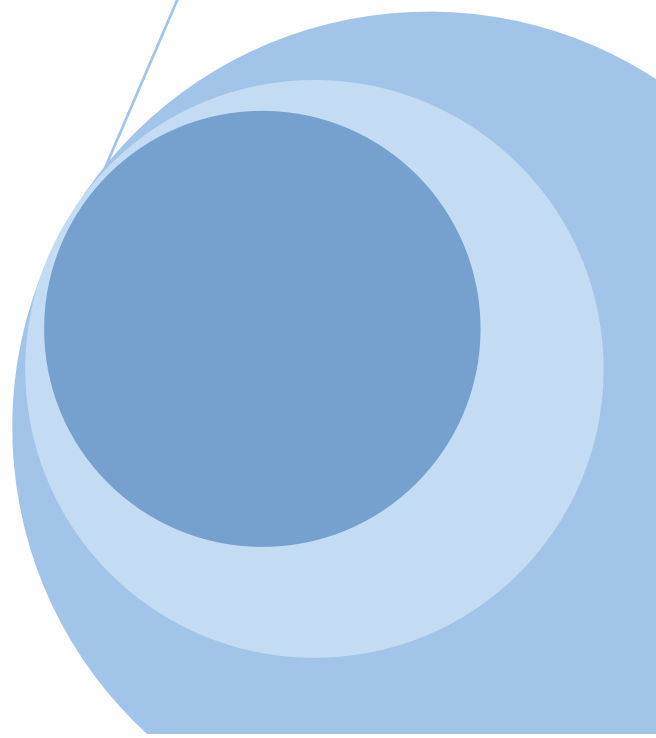


Sonexion™ Administrator's Guide

Software Version 1.3.1

S-2537-131b



© 2014 Cray Inc. All Rights Reserved. This manual or parts thereof may not be reproduced in any form unless permitted by contract or by written permission of Cray Inc.

U.S. GOVERNMENT RESTRICTED RIGHTS NOTICE

The Computer Software is delivered as “Commercial Computer Software” as defined in DFARS 48 CFR 252.227-7014. All Computer Software and Computer Software Documentation acquired by or for the U.S. Government is provided with Restricted Rights. Use, duplication or disclosure by the U.S. Government is subject to the restrictions described in FAR 48 CFR 52.227-14 or DFARS 48 CFR 252.227-7014, as applicable. Technical Data acquired by or for the U.S. Government, if any, is provided with Limited Rights. Use, duplication or disclosure by the U.S. Government is subject to the restrictions described in FAR 48 CFR 52.227-14 or DFARS 48 CFR 252.227-7013, as applicable.

TRADEMARKS

Cray and Sonexion are federally registered trademarks and Active Manager, Cascade, Cray; Apprentice2, Cray; Apprentice2; Desktop, Cray; C++; Compiling; System, Cray; CX, Cray; CX1, Cray; CX1-iWS, Cray; CX1-LC, Cray; CX1000, Cray; CX1000-C, Cray; CX1000-G, Cray; CX1000-S, Cray; CX1000-SC, Cray; CX1000-SM, Cray; CX1000-HN, Cray; Fortran; Compiler, Cray; Linux; Environment, Cray; SHMEM, Cray; X1, Cray; X1E, Cray; X2, Cray; XD1, Cray; XE, Cray; XEm, Cray; XE5, Cray; XE5m, Cray; XE6, Cray; XE6m, Cray; XK6, Cray; XK6m, Cray; XMT, Cray; XR1, Cray; XT, Cray; XTm, Cray; XT3, Cray; XT4, Cray; XT5, Cray; XT5, Cray; XT5m, Cray; XT6, Cray; XT6m, CrayDoc, CrayPort, CRInform, ECOPhlex, LibSci, NodeKARE, RapidArray, The Way to Better Science, Threadstorm, uRiKA, UNICOS/lc, and YarcData are trademarks of Cray Inc.

AMD and Opteron are registered trademarks of Advanced Micro Devices, Inc. Xilinx is a registered trademark of Xilinx, Inc. Linux is a registered trademark of Linus Torvalds. All other trademarks are the property of their respective owners.

All other trademarks are the property of their respective owners.

Direct requests for copies of publications to:

Mail: Cray Inc. Logistics
PO Box 6000
Chippewa Falls, WI 54729-0080
USA

E-mail: spares@cray.com

Fax: +1 715 726 4602

Direct comments about this publication to:

Mail: Cray Inc.
Technical Training and Documentation
P.O. Box 6000
Chippewa Falls, WI 54729-0080
USA

E-mail: ttd_online@cray.com

Fax: +1 715 726 49

Record of Revision

Publication Number	Description
HR5-6093-0	August 2012 Original Printing
HR5-6093-A	September 2012 Revises Appendix A, "CSSM CLI User Documentation" to reflect changes in CSSM 1.2.
HR5-6093-B	April 2013 Revises Appendix A, "CSSM CLI User Documentation" to reflect changes in CSSM 1.2.1.
S-2537-131	March 2014 Updates CSSM and CSCLI description for release 1.3.1. Adds upgrading RPMs. Publication number changed to indicate software document.
S-2537-131a	April 2014 Most of section 3 removed, referring to Field Installation Guide. LNET configuration revised.
S-2537-131b	May 2014 Revised procedure for LNET configuration

Contents

1. Introducing Sonexion.....	11
1.1 Software Architecture	11
1.1.1 Cray Sonexion System Manager (CSSM)	12
1.1.2 Data Protection Layer (RAID).....	12
1.1.3 Unified System Management (USM) firmware.....	12
1.2 Hardware Architecture	13
1.2.1 Metadata Management Unit	14
1.2.2 Scalable Storage Unit	15
1.2.3 Network fabric switches	15
1.2.4 Management switches.....	15
2. What's New?	16
2.1 New software features	16
2.1.1 Enhanced high availability services.....	16
2.1.2 Improved user interface	16
2.1.3 CSSM and MGMT server scalability improvements.....	17
2.1.4 Improved and enhanced health monitoring via Icinga.....	17
2.1.5 Change login password from CSSM GUI	17
2.1.6 Set thresholds for monitoring alerts.....	17
2.1.7 Full high-availability for MGMT nodes	18
2.1.8 Command-line interface and GUI upgrades	18
2.1.9 Support bundle improvements	18
2.1.10 Improved error reporting, progress logging with added SSUs	18
2.1.11 Upgrading to 1.3.1 disables SSDs on MMU	18
2.2 New hardware features.....	19

2.2.1 SSU +1 configuration	19
2.2.2 Expanded Storage Unit (ESU).....	19
2.3 Summary of CLI changes in this release.....	19
2.3.1 New options on alerts_config command	20
3. Custom LNET configuration	21
4. Exploring CSSM	24
4.1 CSSM basics	24
4.1.1 CSSM layout	24
4.2 Node control tab	27
4.2.1 Node filter.....	28
4.2.2 Node interactive.....	28
4.2.3 Status and conditions icons.....	29
4.2.4 All nodes in filter commands.....	30
4.2.5 Selected node commands.....	32
4.3 Performance tab	33
4.3.1 MDT performance data	34
4.3.2 OST performance data.....	34
4.3.3 OST overview plot and controls	35
4.4 Log browser tab.....	37
4.4.1 Filter	38
4.4.2 Logs.....	38
4.5 Support tab.....	39
4.6 Dashboard tab	39
4.6.1 Tactical monitoring overview.....	40
4.6.2 Current network status.....	40
4.6.3 Availability	41
4.6.4 Usage.....	41
4.7 Terminal tab.....	42
4.8 Health tab	42
4.8.1 Header	43
4.8.2 Link groupings.....	44
4.8.3 Status group: Tactical monitoring overview.....	44
4.8.4 Reporting group: Alert Summary	45
4.8.5 State types.....	45
4.9 Configure tab	49
4.9.1 Network.....	50
4.9.2 Storage.....	52
5. Operations	54
5.1 Managing Nodes.....	54

5.1.1 Starting Lustre	54
5.1.2 Stopping Lustre	55
5.1.3 Mounting Lustre	56
5.2 Managing a failover	57
5.2.1 Hand off resources to the HA partner OSS.....	57
5.3 Managing failback.....	58
5.3.1 Command line failback.....	58
5.3.2 Use the CSSM GUI	59
5.4 Using node filters	60
5.4.1 Creating a custom filter	60
5.4.2 Deleting a custom filter	61
5.5 Configuring rebuild rates.....	61
6. Monitoring.....	63
6.1 Viewing the host status	63
6.1.1 Commands	64
6.1.2 Extra host actions	65
6.1.3 Exporting	66
6.1.4 Flap handling	67
6.2 Viewing service status	68
6.2.1 Commands.....	68
6.2.2 Extra service actions.....	69
6.3 Viewing log data	71
6.4 Viewing performance data	71
6.4.1 Viewing MDT performance data.....	72
6.4.2 Viewing OST performance data	72
6.4.3 Customizing the OST performance view.....	74
6.4.4 File preferences	76
6.5 Managing a service issue	77
6.6 Generating host reports	79
6.6.1 Availability reports.....	80
6.6.2 Performance data reports	82
6.6.3 Trend reports	83
6.7 Understanding thresholds	87
6.7.1 Thresholds for general-purpose services	87
7. Support Files	96
7.1 Obtain a support bundle via the CSSM GUI.....	97
7.2 Download a support bundle via CSSM GUI	98
7.3 Get a support bundle via cscli	99

