



Introduction to SUSE HPC Modules

Out-of-Box High Performance Computing Cluster Deployment

Lawrence Kearney
System Administrator Principal
The University of Georgia
lkearney@uga.edu

A Little about Me

- **SUSE evangelist**
- **SUSE Certified Instructor**
- **HPC infrastructure specialist**

- **Presence on SUSE and openSUSE forums**
- **Presence on SUSE Community blogs**
 - Deploying SLURM using SLE HPC patterns
 - Deploying SLURM PAM modules on SLE compute nodes

- **Motorcycle owner and enthusiast**

Agenda

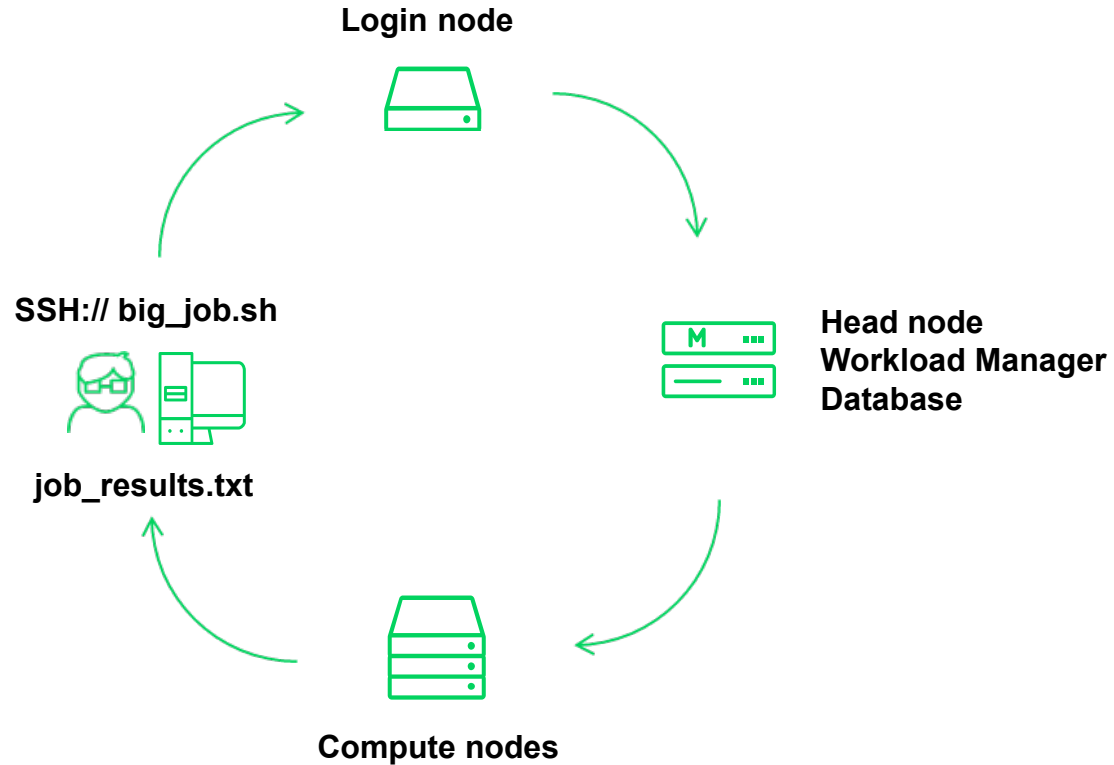
- **Cluster architecture overview**
- **Cluster deployment overview**
- **Workload Manager overview**
- **Workload Manager configuration**
- **Basic interaction with a demonstration cluster**
- **Typical file system overview**
- **Question & Answer**

Typical HPC Environments

Minimal HPC Cluster Components

- **Head node**
- **Compute nodes**
 - Intel, AMD, GPU, ARM
- **Login node**
- **Workload Manager**
 - Resource management and scheduling
 - Job management and scheduling
- **Database or flat file**

Basic HPC Job Flow



Additional HPC Components

Nodes

- **Monitoring node(s)**
- **Metrics collection node(s)**
 - Cluster resources
 - Jobs
 - Users
- **Application delivery node(s)**
- **Application development node(s)**
- **Data transfer nodes**
- **Storage nodes**

