Computing Division

Scientific Computing Services

Town Hall Meeting – Unix Services

Yemi Adesanya, January 14, 2016

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

Scientific Computing Services home page

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

unix-admin@slac.stanford.edu

support/questions

yemi@slac.stanford.edu

650-926-2863

Town Hall Meeting – Unix Services

Objectives:

- communication
- collaboration
- Community of Practice (CoP)

unix-community@slac.stanford.edu

email to: listserv@slac.stanford.edu

subscribe unix-community

Agenda:

- UNIX Storage
- Strategy for Cluster Services
- UNIX Platform
- GPU Computing Support
- Questions/Discussion

UNIX Storage

Scientific Computing Services

Lance Nakata, January 14, 2016

Storage-as-a-Service (StaaS)

- Shared, clustered parallel filesystem using GPFS
- $100/TB/year pricing targeted at programs with limited budgets or capacity requirements (10's of TBs)
- Initially targeting moderate performance needs
- Access via NFS; optional native GPFS access for RHEL
- Looking at possible access via Samba
- Service in production; charging expected in FY17
- All NetApps being moved to StaaS in advance of vendor support phase-out. Groups will need to budget for this. We will provide estimates to those affected.
- 120TB allocated, 54TB in use (out of 320TB)

Tape Storage

SLAC

- Upgrade from 1TB to 8TB tape drives in FY16
  - 8TB drives increase tape library capacity to 100PB
  - 5TB and 8TB tape drives use the same media
  - Would decrease tape purchase cost by 37.5% vs. 5TB tapes
  - Need to find funding
- Retire unfunded astore/mstore service
  - Looking at HPSSfs and GPFS HSM as possible solutions that provide NFS-like interface
  - May require some form of charge-back unless SLAC-funded
  - Questions to ponder:
    - Where does data go when project funding ends?
    - Can we house it cheaply at SLAC? In the cloud?

Storage Tasks and Futures

- Continue work on automated disk-to-tape file migration. Has direct application to Storage-as-a-Service/GPFS use as a way of managing disk costs
- Check current storage building blocks for config changes due to new hardware releases
- Price Spectrum Scale/GPFS appliances to see if there may be cost savings vs. do-it-yourself
- Look at object storage as a possible disk tier
- Look at cloud storage to see where it might fit

Questions?

Strategy for Cluster Services

Scientific Computing Services

Yemi Adesanya, January 14, 2016

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

Strategy for Cluster Services

- SCS is supporting ~19K compute cores across the lab
- Multiple clusters – both shared and dedicated
- Opportunities for consolidation and optimization
- Let’s take a closer look at utilization
- Establish some acceptable policies for lifecycle management
- Option of chargeback for service instead of hardware purchase (we can lease servers)

Strategy for Cluster Services

SLAC

- Many groups share their cluster resources with other users
- Funding sources are combined to purchase hardware
- Stakeholders are usually willing to share as long as their production activities are not negatively impacted
- Can groups buy “service” instead of buying servers?
- Can we establish policy on when servers become End-Of-Life?
- Faster provisioning? Do we have to work on procurement every time a group needs more compute?
- Improve utilization – some work is bursty so why provision based on theoretical maximum?

Strategy for Cluster Services

- Fairshare – a commodity unit for cluster utilization
- Fairshare controls job scheduling priority
- Groups with a fairshare have a guarantee of utilization
- Distribute fairshares based on ownership of shared cluster hardware
- Apply a fairshare tax (15%) to fund non-paying users so they can run on the cluster
- Remove associated fairshares when clusters are retired
- Lease cluster hardware and recover costs by charging per-fairshare