7.4 Reference for cscli support_bundle command	99
7.5 Interpret Sonexion support bundles	100
7.5.1 System-wide logs.....	100
7.6 Node-specific Logs:.....	101
8. Troubleshooting	102
8.1 Lustre performance considerations and tuning	102
8.1.1 Prerequisites	102
8.1.2 Hardware performance testing.....	102
8.1.3 Benchmarking interconnect	103
8.1.4 Benchmarking - RAID tuning	103
8.1.5 Benchmarking - direct MDT and OST testing.....	104
8.1.6 Benchmarking - single client testing	104
8.1.7 Benchmarking - multi-client testing	104
8.2 Management software issues	105
8.2.1 Warning while unmounting Lustre: “Database assertion: created a new connection but pool_size is already reached”	105
8.2.2 Invalid puppet certificate on diskless node boot-up	105
8.2.3 Need to change LDAP settings after GUI/wizard is complete	107
8.2.4 Unclean shutdown of management node causes database corruption	107
8.2.5 Many nodes flapping	108
8.3 Networking issues.....	108
8.3.1 Recovering from a top-of-rack (TOR) Ethernet switch failure.....	108
8.3.2 Reseating a problematic high-speed network cable.....	109
8.4 RAID/HA issues.....	110
8.4.1 RAIDs are not assembled correctly on the nodes	110
8.4.2 Starting Lustre on a given node without mounting the fsys resource	119
8.4.3 Lost one OST during forced failover. Release 1.2.0.....	120
8.4.4 Multiple HDDs spontaneously drop out of RAID arrays (heap overflows) 121	
8.4.5 MD device fails to assemble with the message: mdadm: cannot reread metadata from /dev/disk/by-id/WWN - aborting	123
8.4.6 MD device fails to assemble with the message: “mdadm: cannot reread metadata from /dev/disk/by-id/WWN - aborting.”	125
8.5 Other Issues	126
8.5.1 Nodes are shown in “unknown” state in GUI.....	126
8.5.2 SSUs failed after AC power loss	128
8.5.3 CS-1600 OSS will not power up, BMC out of memory	129
8.5.4 No response when attempting to physically connect to the serial port	129
8.5.5 OSS/MDS nodes go down during FS testing.....	130
8.5.6 Non-responsive server	130
9. Upgrading Lustre RPMs.....	131
9.1 RPM upgrade with Lustre inactive	131

9.2 Lustre live upgrade	134
9.2.1 Requirements	134
9.2.2 Installing HotFix RPMs	134
9.2.3 Using HA to Upgrade MDS, MGS, and OSS	135
10. CSSM CLI User Documentation.....	146
10.1.1 Customer Wizard mode	146
10.1.2 Daily mode	146
10.1.3 CLI command summary	147
10.2 Summary of changes in release CS 1.3.1	149
10.3 Network setup commands	150
10.3.1 Show network parameters	150
10.3.2 Set network parameters	151
10.3.3 Reset network parameters.....	151
10.3.4 Apply network parameters.....	152
10.4 User setup commands	152
10.4.1 Get the file system's AD settings	152
10.4.2 Get the file system's LDAP settings	152
10.4.3 Get the file system's NIS settings	153
10.4.4 Set the file system's AD settings.....	153
10.4.5 Set the file system's LDAP settings	154
10.4.6 Configure the file system's NIS Settings	155
10.4.7 Clear the file system's AD settings	155
10.4.8 Clear the file system's LDAP settings	155
10.4.9 Clear the file system's NIS settings	156
10.5 System alert commands	156
10.5.1 Display current and historic system alerts	156
10.5.2 Manage the alerts configuration	160
10.5.3 Manage the alerts notification	166
10.6 Node control commands	168
10.6.1 Manage node auto-discovery.....	168
10.6.2 Manage node failback and failover.....	168
10.6.3 Mount and unmount Lustre targets.....	169
10.6.4 Manage node power	169
10.6.5 Show node information	170
10.7 Administrative commands	171
10.7.1 Show file system information	171
10.7.2 Retrieve FRU information	171
10.7.3 Change the Sonexion mode	172
10.7.4 List commands.....	173
10.7.5 Display log information.....	173
10.7.6 Set administrator password.....	173
10.7.7 Batching commands	174
10.7.8 Manage IP routing	174

10.8 Configuration commands	175
10.8.1 Configure hosts.....	175
10.8.2 Configure new OSS nodes.....	176
10.8.3 Show information about new OSS nodes	176
10.9 Filter commands	177
10.9.1 Create a filter	177
10.9.2 Show filters.....	177
10.9.3 Delete a filter	178
10.10 Updating system software.....	178
10.10.1 Prepare a software update.....	178
10.10.2 Update software on a system node	179
10.10.3 Split HA partners	179
10.10.4 Show node versions	180
10.10.5 Showing available software versions.....	180
10.11 Managing node position in a Sonexion rack.....	181
10.11.1 Get node position in a Sonexion rack	181
10.11.2 Set node position in a Sonexion rack.....	181
10.11.3 Monitor system health	182
10.11.4 Manage the netfilter level	183
10.11.5 Enable RAID checks	184
10.11.6 Manage the RAID rebuild rate	184
10.11.7 Manage the administrative password.....	185
10.11.8 Manage the system date.....	185
10.11.9 Manage the system timezone.....	186
10.11.10 Manage the InfiniBand Subnet Manager	186
10.11.11 Manage support bundles.....	187
11. GEM CLI Commands	189
11.1 Serial port settings	189
11.2 Supported number bases.....	190
11.3 Supported commands.....	190
11.3.1 ddump.....	190
11.3.2 getboardid.....	191
11.3.3 getmetisstatus	191
11.3.4 getvpd	191
11.3.5 help	192
11.3.6 ipmi_power.....	192
11.3.7 ipmi_setosboot.....	193
11.3.8 logdump.....	193
11.3.9 report_faults.....	194
11.3.10 settime	194
11.3.11 ver.....	195

Tables

Table 1. Customer Wizard Mode - CLI Commands	19
Table 2. Daily Mode - CLI Commands	20
Table 3. Status and Conditions Icons	29
Table 4. Colors for Host and Service Status	41
Table 5. CLI Command Summary	147
Table 6. New CLI Commands, Customer Wizard Mode	149
Table 7. New CLI Commands, Daily Mode	150

1. Introducing Sonexion

This Administration Guide provides step-by-step instructions on how to set up, use, and troubleshoot the Cray Sonexion storage system. This manual is intended for Site Service Providers who maintain Cray Sonexion storage systems.

1.1 Software Architecture

Sonexion software architecture consists of an integrated, multi-layer software stack:

- Cray Sonexion System Manager (CSSM)
- Lustre file system
- Data protection layer (Redundant Array of Independent Disks, RAID)
- Unified System Management (USM)
- Linux OS

Sonexion runs Lustre 2.1.0 software in a standard Linux environment (Scientific Linux 6.2 operating system). The file system is fully integrated with CSSM, USM, and RAID layers in the stack.



1.1.1 Cray Sonexion System Manager (CSSM)

Cray Sonexion System Manager (CSSM) provides a single-pane-of-glass view of the Sonexion solution infrastructure. It includes a browser-based GUI that simplifies cluster installation and configuration, and provides consolidated management and control of the entire storage cluster. CSSM also provides distributed component services to manage and monitor system hardware and software.

CSSM includes wizards to guide you through configuration and node provisioning. Once the cluster is running, you can use the GUI to manage the storage environment with these functions:

- Start and stop file systems
- Manage node failover
- Monitor node status
- Collect and browse performance data

The dashboard reports errors and warnings for the storage cluster and provides tools to aid in troubleshooting, including cluster-wide statistics, system snapshots, and Lustre syslog data.

To maximize availability, CSSM works with USM, the platform's integrated management software, to provide comprehensive system health monitoring, error logging, and fault diagnosis. Users are alerted to changing system conditions and degraded or failed components.

1.1.2 Data Protection Layer (RAID)

Sonexion uses Redundant Array of Independent Disks (RAID) to provide different data protection layers throughout the solution. The RAID subsystem configures each Object Storage Target (OST) with a single RAID 6 array to protect against double disk failures and drive failure during rebuilds. The 8 + 2 RAID sets support hot spares so that when a disk fails, its data is immediately rebuilt on a spare disk and the system does not need to wait for the disks to be replaced.

Sonexion uses write intent bitmaps (WIB) to aid the recovery of RAID parity data in the event of a downed server module or a power failure. For certain types of failures, using WIBs substantially reduces parity recovery time from hours to seconds. In Sonexion OSTs, WIBs are stored on solid-state drives (SSDs), mirrored for redundancy, enabling fast recovery from power and OSS failures without a significant performance impact.

1.1.3 Unified System Management (USM) firmware

Extensive Sonexion system diagnostics are managed by USM management firmware, which runs on each OSS in the scaleable storage unit (SSU). USM monitors and controls the SSU's hardware infrastructure and overall system environmental conditions, providing a range of services including SES and high-availability (HA) capabilities for system hardware and software. USM offers these key features:

- Manages system health, providing random-access services (RAS) that cover all major components such as disks, fans, power-supply units (PSUs), Serial Attached SCSI (SAS) fabrics, PCI (Peripheral Component Interconnect) buses, memories, and CPUs, and provides alerts, logging, diagnostics, and recovery mechanisms
- Power control of hardware subsystems that can be used to individually power-cycle major subsystems and provide additional RAS capabilities – this includes drives, servers, and enclosures
- Fault-tolerant firmware upgrade management
- Monitoring of fans, thermals, power consumption, voltage levels, AC inputs, field-replaceable unit (FRU) presence, and health
- Efficient adaptive cooling keeps the SSU in optimal thermal condition, using as little energy as possible
- Extensive event capture and logging mechanisms to support file system failover capabilities and to allow for post-failure analysis of all major hardware components

1.2 Hardware Architecture

The Sonexion CS-1600 hardware architecture consists of a pre-configured, rack-level storage cluster that can be easily expanded using modular storage node building blocks. The principal hardware components include:

- Metadata Management Unit (MMU)
- Scalable Storage Unit (SSU)
- Network Fabric Switches
- Management Switch

The Sonexion solution is differentiated from other file system solutions by its innovative MMU and SSU architectures.



