Cray Cluster Supercomputers

John Lee
VP of Advanced Technology Solutions
CUG 2013

CRAY
THE SUPERCOMPUTER COMPANY

Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance statements and results have been estimated based on internal Cray analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. Cray does not control or audit the design or implementation of third party benchmarks and product claims in this document.

Cray and Sonexion are registered trademarks of Cray Inc. in the United States and other countries, and Cray XC30, Cray CS300, Cray XK7, Cray XE6, Cray Linux Environment, Cray XE6m, Cray XE6m-200, Cray XT6, Cray XT5, Cray XT4, Cray SHMEM, CrayPat, NodeKARE, YarcData and uRiKA are registered trademarks of Cray Inc.

Other names and brands may be claimed as the property of others. Other product and service names mentioned herein are the trademarks of their respective owners.

Agenda

- **Cray Cluster Supercomputer Presentation**
  - Why Cluster Supercomputer
  - Cray Cluster CS300-AC™, Air-Cooled System
  - CS300™ Cluster Building Block Platforms
  - Networking and Interconnect
  - Cray Cluster Software Stack and ACE™ Management Software

Cray Computing Solutions

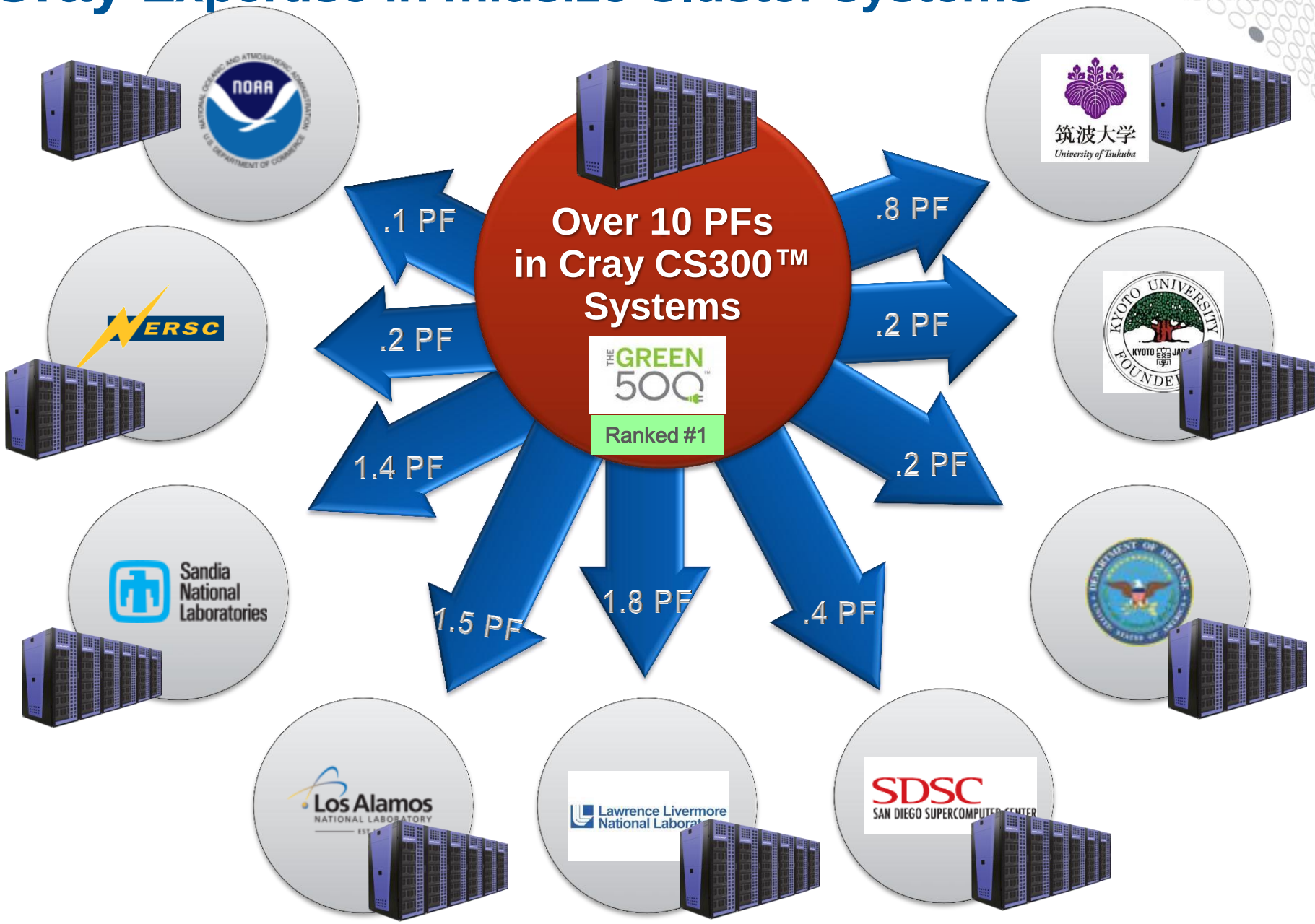
Scalability Focus
(Tightly Integrated Solutions)

Cray XC30 Series: Scalable Performance

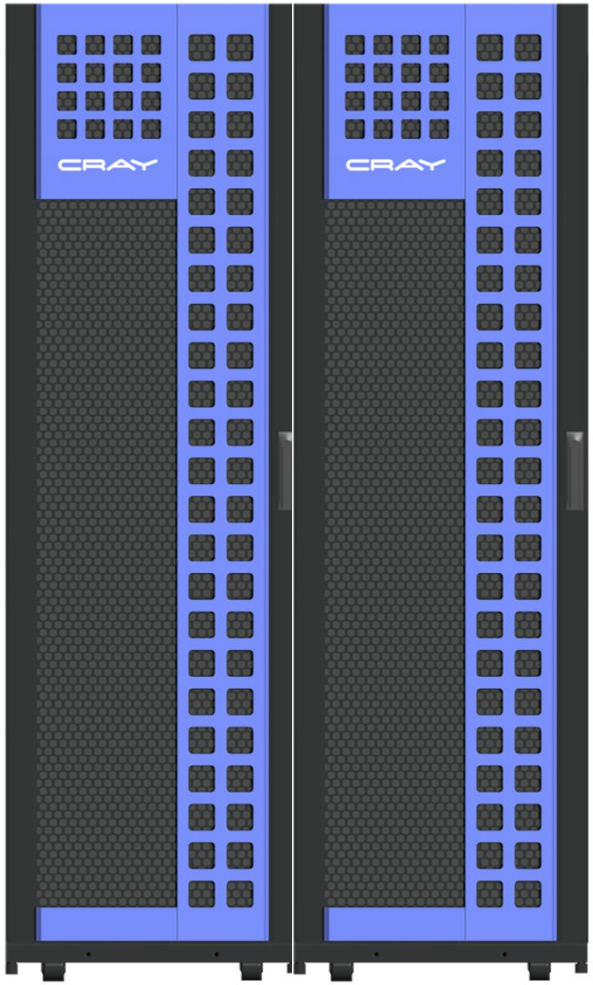
Cray CS300 Series: Flexible Performance

Industry Standards Focus
(Highly Configurable Solutions)

Cray Expertise in midsize Cluster systems

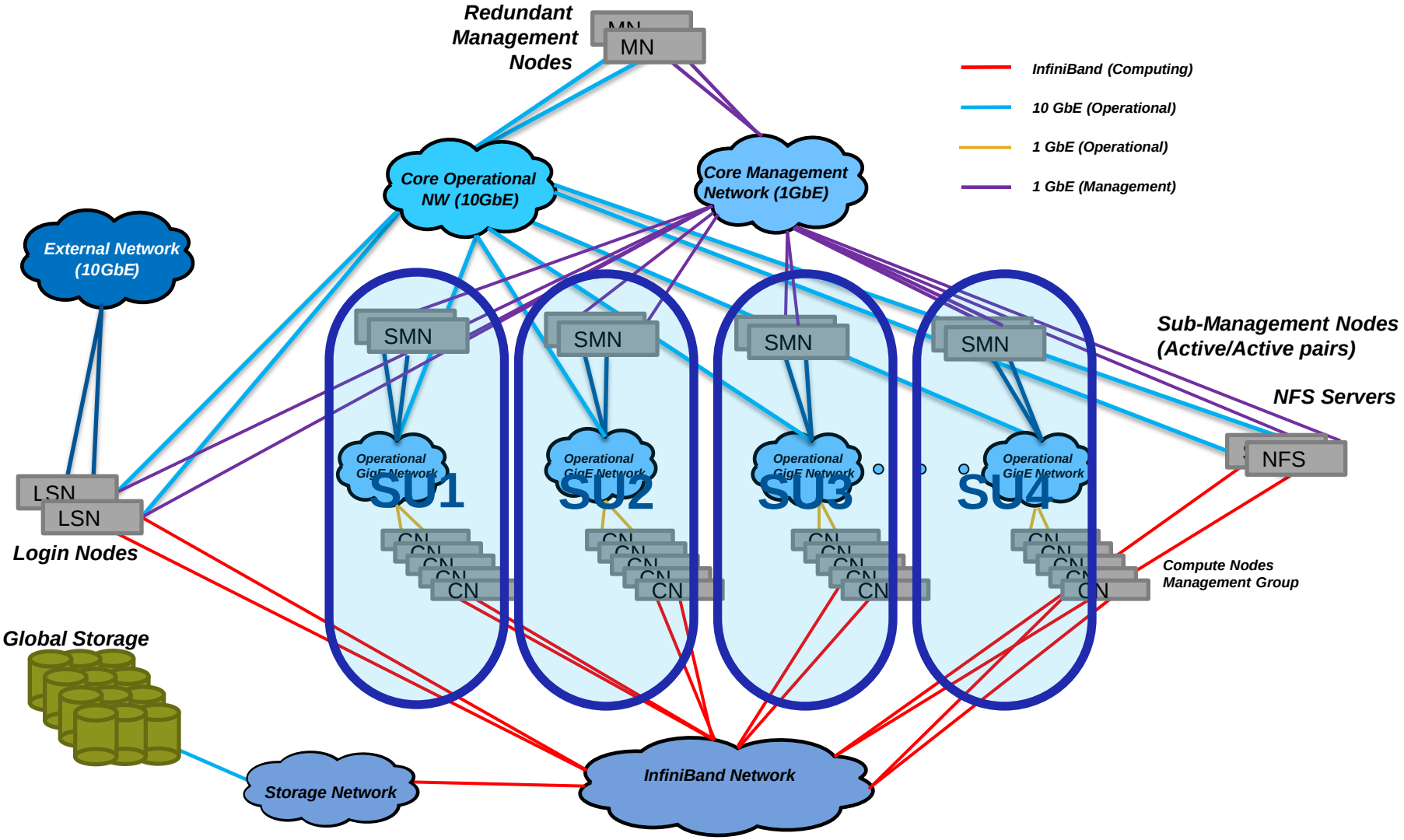


CS300-AC™ Cluster Supercomputer



- **Highly Configurable System Architecture**
  - Cray Solutions Architects work with customers to tailor systems to customer's specifications
- **Purpose designed leveraging best-of-breed open standards technologies**
  - Designed from ground up to be a scalable HPC system
- **Air-cooled energy-efficient design**
  - Shared cooling & power infrastructure to save power
- **Complete turn-key system with integrated HPC software stack powered by ACE**
  - Powerful but easy to manage cluster management software suite
- **Reliable and Serviceable**
  - Designed with hot-swappable, redundant FRUs to maintain a low MTBI

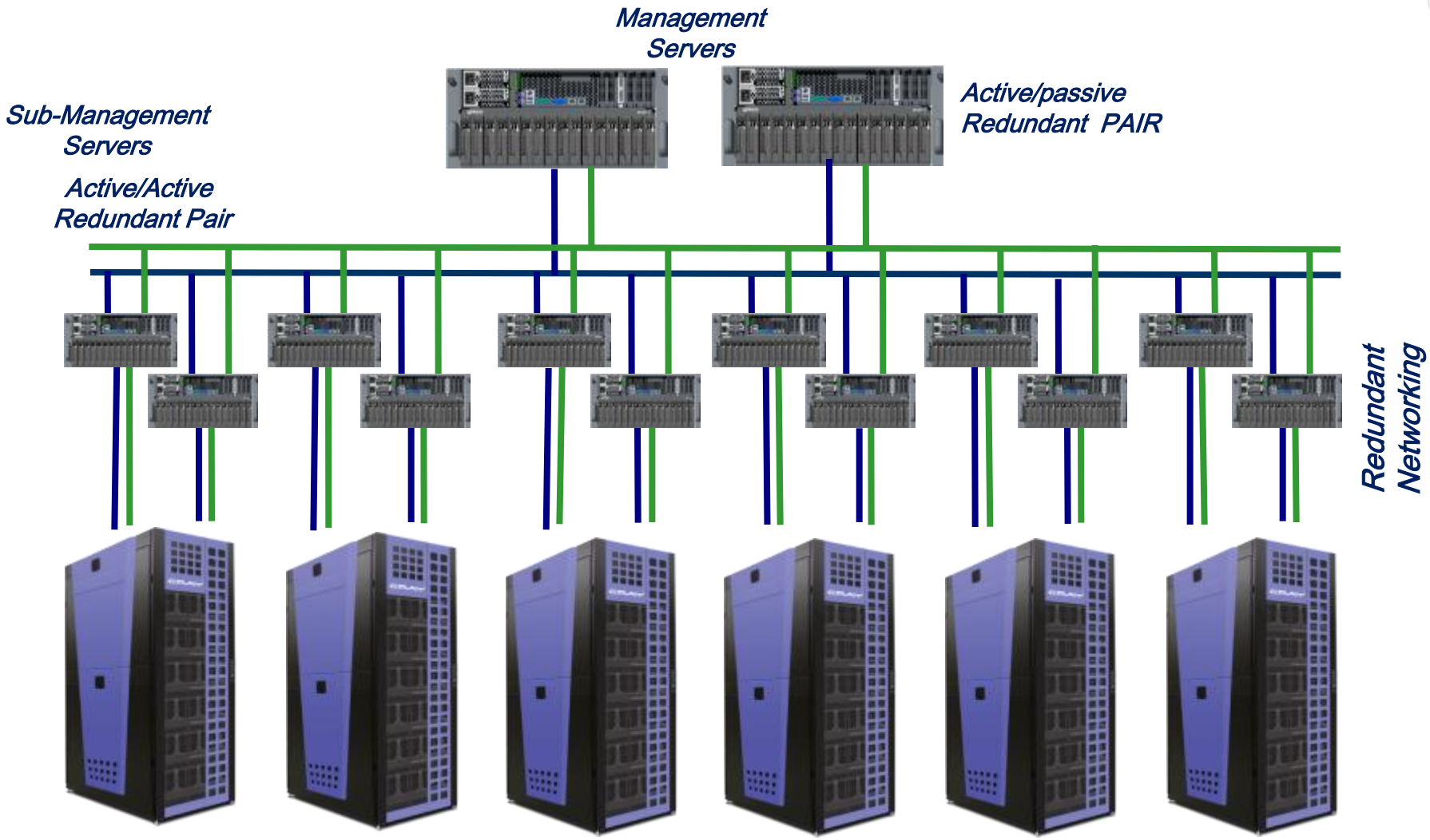
Cray CS300-AC™ System Architecture



Advanced Cluster Engine™ (ACE) Management Software

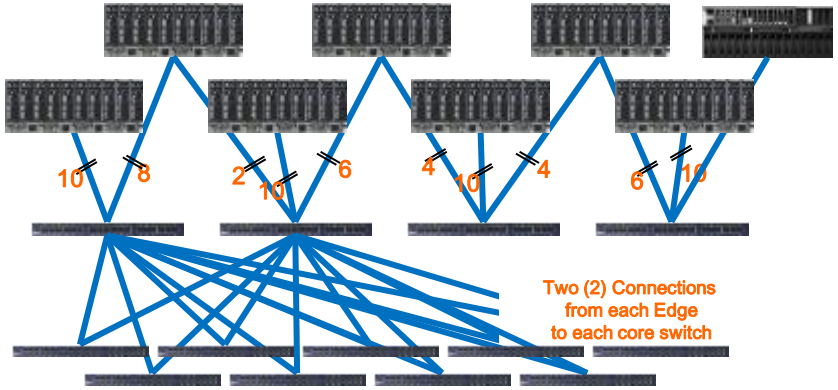
- Easy-to-use Remote System Control Manager with CLI and GUI
- Delivers a reliable, highly available architecture supporting multiple network topologies
- Complete SW suite to include Network, Server, Cluster, and Storage Management
  - ✓ Highly scalable, stateless management
  - ✓ No single failure point in the management stack
  - ✓ Can manage heterogeneous nodes with different OS stacks
  - ✓ System power and temperature monitoring
  - ✓ Version control with ability to rollback changes
  - ✓ Ability to export, import system configurations and images
  - ✓ Ability to detect HW, fabric topology configuration errors

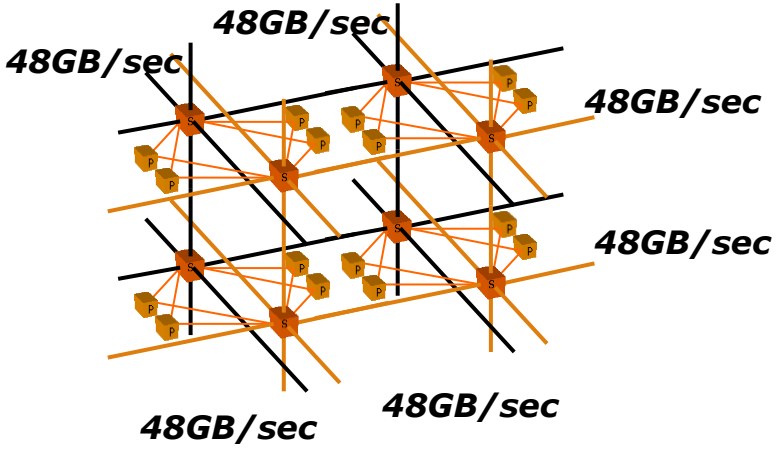
Cray CS300-AC™ System Architecture Scalability & Reliability

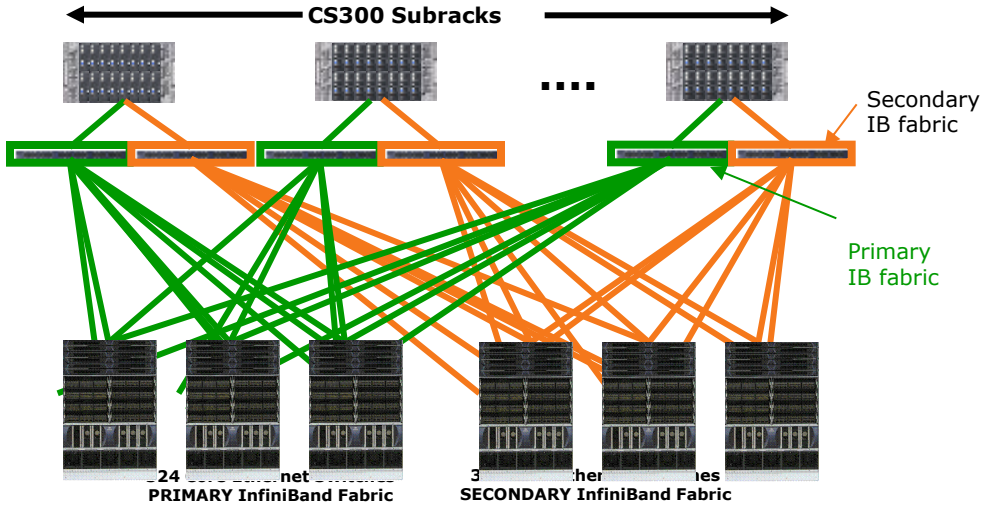


Cray CS300-AC™ System Architecture
InfiniBand Fabric Management

Support for multiple Topologies





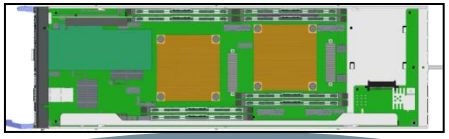


- Scales from hundreds to thousands of nodes
- Single Rail and Dual Rail Fat-Tree
- Distributed Core
- Large Director Class Core
- 2-D and 3-D Torus
- Single and dual rail
- Distributed IO
- Hybrid Topologies
- Large Edge
- Inverted Tree
- 3-D Torus + Tree

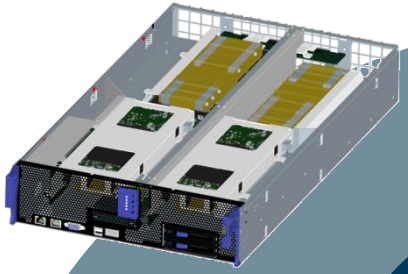
Cray CS300-AC™ Building Block Platform Next Generation, Cray GreenBlade™

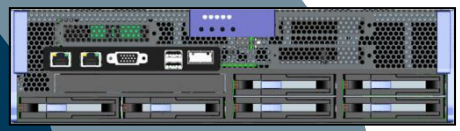
- **Simple.**
  - Singular focus in designing the best HPC-optimized building block
  - Simple and reliable platform-level management appliance (iSCB)
- **Flexible.**
  - Two motherboards & two chassis form factors with modular blade options.
  - Support for multi-vendor accelerators
- **Reliable.**
  - All serviceable FRUs are redundant & hot-swappable
  - High MTBF compute nodes for maximum system reliability

Cray CS300-AC™ Building Block Platform Simple. Flexible. Reliable.