Strategy for Cluster Services

SLAC

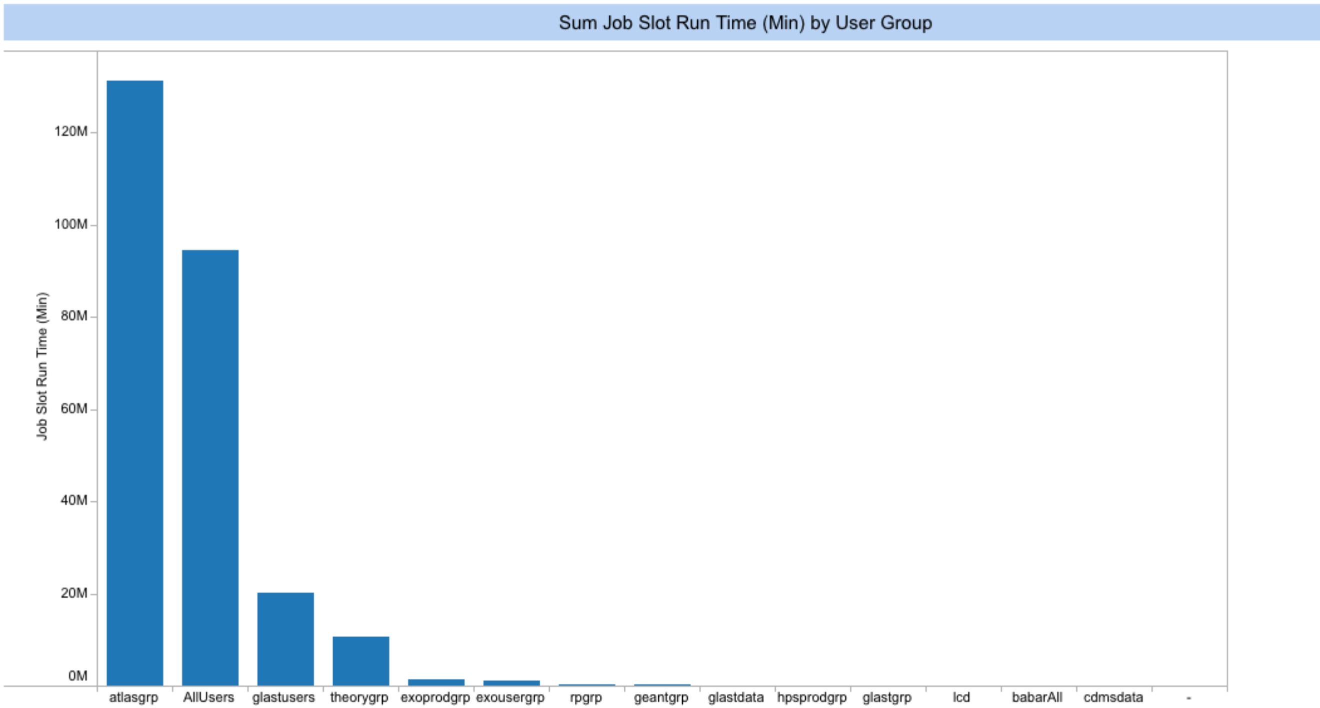
- Run “bqueues -l <short|medium|long>” to view dynamic scheduling priority:

SHARE_INFO_FOR: short/

USER/GROUP	SHARES	PRIORITY	STARTED	RESERVED	CPU_TIME	RUN_TIME	ADJUST
luxlz	3500	1166.667	0	0	0.0	0	0.000
cdmsdata	2000	634.012	0	0	794.6	0	0.000
lcdprodgrp	1100	366.667	0	0	0.0	0	0.000
lcd	600	200.000	0	0	0.0	0	0.000
glastdata	854	187.541	0	0	7990.3	0	0.000
glastgrp	366	103.535	0	0	2751.6	0	0.000
geantgrp	3874	58.937	0	0	322618.0	0	0.000
babarAll	7859	8.351	260	0	419907.8	393007	0.000
hpsprodgrp	1000	5.180	6	0	840340.1	44546	0.000
exoprodgrp	1500	2.189	0	0	3509388.5	0	0.000
rpgrp	500	0.603	6	0	4079954.0	78100	0.000
glastusers	23181	0.579	2422	0	161182064.0	7349614	0.000
atlasgrp	31157	0.211	1486	0	472073344.0	265885455	0.000
exousergrp	550	0.167	49	0	12735401.0	3386393	0.000
AllUsers	14523	0.140	1091	0	362196416.0	152758503	0.000
theorygrp	4257	0.120	510	0	121031384.0	53175958	0.000

https://confluence.slac.stanford.edu/display/SCSPub/Stakeholder+priority+on+the+Shared+Farm

