Computing Division

Scientific Computing Services

Town Hall Meeting – Unix Services

Shirley Gruber, April 10, 2014

Scientific Computing Services

- Scientific Computing Services home page for customers

https://confluence.slac.stanford.edu/display/SCSPub/Scientific+Computing+Services+Home

shirley@slac.stanford.edu

650-926-3146

unix-admin@slac.stanford.edu

Town Hall Meeting – Unix Services

SLAC

Objectives:

- communication
- collaboration
- Community of Practice (CoP)

"Enables Members to expand their own and their organization's capabilities"

unix-community@slac.stanford.edu

69 subscribers

email to: listserv@slac.stanford.edu

subscribe unix-community

Town Hall Meeting – Unix Services

SLAC

Agenda:

- Introductions
  - Requested by attendees of last meeting
  - Around the room, brief, name, department
- Storage Team
  - GPFS status, HPSS plans
- Unix Team
  - RHEL7, RHEL5, Solaris
- High Performance Computing (HPC)
  - Lustre, LSF, cgroups, MPI
- Upcoming – ServiceNow, potential outages
- Plans for future Town Hall meetings
- Discussion

Storage

Scientific Computing Services

Andrew May, April 10, 2014

Scientific Computing Services - Storage

SLAC

- GPFS
  - IBM General Parallel File System (GPFS)
  - Clustered NFS (CNFS): global namespace, scalability, performance, redundancy, manageability
  - Internal test environment
  - MCC production environment
  - ACD environment (in progress)
  - Goal of a shared, tiered storage environment
  - Automatically move GPFS file to TSM tape

Scientific Computing Services - Storage

SLAC

- IBM Tivoli Storage Manager (TSM) 5.5 -> 6.4 upgrade

  TSM 5.5 goes End of Support on 4/30/2014
  - GPFS tiered storage will use TSM 6.4

- HPSS upgrade
  - Multi-step upgrade involving RHEL and HPSS
  - RHEL upgraded from 5.8 to 5.9
  - DB2 database upgraded from 9.5.9 to 9.7.8
  - HPSS upgraded from 7.3.3.p8 to 7.3.3p9a
  - Target of April 22/23 for second step
  - Eventual HPSS 7.3 -> 7.4 upgrade
    - (outage eventually required)

SLAC

Questions?

Unix Support

Scientific Computing Services

Karl Amrhein, April 10, 2014

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

RHEL 5 status

SLAC

- RHEL5
  - On 1-Aug-2014, RHEL 5 machines will be removed from:
    LSF batch farm, build farm (bldlnx), and interactive login pools (rhel5-32,64).

    Exception: Limited login/usage will remain for:

    current version of Fermi L1 processing requires RHEL5 bldlnx and interactive login for another year or so, until next version of L1 goes live. BaBar also wants this access.

    The rhel5-32 and rhel5-64 login pools will be removed, and the machines in those pools will be added to the rhel6-64 login pool, and rhel7 pool when ready.

    noric alias being moved to point to rhel6-64 on 5-May-2014.

RHEL 5 LSF hosts to be reinstalled

~940 LSF hosts still running RHEL5.

We will coordinate the RHEL6 reinstall or retirement with the affected groups:

- 4 bldlnx
- 247 fell
- 1 glastlnx15
- 80 oak
- 80 psana
- 112 psanacs
- 40 sdc
- 5 simes-gpu
- 350 suncat
- 23 yili

Solaris to be retired

- Solaris 10
  - Currently running on ~ 300 hosts
- Infrastructure
  - AFS, NFS, Oracle, NIS, PeopleSoft (HR/Business), web
- Science groups (NFS servers)
  - atlas, babar, glast, harvard, iowa, cd, eb, ec, ee, esa, exo, kipac, lcd, wisconsin

RHEL 7 update

Public beta released Dec-2014. No official GA date.

(GA will be in less than 1 month)

High Touch Beta Program - 8 weeks during Jan–Mar 2014

50 Red Hat customers participated

Once-per week snapshot ISOs released with latest patches

Access to private group portal on red hat customer portal

Weekly technical webinars on RHEL7 changes/topics:

- Anaconda Installer
- Systemd
- Containers
- Firewall Management with firewalld
- Filesystems
- In-place Upgrade
- Performance Tips
- Identity Management
- OpenLMI
- App profiles, Image Creation & Deployment
- Networking
- Containers Update