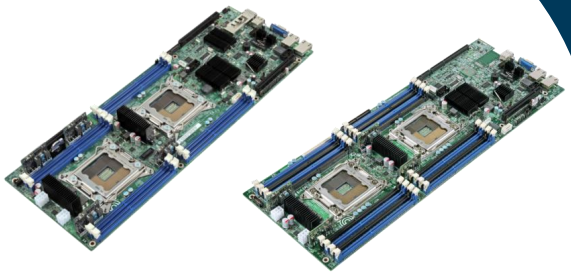


Compute Blades





Hybrid Blades

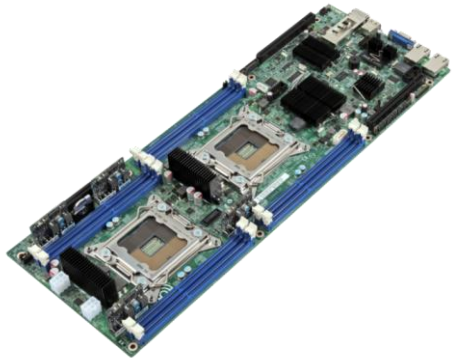


2x Motherboards

Service Blades

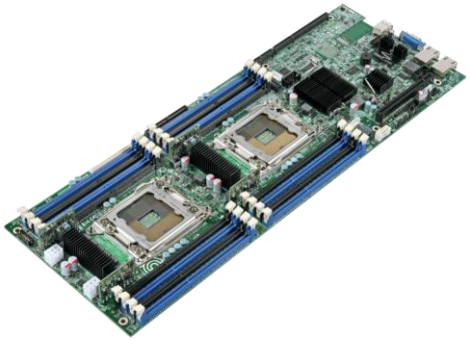
Chassis

Cray CS300-AC™ Building Block Platform System Boards



Compute Boards

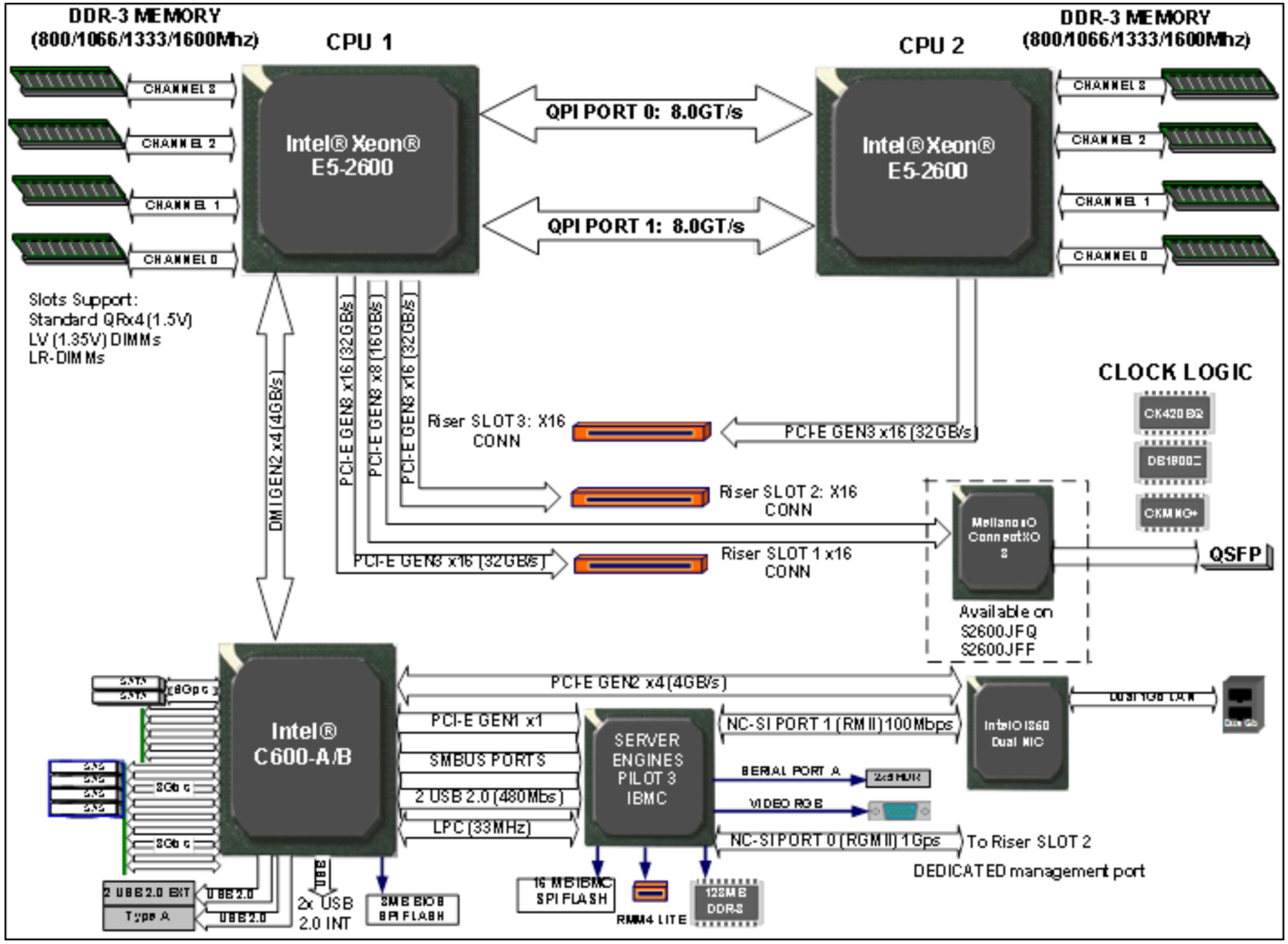
- **Design optimized for Compute platform**
- Support for 2x Intel Sandy Bridge EP CPUs
- Support for 8x DDR3 1600MHz memory modules
- Access to up to **56** PCIe Gen3 lanes
  - 3x PCIe Gen3 x16 expansion slots
- Option for on-board PCIe Gen3 QDR/FDR IB



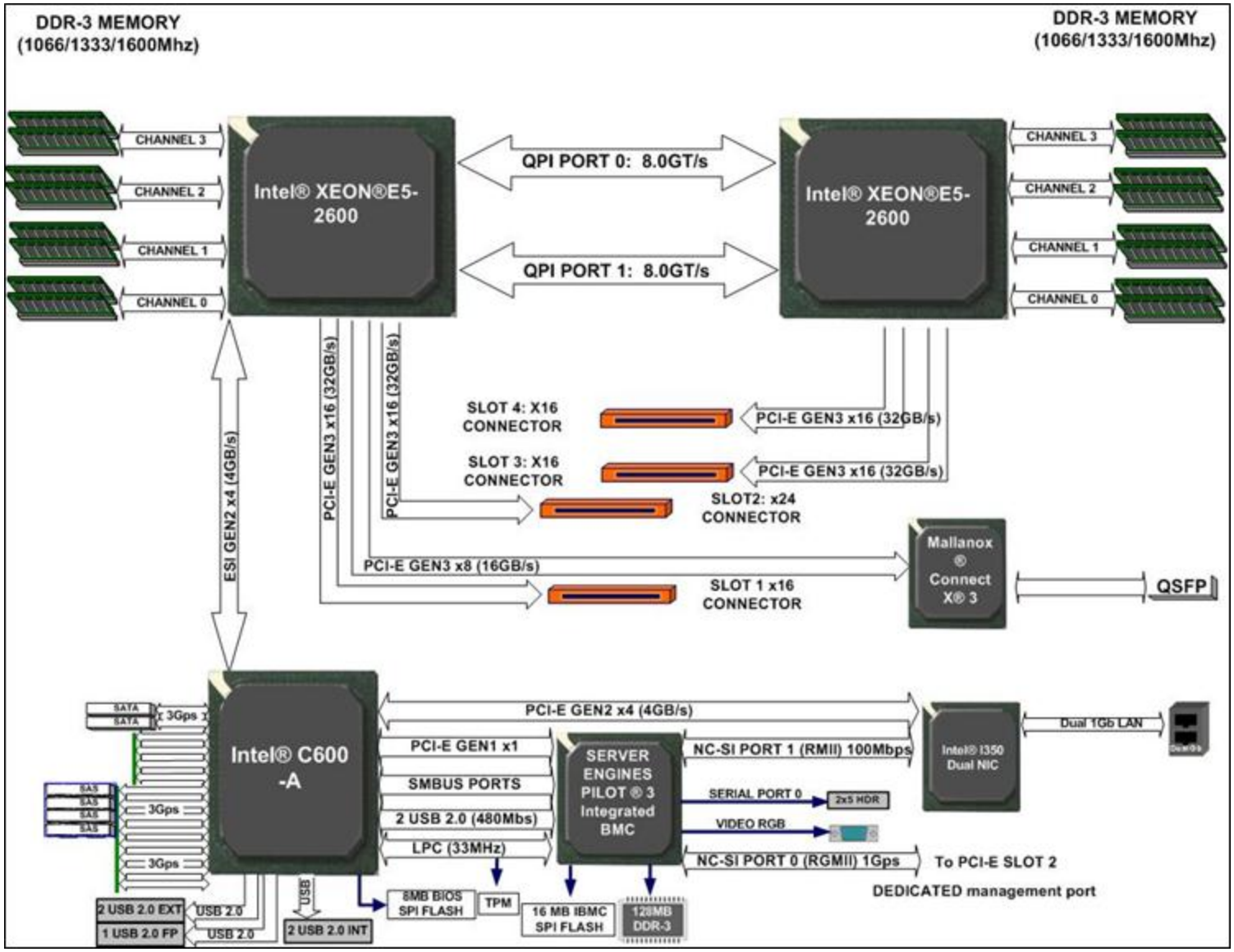
Hybrid Boards

- **Design optimized for Hybrid platform**
- Support for 2x Intel Sandy Bridge EP CPUs
- Support for 16x DDR3 1600MHz memory modules
- Access to up to **72** PCIe Gen3 lanes
  - 4x PCIe Gen3 x16 expansion slots
- Option for on-board PCIe Gen3 QDR/FDR IB

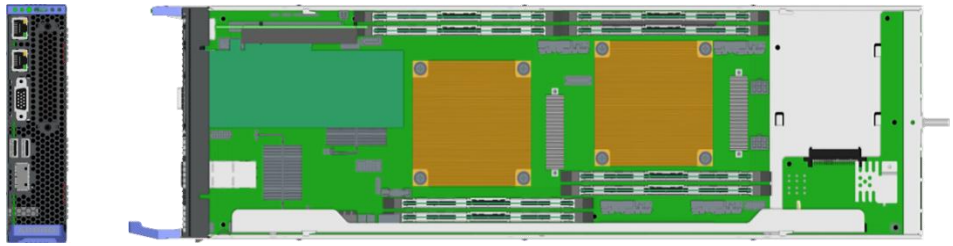
Cray CS300-AC™ Building Block Platform Compute Board Design



Cray CS300-AC™ Building Block Platform Hybrid Board Design

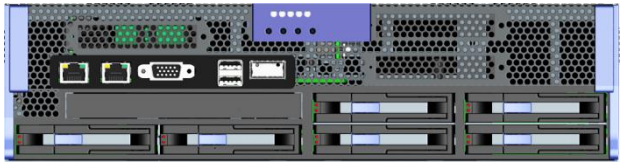


Cray CS300-AC™ Building Block Platform Compute Blade Servers



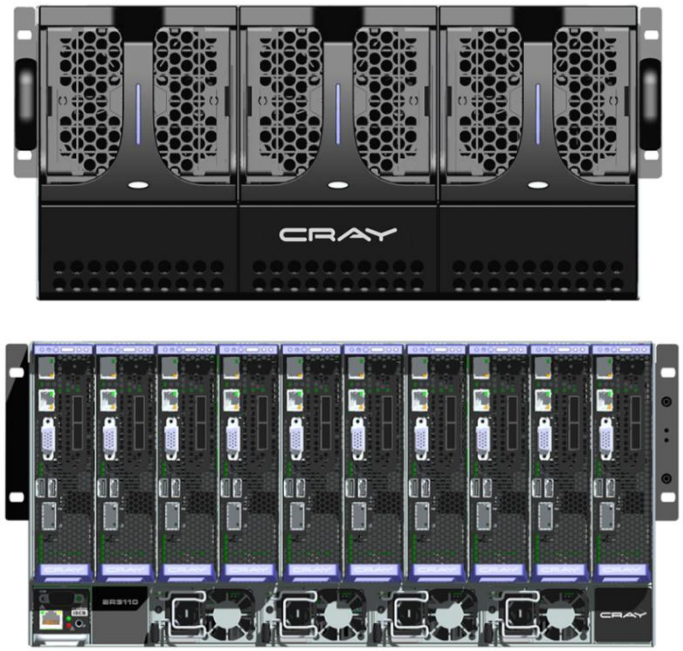
- **0.5RU effective density Compute Blades**
  - Supports Intel Xeon E5 Sandy Bridge CPUs (Ivy Bridge in 2013)
  - Supports up to 128GB DDR3 1600MHz system memory
  - Integrated PCIe Gen3 QDR/FDR IB (optional)
  - PCIe Gen3 x16 expansion slot
  - Support for one internal 2.5” HDD
- **1RU effective density Hybrid Blades**
  - Host + Expansion Blade
  - Supports 2x nVIDIA Keplers
  - Supports 2x Intel KNCs

Cray CS300-AC™ Building Block Platform Service Hybrid Servers



- **2RU effective density**
- **Service Node w/ expanded IO + 3x 3.5” HDDs**
  - Up to three 3.5” SATA/SAS HDDs
  - Up to four PCIe expansion slots
  - Ideal as a GW/Login Node
- **Service Node w/ expanded IO + 6x 2.5” HDDs**
  - Up to six 2.5” SATA/SAS HDDs
  - One DVD ROM Drive
  - Up to four PCIe expansion slots
  - Ideal as a Management/Boot Node
- **Hybrid Node w/ 4x Accelerators + 2x 2.5” HDDs**
  - Up to two 2.5” SATA/SAS HDDs
  - Up to four nVIDIA Keplers(K10 or K20) or Intel KNCs

Cray CS300-AC™ Building Block Platform GreenBlade™ Subrack, SR5000 Chassis



- **SR5000 Chassis supports either:**
  - 10x Compute Blades
  - 5x Hybrid Blades
- Compute Blades are 0.5RU 2P x86 servers
- Hybrid Blades are 1RU 2P x86 combined with either 2x NVIDIA Keplers or 2x Intel KNCs
- Three hot-swappable, redundant Cooling Fan Units
- Up to four hot-swappable, redundant 1630W PS
- Can support one or two redundant iSCB chassis managers

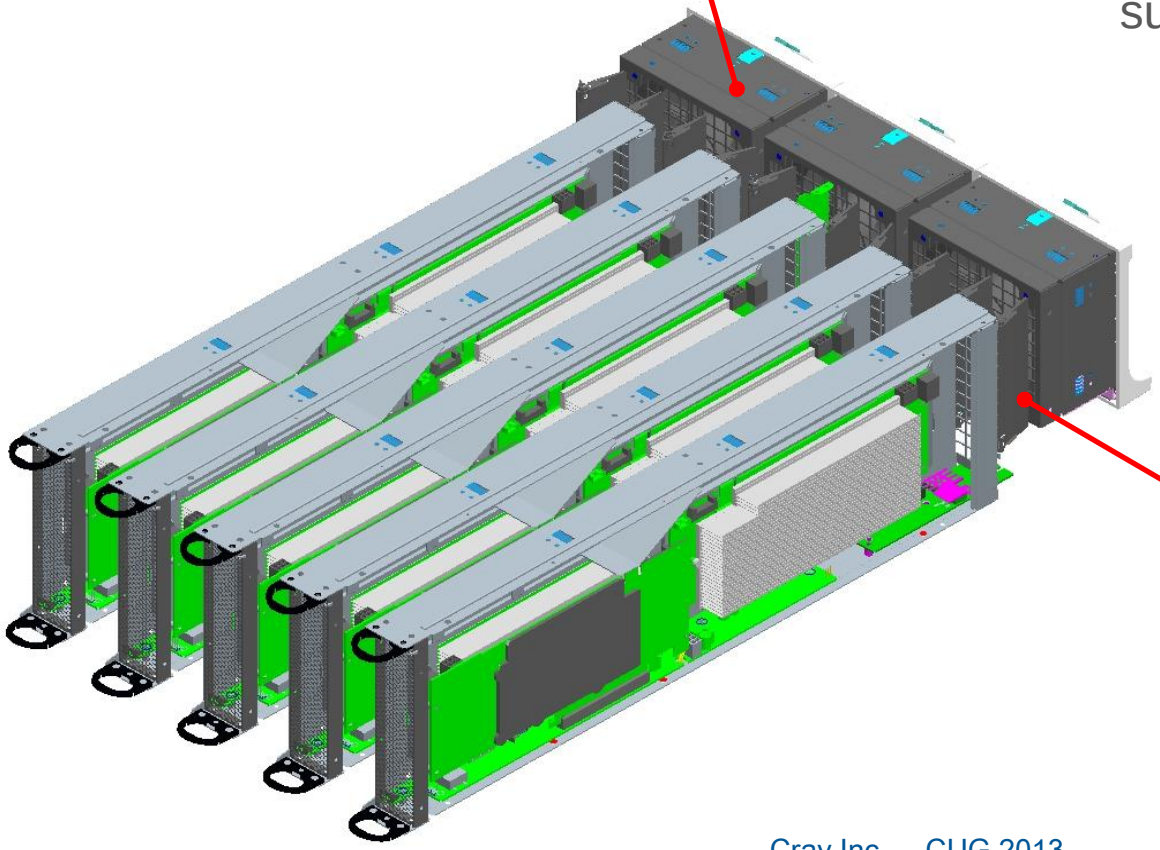
Cray CS300-AC™ Building Block Platform GreenBlade™ Subrack, SR8000 Chassis

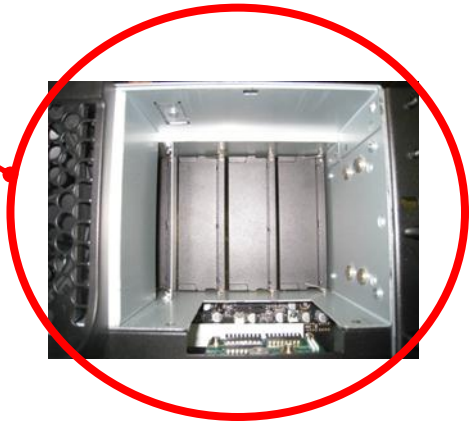
- **SR8000 Chassis supports either:**
  - 16x Compute Blades
  - 8x Hybrid Blades
  - 4x double-wide Service Blades
  - 4x double-wide Hybrid Blades
- Compute Blades are 0.5RU 2P x86 servers
- Hybrid Blades are 1RU 2P x86 combined with either 2x nVIDIA Keplers or 2x Intel KNCs
- Double-wide Service Blades are 2RU 2P x86 servers
- Double-wide Hybrid Blades are 2RU 2P x86 servers with either 4x nVIDIA Keplers or 4x Intel KNCs
- Six hot-swappable, redundant Cooling Fan Units
- Up to six hot-swappable, redundant 1630W PS
- Can support one or two redundant iSCB chassis managers

Cray CS300-AC™ Building Block Platform Cooling

- **Closed-Loop Cooling System**
  - 3x Cooling Fan Unit (CFU)
  - Each CFU has two, redundant 120mm x 120mm x 38mm fans
  - CFU LED: Green for normal and Amber for service
  - iSCB can dynamically control the fan speed or set static speeds
  - iSCB monitors sensors to dynamically change fan speeds to maintain optimal operating temp