Additional HPC Components

File Systems

- **Specialized file systems**
 - Lustre, GPFS
 - Panasas
 - Ceph
 - ZFS
- **Specialized data transfer applications**
 - Globus Online
- **File system auditing**
 - Robinhood

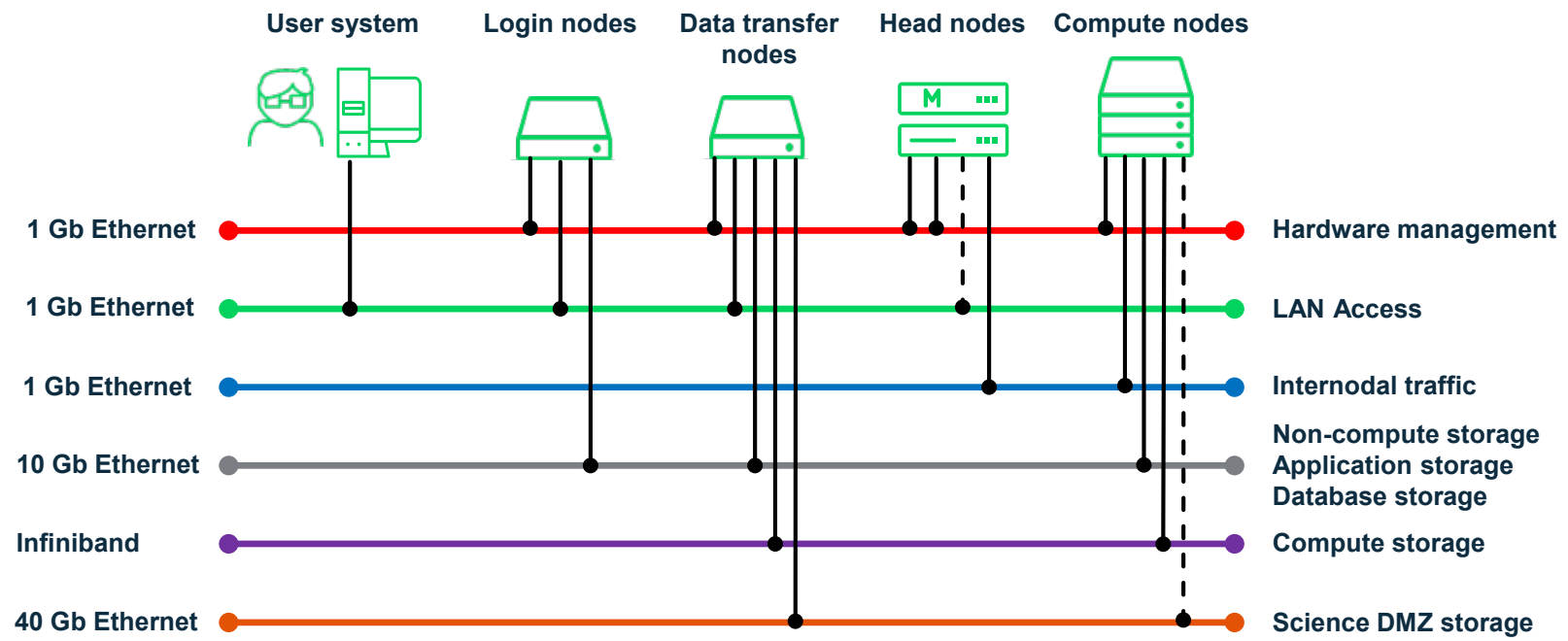
Additional HPC Components

Connectivity

- **High-speed system interconnects**
 - 1, 10, 25, 40, 100 Gb/s Ethernet
 - Buffer vs line speed vs density
 - QDR, FDR, EDR, HDR Infiniband
 - Intel, Mellanox
- **High-speed WAN infrastructures**
 - Internet 2
 - Science DMZ deployment
 - Bursting HPC workloads to cloud infrastructure

Additional HPC Components

Connectivity



Additional HPC Services

Infrastructure Services

- **Identity stores**
 - Active Directory
 - SASL/Kerberos
 - LDAPS
- **DNS/DHCP/NTP**
- **Virtualization infrastructure**
- **Container integration**
- **Perimeter security**
- **Node deployment, management and recovery**