1.2.1 Metadata Management Unit

The Metadata Management Unit (MMU) is a quad-node server which contains the two management (MGMT) nodes, the MGS node, the MDS node, and one shelf of high-availability shared storage. The central point of management for the entire storage cluster, the MMU runs the Sonexion Manager software, manages network request handling, and monitors every storage element within the cluster. Sonexion interface ports support InfiniBand fabric network interface technology connections and 1GbE management network connections.

The MMU is fully redundant and fault-tolerant. Each node is configured for active-passive failover, with an active instance of the server running on one system and a passive instance of the node running on the peer system. If an active node fails, for example, the MDS goes down, then the passive MDS node takes over the MDT operations of the failed MDS. The shared storage of the MMU supports a combination of Small Form Factor (SFF) SAS HDD, protected using RAID 1, for management data, file system data, and journal acceleration.

Sonexion supports InfiniBand connections to the MGMT, MDS, and MGS nodes. Additionally, each server connects, via Ethernet, to dedicated private management networks supporting Intelligent Platform Management Interface (IPMI).

1.2.2 Scalable Storage Unit

The core building block is the Scalable Storage Unit (SSU). Each SSU is configured with identical hardware and software components, and hosts two OSS nodes. The SSU contains two OSSs, with RAID-protected, high availability shared storage, and interface ports to support InfiniBand data networks and 1GbE management network connections.

The OSSs are Storage Bridge Bay (SBB) compliant with SAS expanders that enable both modules to directly access all drives in the enclosure (a differentiator among fully fault-tolerant systems). The OSSs connect through a common midplane, eliminating the need for extra external cables, and share a redundant, high-speed interconnect across the midplane for failover services. This efficient and highly reliable design enables the SSU's SAS infrastructure to deliver robust performance and throughput – 2.5 GB/sec per SSU for reads and writes.

The SSU is fully redundant and fault-tolerant, thus ensuring maximum data availability. Each OSS serves as a Lustre node, accessing the disk as shared OST storage and providing active-active failover. If one OSS fails, the active module manages the OSTs and the disk operations of the failed node. In non-failure mode, the I/O load is balanced between modules. The SSU's shared storage consists of high-capacity SAS disk drives, configured in a RAID 6 array to protect against double disk failures and drive failure during rebuilds.

1.2.3 Network fabric switches

The Network Fabric switches (InfiniBand) manage I/O traffic and provide network redundancy throughout the Sonexion solution. To maximize network reliability, the OSSs in the SSU are connected to redundant network switches. If one switch fails, the second module in the SSU (connected to the active switch) manages the OSTs of the module connected to the failed switch.

To maintain continuous management connectivity within the solution, the network switches are fully redundant at every point and interconnected to provide local access from the MGMT, MDS, and MGS nodes to all storage nodes.

1.2.4 Management switches

The Management switches consists of a dedicated local network on dual 1GbE switches that is used for configuration management and health monitoring of all components in the Sonexion solution. The management network is private and not used for data I/O in the cluster. This network is also used for IPMI traffic to the SSU's OSSs, enabling them to be power-cycled by CSSM.

2. What's New?

The following is a list of new features and improvements for this release of CSSM version 1.3.1.

2.1 New software features

2.1.1 Enhanced high availability services

An improved High Availability (HA) service is integrated with the core OSS platform. Sonexion HA service ensures continuity of management and data services in the event of a single server failure. This lets an enclosure hand off (fail over) resources from a failed server to a paired backup server.

The major improvement for this release concerns the two management servers. Before 1.3.1, server 0 stored system info on a local drive, but node n001 could not access that data. Now MGMT storage is on the 2U24, enabling node n001 to continue operation if node n000 is disabled. The MGMT servers provide boot services plus the CSSM application and associated web server and database.

2.1.2 Improved user interface

Node selection in the **Node Control** tab (daily mode screen) has been made easier by enabling selecting of an individual node by clicking on it, and to select multiple nodes continue selecting the desired nodes in the list. Also, the menu just above the **Hostname** column provides a quick means to select all, none, groups of HA, or by the column types.

2.1.3 CSSM and MGMT server scalability improvements

Significant improvements have been made to CSSM to optimize and decrease the time required to load the **Node Control** tab on cluster configurations. When you 'hover' the cursor over node groups or partners in the **Role** column, nodes are now labeled "primary" or "secondary," where previously they had been labeled "odd" and "even."

A new button on the **Node Control** tab has been introduced to enable quick selection of nodes. Provided with this feature is the ability to filter nodes by their type, power state, Lustre state and HA resources.

2.1.4 Improved and enhanced health monitoring via Icinga

An improved disk summary has been added to **Arrays and Disk Status**, providing the total number of disk slots available and total number of disks found. (This information appears on the **Health** tab > **Status** > **Servicegroup overview**. On that screen, any **View** item in the itemized list must apply to **All Servicegroups**. Click the applicable line to toggle this choice.)

Summary statuses of lower level statistics are collected in a single check, thus reducing the overall number of checks across the system. This improves the ability to navigate and understand the overall status.

Nodes and enclosure components are grouped together in the **Health > Status > Host Detail** screen, using the third and fourth **View** options (Status Overview For All Host Groups and Status Overview For All Host Groups). The enclosure now has a rack label and the location within the rack.

With this release, metrics are now available for fan, thermal, power and voltage readings. This provides essential information for long term periodic historical data and the ability to see the information displayed in graphs.

Service State Information displays collated array and disk status data in a single place. Drive power usage and total power usage for each SSU are now available, and new thermal statistics indicate the physical location of the thermal sensors, providing greater ability to create a pictorial view of the racks and enclosures within thermal zones.

2.1.5 Change login password from CSSM GUI

The CSSM GUI now provides the ability to easily change the administrator password using a simple point-and-click operation. This is described on page 26.

2.1.6 Set thresholds for monitoring alerts

New `cscli` commands have been added to provide support for setting alert thresholds.

2.1.7 Full high-availability for MGMT nodes

2.1.8 Command-line interface and GUI upgrades

- User-configurable minimum rebuild rate control for RAID arrays in the `cscli`
- User-configurable RAID check scheduling in the `cscli`
- NIS Authentication Functionality (`cscli` only)

NIS Authentication Support in `cscli` configures, shows, and clears the filesystem's NIS settings. The commands are:

<code>set_lustre_users_nis</code>	Configures Filesystem NIS settings.
<code>get_lustre_users_nis</code>	Shows configured NIS settings.
<code>clear_lustre_users_nis</code>	Clears the Lustre file system's NIS settings

- Email notifications, configured via the `cscli alerts_config` command.

2.1.9 Support bundle improvements

SysRq-"m" and SysRq-"t" output

Should a system crash occur, the memory contents and a list of current processes will be included in the support bundle. The magic SysRq is a key combination understood by the Linux kernel that allows the user to perform various low level commands regardless of the system's state. This feature is often used to recover from freezes, or to reboot a computer without corrupting the file system.

Collection of firmware version information

Support bundle now collects firmware and BIOS version information.

2.1.10 Improved error reporting, progress logging with added SSUs

Previously the system only reported that something went wrong but provided no details. These improvements for version 1.3.1 add better error reporting and improved progress logging while adding an OSS.

2.1.11 Upgrading to 1.3.1 disables SSDs on MMU

Previous upgrades preserved the enabled/disabled status of the SSDs for external WIBs and journals on the MMU. New installations and any upgrades to 1.3.1 replaces external WIBs and Journals with internal WIBs and journals for the MMU.

2.2 New hardware features

2.2.1 SSU +1 configuration

The SSU+1 configuration refers to an SSU with 1 ESU attached. The SSU+1 configuration consists of the following:

- One Scalable Storage Unit (SSU) and one ESU (Expansion Storage Unit)
- 8 OSTs configured on the SSU and divided between both OSSs
- 8 OSTs configured on the ESU and divided between both OSSs
- Each SSU and its companion ESU must be located adjacently in the same rack

2.2.2 Expanded Storage Unit (ESU)

This is an optional additional component which is offered as an alternative to one or more SSUs (Scalable Storage Unit). An ESU will have the same enclosure hardware as an SSU, including the drives, fans and power supplies, but will have different cabling and I/O controllers. The purpose of this component is to increase capacity by allowing an enclosure to be added to an SSU without Object Storage Servers.

Note: ESU support in v1.3.1 is limited to new builds only. Adding ESU's to an existing system will be supported in a subsequent release.

2.3 Summary of CLI changes in this release

The following commands were added to this release:

Table 1. Customer Wizard Mode - CLI Commands

No.	Addition/ Deletion	Command	Description	Component
1	Added	<code>clear_lustre_users_nis</code>	Clear Filesystem NIS settings	NIS Support
2	Added	<code>get_lustre_users_nis</code>	Show configured NIS settings	NIS Support
3	Added	<code>set_lustre_users_nis</code>	<code>set_lustre_users_nis</code>	NIS Support
4	Added	<code>support_bundle</code>	Manage support bundles and support bundle settings	Support Bundles

Table 2. Daily Mode - CLI Commands

No	Addition/ Deletion	CLI Command	Description	Component
1	Added	<code>configure_hosts</code>	Configure hostname for discovered node.	SSU Addition
2	Added	<code>raid_check</code>	Enable RAID checks on RAID devices	XYRAID
3	Added	<code>support_bundle</code>	Manage support bundles and support bundle settings	Support Bundles

2.3.1 New options on `alerts_config` command

The following options were added to the `alerts_config` command. See “Manage the alerts configuration”, page 160:

- `email_off`
- `thresholds`
- `email_update`
- `email_server_update`
- `email_delete`
- `email_add`
- `email_on`
- `email_server`
- `emails`

3. Custom LNET configuration

Use the procedure in this chapter to configure a custom LNET configuration on the Sonexion system while in “daily mode” (see 10.1.2 Daily mode, page 146).