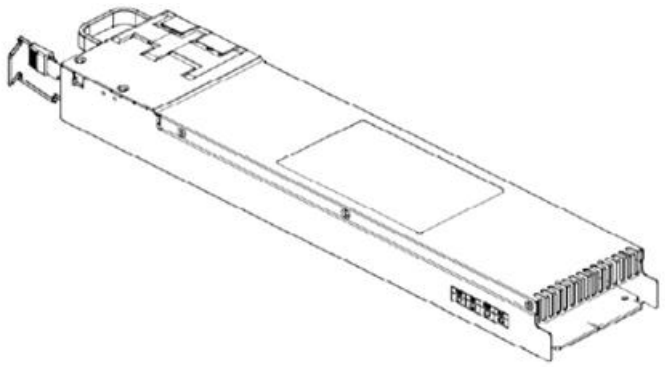
Cray CS300-AC™ Building Block Platform Airflow Management





- Sub-Rack has built-in air shutters that open/close when blades are inserted/removed
- Each CFU cools a zone in the subrack

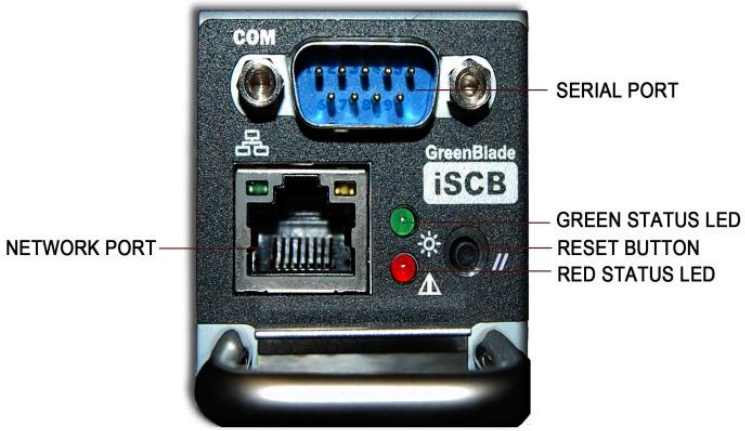
Cray CS300-AC™ Building Block Platform Power



Certification	Result
FCC Part 15 Subpart B, Class A	Compliance
CE/EMC Directive: 2004/108/EC	Compliance
UL 60950-1	Recognized
IEEE Std. 519-1992	Pass
SEMI F47-0706	Pass
V-3/2012.04 Class A, V-4/2012.04	Compliance
CBEMA	Pass

- **Designed specifically for the Cray GreenBlade™ Platform**
  - Designed to support 5+1 load-sharing design
  - Wide 200-277V AC input range
  - 1630W Gold-rated Power Supply
  - Typical efficiency(50% load) of 93%
  - Average efficiency of 92%
- **Designed to meet future RFQ requirements**
  - Designed to meet stringent SEMI F47 and CBMEA requirements for power sag/surge

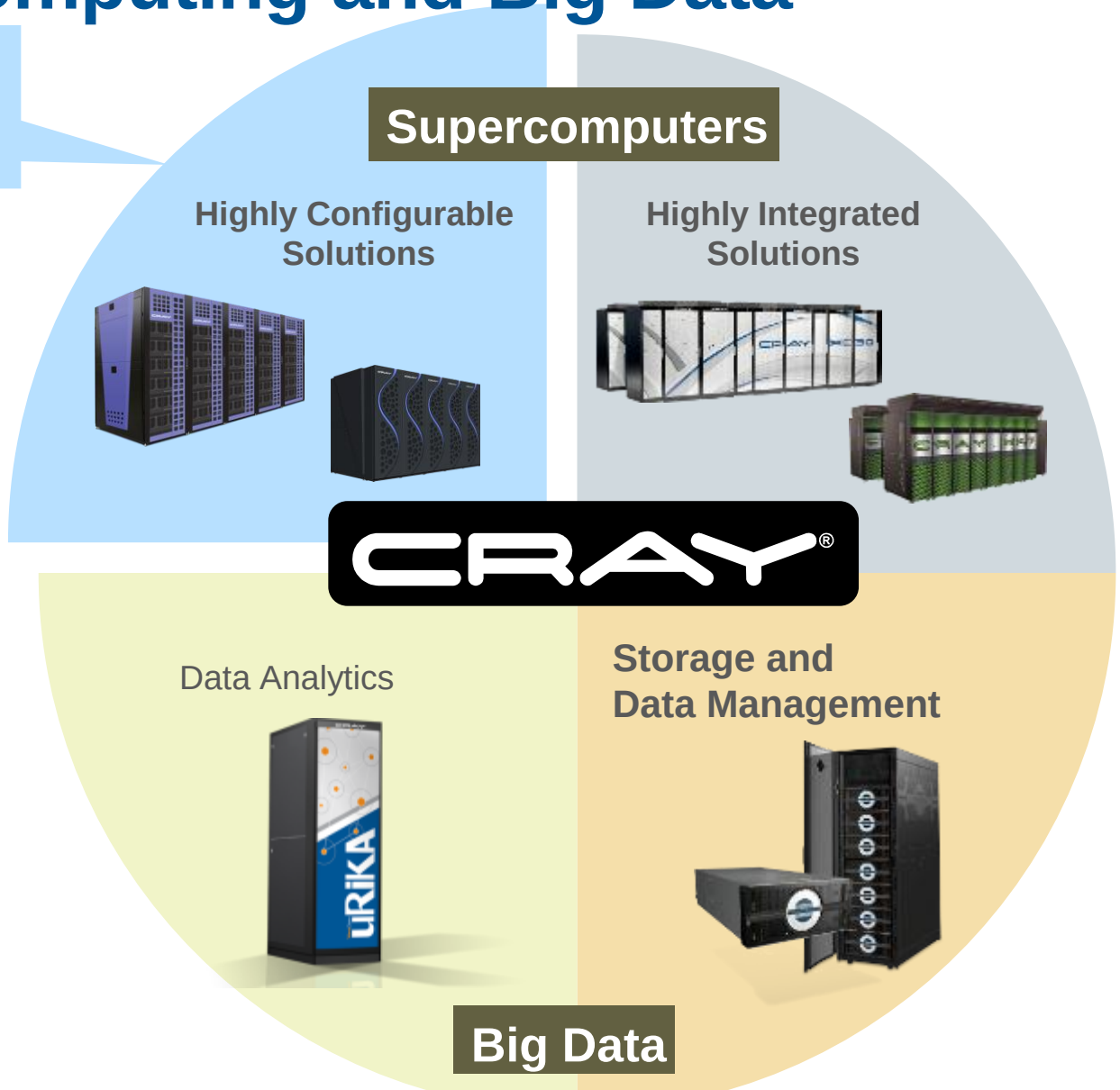
Cray CS300-AC™ Building Block Platform Management



- **Designed specifically for the Cray GreenBlade™ Platform**
  - Common interface to all GreenBlade™ product series
- **ARM-based appliance running embedded linux**
  - Node health monitoring
  - Supports concurrent console sessions to each node (terminal concentrator)
  - Reliable power control
  - Active dynamic fan control
  - Power monitoring
  - Dedicated powerman/conman interface
  - GPU/MIC power control
- **Can easily upgrade features**

Cray Focus Areas
Supercomputing and Big Data

The Missing Piece is Here!



Cray Cluster Supercomputers

Thank You
Questions?

CRAY
THE SUPERCOMPUTER COMPANY

Cray Cluster Software Stack

Susan Kraus
Sr. Director
Software Engineering
skraus@cray.com

CUG 2013

CRAY
THE SUPERCOMPUTER COMPANY

Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance statements and results have been estimated based on internal Cray analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. Cray does not control or audit the design or implementation of third party benchmarks and product claims in this document.

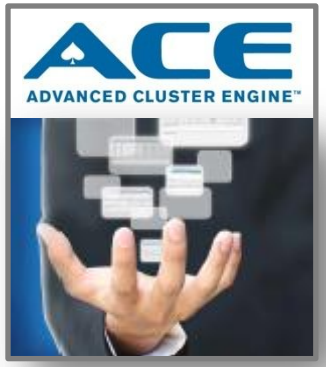
Cray and Sonexion are registered trademarks of Cray Inc. in the United States and other countries, and Cray XC30, Cray CS300, Cray XK7, Cray XE6, Cray Linux Environment, Cray XE6m, Cray XE6m-200, Cray XT6, Cray XT5, Cray XT4, Cray SHMEM, CrayPat, NodeKARE, YarcData and uRiKA are registered trademarks of Cray Inc.

Other names and brands may be claimed as the property of others. Other product and service names mentioned herein are the trademarks of their respective owners.

 Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

Contents

- Cray Cluster Software Stack Components
- ACE Details
  - ✓ What is ACE?
  - ✓ Total Cluster Management Solution
  - ✓ Scalability and Reliability
  - ✓ Architecture
  - ✓ Features, Benefits
  - ✓ Components
  - ✓ Data and File Systems
  - ✓ Interfaces
  - ✓ ACE-Managed Objects
  - ✓ ACE Clusters
  - ✓ Server, Co-processor, Cluster, System Management
  - ✓ Configuration and Miscellaneous
  - ✓ Network Management
  - ✓ Job Management
  - ✓ Plugins
- ACE Demo System



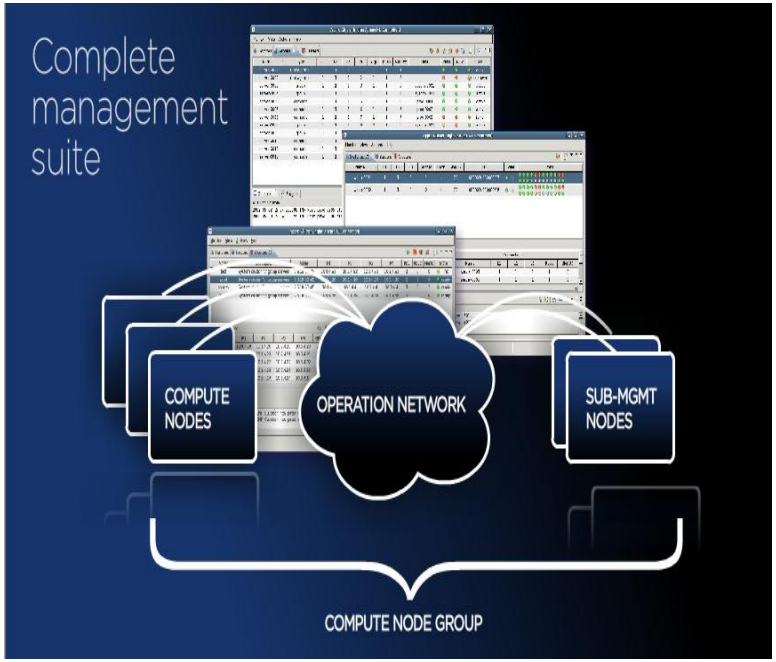
Cray Cluster Software Stack
Components

The Essential Software and Management Tools Needed to Build a Powerful, Flexible, and Highly Available Supercomputer.

What is ACE?

Cray Advanced Cluster Engine™ Management Software

ACE stands for Advanced Cluster Engine™, which is a Cray Management Software designed from the ground up to provide a highly-available, scalable, lights-out, remote management system; to obscure the complexity of a large HPC cluster; and to make managing ten thousand nodes as easy as managing one.



- **Performance** - ACE transforms state of the art commodity hardware into a unified complete HPC system.

- **Scalability** – ACE enables simple common management, monitoring, administration, and operation of HPC clusters with 10 to 10,000 nodes.

- **Reliability** – ACE provides a highly-available, scalable HPC cluster. ACE maintains the state of the cluster in a fault tolerant management database.

- **Flexibility** – ACE's Dynamic Cluster Provisioning allows multiple clusters to share the physical resources

Cray Advanced Cluster Engine™ (ACE) Management Software

Server Management
Configure & Manage
(Provisioning, BIOS)
Report

Storage
Scalable Root File System
High Bandwidth to Secondary Storage

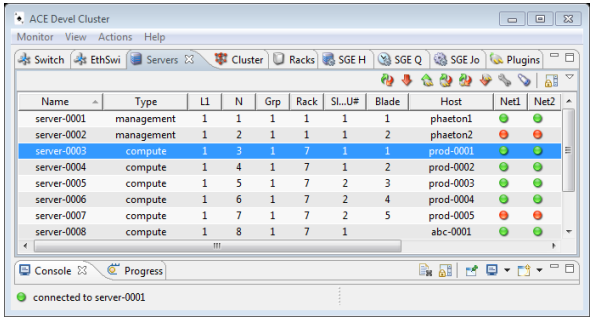
ACE
ADVANCED CLUSTER ENGINE™

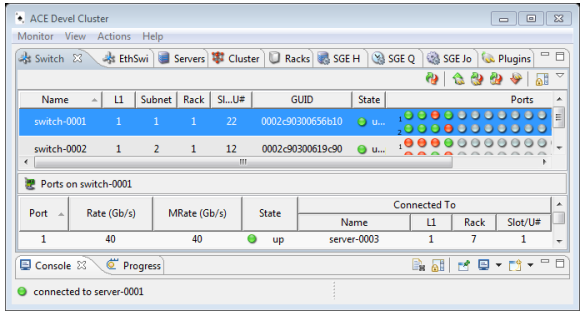
GUI and CLI
- View/Change/Control
- Monitor Health
- Plugin Interface

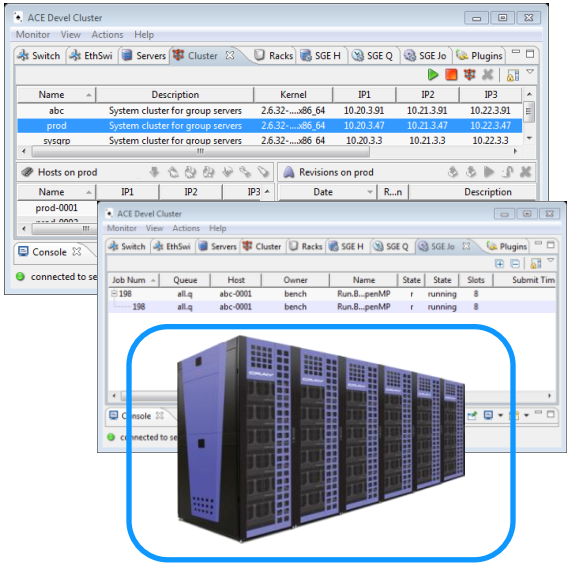
Cluster Management
Job Management
Provision & Partition
Configure Monitor & Report
Resource Management/Scheduling

Network Management
Configure & Manage
Error Detection & Recovery
Report

Management Servers | **Compute Servers** | **Interconnects** | **Storage**



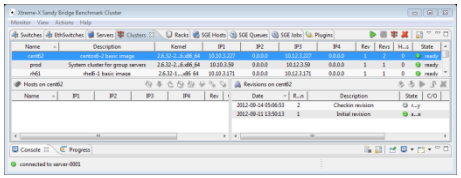




Cray Advanced Cluster Engine™ (ACE)

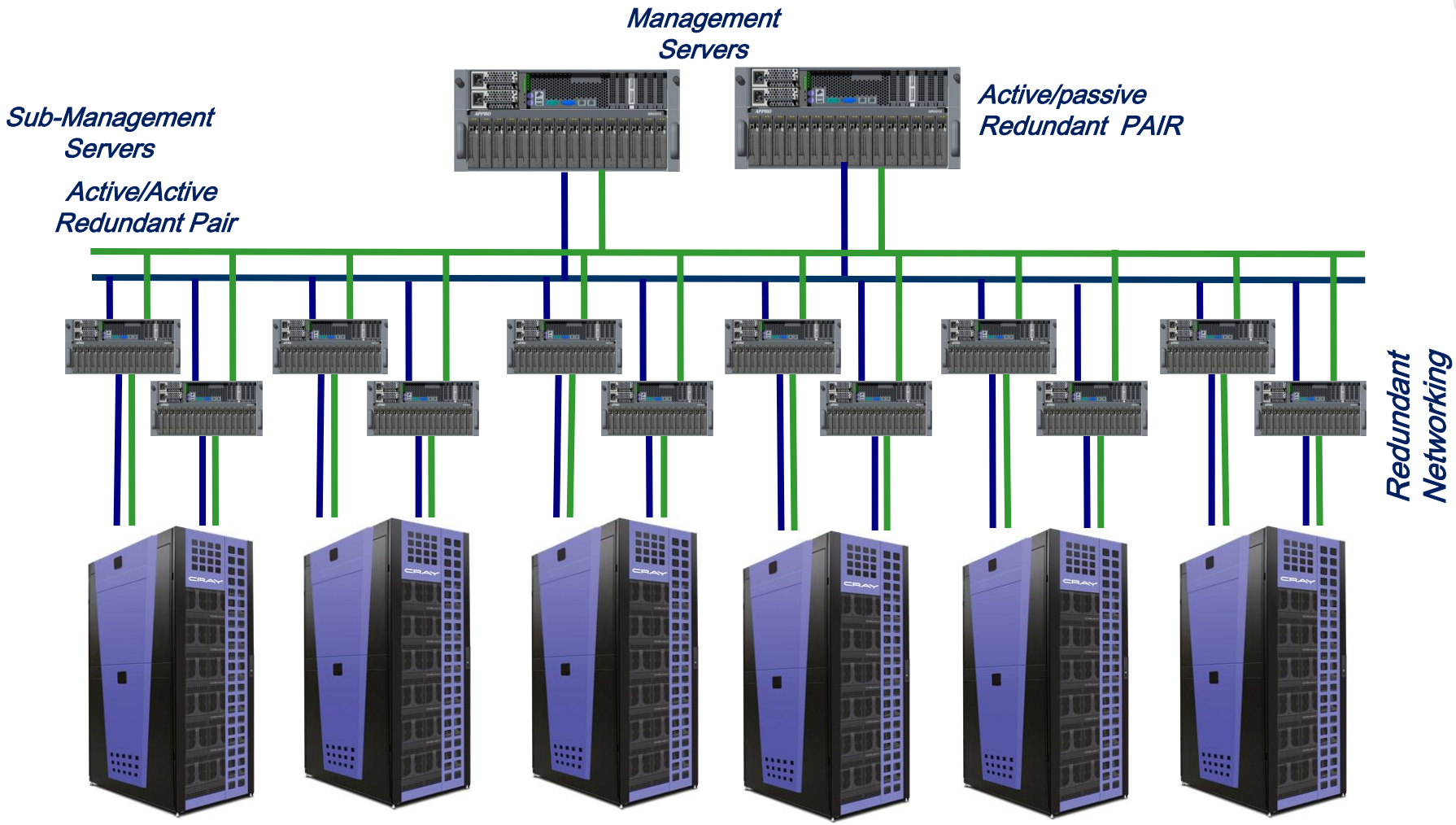
- **Diskless/Stateless Operation**
  - Simplifies System Administration
  - Improves Performance
  - Multi-Level Cached Root File System
  - Does not require a Light Weight Kernel
  - Local Storage also Supported
  - State maintained in two HA management servers

- **“Instant” Provisioning**
  - Multiple Logical Clusters
  - Multiple OS Configurations
  - Provisioning Time = Reboot Time = Less than 10 minutes