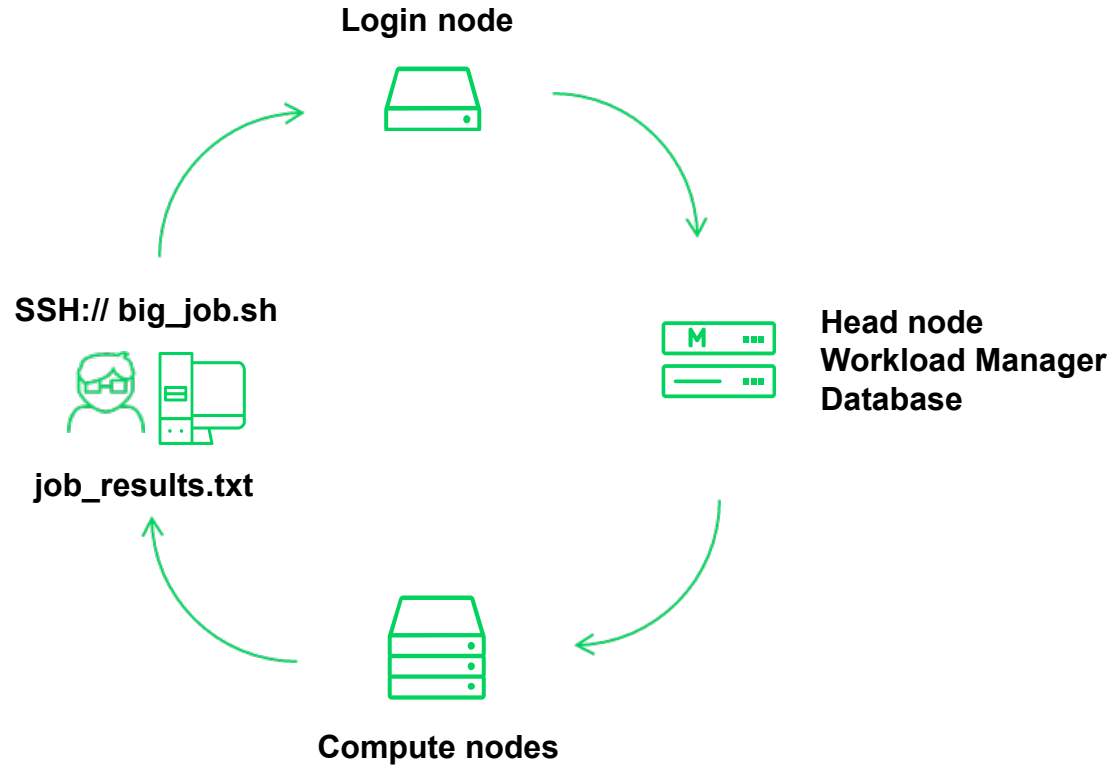
Additional HPC Components

System and Application Software

- **Custom repositories or mirrors**
 - No direct consumption of public repositories
 - Cluster dependencies
 - Kernel dependencies
 - Application dependencies

Where do HPC applications come from, anyway?

Back to Basics HPC Workflow



SUSE Linux Enterprise Server for HPC

SUSE Linux Enterprise for HPC

What Is SLE HPC?

Open Source advantage with superior support

- Built on standard SLE Server base
- Reduced subscription costs
- Extended Service Pack Overlap Support
- Long Term Service Pack Support for HPC
- Package Hub for HPC

<https://www.suse.com/products/server/hpc/>

SUSE Linux Enterprise for HPC

Deploying SLE HPC

- **Begin with the SLES Unified Installer**
- **Base product selection**
- **Extension and Module selections**
- **System role selections**
- **Configuration and testing**

System Role Selection

System Role

○ HPC Management Server (Head Node)

- Uses xfs as the default root filesystem
- Includes HPC-enabled libraries
- Disables firewall and kdump services
- Installs controller for the Slurm Workload Manager
- Mounts a large scratch partition to /var/tmp

○ HPC Compute Node

- Uses xfs as the default root filesystem
- Includes HPC-enabled libraries
- Disables firewall and kdump services
- Based from minimal setup configuration
- Installs client for the Slurm Workload Manager
- Does not create a separate home partition
- Mounts a large scratch partition to /var/tmp