For a new system, first follow the setup procedures described in Cray publication HR5-6124, *Sonexion Field Installation Guide*. Then execute the following installation.

1. Log in to the primary management mode.

2. Change to root:

```
$ sudo su -
```

3. Stop the Lustre file system by running the command:

```
# cscli unmount -f file_system_name
```

4. For version 1.2:

a. Log into the MGS node via SSH.

b. If you do not know the MGS group, run this command:

```
crm_mon -l -r
```

The group with **md65** in its name is the MGS group.

c. Stop the MGS service, by entering:

```
# stop_xyraid mgs_group
```

d. Log out of the MGS node and back into primary MGMT server node.

5. To change the **o2ib** index follow the steps below:

a. Start the MySQL client and connect to the **t0db** database by entering:

```
# mysql t0db
```

- b. Display the **mgsNID**, **nidFormat**, and **nidIndex** entries by entering:

```
mysql> select * from property where name in ('nidFormat',
      'mgsNID', 'nidIndex');
```

This is a sample output for version 1.3.1:

```
mysql> select * from property where name in ('nidFormat', 'mgsNID',
      'nidIndex');
```

id	context	name	value	attr_type
22	snx11033n:beConfig	nidFormat	l%s@o2ib%d	str
106	snx11033n:beConfig	nidIndex	3	int
109	snx11033n:beConfig	mgsNID	lsnx11033n002@o2ib3	str

- c. Set the **o2ib** index by modifying the **nidIndex** entry by specifying:

```
mysql> update property set value=desired_odib_index where
      name='nidIndex';
```

Example:

```
mysql> update property set value=2 where name='nidIndex';
Query OK, 1 row affected (0.02 sec)
Rows matched: 1 Changed: 1 Warnings: 0
```

- d. Set the **mgsNID** entry to match the **o2ib** index by entering:

```
update property set value='original_value@o2ibdesired_o2ib_index'
      where name='mgsNID';
```

Sample output:

```
mysql> update property set value='lsnx11033n002@o2ib2'
      where name='mgsNID';
Query OK, 1 row affected (0.04 sec)
Rows matched: 1 Changed: 1 Warnings: 0
```

- e. Verify the changes by repeating step 5b.

Sample output:

```
mysql> select * from property where name in ('nidFormat', 'mgsNID',
      'nidIndex');
```

id	context	name	value	attr_type
22	snx11033n:beConfig	nidFormat	l%s@o2ib%d	str
106	snx11033n:beConfig	nidIndex	2	int
109	snx11033n:beConfig	mgsNID	lsnx11033n002@o2ib2	str

3 rows in set (0.00 sec)

- f. Close the MySQL session by specifying:

```
mysql> quit
```

- g. Run puppet:

```
/opt/xyratex/bin/beUpdatePuppet -sa
```

6. Run the script on the primary management node and wait for it to finish.

```
# /opt/xyratex/bin/beSystemNetConfig.sh -c file_location/lnet.conf  
-r file_location/routes.conf -i file_location/ip2nets.conf clustername
```

7. Verify that the customized LNET configuration has been applied.

- a. List the node NIDs by entering:

```
# pdsh -g lustre lctl list_nids | sort
```

- b. List the nodes and targets, by entering:

```
# cscli fs_info
```

8. For version 1.2, run:

```
# pdsh -g lustre mdraid-deactivate  
# pdsh -g lustre manage_all_xyraid
```

9. Start the Lustre file system by entering:

```
# cscli mount -f file_system_name
```

Wait for the targets to mount on all system nodes.

This completes the procedure.

4. Exploring CSSM

4.1 CSSM basics

CSSM is a browser-based GUI that provides a consolidated management control and troubleshooting interface for the storage cluster. In addition, CSSM provides distributed component services to manage and monitor system hardware and software.

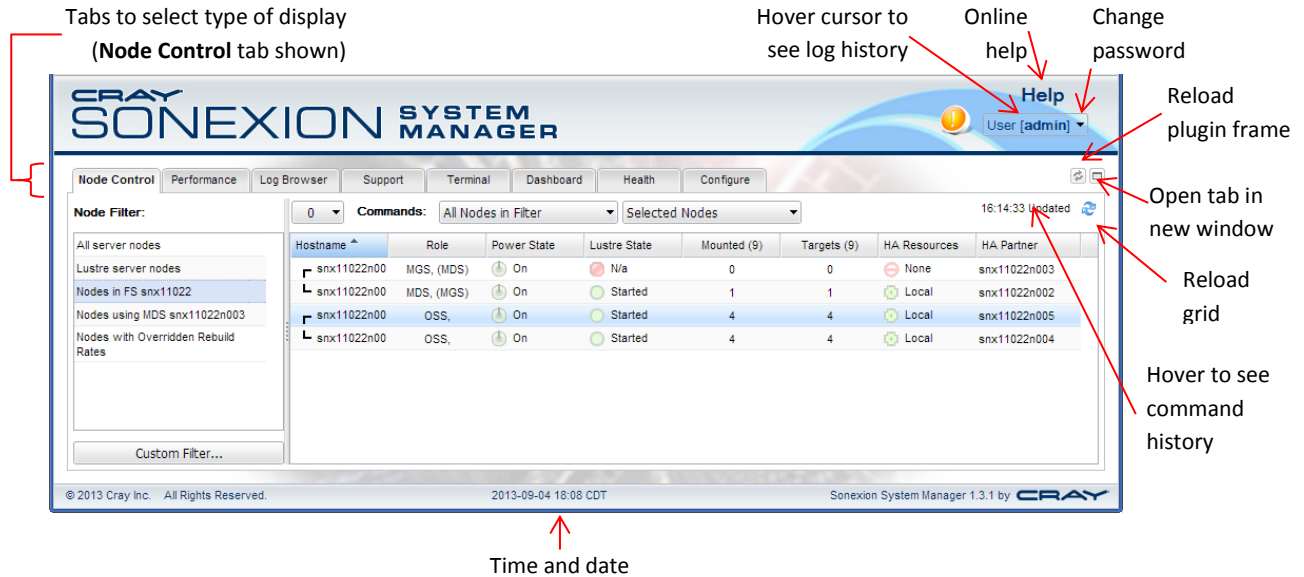
CSSM provides a wizard to help you configure and provision nodes and clusters in your system. Once the Sonexion environment is operational, use CSSM to administer the system.

- Start and stop file systems
- Manage node resources
- Monitor node status
- Collect and browse performance
- Enter terminal commands

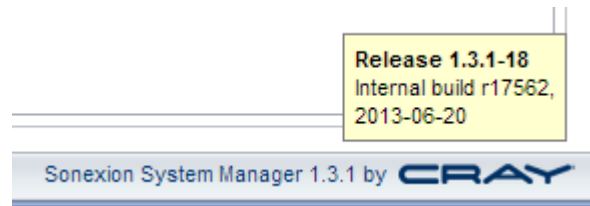
Additionally, CSSM reports errors and warnings for the storage cluster and provides aid in troubleshooting, including cluster-wide statistics, system snapshots and log files data. You are also alerted to changing system conditions and degraded or failed components.

4.1.1 CSSM layout

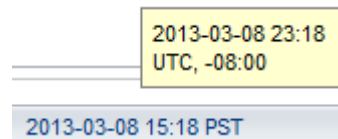
When CSSM is started, it defaults to the **Node Control** tab, also known as the *daily operation mode* screen. The tab bar displays a tab for each functional area of the interface. These screens are summarized below. The following screen shot points out features common to all screens. The **Node Control** tab is summarized on page 27.



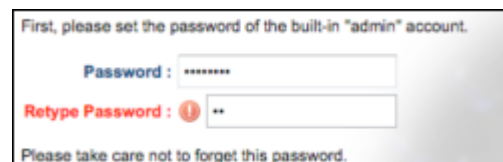
In the footer area are the following components: copyright information, time and software build information. Passing the mouse pointer over the version number pops up a display of build information, version number of the system release, software build number and date.



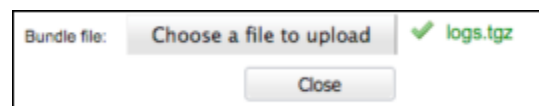
If you hover the cursor over the time display, it opens a pop-up text showing the universal coordinated time and the conversion factor. The time displayed by default is the time of the management server. If a **Z** is displayed for the conversion, this indicates that the local time zone is also the UTC time.



Red alert messages appear when various warnings and alerts occur.



A green check mark that appears would indicate a successful operation.



When two nodes have just been asynchronously updated, they are temporarily highlighted with a yellow background.

snx11022n00	MGS, (MDS)	On	N/a	0	0	None	snx11022n003
snx11022n00	MDS, (MGS)	On	Started	1	1	Local	snx11022n002

HA Group node rows are highlighted with either blue or green background for easier identification.

Alert icon

Located at the top of the screen near the Help link, is an alert icon. Shown as a circular icon with an exclamation point.

The following icon is displayed when a critical alert is detected by Icinga.