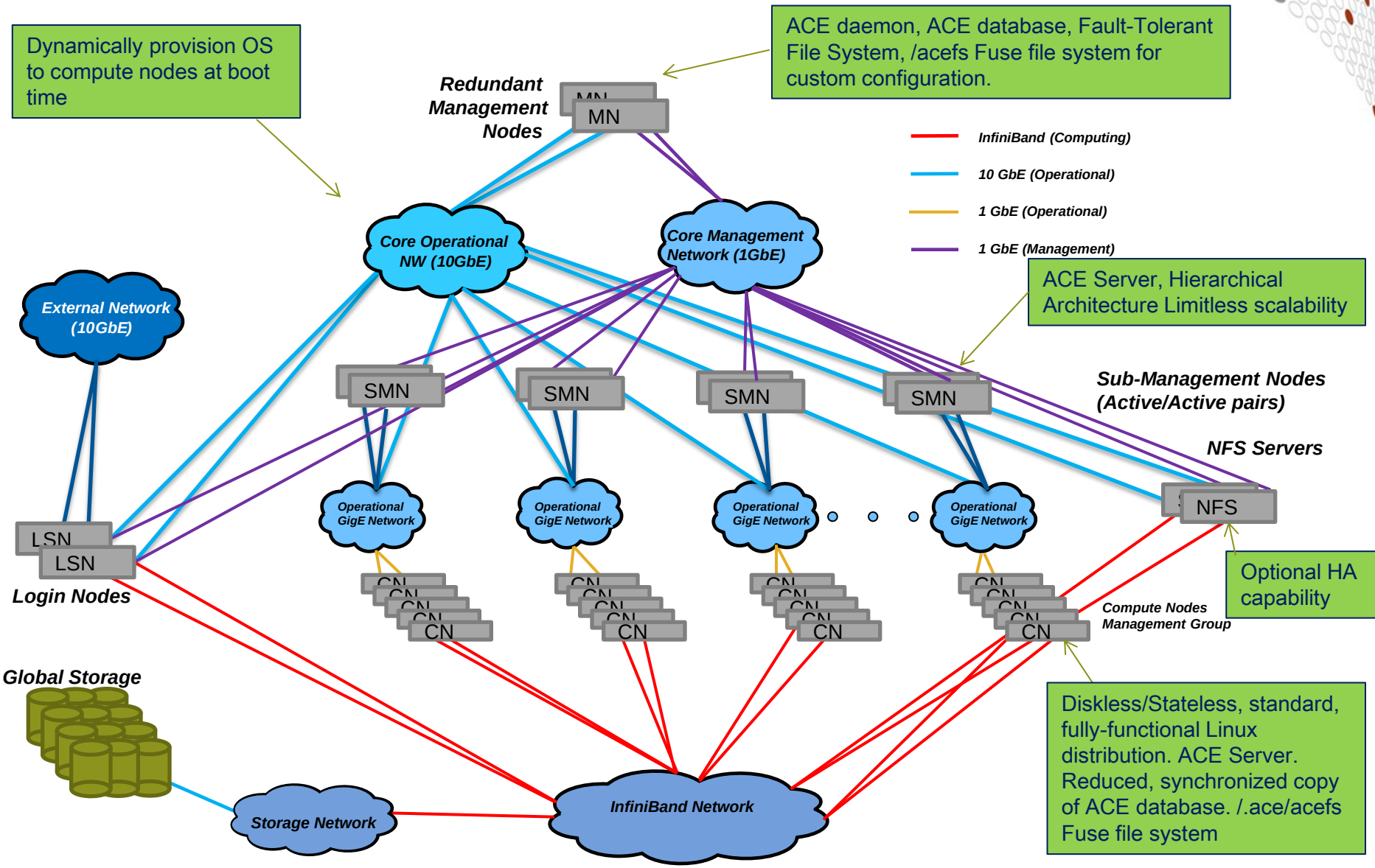


- **Configuration Management**
  - Multiple Revisions with Roll-Back Capability
  - Rolling Upgrades on Individual servers between jobs

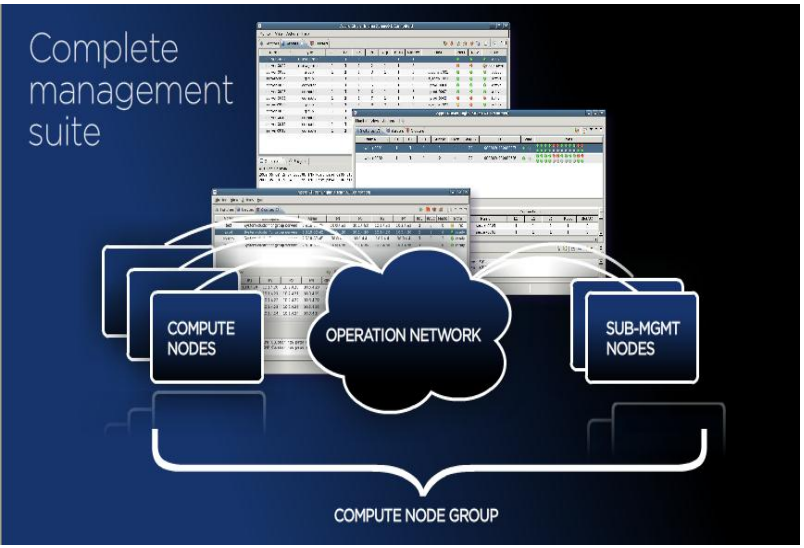
Cray ACE™ System Architecture
Scalability & Reliability



Cray ACE™ System Architecture



Cray ACE™ Software Management Features

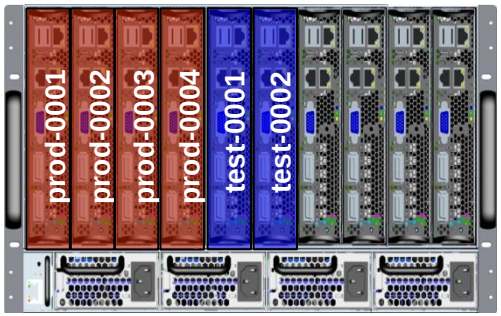


ACE's Scalable Hierarchical Architecture Provides the Following Services:

- Provisioning
- Launching
- Scalable File Services
- System Management
- Server Management
- Network Management
- Cluster Management & Monitoring
- Storage Management

Cray ACE™ Software Management Features

ACE
Advanced Cluster Engine

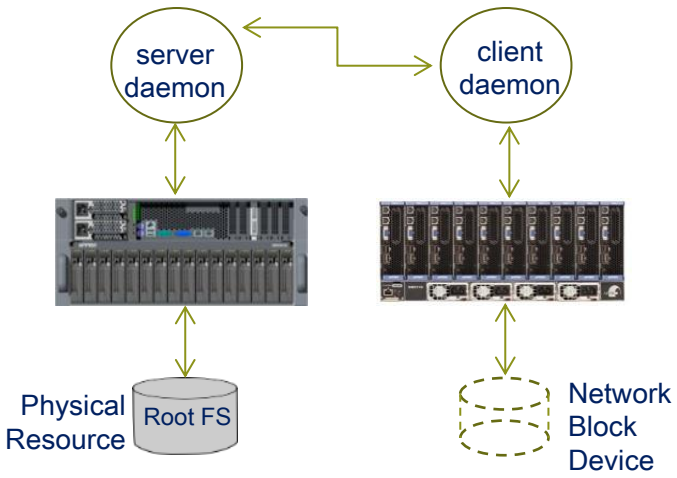


- Provisioning
  - Supports partitioning a Supercomputer into multiple logical computers
  - Maps logical computers (clusters) onto servers (nodes)
  - Supports multiple independent OS configurations, each with up to 10 revisions and rollback capability
  - Manages and monitors logical computer (cluster) status
  - Integrates Supercomputer status into the management system

Cray ACE™ Software Management Features

SLURM

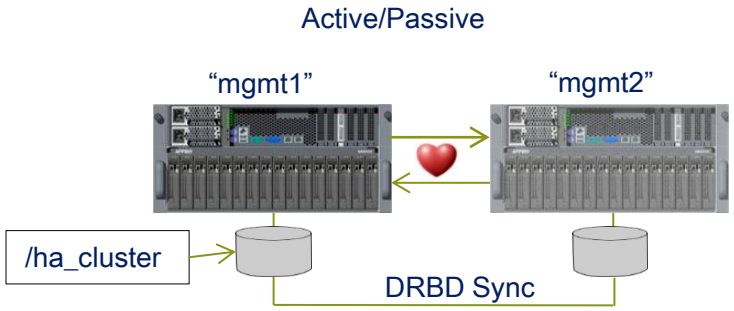
GRID ENGINE



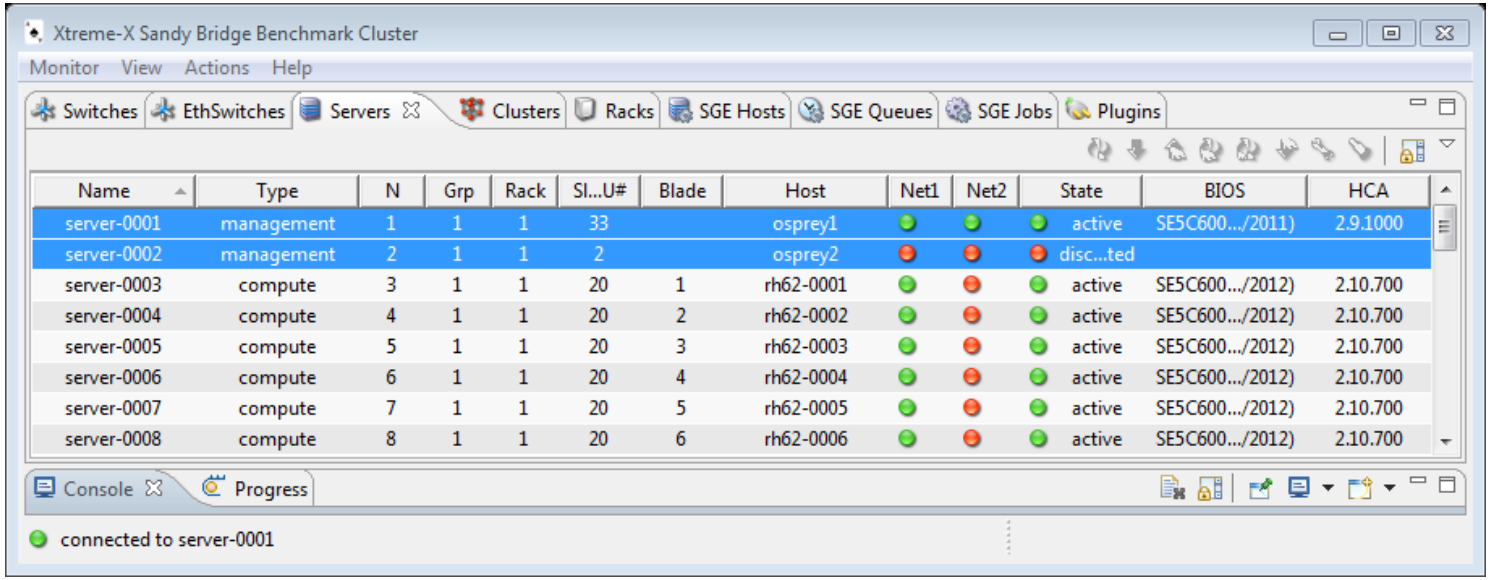
- Launching
  - Jobs
  - Job environment configuration
  - Job pre- and post-processing

- Scalable File Services
  - Root File System
  - Supports scalable root file systems for diskless nodes
  - Integrates server status into management system

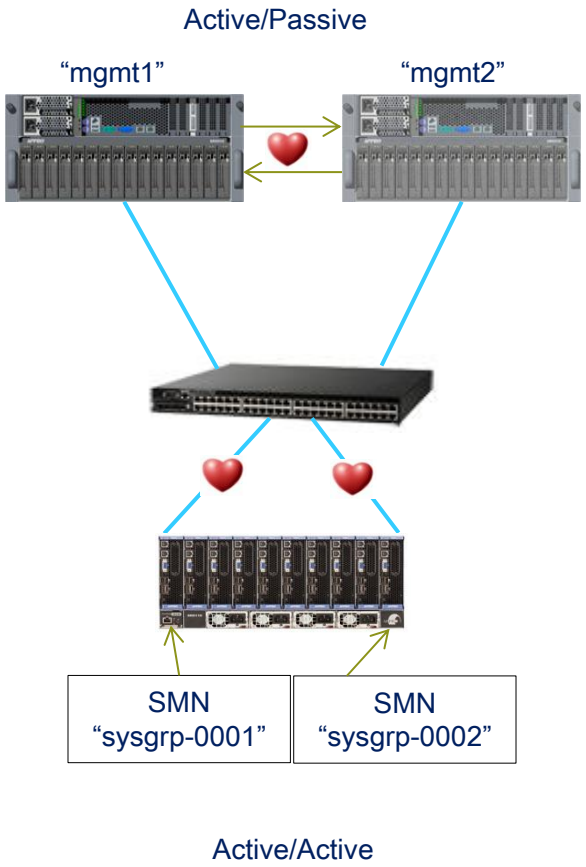
Cray ACE™ Software Management Features



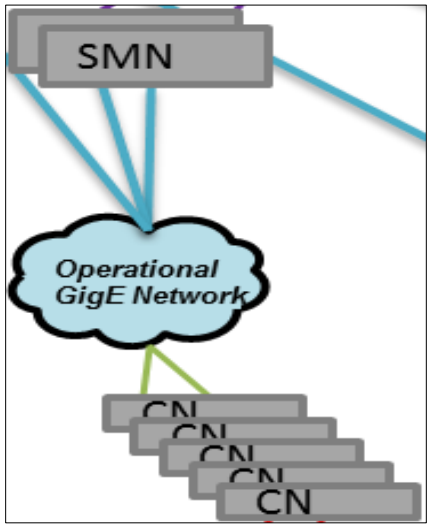
- System Management
  - Management of overall system configuration
  - Redundant Management Servers
  - Automatic failover



Cray ACE™ Software Management Features

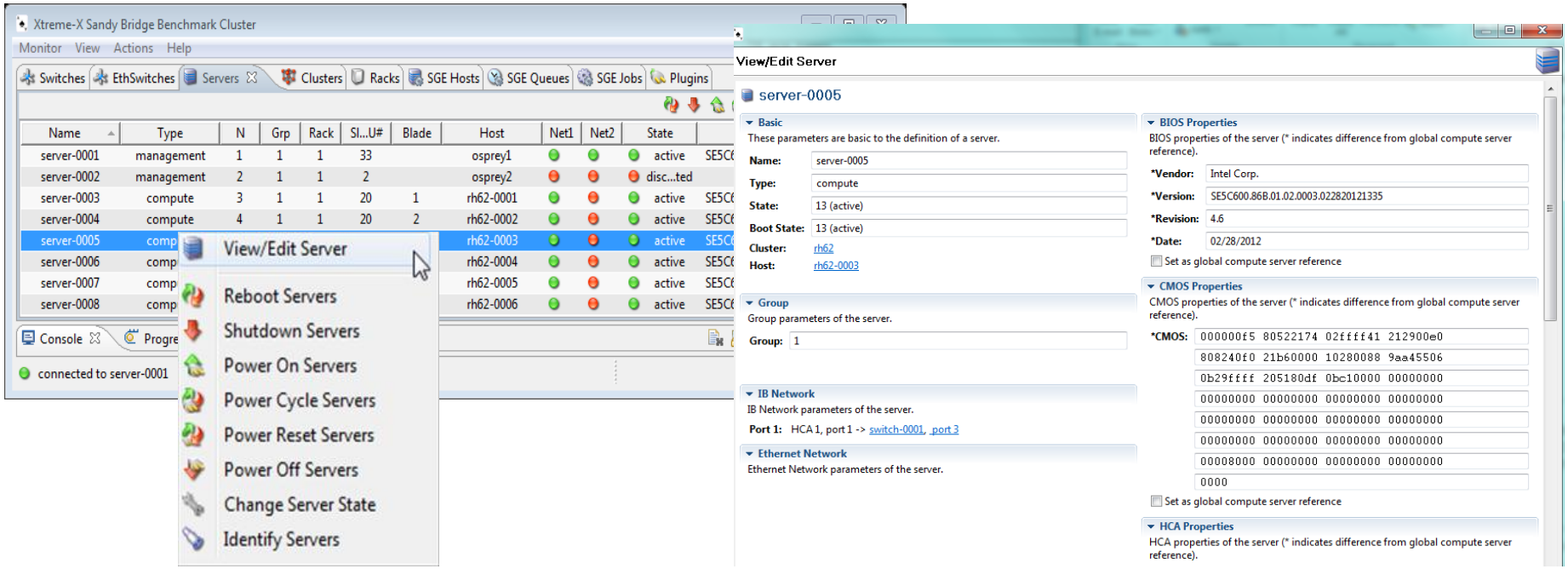


- System Sub-Management
  - Management of sub-management groups
  - Redundant Sub-Management Servers (standard diskless compute nodes)
  - Automatic failover



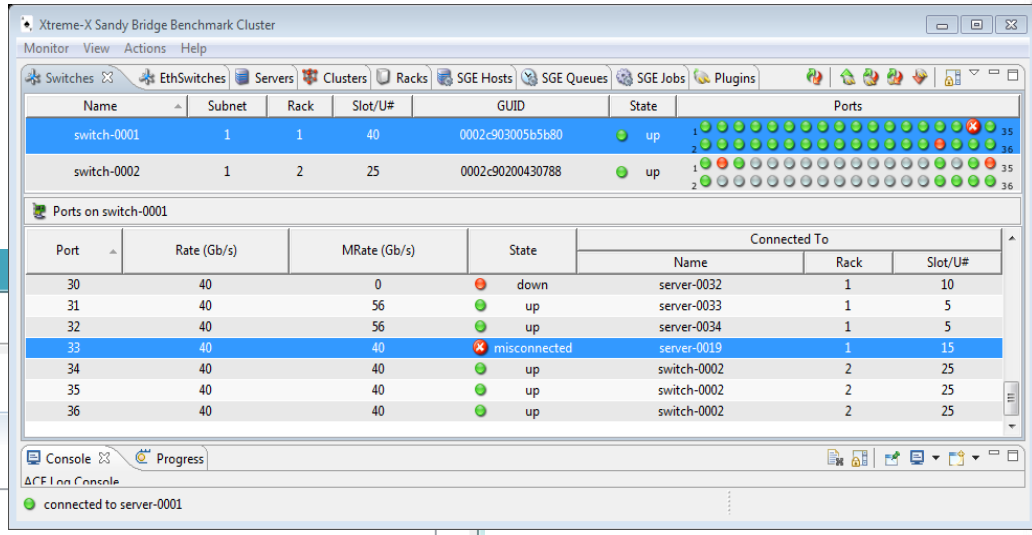
Cray ACE™ Software Management Features

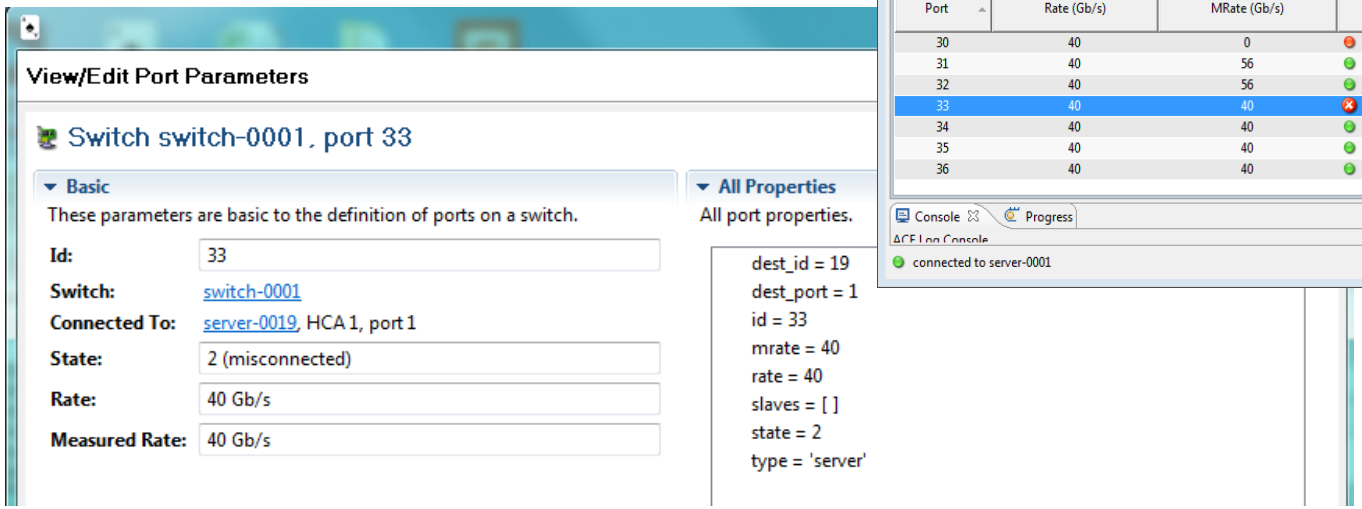
- Server Management
  - Automatic discovery of server hardware
  - Remote server control (Power On, Off, Cycle)
  - Remote server initialization (Reset, Reboot, Shut Down)
  - Scalable, fast, diskless booting for large node count systems
  - Server redundancy and failover (management & sub-management)
  - Integrates server status into management system



Cray ACE™ Software Management Features

- Network Management
  - Validates network topology
  - Monitors switches and cables
  - Notification of mis-connected and slow links
  - Route around network failures on dual-rail management networks