High Performance Computing

Scientific Computing Services

Yemi Adesanya, April 10, 2014

U.S. DEPARTMENT OF ENERGY
Office of Science

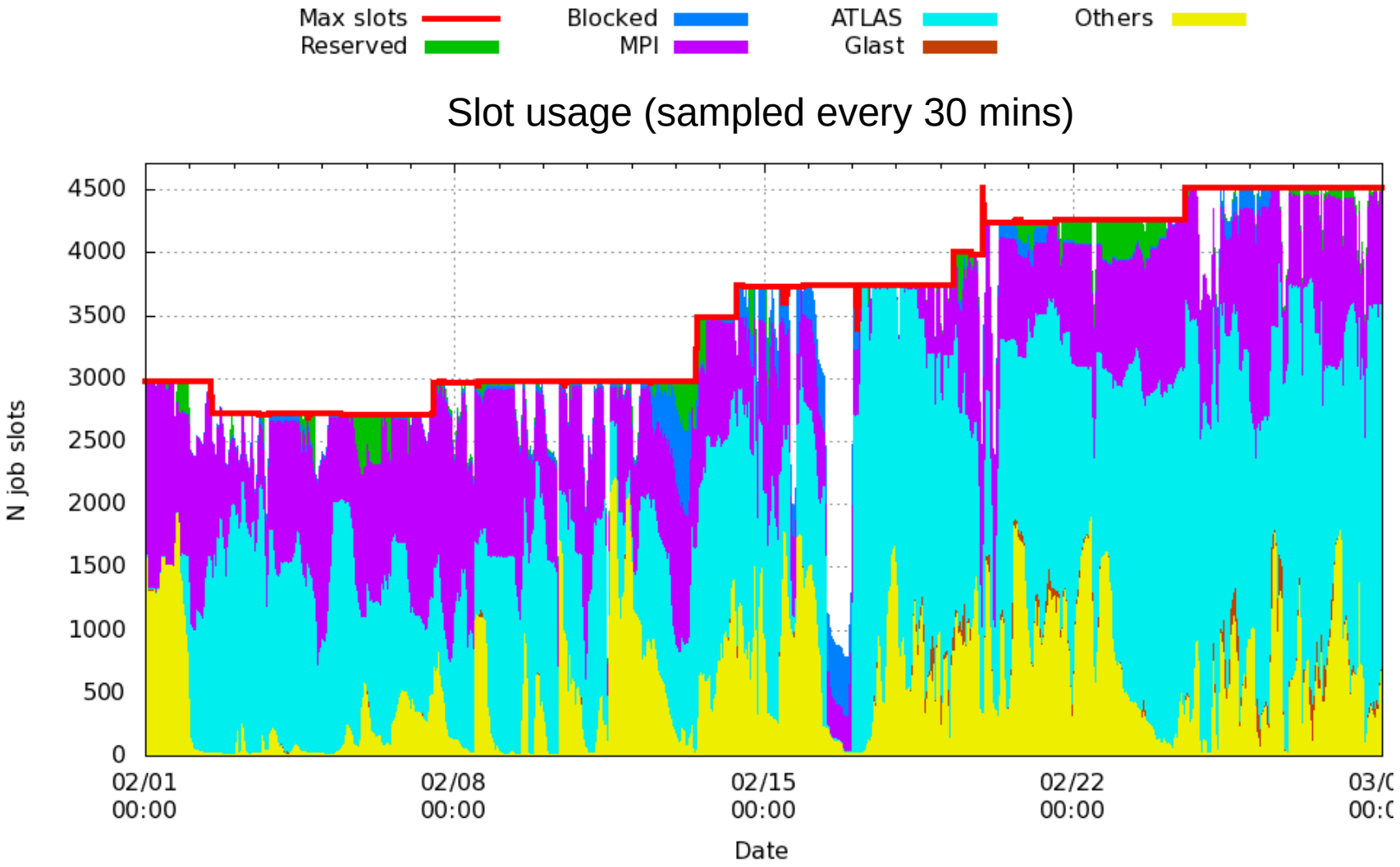
SLAC NATIONAL ACCELERATOR LABORATORY

Bullet cluster upgrade

SLAC

- From 2864 (production) cores to 4512 cores in Q1 2014
- Identical configuration across all 288 hosts:
  - Dual E5-2660 8-core CPUs @ 2.2GHz
  - 64GB RAM (4GB per core)
  - 10Gb ethernet to SLAC network (oversubscribed)
  - 40Gb Infiniband for MPI (50% blocking)
- Additional Infiniband switch ports and recabling work
- Homogeneous expansion is possible via fiber Infiniband