○ HPC Login and Development Node

- Includes HPC-enabled libraries
- Adds compilers and development toolchain

Minimal System Package Examples

A Little More about System Roles

- **Compute nodes**
 - Minimal OS
 - Modularized libraries
 - Compute node pattern
- **Login nodes**
 - Minimal OS, system, and SLURM utilities
 - Modularized libraries
- **Data transfer nodes**
 - Minimal OS and sparse utilities

Minimal System Package Examples

A Little More about System Roles

- **Development (software build) nodes**
 - More complete OS
 - Modularized libraries
 - Development tool chains
 - Compilers
- **Also consider**
 - Multiple hardware architectures
 - Administrator and user access

Minimal System Partitioning Examples

A Little More about System Roles

- **Head and compute nodes**
 - Create partition (6.00 GiB) for / with xfs
 - Create partition (12.00 GiB) for /var/tmp with xfs
 - Create partition (2.00 GiB) for / swap
- **Application development nodes**
 - Create partition (12.00 GiB) for / with btrfs
 - Create partition (6.00 GiB) for /home with xfs
 - Create partition (2.00 GiB) for / swap

Installation Summary

- **Product: SUSE Linux Enterprise High Performance Computing 15**
- **Product: HPC Module**
- **Patterns:**
 - HPC Workload Manager
 - HPC Basic Compute Node
 - HPC modularized Libraries
 - HPC Development Packages
 - Infiniband (OFED)

Workload Management Stack

Simple Linux Utility for Resource Management Overview

SLURM Control daemon (slurmctld)

- Head node
- Resource and compute job management

SLURM Database daemon (slurmdbd)

- Head node or dedicated node
- SLURM cluster configuration and compute job metric store

SLURM daemon (slurmd)

- Compute node
- Compute job execution

Workload Management Stack

Recommendations

- **SLURM Control daemon (slurmctld)**
 - Instances on primary and secondary head nodes
 - Use a remote database
- **SLURM Database daemon (slurmdbd)**
 - Dedicated node using SSDs for database storage
 - Consider using a single database for multiple clusters
- **SLURM daemon (slurmd)**
 - Implement SLURM Pluggable Authentication Modules (PAM)

Workload Management Stack

Configuration Files

/etc/slurm/slurm.conf

- All hosts running slurmctld and slurmd
- All login nodes with select SLURM binaries

/etc/slurm/slurmdb.conf

- slurmdbd configuration
- Database hosts only

/etc/my.cnf.d/innodb.cnf

- Database optimizations and configuration

Workload Management Stack

Database Configuration Overview

mariadb: what needs to be configured

- SLURM user and/or table access
- SLURM tables and/or table access
- Buffer optimizations