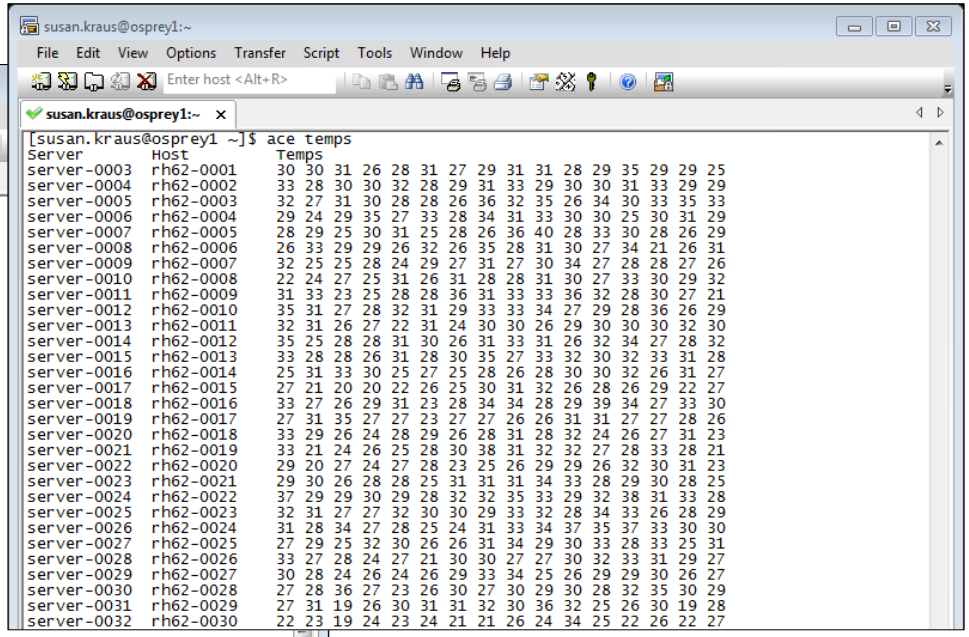


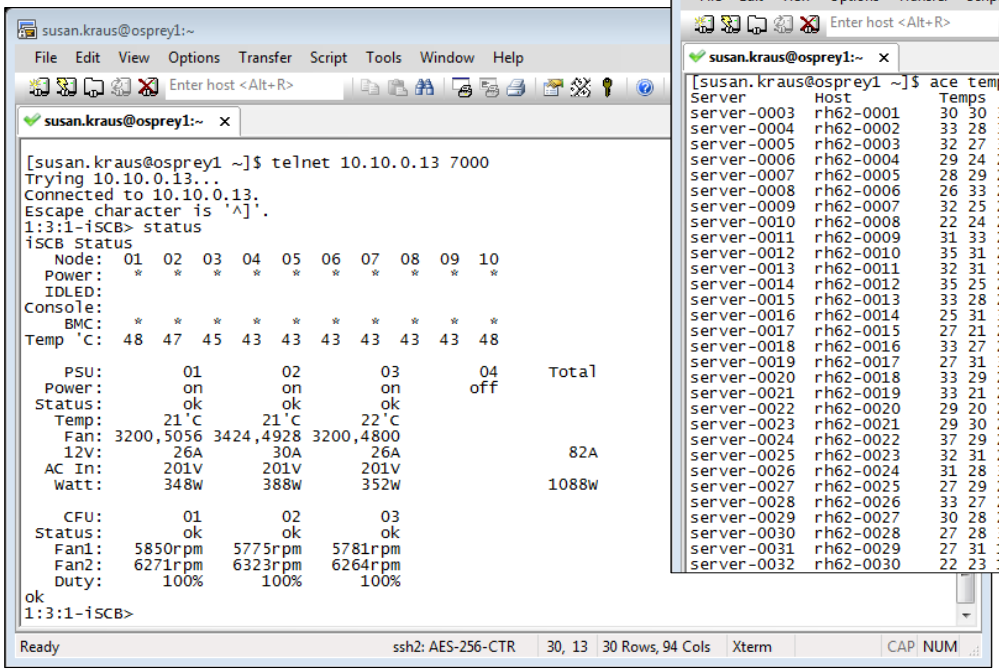


Cray ACE™ Software Management Features

- Monitoring (Server/Sub-rack)
  - CPU temperatures
  - Power
  - Fan Speeds

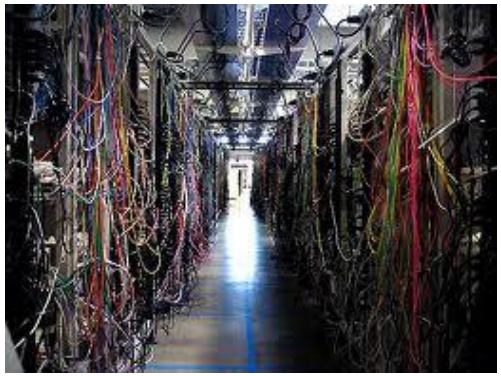
"ace temps" command on management node





"status" command on iSCB

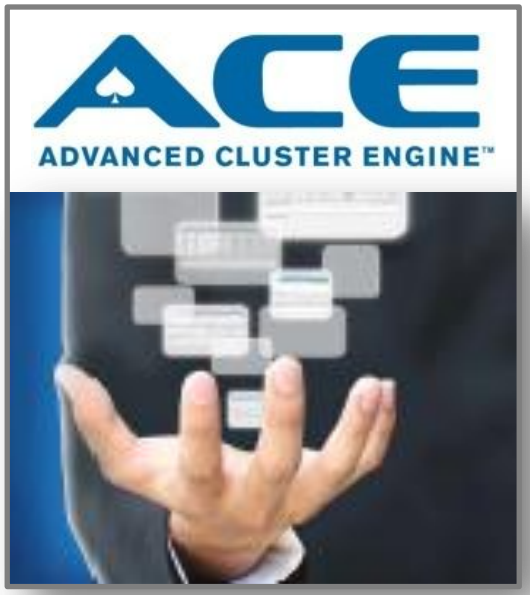
Cray ACE™ Software Management Features



- Network Management (InfiniBand and Ethernet)
  - Automatic discovery of interconnect hardware
  - Redundant paths and networks (dual-rail optional)
  - Failover for dual-rail networks
  - Integrates network status into management system

- Storage
  - Supports Lustre, NFS, Panasas
  - High bandwidth to secondary storage

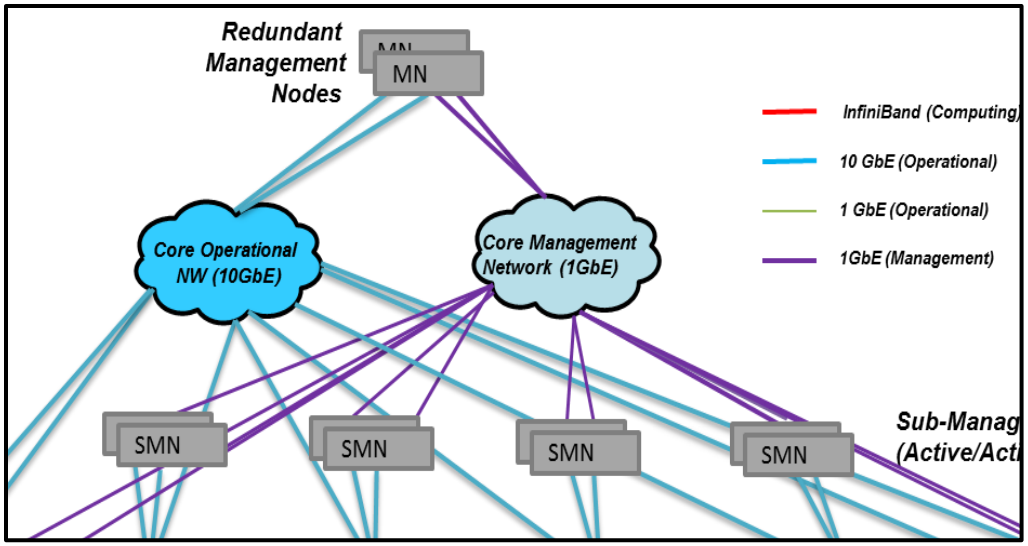
Cray ACE™ Software Management Benefits



- Hierarchical management system
- Diskless/Stateless Computing Environment
- Reduced System Administration

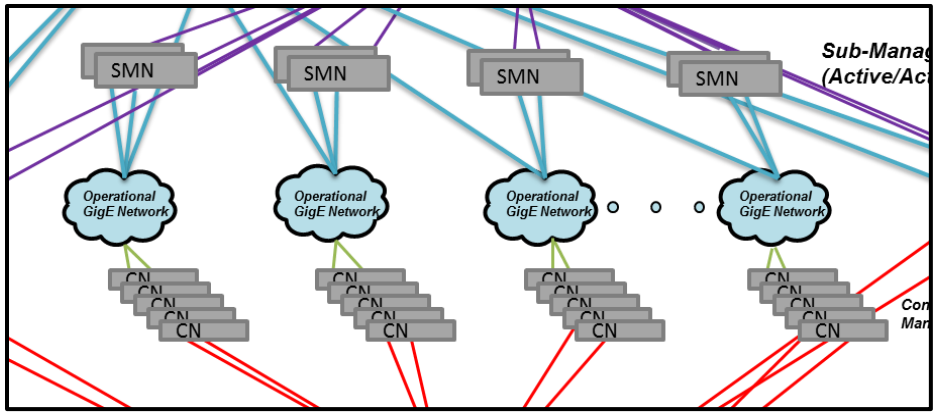
Cray ACE™ Software Management Benefits

- **Hierarchical Management System**
  - Hardware scalability through the use of sub-management servers
  - Management scalability through dynamic provisioning of cluster images – if you can manage a one-node system, you can manage a 10,000-node system



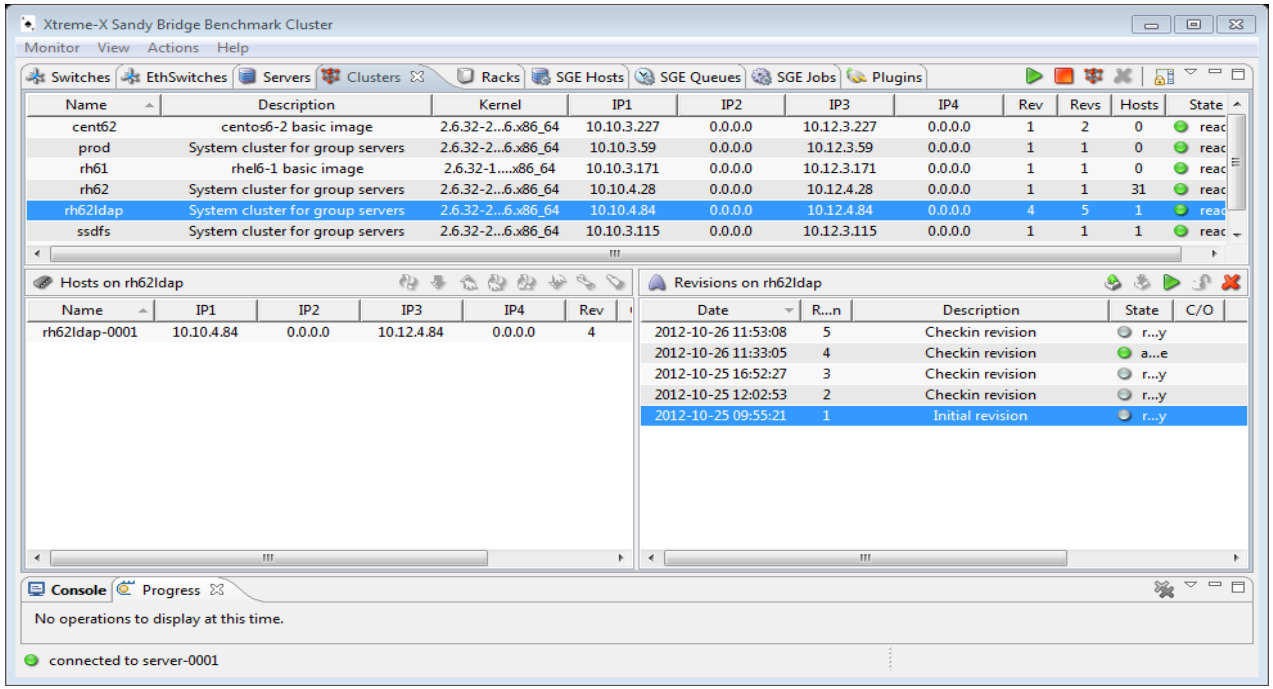
Cray ACE™ Software Management Benefits

- **Diskless/Stateless Computing Environment**
  - Dynamic provisioning of OS and run-time environment
  - Scalable network booting
  - Ability to use a full kernel and OS without requiring a large amount of memory on each node
  - Removes need for local hard drives, reducing MTBI for the compute node
  - Reduces complexity of managing and synchronizing thousands of states at the compute node level
  - Scalable access to root file system using a network block device
    - Only required files and data are pulled from the management servers, and sub-management servers cache all required data

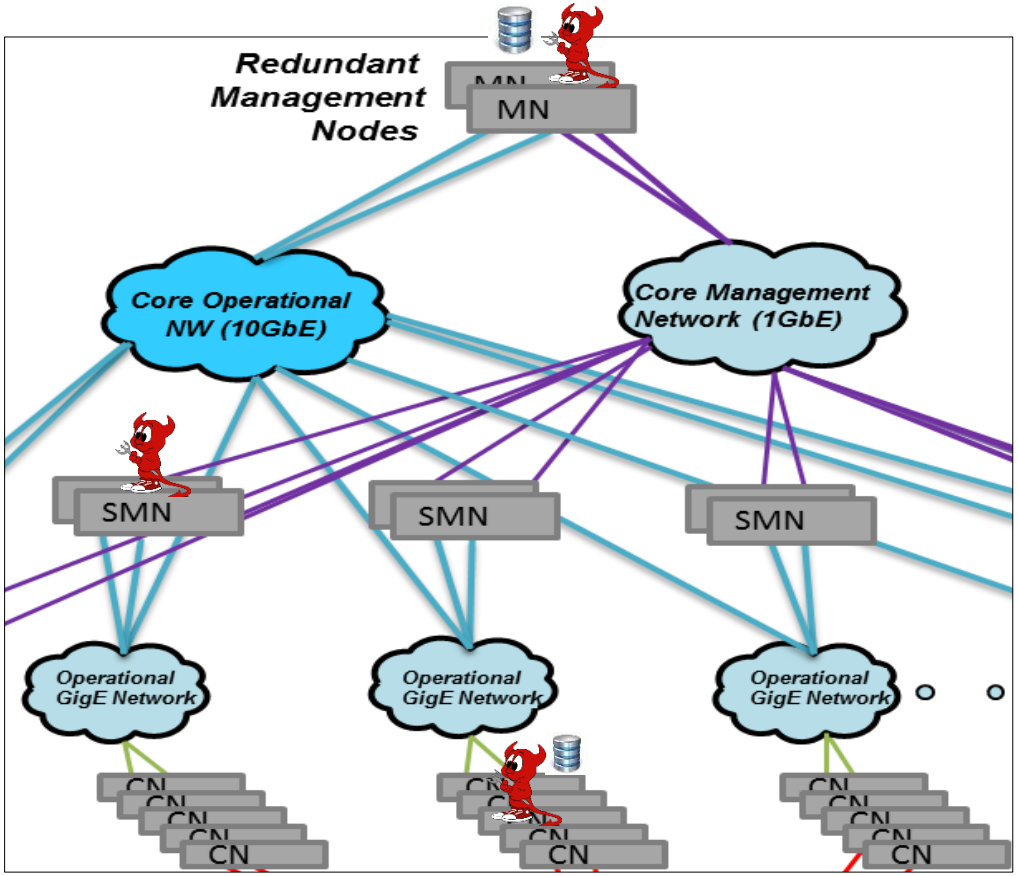


Cray ACE™ Software Management Benefits

- **Reduced System Administration**
  - OS, run-time, and support programs all stored in one location
    - Streamlines system administration
    - Minimizes security risks
  - System Administration time is not impacted by the number of nodes
  - Revision system allows sys admin to quickly provision nodes with different OS and run-time environment

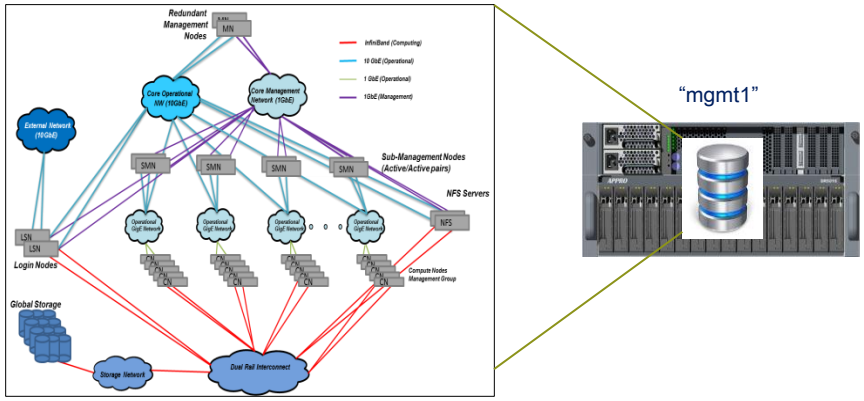


Cray ACE™ Software Management Components

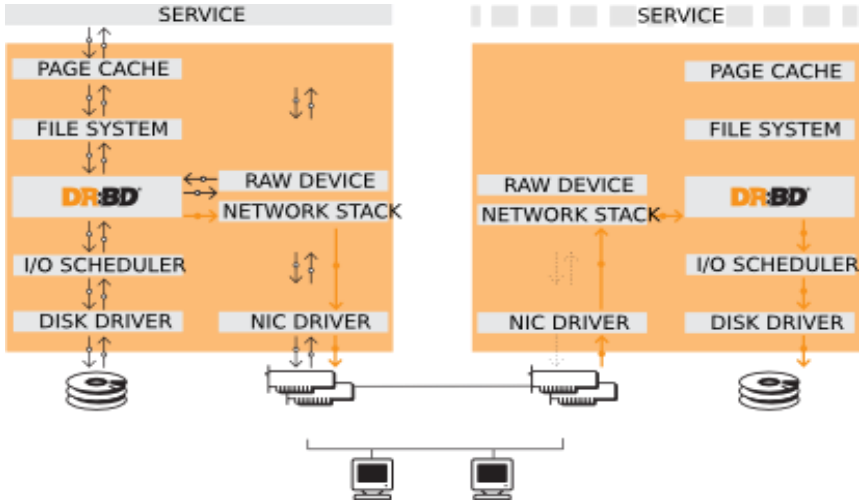


- **Daemons**
  - ACE Daemon (aced) runs on Management Server
  - ACE Server (ace_server) runs on sub-management and compute servers.
- **Data**
  - ACE database and "/ha_cluster" fault-tolerant file system on Management Servers
  - ACE database on Compute Server (reduced, synchronized copy)

Cray ACE™ Software Management Data and File Systems



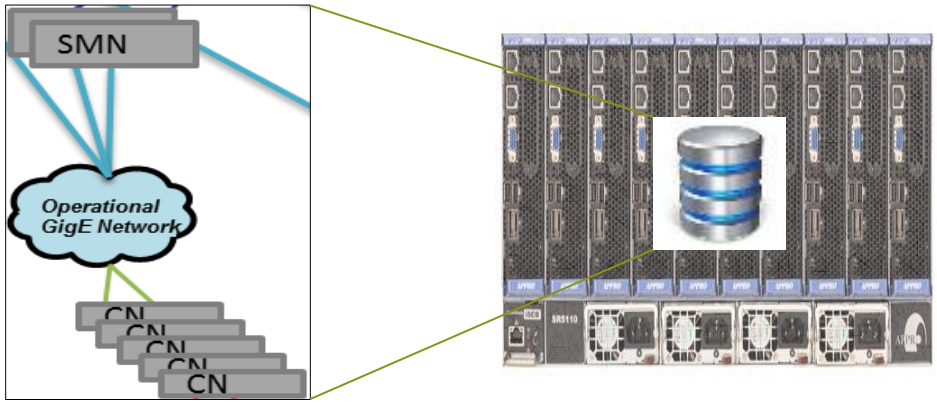
- **Management Servers**
  - **ACE database**
    - System Configuration
    - Server Configuration
    - Cluster Configuration
    - Network Connections
    - Status of Servers, Switches, and Ports



  - **/ha_cluster file system**
    - Fault tolerant with HA configuration
      - DRBD Sync
    - Cluster storage (/home and /global)

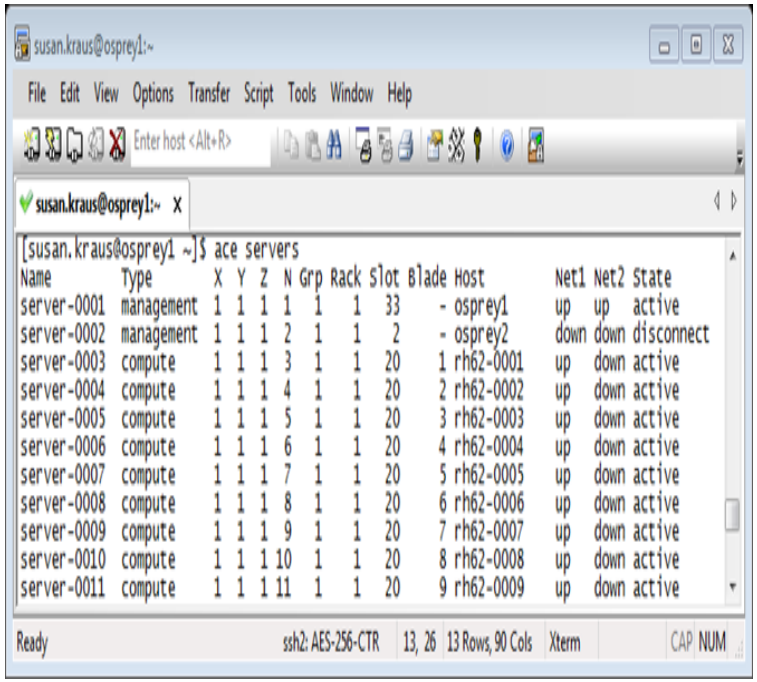
Cray ACE™ Software Management Data and File Systems

