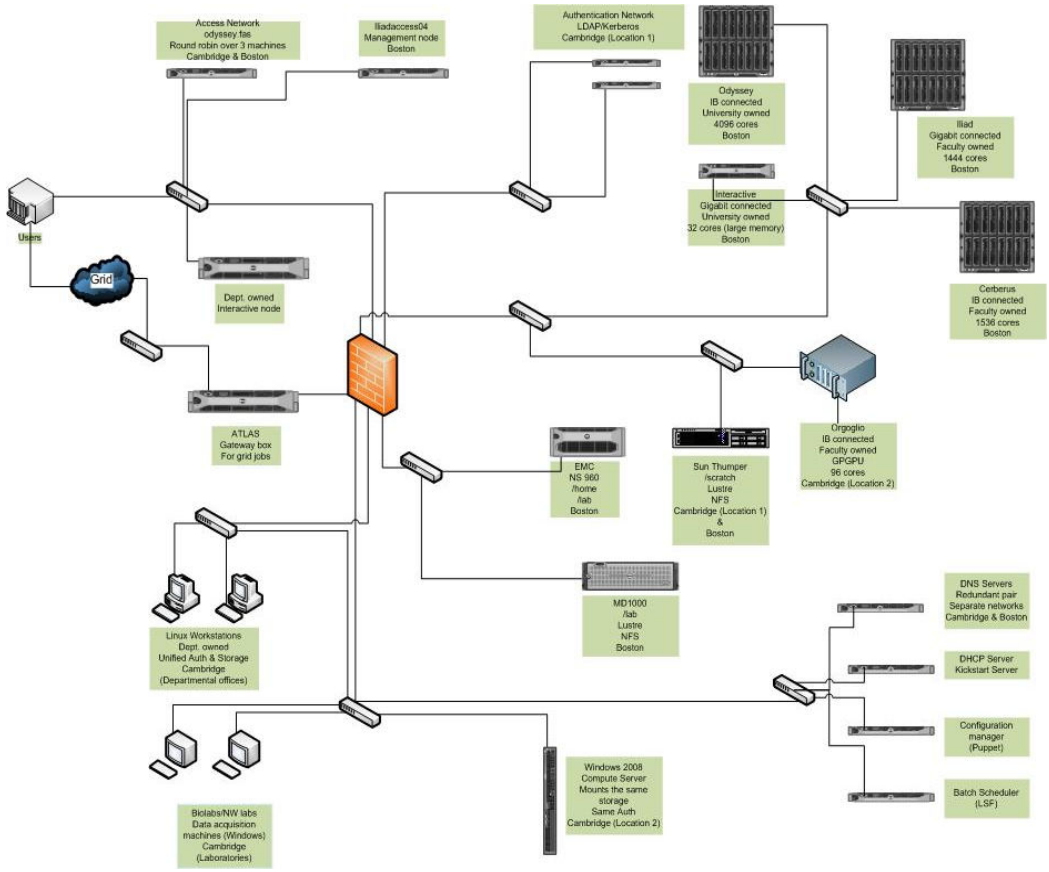
FAS Research Computing

- Old model
  - 1.2MW power in over 15 “data centers” (closets)
  - Many “system administrators”
  - No centralized license management
  - Security nightmare
- New model
  - Faculty purchase hardware
  - We provide
    - System administration
    - Scheduling support
    - Algorithm/programming
    - Subject matter expertise
    - Data center
    - Site Licensing

- **14 staff total, 9 doing HPC**
- **We depend heavily on IT networking and infrastructure team**