slurmdbd: what needs to be configured

- SLURM database plugin
- Database record maintenance/expiration

Workload Management Stack

Workload Manager Configuration Overview

slurmctld: what needs to be configured

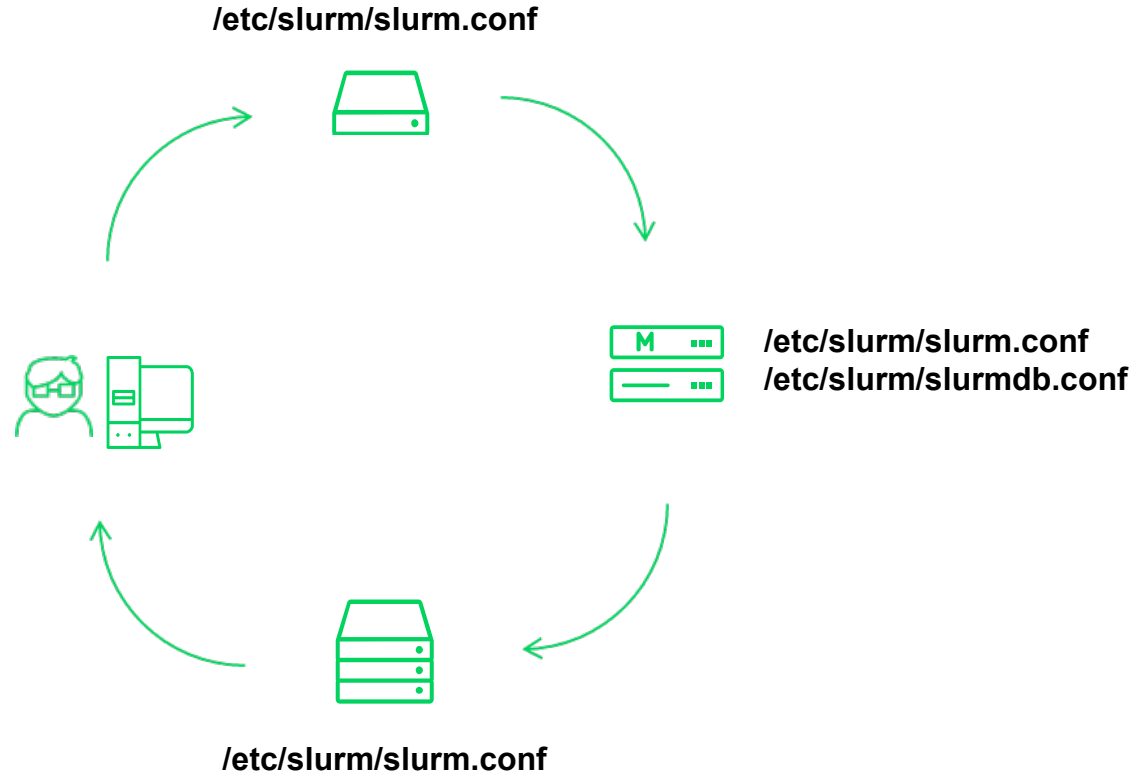
- Cluster name
- Cluster control daemon hosts
- Cluster compute partitions
- Daemon users and ports
- Scheduling
- Interconnects
- Process tracking
- Resource selection
- Logging
- Accounting

Workload Management Stack

Configuration Tools

- **SLURM Documentation**
 - Installed when the Workload Manager pattern is installed
 - `/usr/share/doc/slurm-<version>`
- **Consider making them available over HTTP**
 - `/usr/share/doc/slurm-<version>/html`
- **SLURM configuration forms**
 - `/usr/share/doc/slurm-<version>/html/configurator.easy.html`
 - `/usr/share/doc/slurm-<version>/html/configurator.html`
 - <https://slurm.schedmd.com/configurator.html>

Configuration File Placement



Workload Management Stack

Daemon Start Order

1. Maria DB MySQL server (mariadb.service)
2. SLURM Database daemon (slurmdbd.service)
3. SLURM Control daemon (slurmctld.service)
4. SLURM daemon (slurmd.service)

Workload Management Stack

Useful Administrator Commands

- **sinfo:** View information about SLURM nodes and partitions
- **scontrol:** View and modify SLURM configuration and state
- **sacctmgr:** View and modify SLURM account information

Workload Management Stack

Useful Administrator Commands

- On any cluster system to view cluster node and partition information:
 - `~# sinfo -Nel`
- On a system running slurmctld to reconfigure a cluster:
 - `~# scontrol reconfigure`
- On a system running slurmdbd to create a user/group/access account:
 - `~# sacctmgr create user Name=msteele Account=bioinf_lab_g
DefaultAccount=bioinf_lab_g Cluster=hangar_hpc Partition=normal_q`

Workload Management Stack

Useful Administrator Commands

- **pdsh**: Issue commands to groups of hosts in parallel
- **pdcp**: Copy files to groups of hosts in parallel
- **rdcp**: Copy files from groups of hosts in parallel
- **dshbak**: Format output from pdsh command, for humans

(Future blog post coming)

Workload Management Stack

Useful Administrator Commands

Running a command on all nodes in a partition:

```
~# pdsh -R mrsh -P normal_q uptime | dshbak
```

```
-----
```

```
node1
```

```
-----
```

```
06:04:48 up 24 days 23:26, 0 users, load average: 0.00, 0.00, 0.00
```

```
-----
```

```
node2
```

```
-----
```

```
06:04:48 up 24 days 22:50, 0 users, load average: 0.00, 0.00, 0.00
```

Workload Management Stack

Useful End User commands

- **sbatch**: Submit batch script jobs to SLURM
- **srun**: Submit parallel and interactive jobs to SLURM
- **squeue**: View information about jobs located in the SLURM queue
- **sview**: Graphical user interface to view and modify SLURM state