Analytics: Run Time Usage on shared farm



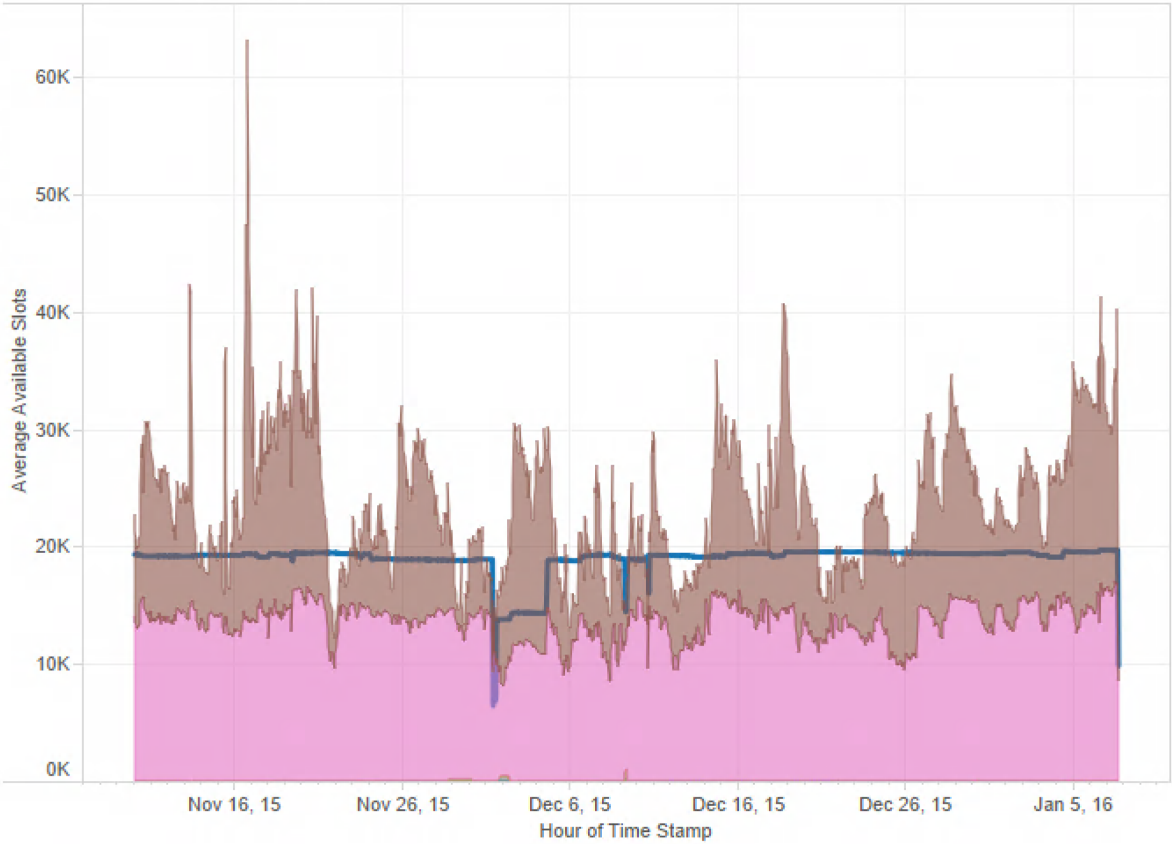
User Group	2015	2016	Roll-up
atlasgrp	88,275,399	42,954,581	131,229,980
AllUsers	69,253,043	25,118,558	94,371,601
glastusers	14,812,457	5,341,170	20,153,627
theorygrp	4,618,874	6,132,235	10,751,109
exoprodgrp	1,190,326	165,125	1,355,452
exousergrp	959,419	176,160	1,135,579
rpgrp	107,952	175,288	283,240
geantgrp	192,633	1,410	194,043
glastdata	48,913	11,561	60,474
hpsprodgrp	16,799	16,237	33,037
glastgrp	20,564	4,863	25,427
lcd	818		818
babarAll		258	258
cdmsdata	62	77	138
-	0	0	0
Grand Total	179,497,261	80,097,521	259,594,783

Analytics: Cluster Slot Utilization

Slots by Dimension

Selected Dimension	Average Pend Slots	Average Run Slots
-AVAILABLE SLOTS		
PEND	10,272.04	
RUN		13,554.63
SSUSP		
UNKWN+RUN		
UNKWN+SSUSP		
USUSP		
WAIT		

Workload Utilization Chart



Durations

Available Hours	26,780,068
Pend Hours	14,466,454
Run Hours	19,089,437

Pend to Run Ration

Average Pend Slots	10,272
Average Run Slots	13,555
Pend to Run Ratio	1

Run Ratio

71.28%

Virtualization with OpenStack

- Infrastructure-as-a-Service (IaaS)
  - Private cloud interface for spinning up VMs
  - Ideal for test environments
  - Production replacement for Nebula environment
- Batch clusters
  - OpenStack VMs as batch nodes
  - LSF farms that grow/shrink dynamically
  - Spin up batch nodes to meet current demand
  - Provision virtual clusters immediately
- Common hypervisor hardware (blade servers)
- Optional Chargeback models (TBD)

Questions?