Odyssey Cluster

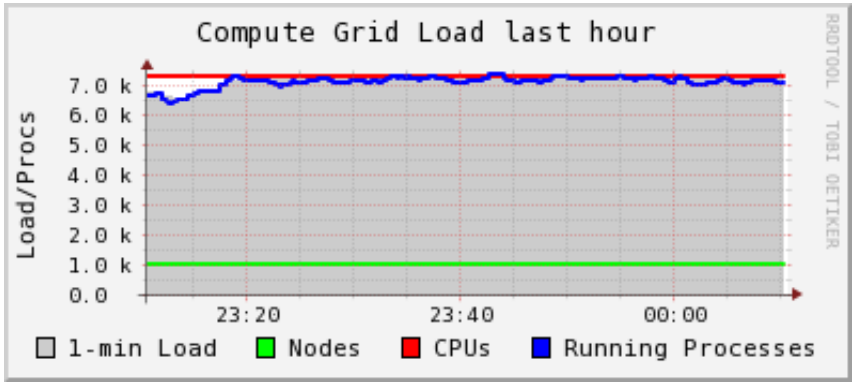
- >7000 cores
- >1 PB storage
- 500 users
- heterogeneous
  - GPGPU
  - IB
  - Gigabit
  - Windows boxes
- Multiple locations
  - Cross town sites
- Shared and exclusive storage (Lustre, NFS)
- RHEL 5, Platform LSF



FAS INFORMATION TECHNOLOGY

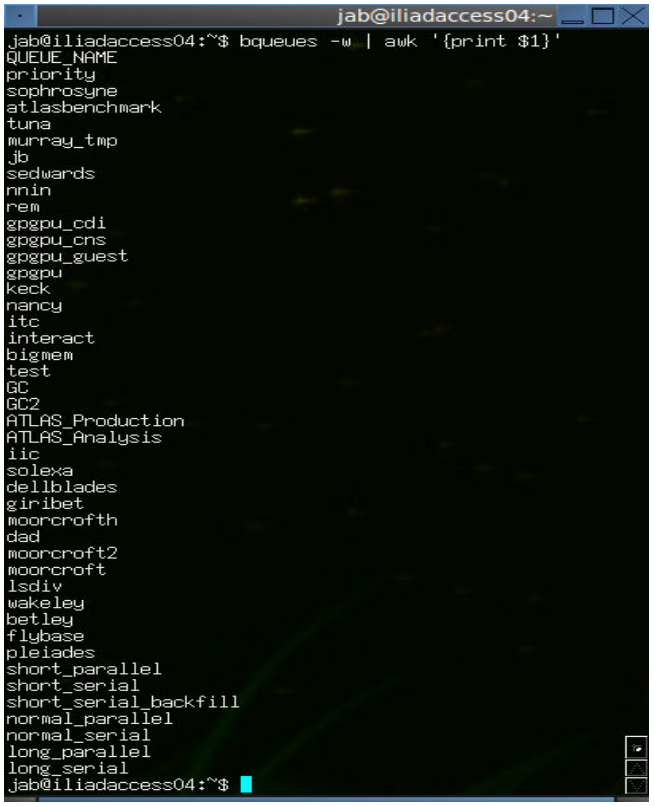
Throughput or On Demand?

- Conflicting mandates
  - Run machines at capacity
  - Faculty owned compute nodes available on demand

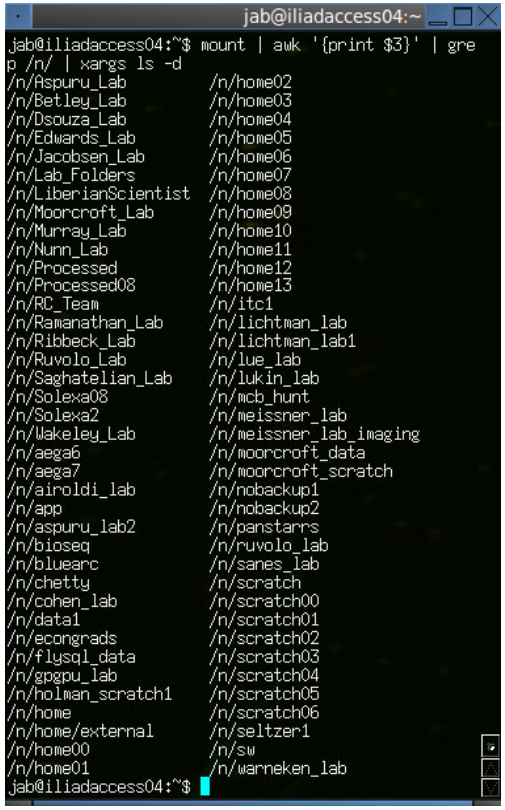


Throughput or On Demand?