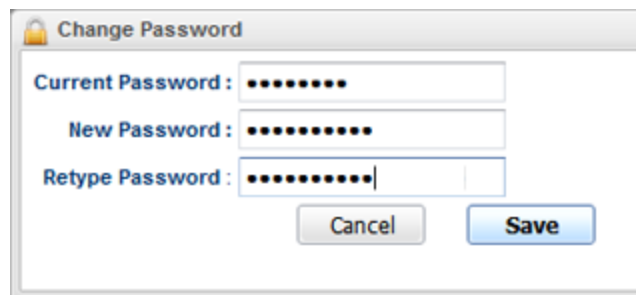
Changing the administrator password

Normally the Administrator password is established when the system is started the first time. Afterwards, the administrator has the option to change the password at any time.

1. Click the down arrow to the right of the logged-in user **User [admin]** and select **Change Password....**



2. In the dialog window, enter the **Current** password, followed by the **New** password.
3. Retype the **New** password and click the **Save** button.



You can cancel the operation at any time by clicking the **Cancel** button.

Open tab in new window

This feature allows the administrator to open tabs in new browser windows. You may prefer to have multiple windows open. For example, the **Node Control** tab in one window,

Health monitoring tab in another and, perhaps, the **Log Browser** and **Dashboard** tabs in other windows. This feature gives you a view of the entire system on one display.

You may perform this operation in one of two methods:

- Click on any tab, and once the tab's screen is displayed, right-click and select **Open Tab in New Window**.
- Click on any tab, and once the tab's screen is displayed, click on the **Open tab in new window** icon on the right hand side of the screen.

Adjust and size your browser windows to optimize your screen's display of the desired CSSM views.



Right-click any tab to view these

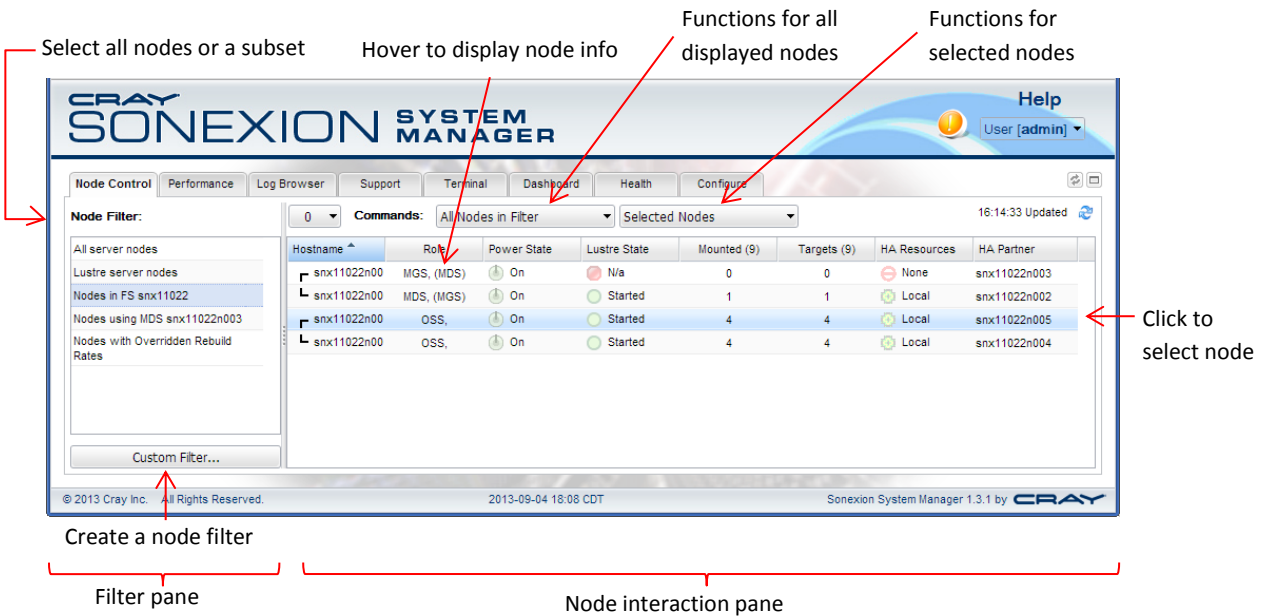
NOTE: The “Open tab in new window” feature is not applicable to the Configure tab.

4.2 Node control tab

The **Node Control** tab presents a high-level view of the Lustre file system and node status providing control of the nodes and their high availability partners and resources.

For more information on these **Node Control** tab commands and operations:

- To start and stop the file system, see “Managing Nodes”, page 54.
- To manually hand off resources, see “Managing a failover”, page 57.
- To perform the manual take-back or hand back resources, see “Managing failback”, page 58.
- To monitor Lustre, see Table 3, page 29.



In the left pane (**Filter Pane**), each file system or node is listed. Selecting a node, file system or filtered file system, will display the nodes or filtered nodes in the right pane (Node Interaction Pane), from there you can manage or monitor the nodes.

Commands that are relative to defined nodes are listed under the first dropdown menu button, **All Nodes in Filter**. An adjacent or second dropdown menu button, **Selected Nodes**, lists the commands relative to the selected node(s).

To select a node, simply point and click. The feature is similar to a toggle, the node is either selected (blue) or not. To make multiple selections, continue to click on that additional nodes. To de-select a node, simply point and click again on the node. The blue highlight will be removed once de-selected.

4.2.1 Node filter

The left side of the **Node Control** screen is the Node Filter list. You can select from all server nodes, Lustre nodes, or one of the pre-defined nodes or create your own using the "Custom Filter..." button located at the bottom of the left pane. For more information, see "Using node filters", page 60.

4.2.2 Node interactive

The right side of the screen is the Node Interactive pane which shows all nodes that match the filter selected in the left side Node Filter pane. This is where you will select specific nodes, perform actions on all nodes for the filter using the "All Nodes in Filter" menu. The following provides an explanation for each field.

- **Hostname** - Name assigned to the node.
- **Role** -The node's type: MGS, MDS, MGMT, OSS.
- **Power State** - The node's power state. Refer to the Status and Conditions Icons chart below for the possible power states.
- **Lustre State** - The node's Lustre state. Refer to the Status and Conditions Icons chart below for the possible Lustre states.
- **Mounted** - Number of mounted targets on the node. Parenthetical value is the total number of mounted targets across all nodes in the file system, excluding the MGS node.
- **Targets** - Number of targets for the node. Parenthetical value is the total number of targets across all nodes in the file system.
- **HA Resources** - Indicates whether HA resources (Lustre targets) are local to the node or assigned to its HA partner (paired node).






NOTE: In Sonexion usage, the term "HA resources" refers to Lustre targets (OSTs and MDTs). OSTs are managed by odd-/even-numbered pairs of OSS nodes. MDTs are managed by odd-/even-numbered pairs of MDS nodes. CSSM uses this even/odd numbering scheme to control the assignment of HA resources between paired nodes in "Hand Off" and "Take Back" operations (ensuring uninterrupted cluster operations and data redundancy).











- HA Partner - The OSS or MDS's partner node.

4.2.3 Status and conditions icons

The Power State, Lustre State and HA Resources fields in the 'Nodes in FS ...' filter view can display a status or condition icon shown in the table below.

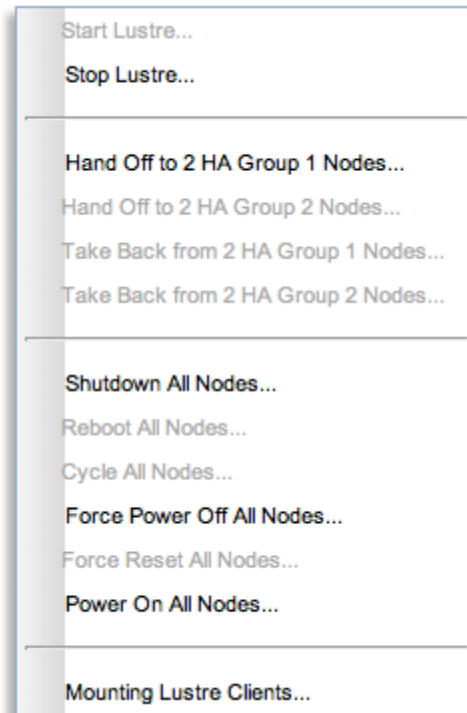
Table 3. Status and Conditions Icons

Icon and state	Icon indicates
 Power - OFF	The power state of the node is OFF.
 Power - ON	The power state of the node is ON.
 Power - Unknown	The power state of the node is unknown. This is an abnormal condition that could indicate a cabling problem with the SSU, or possibly an issue with the BMC.
 Lustre - Started	The node is servicing the Lustre requests and has started (the targets are mounted).
 Lustre - Stopped	The node has stopped servicing the Lustre requests and is stopped (the targets are unmounted).

Icon and state	Icon indicates
 Lustre - Stopped	The node has stopped servicing the Lustre requests and is stopped (the targets are unmounted) and is displayed when the node is powered off.
 Lustre - Partial	The node has partially started servicing the Lustre requests (some targets are mounted). This is a rare condition. In some cases, the targets missing from one node are mounted on the partner node (check the Targets column value for the HA partner node).
 Lustre - Error	There is an error with the node servicing the Lustre requests. This is a rare and abnormal condition. Check the cabling to the SSU containing the node. Contact technical support.
 HA Resource - Local	The node's resources are assigned locally to the respective node.
 HA Resource - All	The node has both its resources and the resources of its partner assigned to it.
 HA Resource - None	The node has no resources assigned to it.
 HA Resource - None	The node has no resources assigned and the node is in a power off state.
 HA Resource - Pending	The node's resources are in the process of moving to its partner node.
 HA Resource - Foreign	The node is controlling the partner node assigned resources, but not its own resources. This is a rare and abnormal condition.
 N/A	Not applicable as the MGS node serves all file systems. The Lustre state is not relevant.