- **Compute Servers – ACE Database**
  - **ACE Database**
    - Reduced synchronized copy
    - A subset related to only that Sub-Management Server's group

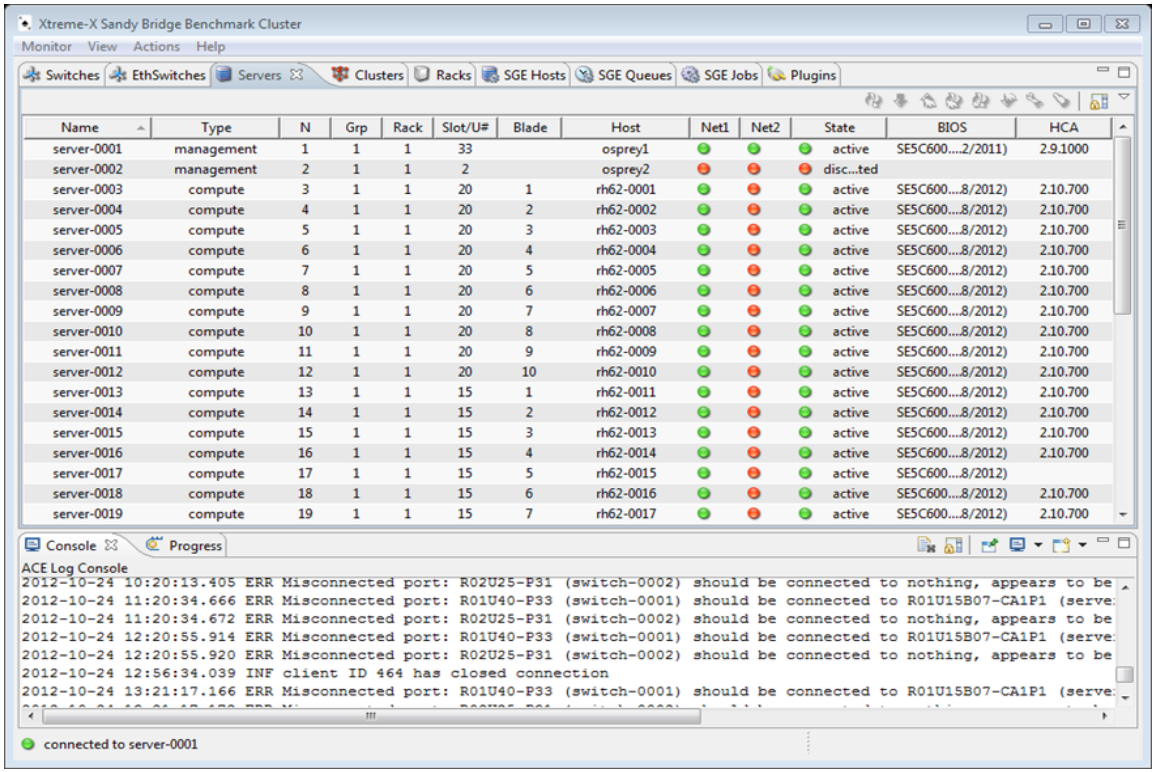


Cray ACE™ Software Management Interfaces

ACE™ Command Line Interface



ACE™ GUI Graphical User Interface



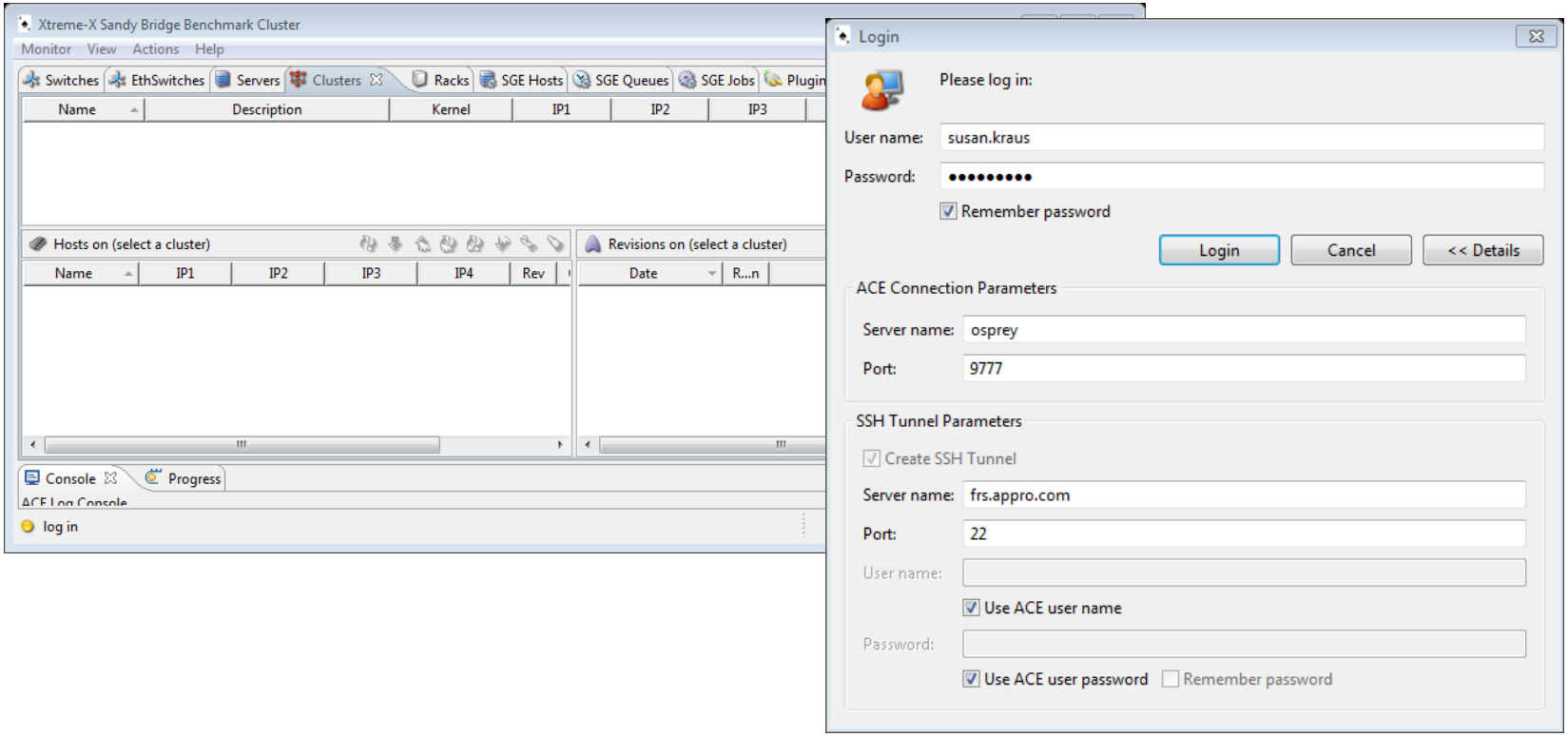
Cray ACE™ Software Management Interfaces - CLI

- ACE Command Line Interface (CLI) invoked using “ace” command

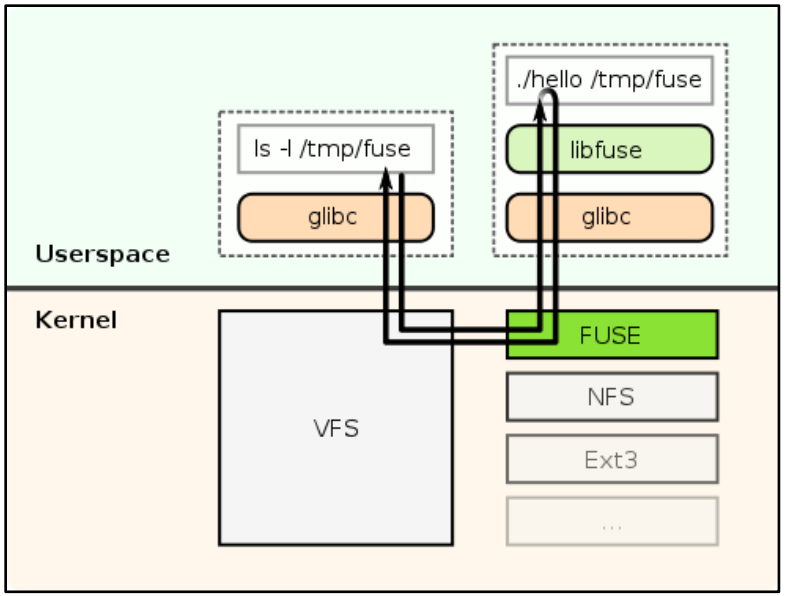
```
[susan.kraus@osprey1 ~]$ ace help

The Appro Cluster Engine supports the following commands:

General Commands

    help                Print this help message
    ping                Check connectivity
    date                Show UTC date
    time                Show time
    log_flush           Flush logs to disk
    debug_on            Turn debug logs on
    debug_off           Turn debug logs off
    version             Show ACE version

Status Commands

    switches            Show switch status
    ports               Show detailed switch port status
    ethswitches         Show Ethernet switch status
    servers             Show server status
    clusters            Show cluster status
    hosts               Show host status
    revisions           Show revision status
    get_logs            Show logs
    loads               Show load status per server
    memory              Show memory usage status per server
    temps               Show CPU temperature status per server

Server Commands

    poweron             Power on one or more servers
    poweroff            Power off one or more servers
    cycle               Power cycle one or more servers
    reset               Power reset one or more servers
    shutdown            Shutdown one or more servers
    reboot              Reboot one or more servers
    identify            Identify one or more servers

Cluster Commands

    clone               Clone a cluster
    delete              Delete a cluster
    start               Start hosts on a cluster
    stop                Stop hosts a cluster
    update              Update the boot images for a cluster
    checkin             Check in a cluster revision
    checkout            Check out a cluster revision
    release             Release a cluster revision
    remove              Remove a cluster revision
    activate            Activate a cluster revision
    create              Create a cluster
    export_image        Export cluster image to File
    import_image        Import File(Cluster Image) to ACE

Grid Engine Commands

    add_sge_host        Add hosts to Grid Engine
    delete_sge_host     Delete hosts from Grid Engine
    add_sge_cluster     Add clusters to Grid Engine
    delete_sge_cluster  Delete clusters from Grid Engine

    sge                 Show hosts of Grid Engine


Plugin Commands

    plugins             Show the Plugins
    plugin_add          Add a Plugin
    plugin_delete       Delete a Plugin
    run                 Run a Plugin


Database Commands

    get                 Get database information
    put                 Put data in to database
    sync                Synchronize database with secondary storage

Network Commands

    restart1            Restart infiniband network #1
    restart2            Restart infiniband network #2
[susan.kraus@osprey1 ~]$
```

Cray ACE™ Software Management Interfaces - GUI

- **ACE Graphical User Interface (GUI)**
  - Efficient, responsive, eclipse-based GUI application
    - Updates automatically sent to GUI
  - Secure access through SSH tunnel
  - Supports Windows, Linux, Mac OS



Cray ACE™ Software Management Interfaces - ACEFS



- **ACEFS – ACE Fuse File System (File System In User Space)**
  - A file system interface into the ACE database
  - /acefs – on management servers
    - /acefs/global – globally applied to all clusters
    - /acefs/clusters – cluster specific which overrides global configuration
  - /.ace/acefs – on compute servers
    - Read only file system on the compute servers

Cray ACE™ Software Management Interfaces - ACEFS

```
[susan.kraus@osprey1 ~]$ cd /acefs
[susan.kraus@osprey1 acefs]$ ls -la
total 8
drwxr-xr-x  10 root root    0 Oct 24 16:06 .
dr-xr-xr-x. 28 root root 4096 Oct 11 03:06 ..
drwxr-xr-x   8 root root    0 Oct 24 16:06 clusters
drwxr-xr-x   2 root root    0 Oct 24 16:06 ethswitches
drwxr-xr-x  11 root root    0 Oct 24 16:05 global
drwxr-xr-x  12 root root    0 Oct 24 16:06 plugins
drwxr-xr-x   2 root root    0 Oct 24 16:06 self
drwxr-xr-x  42 root root    0 Oct 24 16:06 servers
drwxr-xr-x   5 root root    0 Oct 24 16:06 sgedb
drwxr-xr-x   4 root root    0 Oct 24 16:06 switches
[susan.kraus@osprey1 acefs]$ ls -la clusters
total 0
drwxr-xr-x  8 root root 0 Oct 24 16:06 .
drwxr-xr-x 10 root root 0 Oct 24 16:06 ..
drwxr-xr-x  8 root root 0 Oct 24 16:06 cent62
drwxr-xr-x  8 root root 0 Oct 24 16:06 prod
drwxr-xr-x  8 root root 0 Oct 24 16:06 rh61
drwxr-xr-x  8 root root 0 Oct 24 16:06 rh62
drwxr-xr-x  8 root root 0 Oct 24 16:06 ssdfs
drwxr-xr-x  8 root root 0 Oct 24 16:06 sysgrp
[susan.kraus@osprey1 acefs]$
```

```
[susan.kraus@osprey1 acefs]$ ls -la clusters/rh62
total 0
drwxr-xr-x  8 root root   0 Oct 24 16:07 .
drwxr-xr-x  8 root root   0 Oct 24 16:07 ..
-r--r--r--  1 root root   2 Oct 24 16:07 clone
drwxr-xr-x 12 root root   0 Oct 24 16:07 control
-r--r--r--  1 root root  33 Oct 24 16:07 desc
drwxr-xr-x  2 root root   0 Oct 24 16:07 files
-r--r--r--  1 root root  59 Oct 24 16:07 hostfs_files
drwxr-xr-x 34 root root   0 Oct 24 16:07 hosts
-r--r--r--  1 root root   2 Oct 24 16:07 id
-r--r--r--  1 root root   2 Oct 24 16:07 initramfs
-r--r--r--  1 root root  11 Oct 24 16:07 install
-r--r--r--  1 root root  11 Oct 24 16:07 ip1
-r--r--r--  1 root root   8 Oct 24 16:07 ip2
-r--r--r--  1 root root  11 Oct 24 16:07 ip3
-r--r--r--  1 root root   8 Oct 24 16:07 ip4
-r--r--r--  1 root root  22 Oct 24 16:07 kernel
-r--r--r--  1 root root  34 Oct 24 16:07 kernel_args
-r--r--r--  1 root root   2 Oct 24 16:07 kickstart
-r--r--r--  1 root root   0 Oct 24 16:07 ks_eth0
-r--r--r--  1 root root   0 Oct 24 16:07 ks_eth1
-r--r--r--  1 root root   0 Oct 24 16:07 ks_netmask
-r--r--r--  1 root root   2 Oct 24 16:07 localboot
-r--r--r--  1 root root   3 Oct 24 16:07 max_hosts
drwxr-xr-x  5 root root   0 Oct 24 16:07 mount
ot root   5 Oct 24 16:07 name
ot root   3 Oct 24 16:07 num_hosts
ot root   0 Oct 24 16:07 pxe
ot root   2 Oct 24 16:07 revision
ot root   0 Oct 24 16:07 revisions
ot root   2 Oct 24 16:07 state
ot root   0 Oct 24 16:07 storage
ot root   0 Oct 24 16:07 test-dir
ot root 275 Oct 24 16:07 tmpfs_files
rey1 acefs]$
```

```
[susan.kraus@osprey1 acefs]$ ls -la clusters/rh62/mount
total 0
drwxr-xr-x 5 root root 0 Oct 24 16:08 .
drwxr-xr-x 8 root root 0 Oct 24 16:08 ..
drwxr-xr-x 6 root root 0 Oct 24 16:08 boot_commands.d
drwxr-xr-x 5 root root 0 Oct 24 16:08 etc
drwxr-xr-x 2 root root 0 Oct 24 16:08 root
[susan.kraus@osprey1 acefs]$
```

Cray ACE™ Software Management Managed Objects

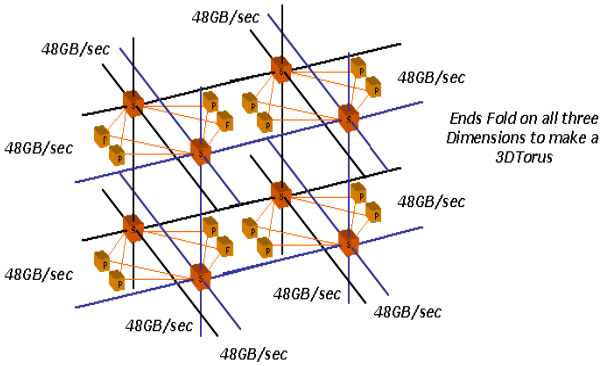
Servers

Management

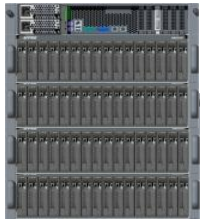
Compute

iSCB

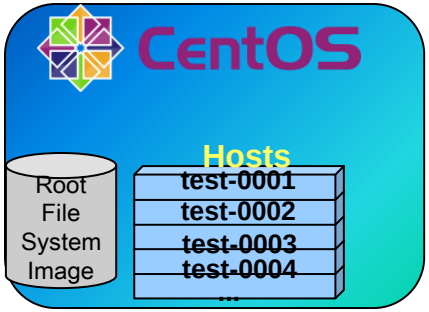
Networks

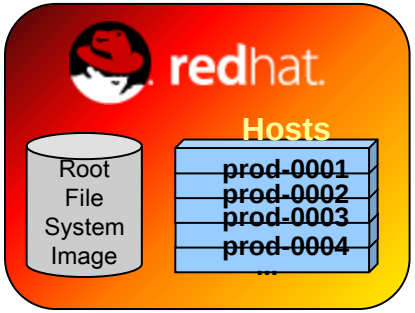


Storage



Clusters



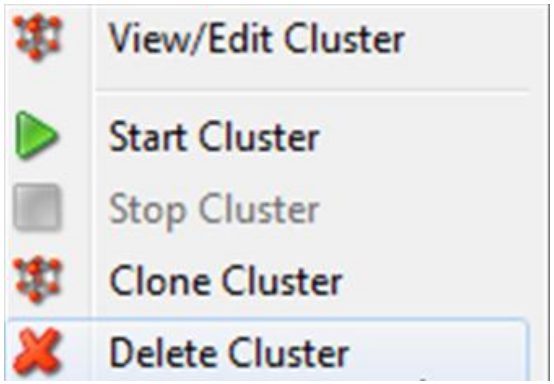


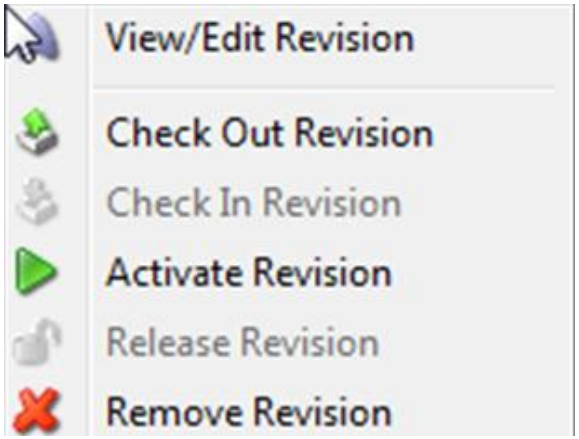
Cray ACE™ Software Management
Cluster Operating Systems

ACE Cluster Operating Systems

- Full CentOS, Red Hat, or SUSE installation
- Shared read-only root file system image
- Configuration files can be overridden or updated as necessary
- Cluster hosts with automatically configured IP addresses can be started on any compute node