- Many Queues



- Many Filesystems



Example Challenge – Storage

- Stability and Quality of service
- Solution – No more shared storage
- Pros
  - Problems are localized
  - Affects smaller groups of users
- Cons
  - Management overhead
  - Lose good value gained through consolidation

Cluster Management: Puppet

- http://reductivelabs.com/products/puppet/
- Declarative language for cluster configuration (based on ruby)
- Clients (cluster nodes) pull their configuration from the server, the *PuppetMaster*
- Great for on-the-fly changes and configuration too dynamic for node image / kickstart

Puppet Example: FS Mount

```
# /etc/puppet/external_node_selector.pl
if ($fullhost =~ m/hero[0-3]\d{3}\.rc.fas.harvard.edu/) {
    @classes = ( 'hero_compute' );
}

# /etc/puppet/nodes/hero_compute/manifests/init.pp
include fstab::aegalfs_ib_mount

# /etc/puppet/modules/fstab/manifests/aegalfs_ib_mount.pp
class fstab::aegalfs_ib_mount {
        file { "/n/scratch":
                ensure => directory,
                backup => false
        }
        mount { "/n/scratch":
                device  => "aegamds1-ib@o2ib0:/aegalfs",
                fstype  => "lustre",
                ensure  => mounted,
                options => "defaults,_netdev,localflock",
                atboot  => true,
                require => File["/n/scratch"]
        }
}
```

Puppet Issues

- Stability
  - *PuppetMaster* instability (occasional huge load making the box unresponsive) – solved by upgrading from ruby 1.8.5 to 1.8.7
  - *Puppetd client* instability (crashing) – worked-around by adding a daily cronjob that restarts the service
- Scalability
  - We have 1000+ nodes with 40+ puppet modules (fstab is just one… auth, modprobe, kdump, ganglia, etc.) and running quite smoothly lately

Cluster Monitoring

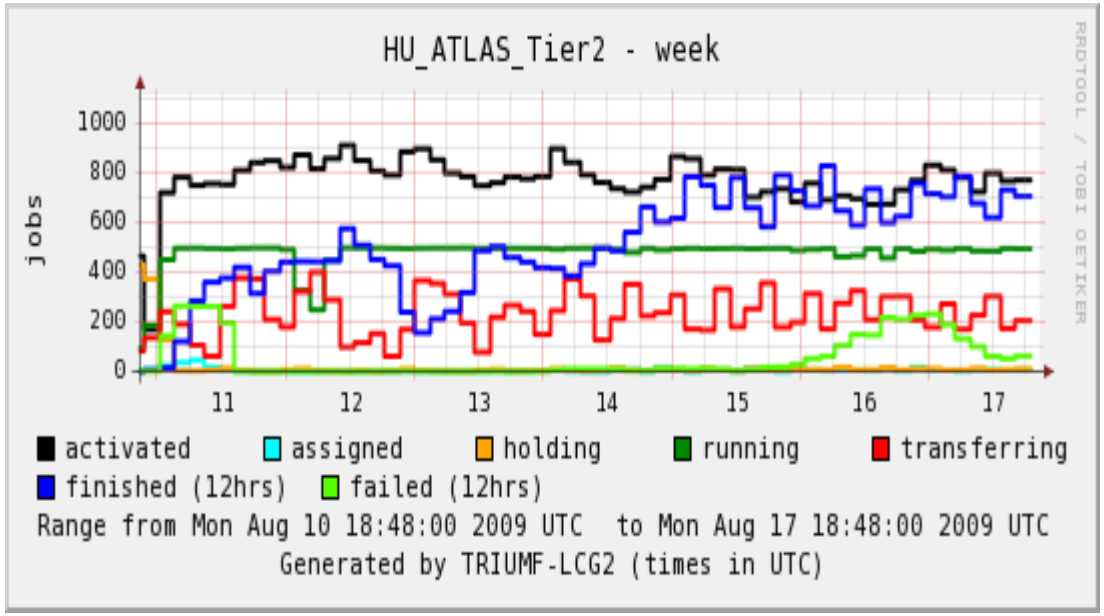
- Ganglia: http://ganglia.sourceforge.net/
  - Nice RRD plots of system stats, individually and in aggregate
  - Easy to setup and use, including plugin architecture for custom status
- Nagios: http://www.nagios.org/
  - Host-alive checks, SNMP, plugin architectures for custom checks
  - Web front-end, email and TXT alerts, etc.