4.2.4 All nodes in filter commands

This command menu lists global commands that act upon multiple (sometimes all) nodes in the file system.



NOTE: Any commands that are not available because of the file system state are grayed out.

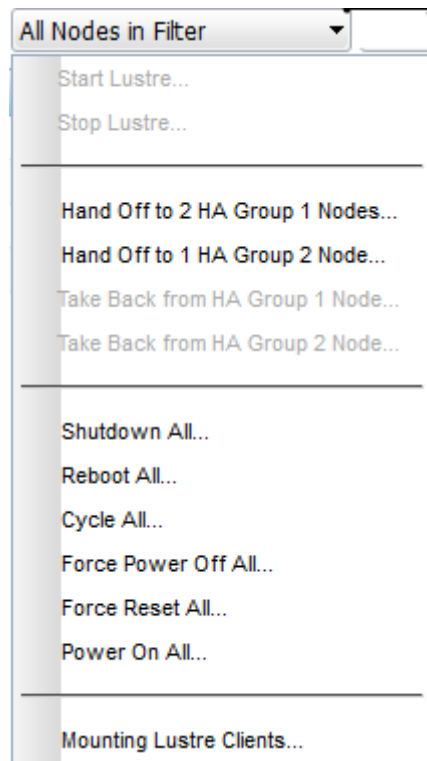
- **Start Lustre...** - Mounts all targets on the file system nodes and starts servicing Lustre clients.
- **Stop Lustre...** - Unmounts all targets on the file system nodes and stops servicing Lustre clients.
- **Hand Off to *HA_group_name*...** - Hands off (fails over) the assigned HA resources to its HA partner nodes.
- **Take Back from *HA_group_name*...** - Takes back (fails back) the assigned HA resources to its HA partner nodes.
- **Shutdown All Nodes...** - Initiates a graceful shutdown of the file system's nodes in an order that minimizes Lustre errors. When this command is run, a dialog box prompts you to confirm the node shutdown and indicate whether the Lustre management node should also be shut down (making the GUI unavailable).
- **Reboot All Nodes...** - Initiates a graceful reboot of the all nodes. When this command is run, a dialog box prompts you to confirm the nodes reboot. When the nodes reboot, their targets (MGT, MDT, and OSTs) goes offline temporarily.
- **Cycle All Nodes...** - Equivalent to turning the hardware power off to the nodes, followed by turning power back on immediately. This may cause data loss.
- **Force Power Off All Nodes...** - Forces the file system's nodes to shut down immediately, regardless of their state.

CAUTION: When this command is run, unsaved data in flight is lost.

- **Power On All Nodes** - Powers on the file system's nodes.
- **Mounting Lustre Clients** - When this item is selected, a popup window displays a mount command line that can be used for mounting this fs on a client displays with a selectable "mount lustre" command string. Copying and pasting the string into a terminal window on the Lustre client mounts the file system on that client.

4.2.5 Selected node commands

The selected node commands menu lists commands that act on the selected node(s). Select one or more nodes by checking the box(es) to the left of the Hostname column.



NOTE: Any commands that are not available because of the node state, are grayed out.

- **Start Lustre...** - Mounts all targets on the selected file system node(s) and starts servicing Lustre clients.
- **Stop Lustre...** - Unmounts all targets on the selected file system node(s) and stops servicing Lustre clients.
- **Hand Off to *HA_group_name*...** - Hands off (fails over) the selected node's assigned HA resources to its HA Group nodes.
- **Take Back from *HA_group_name*...** - Takes back (fails back) the selected node's assigned HA resources from its HA Group nodes.
- **Shutdown...** - Initiates a graceful shutdown of the selected node. When this command is run, a dialog box prompts you to hand off the selected node's assigned HA resources

to its failover partner (node) or skip failover and take the HA resources offline before the node shuts down.

- **Reboot...** - Initiates a graceful reboot of the selected node. When this command is run, a dialog box prompts you to confirm the node reboot. When the node reboots, its targets (OSTs) goes offline temporarily.
- **Cycle...** - Equivalent to turning the hardware power off to the nodes, followed by turning power back on immediately.

CAUTION: When this command is run, unsaved data in flight is lost.

- **Force Power Off...** - Forces the selected node to shut down immediately, regardless of its state.

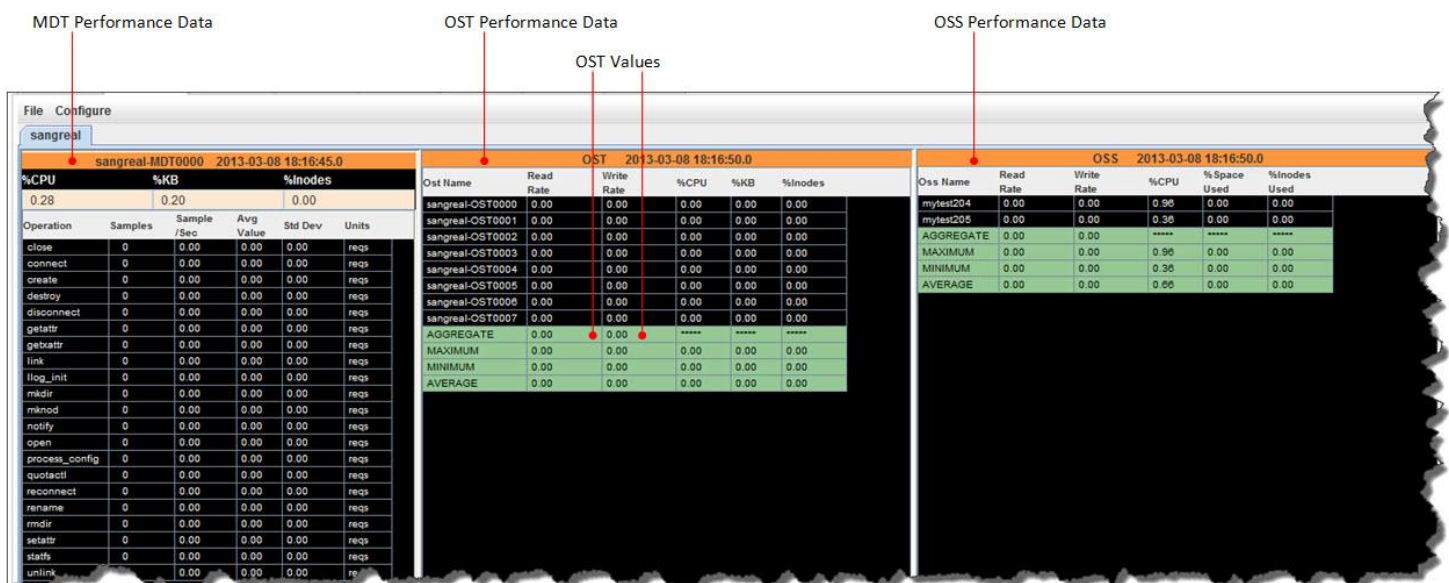
CAUTION: When this command is run, unsaved data in flight is lost.

- **Force Reset...** - Forces the selected node to reboot immediately, regardless of its state.
- **Power On...** - Powers on the selected node.
- **Minimum Rebuild Rates...** - Provides the option to set rebuild rates for selected nodes for single drive or multiple drive failures.

4.3 Performance tab

The **Performance** tab provides a high-level performance monitoring view of the Lustre file system and individual nodes metrics. Detailed descriptions are provided below.

To view performance data and manage data displays in a plotting graph, see “OST overview plot and controls”, page 35.



The **Performance** tab lists data for different file system components (MDT, OSTs and OSSs). The data is collected from the Lustre Monitoring Tool (LMT).

The following color are used to represent ordinary cells in the OST/OSS performance table:

Color	Description
Green	OSS/OST is OK.
Amber	OSS/OST is in a warning state.
Red	OSS/OST value is in a critical state.
Light Gray	OSS/OST is in a stale state.
Dark Gray	OSS/OST value is empty.
Yellow	OSS/OST value is in a warning state.
Black	OSS/OST value is normal.

4.3.1 MDT performance data

The left side column lists performance data for the file system's MDT.

- % CPU - Percent of CPU utilization.
- % KB - Percent of kilobytes used.
- % Inodes - Percent of inodes used.

4.3.2 OST performance data

The center column lists performance data for individual OSTs, and the right pane lists performance data for individual OSS's. To chart OST performance data in an overview plot (plotting graph) and specify variables, see “OST overview plot and controls”, page 35.

The following parameters are displayed in columns:

- Read Rate - Read rate analyzed.
- Write Rate - Write rate analyzed.
- % CPU - Percent of CPU utilization.
- % KB or % Space Used - Percent of kilobytes used.
- % Inodes - Percent of inodes used.

Summary rows at the bottom:

- Aggregate - Combined totals.
- Maximum - Maximum rates.