Cray ACE™ Software Management Clusters

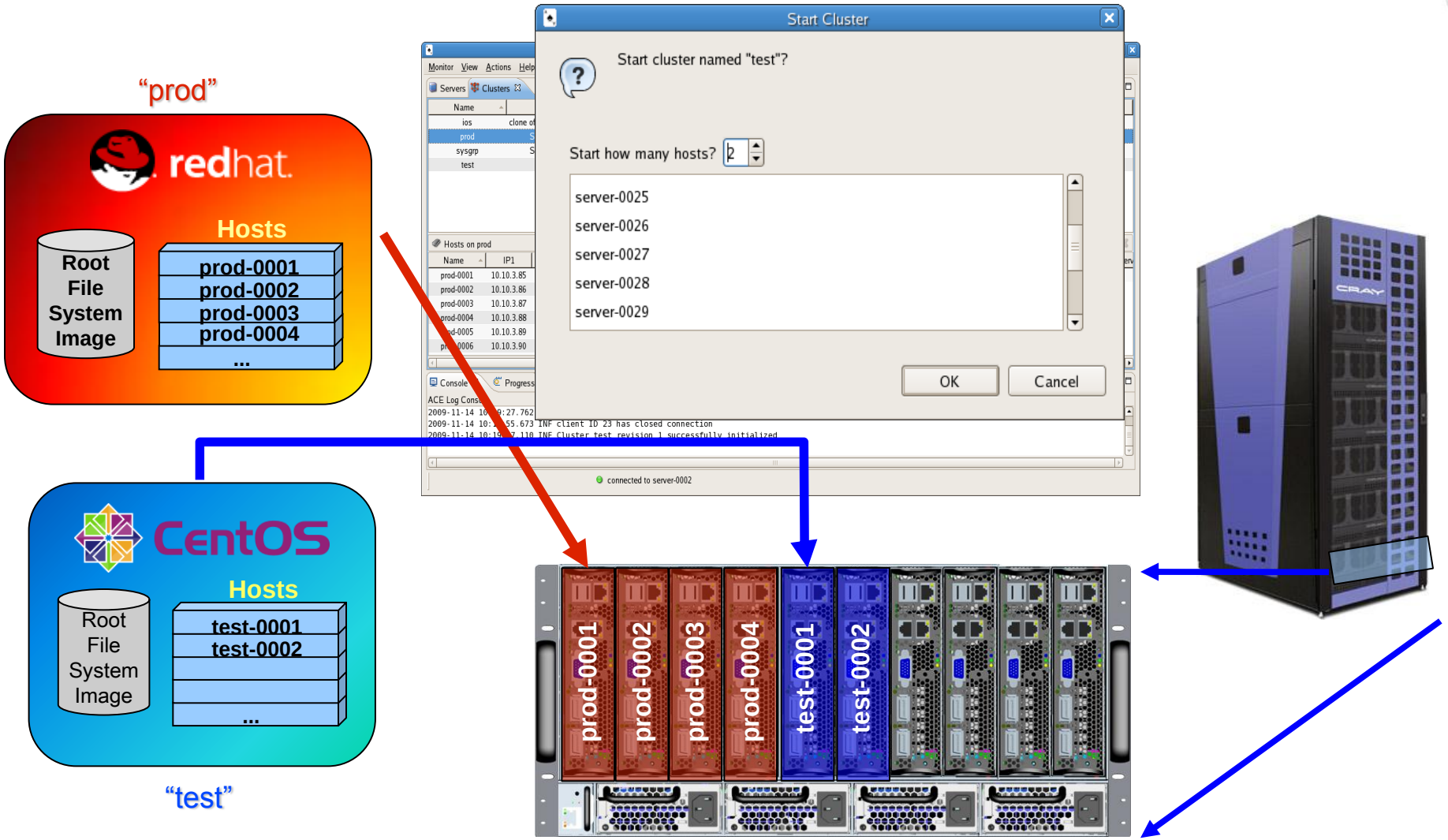




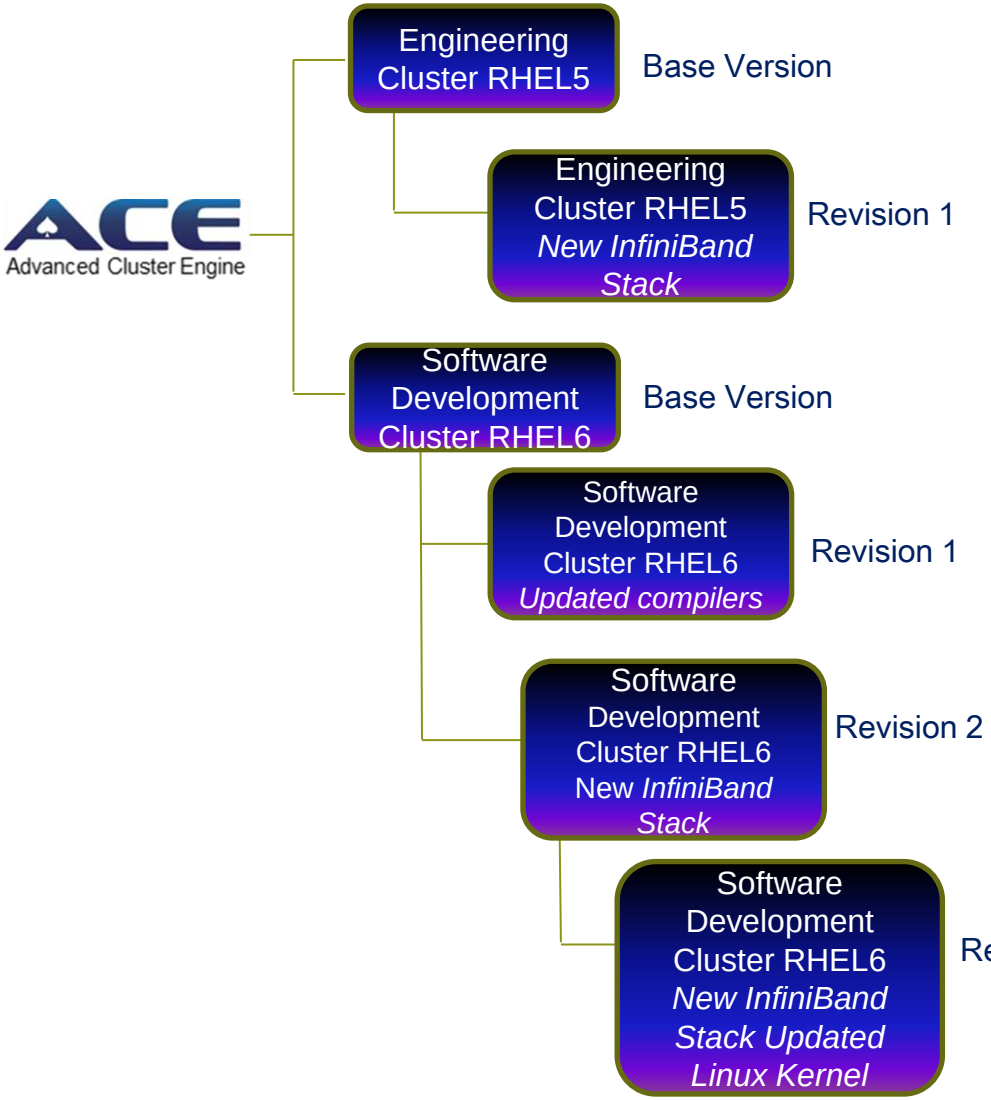
ACE cluster characteristics

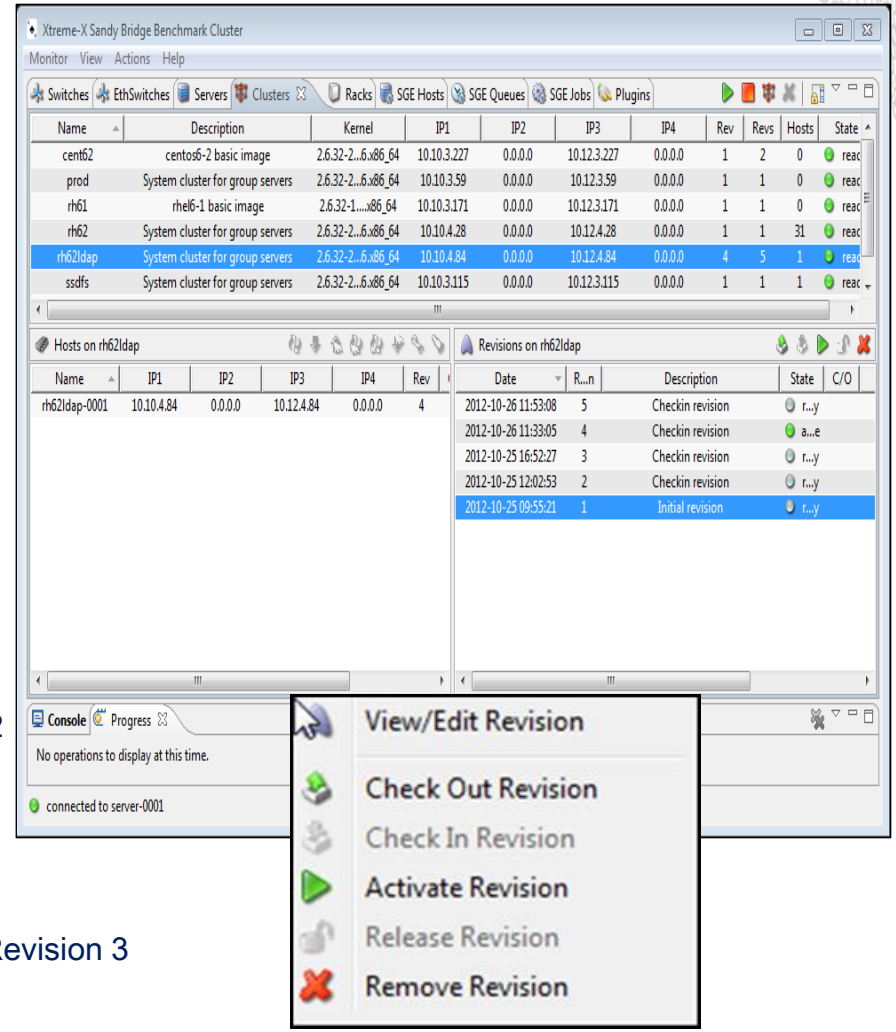
- Can be dynamically started on the physical servers with a simple start command specifying the number of hosts
- Root file system can be updated by checking out a “revision host”
- Supports up to 10 revisions of the root file system image
- Can be cloned to create new clusters or exported to be shared with other sites

Cray ACE™ Software Management
Cluster Provisioning with Diskless Nodes



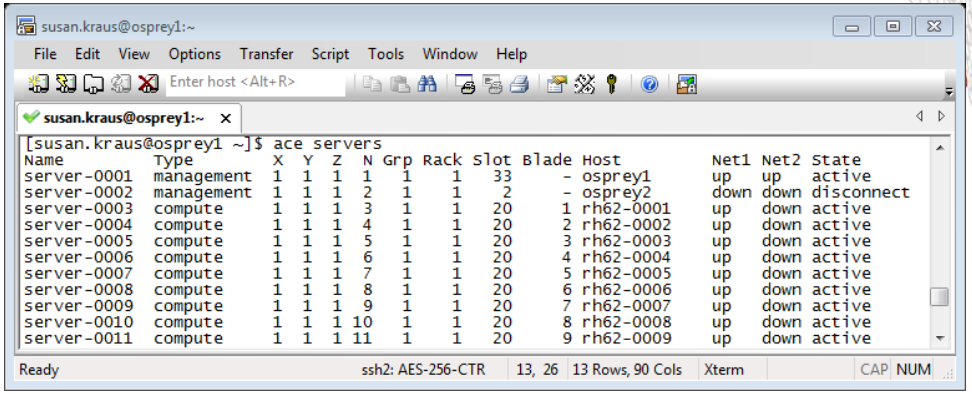
Cray ACE™ Software Management Cluster Revisions

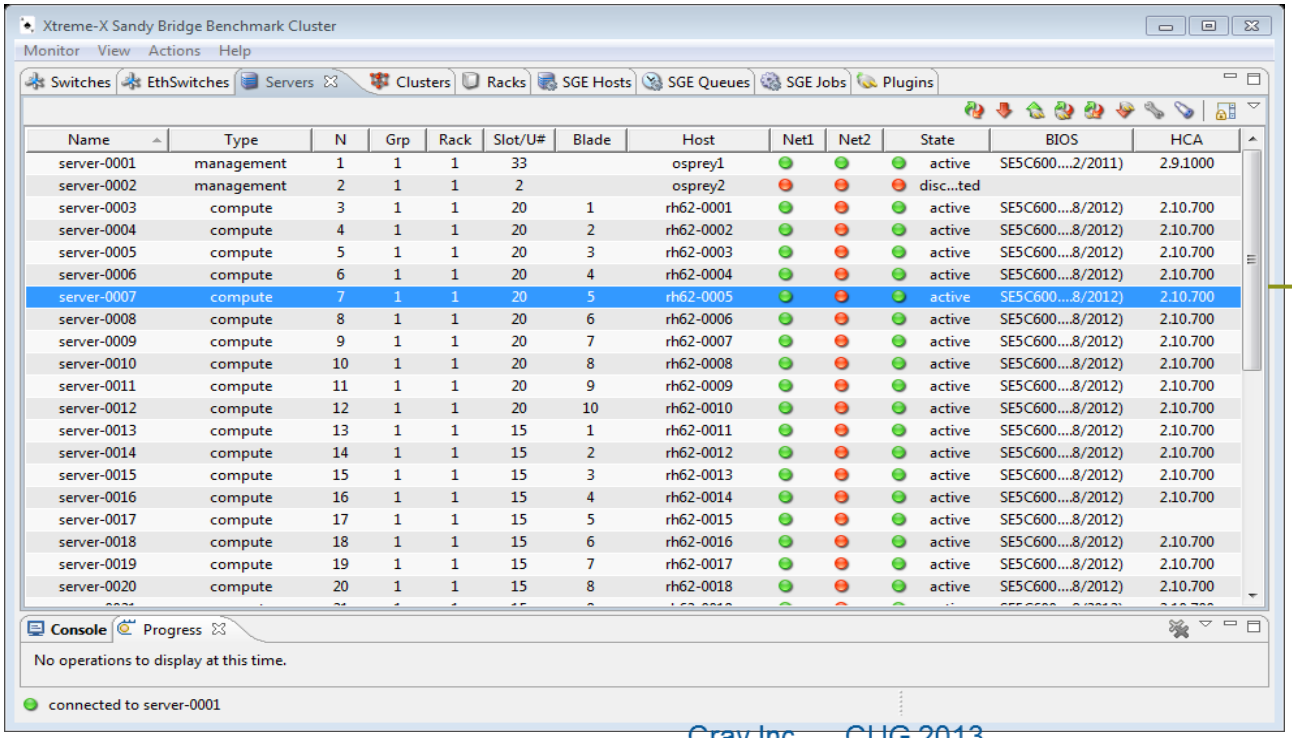


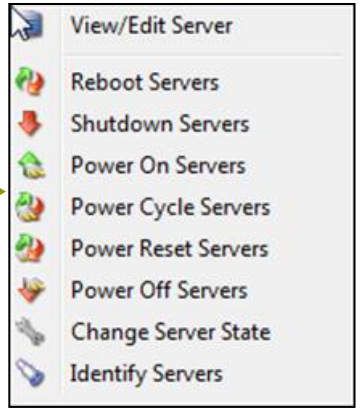


Cray ACE™ Software Management
Server Management

- Monitoring Servers
- Powering servers on and off via IPMI & iSCB
- Serial Console via SOL & iSCB
- Updating BIOS and system firmware



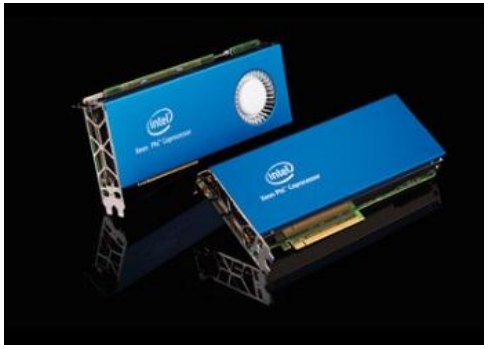


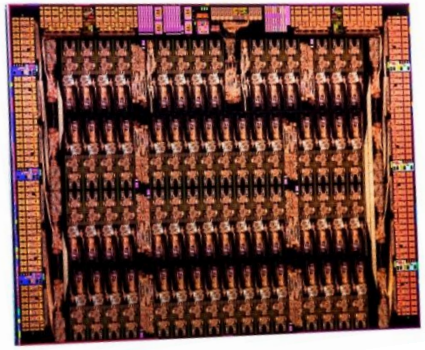


Cray ACE™ Software Management
NVIDIA GPU Accelerator Management

- Ability to power on and off GPUs via the iSCB
- Pre-installed, pre-configured, device drivers plus CUDA environment installed in the compute node image

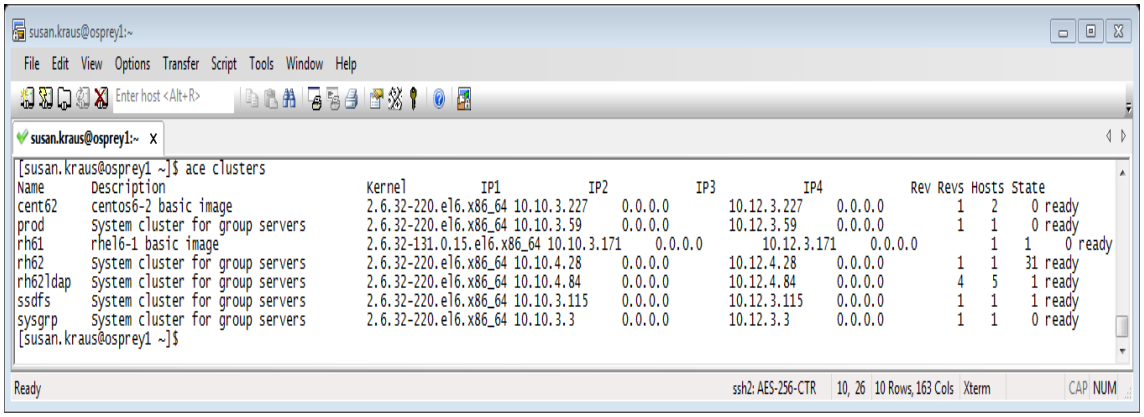
Cray ACE™ Software Management
Intel® Xeon Phi™ Coprocessor Management

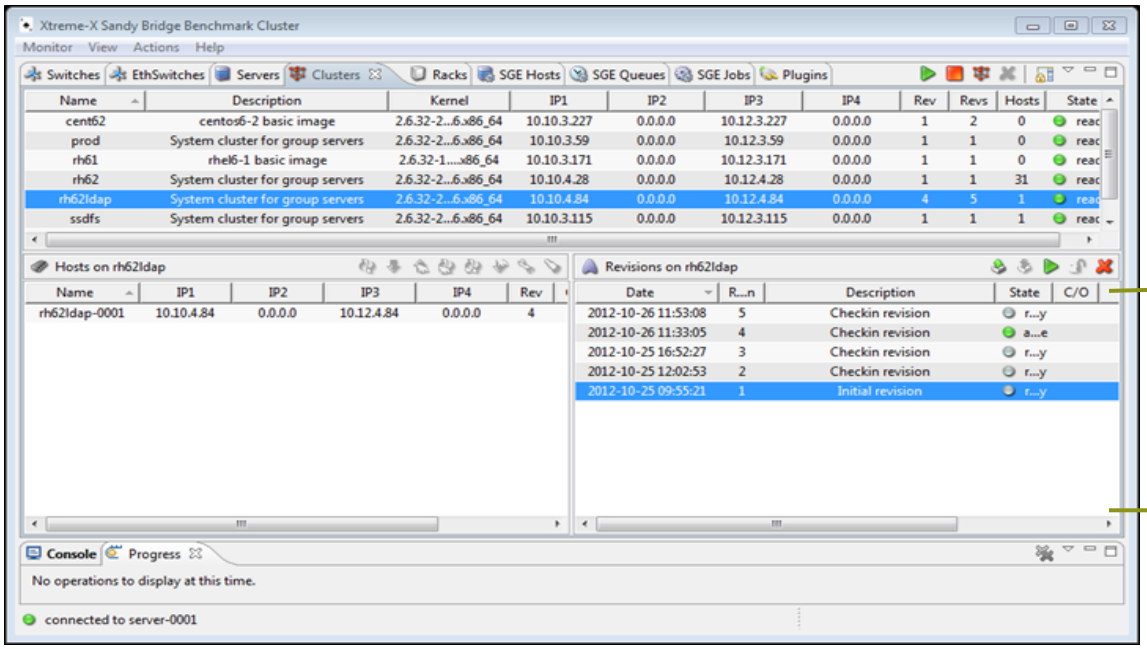




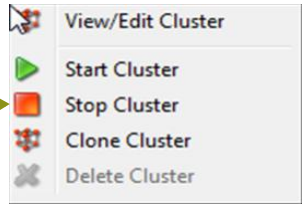
- MIC support is automatically installed and configured by ACE
- MPSS software resides on small SSD or disk local to the host node
- Ethernet bridging supported – MICs can be accessed directly from cluster using hostname – “prod-0001-mic0”
- ACE system administrator has full control of the MICs from ACE – MIC status, boot, reboot, reset, and shutdown
- User accounts automatically propagate to MIC
- MIC status available from the ACE CLI and GUI
- ACE plugins will include the MIC tests supplied with the software stack
- Friendly user environment
  - Global file system available from the MIC
  - /opt/intel available from the MIC
    - Avoids having to copy libraries and binaries to the MIC