Other Cluster Tools

- pdsh: http://sourceforge.net/projects/pdsh/
  - Parallel, distributed shell code execution
- nfswatch: http://nfswatch.sourceforge.net/
  - NFS server statistics at the command line
  - Breakdown by RPC authenticator (hit a) is the easiest way to see by who's pounding the filesystem
- kdump: http://lse.sourceforge.net/kdump/
  - kexec-based kernel crash dump mechanism, supporting custom scripts to be run at crash time

ATLAS Distributed Computing

Harvard is part of the Northeast Tier 2 Center, which is centered at Boston University



FAS INFORMATION TECHNOLOGY

NET2 Setup

- Boston University, Harvard
- Gatekeepers and LFCs at both BU and HU sites
- BU has the storage, and runs the single site-services instance
- BU dedicated cluster, HU shared cluster
- Tufts involved, too (a friendly Tier3)

Our experiences are relevant to a Tier3-gs, esp. integrating into an existing cluster

FAS
INFORMATION TECHNOLOGY

Firewall / Network ACL Issues

- Authentication to the gatekeeper requires a reverse DNS lookup of the gatekeeper IP to match the gatkeeper hostname… cannot use NAT!
- ATLAS uses Globus two-phase commit: two-way handshakes where the gatekeeper dynamically binds to ephemeral, high-numbered ports and expects clients to initiate connections
- Port range be limited with GLOBUS_TCP_PORTRANGE, but in any case will need a large block (possibly thousands) of ports open

Firewall / Network ACL Issues (cont.)

- Pilot traffic itself is over HTTP(S) and has no problem using a proxy
- Authentication to LFC uses GSI GSSAPI – SOCKS proxy or simple SSH tunnel will not work (iptables nat REDIRECT *may* work)
- Conditions Data – direct access to BNL Oracle should now be replaced by local SQUID/Frontier setup
- Data stage-in / stage-out

Firewall / Network ACL Solutions

- Gatekeeper is outside the cluster network and heavily locked down
  - open only to specific sets of internet subnets and grid users
  - *Not viable for Analysis queue setup*
- Data stage-in / stage-out done with custom Local Site Mover using ssh/scp: http://www.usatlas.bnl.gov/twiki/bin/view/Admins/LocalSiteMover
- Cluster nodes uses Policy NAT to access the internet
  - Before this, we were getting along okay using various proxy schemes, two LFCs, etc., but it was a large headache

Filesystem Needs

- Home directory performance is crucial
  - Many small files, many i/o operations
  - Especially if jobs get backed up or killed for some reason
- Worker nodes need large local scratch disks
- ATLAS kits on lustre kept triggering a lustre bug (hardlinked files were at the heart of it)
- Full disks cause lots of problems!

  Starting/stopping services can swap in empty files (like /etc/services); fetch-crl will make empty files and then refuse to update them when space frees up

FAS
INFORMATION TECHNOLOGY

Other Gotchas

- PanDA Jobs cannot be pre-empted
- Starting/stopping services modifies root's crontab, system files, etc. (beware of conflicts with puppet, cfengine, etc)
- Must manage services from a clean environment – e.g. no X-forwarding associated with your shell session!
- xinetd per_source limits
- Changing installation directory path is difficult

Local ATLAS Users

- We also have about 10-20 local ATLAS people using the Odyssey cluster at any given time
- Local users so far prefer direct cluster usage rather than pAthena
- DQ2_LOCAL_SITE_ID, dq2-get --threads
  - So far, just handling issues as they arise (mainly storage throughput), and its not a large burden
  - dq2-get is starting to show more and more issues (contention on local resources)

The Future

- SRM at Harvard (if necessary; GridFTP is already there)
- Frontier/Squid setup, for local users

Thank you