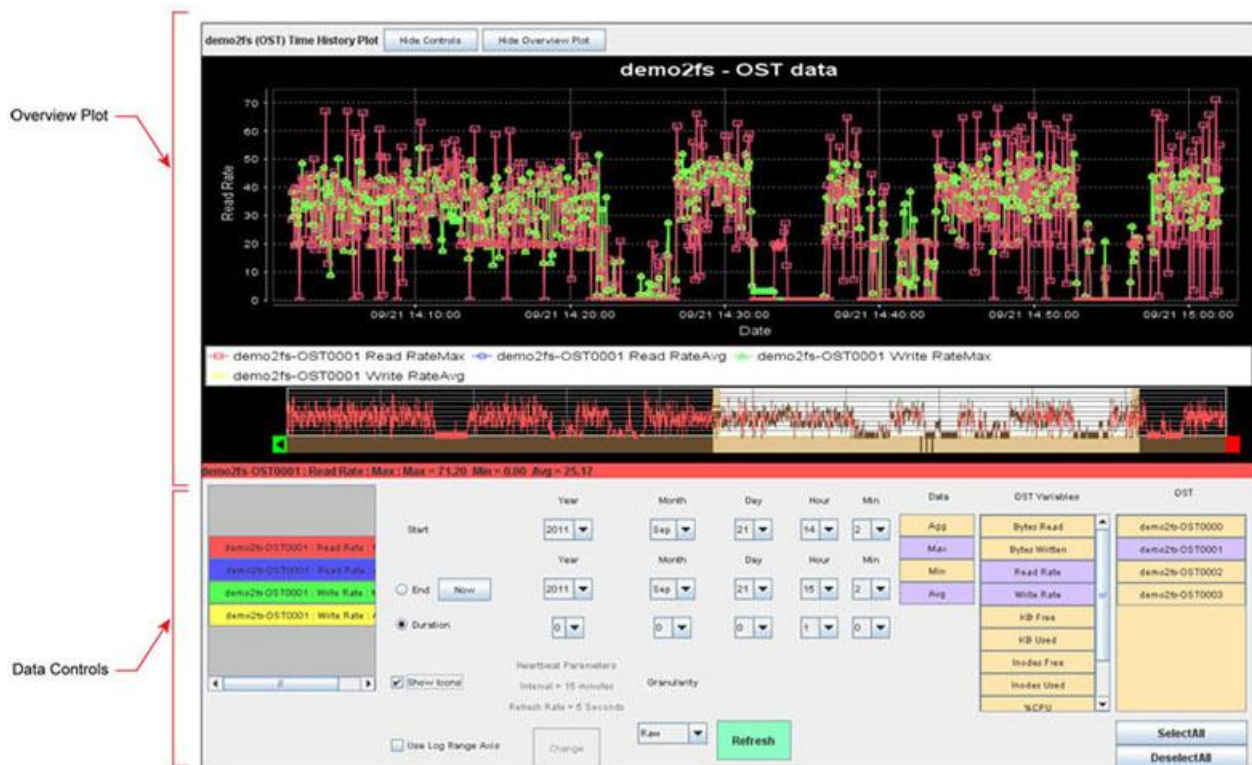
- Minimum - Minimum rates.
- Average - Average or mean rates.

4.3.3 OST overview plot and controls

OST performance data can be charted in an overview plot (plotting graph) on a specified time line.

To access the chart, double-click on one of the OST values (calculated), as shown in the previous illustration, in cells with black backgrounds.

When the graph opens in the upper pane, data is plotted for the selected OST and variable over a one-hour period (default time frame). In the lower pane are controls used to customize the data shown.



The following colors are used to represent the OST/OSS overview plot and controls:

Color	Description
Purple	Selected item
Amber	Unselected item

Multi-color rows in the data controls section represents the legend of the plot, and each is uniquely identified with the OST name.



Overview plot

The upper pane displays the name of the selected OST, the plotting graph, and show/hide toggle buttons.

- Controls (Show/Hide) - Shows/hides the OST performance variables (toggle button).
- Overview Plot (Show/Hide) - Shows/hides the plotting graph (toggle button).

Data controls

The lower pane lists performance variables that can be graphed and controls for start/end time or duration time.

1. OST - OSTs for which performance data can be graphed.
2. OST Variables - Performance variables that can be graphed.
 - Bytes Read/Written - Number of bytes that have been read/written by the OST.
 - Read/Write Rate – KB/sec written or read.
 - KB Free - Number of free KB free on the OST.
 - KB Used - Number of free KB used by the OST.
 - Inodes Free - Number of free inodes on the OST.
 - Inodes Used - Number of inodes used by the OST.
 - % CPU - CPU percentage used by the OST.
 - % KB - KB percentage used by the OST.
 - % Inodes - Inode percentage used by the OST.
3. Start (Year/Month/Day/Hour/Min) - Specified start for the plotting graph timeline. Default setting is the day/hour/min when the plotting graph is launched from the Performance tab.
4. End (Year/Month/Day/Hour/Min) - Specified end time for the plotting graph timeline (used instead of Duration variables). The Now button stops plotting performance data immediately.
5. Duration (Year/Month/Day/Hour/Min) - Specified duration for the plotting graph timeline (used instead of End variables). Default setting is one hour, measured from the day/hour/min when the plotting graph is launched from the Performance tab.
6. Show Icons - Will display or remove the icons in the chart.
7. Use Log Range Axis - Is a way of visualizing data that is changing with an exponential relationship.

8. Granularity - The amount of computation in relation to communication, i.e., the ratio of computation to the amount of communication. Fine-grained parallelism means individual tasks are relatively small in terms of code size and execution time. Available values are: Raw, Hour, Day, Week, Month, Year, and Heartbeat.
 - Heartbeat Interval - The amount of time between system checks for failed components. Available values are: Display Internal, 15 minutes, 30 minutes, 1 hour, 2 hours, or 4 hours.
 - Refresh Rate - Refers to how frequently a datum is updated with a new external value from another source. Available values are: 5 seconds, 10 seconds, 15 seconds, 30 seconds, 1 minute, 2 minutes, and 5 minutes.
9. Refresh - Updates the plotting graph for the selected performance variables.

4.4 Log browser tab

From the **Log Browser** tab, users can view combined all log files. The logs capture events from the kernel and any daemons that use syslog while they are running. As the logs may contain thousands of entries, filter and sort tools are available to more effectively search the log data.

The Log Browser page shows syslog messages. Other kinds of logs, such as GEM event logs and IPMI logs, get converted to syslog and are displayed here as well. Some logs do not appear, such as web server logs; as these are considered to be strictly internal and not an essential part of the cluster monitoring solution. However, web server logs and other internal logs are included in a support bundle, which can be created and viewed as described under “Interpret Sonexion support bundles”, page 100.

Detailed field descriptions are provided below.

Hover cursor over field for the down arrow to display and click on the arrow for drop-down menu to display.

Log Filtering (all drop-down menus displayed)

Log Events Display

The screenshot shows the Log Browser interface with the following components:

- Filter (56342 rows):** A search bar with a filter name field.
- Calendar:** A calendar for March 2013 with a date selector.
- Filter Fields:** From, To, Message, Facility, Program, Priority, Subsystem, and Apply buttons.
- Table:** A table with columns: Host, Facility, Priority, PID, Date, Message. The table contains 21 rows of log entries.
- Log Events Display:** A detailed view of a log entry showing the full message text.

Host	Facility	Priority	PID	Date	Message
1	daemon	info	0	Mar 8, 2013 07:21:03	syslog
2	daemon	info	0	Mar 8, 2013 07:21:03	dhcpcd
3	daemon	info	0	Mar 8, 2013 07:21:03	dhcpcd
4	daemon	info	49370	Mar 8, 2013 07:21:04	Loading cron
5	daemon	info	49370	Mar 8, 2013 07:21:04	Loading puppet-agent
6	daemon	info	49370	Mar 8, 2013 07:21:04	Loading puppet-agent
7	daemon	info	31329	Mar 8, 2013 07:21:04	Loading puppet-agent
8	daemon	info	31329	Mar 8, 2013 07:21:04	Loading puppet-agent
9	mytes201	daemon	info	Mar 8, 2013 07:21:04	Loading puppet-agent
10	mytes200	daemon	info	Mar 8, 2013 07:21:07	Dynamic lookup of Sprivate_interface at /etc/puppet/modules/dhcp/manifests/files.pp:26 is deprecated. For more information, see http://docs.puppetlabs.com/puppet/2.8/reference.html#dynamic_lookups
11	mytes200	daemon	warn	Mar 8, 2013 07:21:07	Dynamic lookup of Sprivate_interface at /etc/puppet/modules/dhcp/manifests/files.pp:26 is deprecated. For more information, see http://docs.puppetlabs.com/puppet/2.8/reference.html#dynamic_lookups
12	mytes200	daemon	warn	Mar 8, 2013 07:21:07	Dynamic lookup of Sprivate_interface at /etc/puppet/modules/dhcp/manifests/files.pp:26 is deprecated. For more information, see http://docs.puppetlabs.com/puppet/2.8/reference.html#dynamic_lookups
13	mytes200	daemon	warn	Mar 8, 2013 07:21:07	Dynamic lookup of Sprivate_interface at /etc/puppet/modules/dhcp/manifests/files.pp:26 is deprecated. For more information, see http://docs.puppetlabs.com/puppet/2.8/reference.html#dynamic_lookups
14	mytes200	daemon	warn	Mar 8, 2013 07:21:07	Dynamic lookup of Sprivate_interface at /etc/puppet/modules/dhcp/manifests/files.pp:26 is deprecated. For more information, see http://docs.puppetlabs.com/puppet/2.8/reference.html#dynamic_lookups
15	mytes200	daemon	warn	Mar 8, 2013 07:21:07	Dynamic lookup of Sprivate_interface at /etc/puppet/modules/dhcp/manifests/files.pp:26 is deprecated. For more information, see http://docs.puppetlabs.com/puppet/2.8/reference.html#dynamic_lookups
16	mytes200	daemon	notice	Mar 8, 2013 07:21:07	Compiled catalog for mytes201 in environment production in 1.47 seconds
17	mytes200	daemon	notice	Mar 8, 2013 07:21:07	Compiled catalog for mytes202 in environment production in 1.14 seconds
18	mytes200	daemon	notice	Mar 8, 2013 07:21:07	Compiled catalog for mytes200 in environment production in 1.55 seconds
19	mytes200	daemon	notice	Mar 8, 2013 07:21:07	Compiled catalog for mytes203 in environment production in 1.24 seconds
20	mytes200	daemon	notice	Mar 8, 2013 07:21:08	Compiled catalog for mytes204 in environment production in 1.13 seconds
21	mytes200	daemon	notice	Mar 8, 2013 07:21:08	Compiled catalog for mytes205 in environment production in 1.15 seconds

The **Log Browser** pane contains an upper section with filtering criteria and a lower section that displays the log data.