UNIX Platform

Scientific Computing Services

Karl Amrhein, January 14, 2016

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

Unix Platform Update

- Red Hat Enterprise Linux 7
- Data center virtualization and lifecycle management
- OpenStack private cloud
- AWS and Azure public cloud
- Vision for virtualization and cloud services

Red Hat Enterprise Linux (RHEL) 7

- Chef configuration management status
- Red Hat Enterprise Linux 5 – EOL
- Server, Interactive Login, Batch, Desktop
- Desktop Support

Data center virtualization and lifecycle management

SLAC

- Aging hardware in data center
- No budget to replace all bare metal servers
  - Baremetal footprint reduction
- VMware infrastructure in place
- Physical to Virtual (p2v) efforts underway
- Vision: software defined and API driven datacenter

OpenStack private cloud

- Test environment – Nebula
- Production environment – RDO (Iaas and Batch)
  - Using automated deployment and config mgmt tools
- Working with OpenStack community:
  - Tim Bell, CERN
  - New Scientific OpenStack working group
  - OpenStack user community group

AWS and Azure public cloud

- AWS testing underway
  - Data archiving
  - Batch compute via spot pricing (BNL is already doing this)

- Working with NuSpective – professional services
  - Technical support and ongoing advice

Vision for virtualization and cloud services

- Cloud Management Platform (CMP)
- Avoid large, unstructured ec2 instance sprawl
- CMP can provide an automated, secure, auditable, cloud computing environment
- Put workloads on appropriate platform:
  - Baremetal if necessary (on-prem compute clusters)
  - VMware for traditional legacy virtualization
  - Private cloud (OpenStack) – on-prem, horizontally scalable
  - Public cloud – bursty workloads, provide capability for peak workloads without the requirement for bare metal purchase. Take advantage of AWS products and off site Availability Zones.

Questions?

GPU Computing Support

Scientific Computing Services

Yemi Adesanya, January 14, 2016

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

The Future of HPC

SLAC

- DOE is committed to HPC (High Performance Computing) innovation
- The future of HPC depends on massive parallelism
- CPU clock speeds are not getting faster - Moore's law does not apply
- Get ready for parallel computation with hybrid CPU/GPU and many-core clusters
- DOE is funding next-generation hybrid clusters for Livermore and Argonne
- GPU programming is not trivial; scientists will need training and access to subject matter experts
- Future SLAC scientific computing will have to leverage GPU and many-core in order to scale

GPU Strategy

- Provide shared resources whenever possible to minimize lab costs and maximize utilization
- Provide a shared GPU resource available to all SLAC users
- Use indirect-funding when possible
- Optional chargeback for high-priority projects/users
- Provide access to GPU training and facilitate development with the SLAC/Stanford community

GPU Outreach and Training

- SCS has already gathered feedback and requests for a shared GPU cluster:
- KIPAC, LCLS, SSRL, EPP Theory/Simulation, Biosciences, LSST
- Costs prohibit any single project from funding an entire shared cluster ($30K upwards per server)
- SCS co-hosted Intel Xeon Phi programming workshop in 2015

NVIDIA training for 2016

- Partner with NVIDIA to host GPU training
- Mutual interest in Scientific Computing
- SLAC HPC codes could shape future GPU architectures
- Identify popular algorithms/frameworks at SLAC
- NVIDIA maintain a repository of GPU-optimized libraries and functions
- Identify key software developers to spearhead GPU adoption

Zoox GPU cluster collaboration

- Discussions between OCIO and Zoox began on 2/2015
- DOE legal agreement (CRADA): Zoox relationship must provide value to SLAC's science mission
- Zoox buys GPU cluster hardware, Computing Division will host the cluster and provide service
- SLAC GPU developers will have access to the cluster
- Zoox will fund Computing Division effort (labor) for cluster
- Planning for 5 racks of GPU cluster hardware and storage in building 50
- SCS is developing the cluster specs with Zoox:
  - Dense config with 8 Titan-X GPUs per server

Questions?