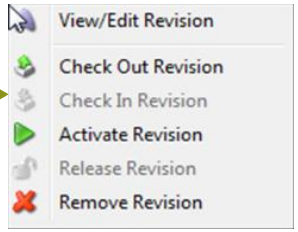
Cray ACE™ Software Management
Cluster Management



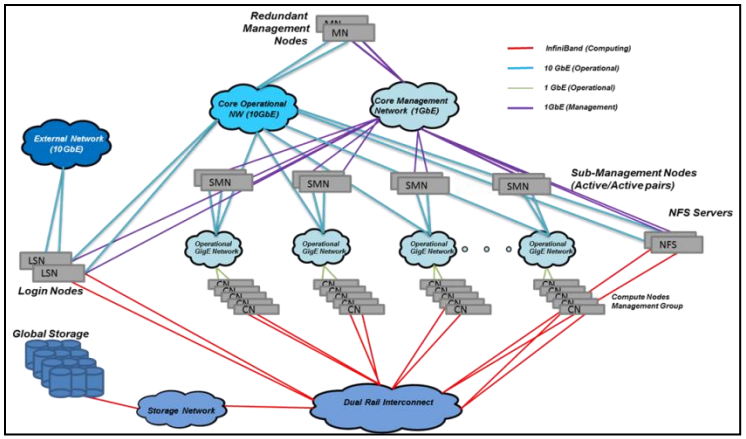


- Starting Cluster Hosts
- Stopping Cluster Hosts
- Cloning Clusters
- Updating Clusters
  - Checking out/in
  - Activating





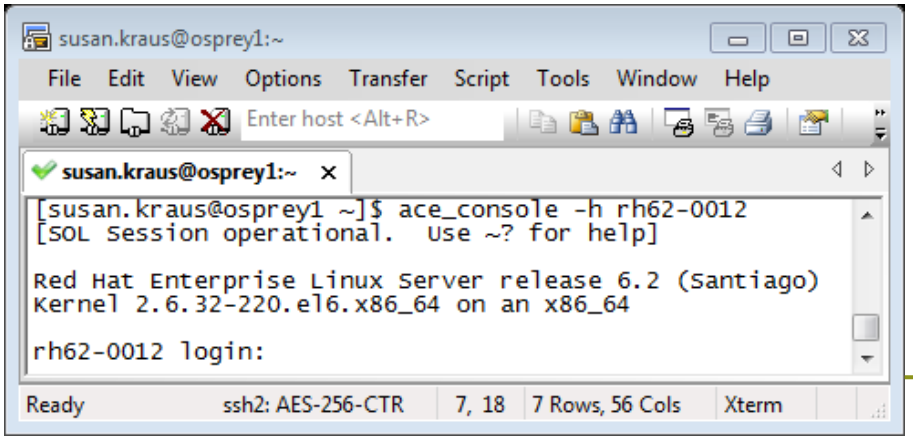
Cray ACE™ Software Management
Cluster Management

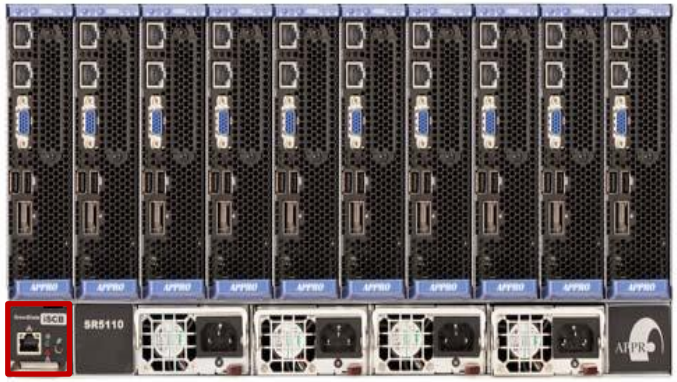


- Redundant Hierarchical System Management
- Dual networks provide protection from network component failures - switches, cables, NICs
- All system states are managed on the redundant management servers
- All the system configuration data is kept on a fault tolerant data base and file system on the management servers - system configuration files, root file systems
- If the primary management server dies, the secondary management server takes over and the system stays up and running
- Redundant sub-management servers provide scaling of network services and caching of operating system.
- Scalable booting of compute servers - allows thousands of compute nodes to boot quickly and simultaneously

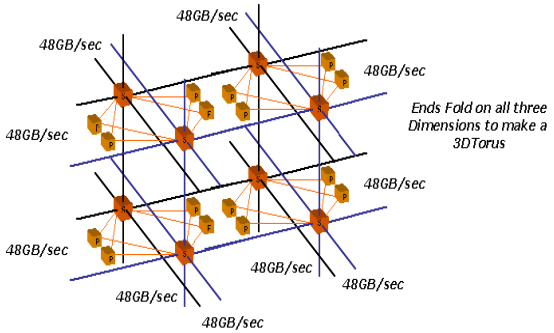
Cray ACE™ Software Management Configurations and Miscellaneous

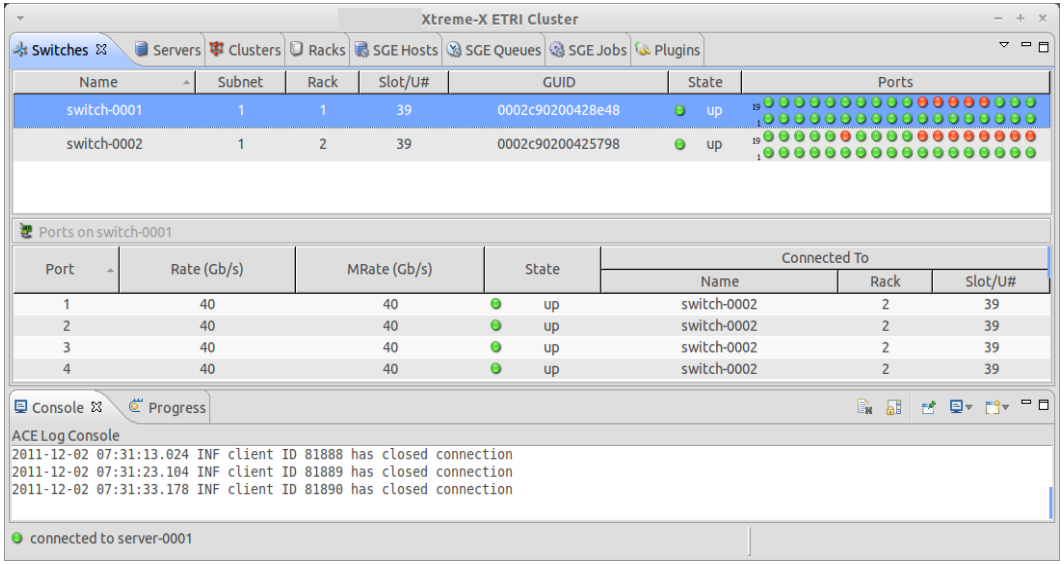
- Hosts management
  - /acefs/global/hosts – user controllable custom hosts
  - /acefs/global/all_hosts – ace controlled master hosts file includes /acefs/global/hosts
- User management – global password, group, shadow files
  - ace_useradd, ace_userdel, etc.
  - Cluster-unique files can be maintained under /acefs
- Serial Console via iSCB





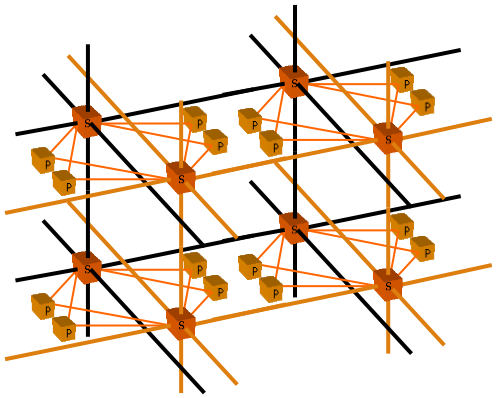
Cray ACE™ Software Management
Network Management





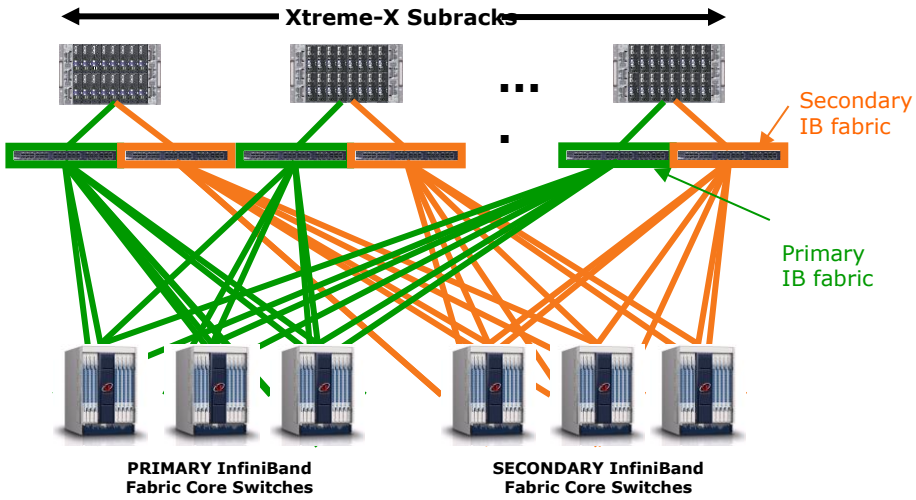
- Monitoring Infiniband Failures
- Monitoring Ethernet Failures
- Running Network Diagnostics

Cray ACE™ Software Management Support for Multiple Topologies



Scales to Thousands of Nodes

Single Rail and Dual Rail Fat-tree
- Distributed Core
- Large Director Class Core



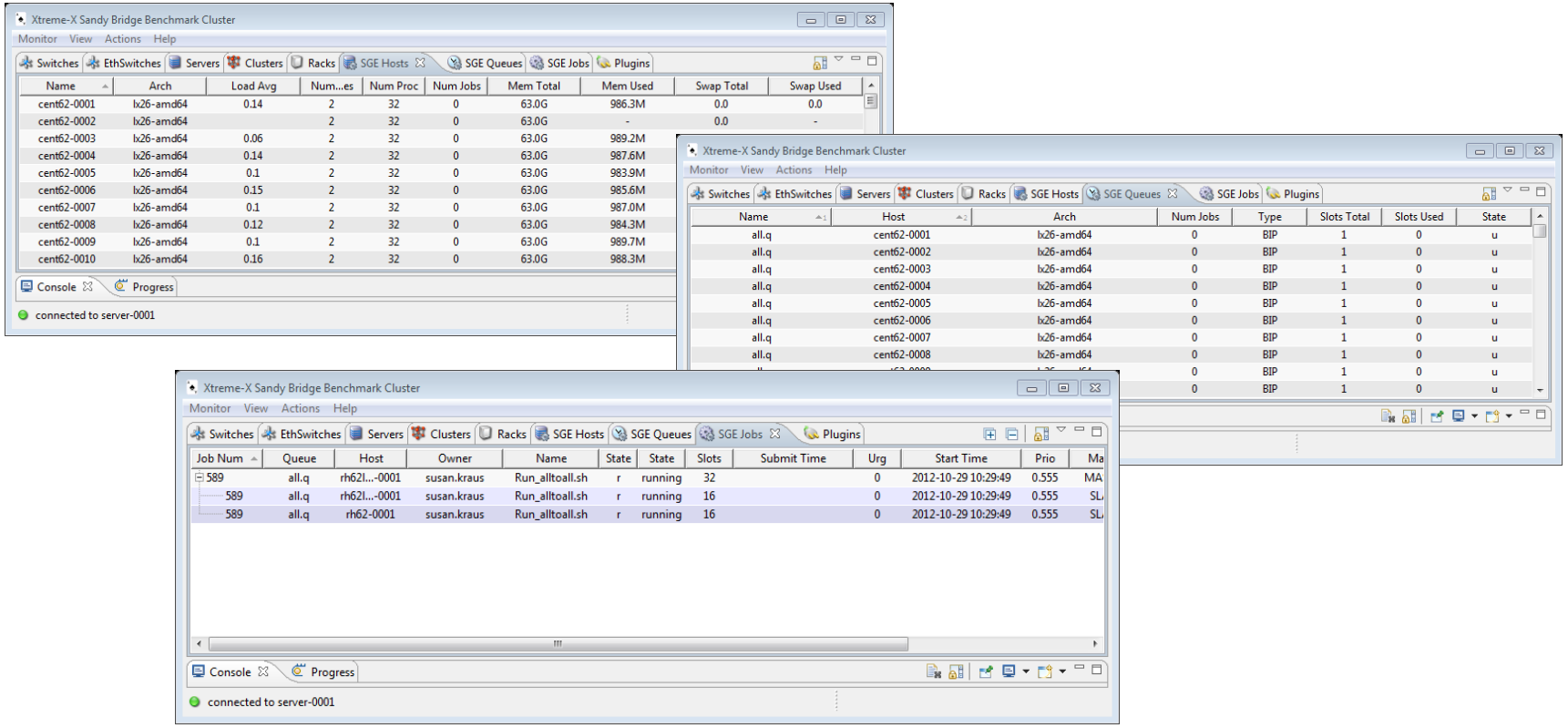
2-D and 3-D Torus
- Single and dual rail
- Distributed IO

Cray ACE™ Software Management
Job Management

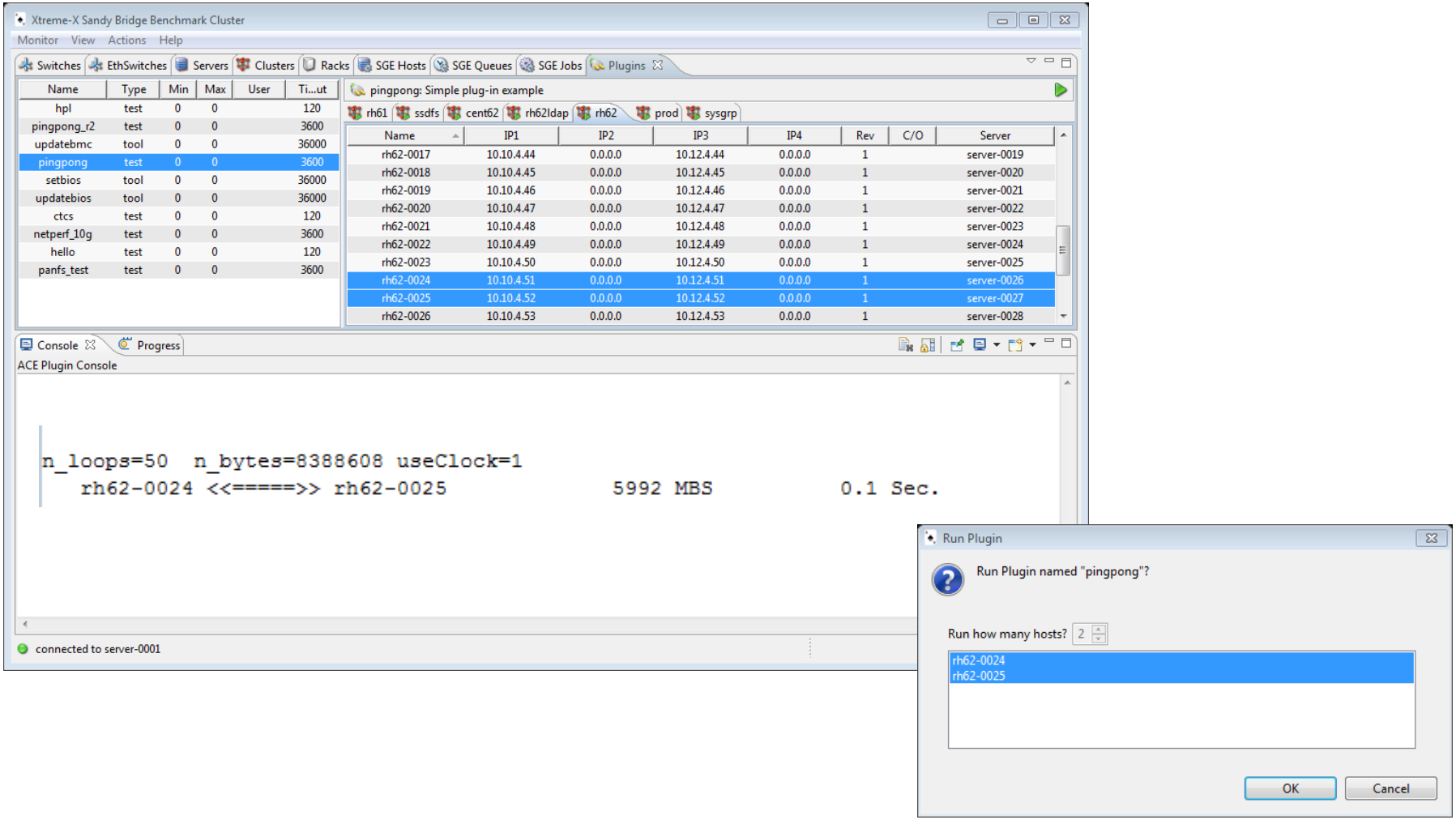
SLURM

GRID ENGINE

- Queuing System Configuration
- Submitting Jobs
- Monitor Jobs



Cray ACE™ Software Management Plugins



Cray ACE™ Software Management Demo

ACE
Advanced Cluster Engine

An ACE cluster is available for remote or on-site testing at Cray Cluster Solution's Advanced Computing Center in The Woodlands, TX

ACE Devel Cluster

Monitor View Actions Help

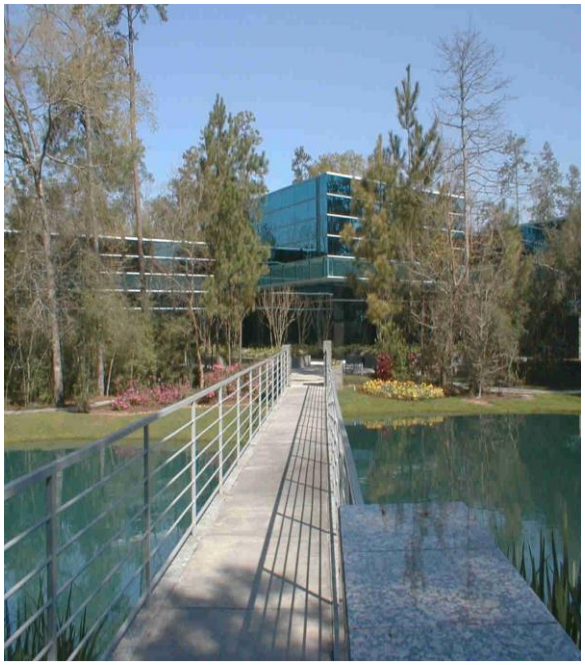
Switches | EthSwitches | Servers | Clusters | Racks | SGE Hosts | SGE Queues | SGE Jobs | Plugins

Name	Type	L1	N	Grp	Rack	Slot/U#	Blade	Host	Net1	Net2	State	BIOS	HCA
server-0001	management	1	1	1	1	1	1	phaeton1			active	SE5C600....9/2011)	2.9.1000
server-0002	management	1	2	1	1	1	2	phaeton2			unknown		
server-0003	compute	1	3	1	7	1	1	prod-0001			active	SE5C600....2/2011)	2.11.500
server-0004	compute	1	4	1	7	1	2	prod-0002			active	SE5C600....2/2011)	2.11.500
server-0005	compute	1	5	1	7	2	3	prod-0003			active	SE5C600....2/2011)	2.11.500
server-0006	compute	1	6	1	7	2	4	prod-0004			active	SE5C600....2/2011)	2.11.500
server-0007	compute	1	7	1	7	2	5	prod-0005			off	SE5C600....8/2012)	
server-0008	compute	1	8	1	7	1		abc-0001			active	SE5C600....1/2012)	2.11.500
server-0009	compute	1	9	1	7	3		abc-0002			active	SE5C600....5/2012)	
server-0010	compute	1	10	1	1	2	6				unknown		

Console | Progress

connected to server-0001

Cray Cluster Solutions Division
Advanced Computing Center



The ACC supports compute resources for software development, benchmarking, testing, and training

ACC
The Woodlands, TX

Cray Cluster Software Stack

Thank You
Questions?

THE SUPERCOMPUTER COMPANY