Workload Management Stack

Looking at a Sample Batch Job Script

```
#!/bin/bash
#SBATCH --job-name=randomNum
#SBATCH --partition=normal_q
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --mem=200M
#SBATCH --time=08:00:00
#SBATCH --output=/work/msteele/job_name-%x.job_number-%j.nodes-%N.out
#SBATCH --error=/work/msteele/job_name-%x.job_number-%j.nodes-%N.err
for i in {1..100000}; do
echo $RANDOM >> /work/msteele/SomeRandomNumbers.txt
done
sort -n /work/msteele/SomeRandomNumbers.txt
```

Workload Management Stack

Useful End User Commands

- **Submit a batch job:**

```
msteele@darkvixen100:~> sbatch ./scripts/big_job.sh
```

- **View information about jobs in a SLURM partition:**

```
~# squeue --clusters=hangar_hpc --partition=normal_q --user=msteele
```

```
CLUSTER: hangar_hpc
```

JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODE LIST(REASON)
69	normal_q	randomNu	msteele	R	0:24	1	node1

Workload Management Stack

Useful End User Commands, Using svview

Use the graphical svview utility to view and manage jobs:

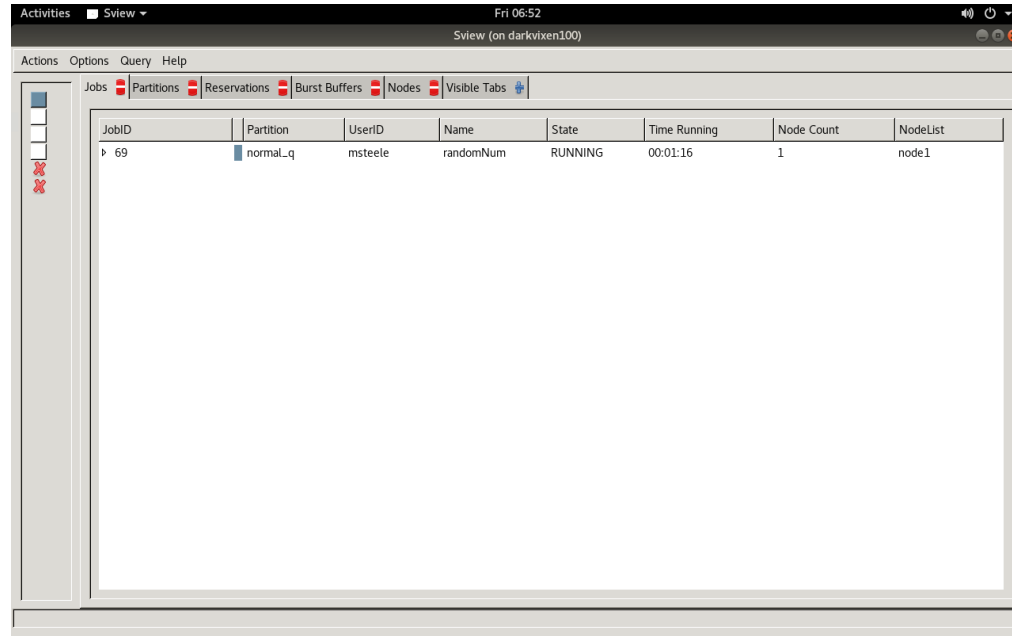
```
msteele@darkvixen215:~> ssh -Y msteele@darkvixen100.dvc.darkvixen.com
```

```
msteele@darkvixen100:~> svview &
```

(Future blog post coming)