Bullet Cluster usage



Storage must scale with compute capability

SLAC

- Increased compute power = Increased parallelism = greater potential for I/O bottlenecks
- NFS server crashes often coincide with expansion of batch clusters
- Scalable, striped filesystems are a common feature of modern HPC environments
- Reduce the reliance on local scratch disk space

Lustre filesystem upgrade

SLAC

- Doubled the storage capacity from 170TB to 340TB
- Doubled the number of RAID controllers
- Upgraded ethernet from 3x1Gb to 10Gb (per-server)
- 40Gb Infiniband I/O for bullet cluster clients
- Ethernet I/O for KIPAC hosts on SLAC network
- Reinstalled and reformatted with version 2.5.1
- Completed in a 5-day outage window including 110TB data migration

Lustre filesystem upgrade

SLAC

- Benchmarks for parallel I/O via Infiniband
  - ~400MB/s max for 1 stream on 1 bullet node
  - ~3000MB/s max for 16 streams on 1 bullet node
  - ~6500MB/s max for 32 streams on 8 bullet nodes

- Lustre performs best with transfer operations >=1MB
- Mounted under /lustre/ki/pfs
- Per-user KIPAC directories
- Scratch space area for Fermi jobs

https://confluence.slac.stanford.edu/display/SCSPub/PPA+Lustre+filesystem+2014+upgrade

LSF batch system progress

SLAC

- LSF upgraded to 9.1.2 (3/2014)

- Shared cluster model is working
  - MPI jobs have priority on the bullet hosts
  - Jobs from general queues, ATLAS and Fermi running alongside MPI tasks
  - Please join the OPENMPI mailing list (listserv.slac.stanford.edu)

- Continue to research new features to improve/optimize LSF performance
  - Minimize “slot fragmentation” by packaging single-slot jobs
  - VM-based Job migration?
  - Improve robustness with rigid resource limits on jobs

LSF enforcement using cgroups

- Default CPU core and memory restrictions are “soft”

- A single-slot (SS) job is not confined to one CPU core
  - SS job’s threads and child processes can run on all cores
  - User may be unaware of threads or forking in APIs

- Specifying memory limits
  - We have no memory limit defaults
  - Apps with memory leaks can take down hosts
  - LSF checks are periodic so apps could exceed limits between checks

LSF enforcement using cgroups

SLAC

- Linux Control Groups AKA “cgroups”
  - Introduced in RHEL6
  - Kernel-level resource restrictions that apps cannot bypass
  - LSF ties job’s PGID to cpuset and mem cgroups
  - One CPU core per job slot
  - LSF kills job immediately once memory limit is reached

- Single-slot job bound to 1 CPU core with 200MB physical RAM limit:

```
“bsub -q testq –R ‘affinity[core:membind=localprefer]’
–M 200 ./myjob”
```

LSF enforcement using cgroups

SLAC

- CPU enforcement
  - Now in production for ATLAS queue (atlasq)
  - Under consideration for all general queue jobs

- Memory enforcement
  - Can we define a default memory (RAM) limit for all single-slot jobs?

- Enforcement *might* work for parallel jobs too:

```
"bsub -q testq -n 64 -R 'span[ptile=16]
affinity[core:membind=localprefer]'
-M 200 ./myjob"
```

Questions?

OCIO – Upcoming plans

- ServiceNow rollout in June
  - web and email interface
  - Incident management
    - "Something is broken", e.g., host X is not responding
  - Request management
    - "I need something, e.g., request for access, storage, backup

- Outages to move from old EOL switches to newer ones
  - needs discussion

- Outages for project to build out rack space in Building 50
  - Interim server moves
  - Will coordinate with customers over the next few weeks
  - Potential outage in June
    - Bullet cluster
    - Suncat cluster (new)

FutureTown Hall meetings

- *Some proposed dates for 2 upcoming meetings*
  - *June*
    - *Thursday, 6/5, 2pm*
    - *Thursday 6/12, 2pm*
  - *July*
    - *Thursday 7/03, 2pm*
    - *Thursday 7/10, 2pm*
    - *Tuesday 7/22, 2pm*
  - *October*
    - *Tuesday 10/7, 2pm*
    - *Thursday 10/9, 2pm*
    - *Tuesday 10/14, 2pm*
    - *Tuesday 10/21, 2pm*

SLAC

Discussion

Scientific Computing Services

- Scientific Computing Services home page for customers

https://confluence.slac.stanford.edu/display/SCSPub/Scientific+Computing+Services+Home

shirley@slac.stanford.edu

650-926-3146

unix-admin@slac.stanford.edu