Workload Management Stack

Useful End User Commands, Using sview

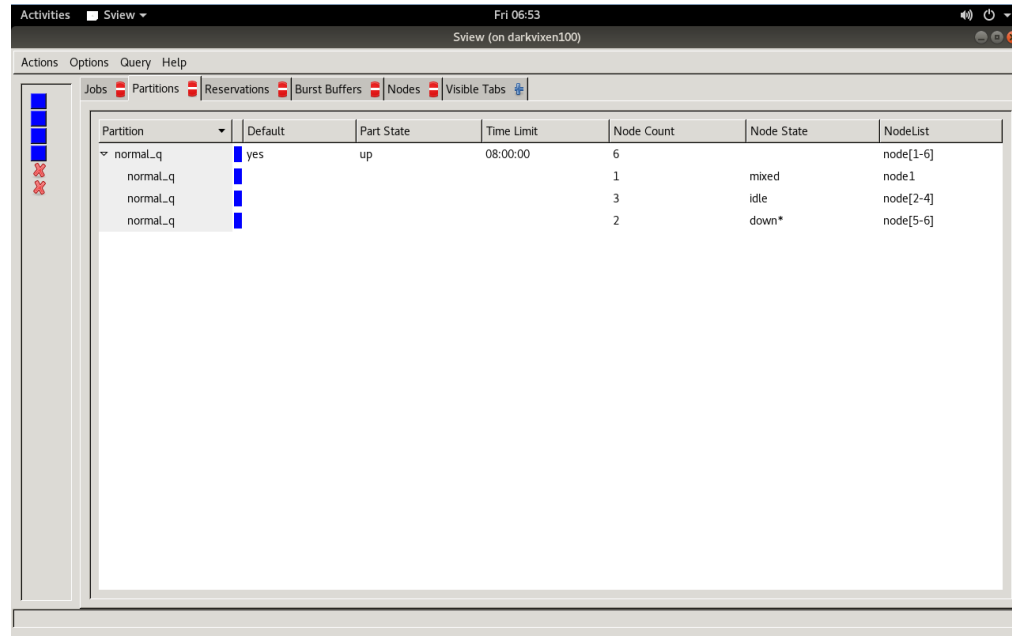


The screenshot shows the sview application window. The title bar indicates the system is on 'Fri 06:52' and the window is titled 'Sview (on darkvixen100)'. The application has a menu bar with 'Actions', 'Options', 'Query', and 'Help'. Below the menu bar is a toolbar with icons for 'Jobs', 'Partitions', 'Reservations', 'Burst Buffers', 'Nodes', and 'Visible Tabs'. The main content area displays a table with the following data:

JobID	Partition	UserID	Name	State	Time Running	Node Count	NodeList
▶ 69	normal_q	msteele	randomNum	RUNNING	00:01:16	1	node1

Workload Management Stack

Useful End User Commands, Using sview

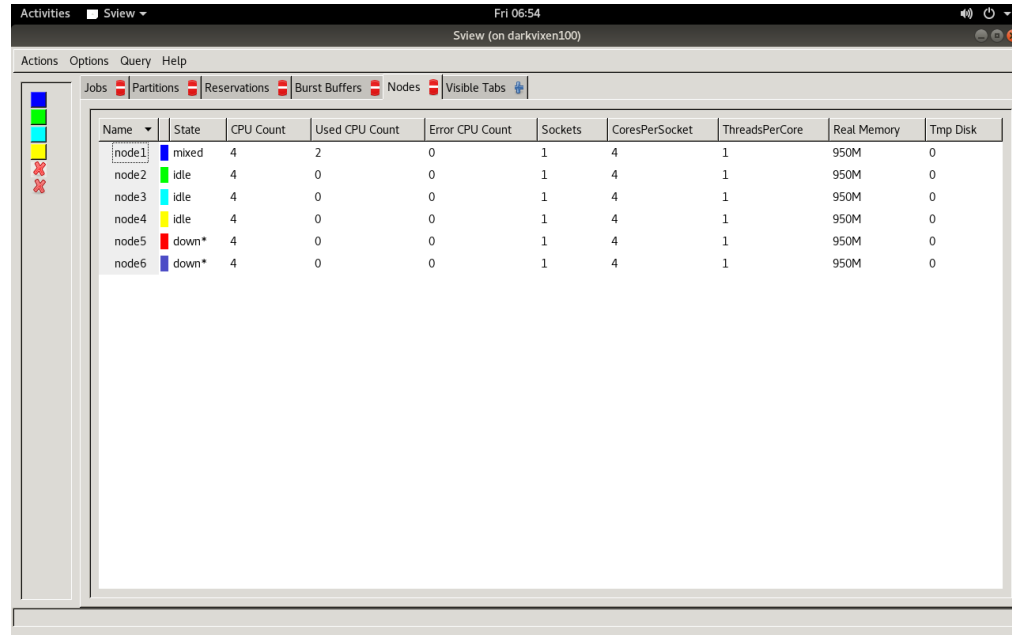


The screenshot shows the sview application window. The title bar indicates the system is on 'darkvixen100' at 'Fri 06:53'. The menu bar includes 'Actions', 'Options', 'Query', and 'Help'. Below the menu bar are several tabs: 'Jobs', 'Partitions', 'Reservations', 'Burst Buffers', 'Nodes', and 'Visible Tabs'. The 'Partitions' tab is active, displaying a table with the following data:

Partition	Default	Part State	Time Limit	Node Count	Node State	NodeList
normal_q	yes	up	08:00:00	6		node[1-6]
normal_q				1	mixed	node1
normal_q				3	idle	node[2-4]
normal_q				2	down*	node[5-6]

Workload Management Stack

Useful End User Commands, Using sview



The screenshot shows the sview application window. The title bar indicates the system is on 'Fri 06:54' and the window is titled 'Sview (on darkvixen100)'. The application has a menu bar with 'Actions', 'Options', 'Query', and 'Help'. Below the menu bar is a toolbar with icons for 'Jobs', 'Partitions', 'Reservations', 'Burst Buffers', 'Nodes', and 'Visible Tabs'. The main area contains a table with the following data:

Name	State	CPU Count	Used CPU Count	Error CPU Count	Sockets	CoresPerSocket	ThreadsPerCore	Real Memory	Tmp Disk
node1	mixed	4	2	0	1	4	1	950M	0
node2	idle	4	0	0	1	4	1	950M	0
node3	idle	4	0	0	1	4	1	950M	0
node4	idle	4	0	0	1	4	1	950M	0
node5	down*	4	0	0	1	4	1	950M	0
node6	down*	4	0	0	1	4	1	950M	0

Typical Cluster File Systems

Scratch, Nearline and Environment

- **Scratch: Temporary storage for active job data**
 - Performant and capacious parallel file systems that are inline with compute resources
 - Lustre and GPFS are prominent (Performant local disks can also be used)
- **Nearline: Longer term storage for active job data**
 - Capacious resilient file systems that are not accessible to compute resources
 - Ceph, Panasas and ZFS are prominent
- **Environment: Home directories**
 - Standard Linux file systems on commodity hardware

Typical Cluster File Systems

Mounting and Backup Services

- **NFS versions**

- NFS version 3 is more performant and connection tolerant, but less secure
- Environment must provide required security

- **Backup services**

- Home directories and select nearline storage
- File systems > 500 TB not practical
- Reproduce data by re-executing jobs

Typical Cluster Security

Authentication and Firewalls

- **Authentication**

- Secure directory service-based authentication to edge and administrative systems
- Multi-factor authentication to edge and administrative systems where possible
- Less security is generally permissible and operationally favored within the cluster

- **Firewalls**

- Perimeter devices provide primary security for edge and administrative systems
- Edge and administrative systems should be hardened and optimized, and implement local firewalls

Questions

- **SUSE Community blogs**
 - Deploying SLURM using SLE HPC patterns
 - Deploying SLURM PAM modules on SLE compute nodes
- **SUSE Documentation**
- **SLURM Workload Manager**
 - <https://slurm.schedmd.com>



We adapt. You succeed.

Unpublished Work of SUSE LLC. All Rights Reserved.

This work is an unpublished work and contains confidential, proprietary and trade secret information of SUSE LLC. Access to this work is restricted to SUSE employees who have a need to know to perform tasks within the scope of their assignments. No part of this work may be practiced, performed, copied, distributed, revised, modified, translated, abridged, condensed, expanded, collected, or adapted without the prior written consent of SUSE. Any use or exploitation of this work without authorization could subject the perpetrator to criminal and civil liability.

General Disclaimer

This document is not to be construed as a promise by any participating company to develop, deliver, or market a product. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. SUSE makes no representations or warranties with respect to the contents of this document, and specifically disclaims any express or implied warranties of merchantability or fitness for any particular purpose. The development, release, and timing of features or functionality described for SUSE products remains at the sole discretion of SUSE. Further, SUSE reserves the right to revise this document and to make changes to its content, at any time, without obligation to notify any person or entity of such revisions or changes. All SUSE marks referenced in this presentation are trademarks or registered trademarks of SUSE LLC. in the United States and other countries. All third-party trademarks are the property of their respective owners.