Clicking the **Filter** down-arrow, in the title bar of the Filter section, collapses the filter options, clicking it again will display them. The value expressed in the parenthesis in the title bar indicates the number of rows of total log messages for the filter applied.

4.4.1 Filter

The upper section lists filter criteria to sort the log entries.

- **From** - Starting date period.
- **To** - Final date period.
- **Hosts** - The hostname field can accept a single hostname, a list of comma-separated hostnames, a host range format ("hostname[00-03]" and a gender query such as "mds=primary", etc.
- **Facility** - This identifies who submitted the message. There are a small number of facilities defined. The kernel, the mail subsystem, FTP server, are just some examples of recognized facilities.
- **Priority** - Filter the entries based on this criteria.
- **PID** - Process ID.
- **Program** - Name of program or module that produced the message.
- **Subsystem** - Filters Lustre, LustreError and some other classes of messages.
- **Message** - This is the text of the syslog message, along with some additional information about the process that generated the message.

4.4.2 Logs

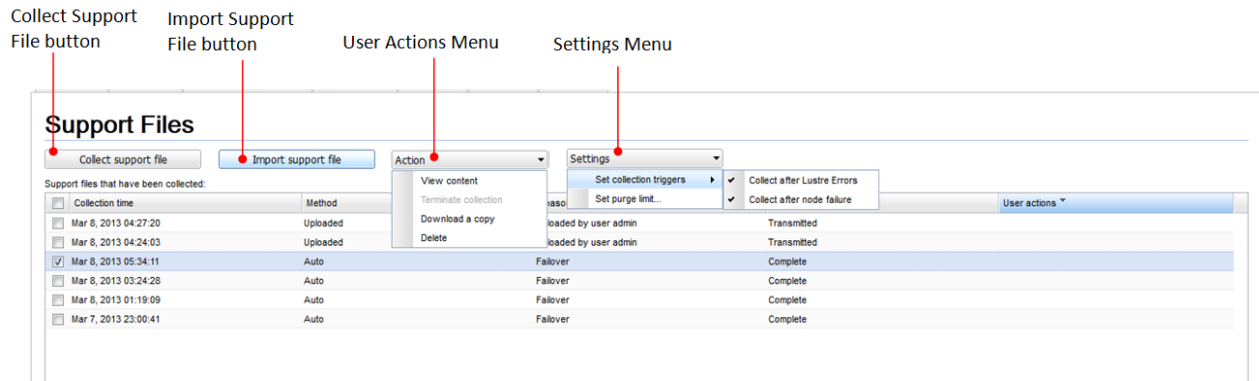
The lower section lists individual log entries.

- **Host** - The hostname field consists of the host name (as configured on the host itself).
- **Facility** - This identifies who submitted the message. There are a small number of facilities defined. The kernel, the mail subsystem, FTP server, are just some examples of recognized facilities.
- **Priority** - Filter the entries based on this criteria.
- **PID** - Process ID.
- **Program** - Name of program or module that produced the message.
- **Subsystem** - Filters Lustre, LustreError and some other classes of messages.
- **Message** - This is the text of the syslog message, along with some additional information about the process that generated the message.

4.5 Support tab

The **Support** tab enables users to collect diagnostic information from the storage cluster, including logs and configuration settings. If a Lustre error or event occurs, such as node failover, the process to collect system information in support files is triggered automatically. This process can also be started on a manual basis. Detailed field descriptions are provided below.

To view support files, create bundles of support files, and delete bundle, see “Support Files”, page 96.



The **Support Files** tab screen displays all of the support files that have been collected, on either an automatic or manual basis. For each support file, several user actions are available:

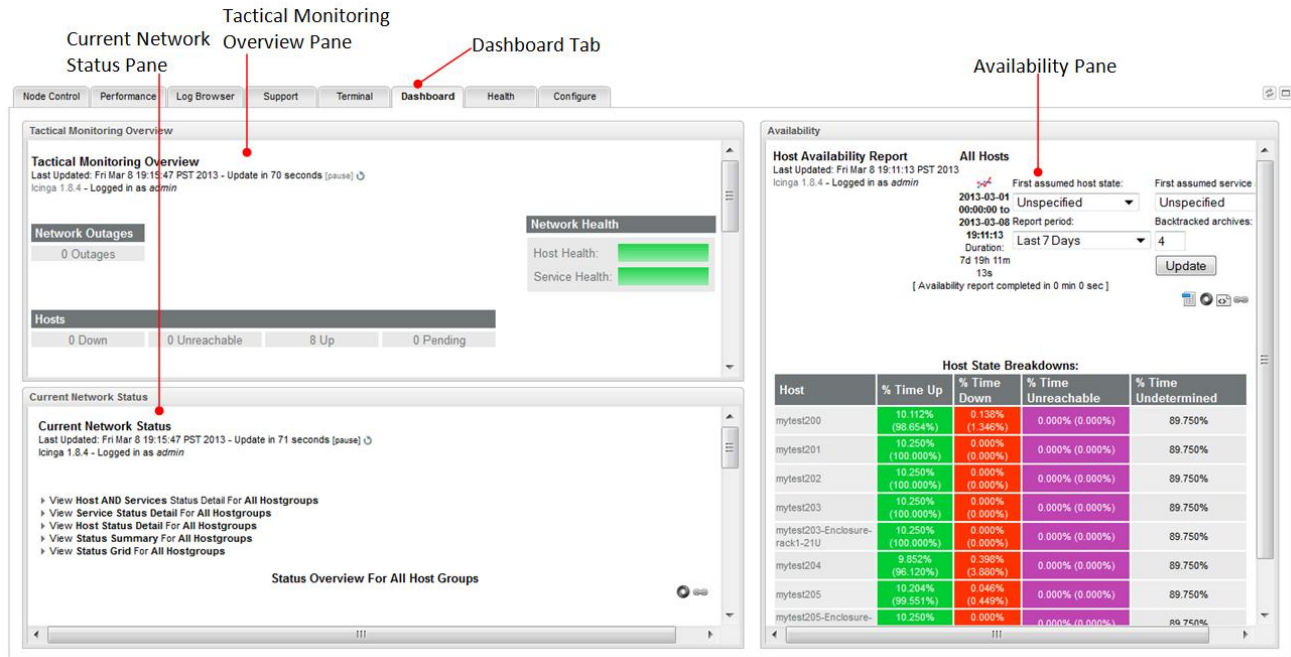
- **View content** - View system and web logs, node and application status information for the support file.
- **Terminal collection** - Stops the process of a support file collection.
- **Download a copy** - Download a copy of the support file to your local system.
- **Delete** - Delete the support file.

4.6 Dashboard tab

The **Dashboard** tab provides a quick glance of the status of the system components. To monitor the status of Lustre, refer to the **Node Control** tab (page 27).

On the Dashboard you will find three pre-selected panes from the Health monitoring that are displayed in these panes. These panes provide a quick assessment of the status of the monitoring of all components, the network status, and resource availability.

Detailed field descriptions are provided below.



4.6.1 Tactical monitoring overview

The Tactical monitoring overview pane is designed to serve as a “birds-eye” view of all network monitoring activity. It allows you to quickly see network outages, host status, and service status. It distinguishes between problems that have been handled in some way (that is, they have been acknowledged, had notifications disabled, and so forth) and those that have not been handled, and thus need attention. It is useful if you are monitoring many hosts and services and need to keep a single screen up to alert you of problems.

Each category is a hot link to view-specific detailed data. As you pass the mouse pointer over them, the cursor changes to a hand pointer tool, indicating the link. This is also applicable to the service checks and host checks, where clicking on the item in the green field opens more details in the window.

To return to previous information use standard web browser navigation functions such as the back button, right-click and select Back, or on a Windows system press the Delete key, or on some Unix systems use the Control (Command) - (left bracket) key.

4.6.2 Current network status

This pane creates a service overview for all host groups, network outages, network health status, service checks, host checks, monitoring features. Enclosure component issues that appear as alerts can be viewed from this pane.

4.6.3 Availability

This pane is used to report on the availability of hosts and services over a user-specified period of time.

4.6.4 Usage

By providing a quick overview, Dashboard enables users to spot issues quickly, and use the Health tab monitoring features to drill down to determine the details of the issue and possible cause/resolution. Typically, these are presented as “unhandled problems” and you do have the ability to click the active link in Dashboard to view the summary, however due to the small size of the windows, the preferred method would be to click the Health tab, locate the unhandled problem and click the link. Then in a larger window, details are displayed.

Table 4. Colors for Host and Service Status

Color	Description
Green	State is OK, host is up, service is ok, not flapping or not in scheduled downtime.
Amber	State is in a warning state.
Light Amber (2 shades)	State is warning but acknowledged.
Red	State is in a critical state.
Light Red (2 shades)	State is critical but acknowledged. Also can indicate a host is down, a service is critical, host or service properties have been disabled, is flapping or in scheduled downtime, or various other possible issues related to host and service problems.
Purple (3 shades)	State is in an unknown or unreachable state.
Blue	Indicates a pending action.

4.7 Terminal tab

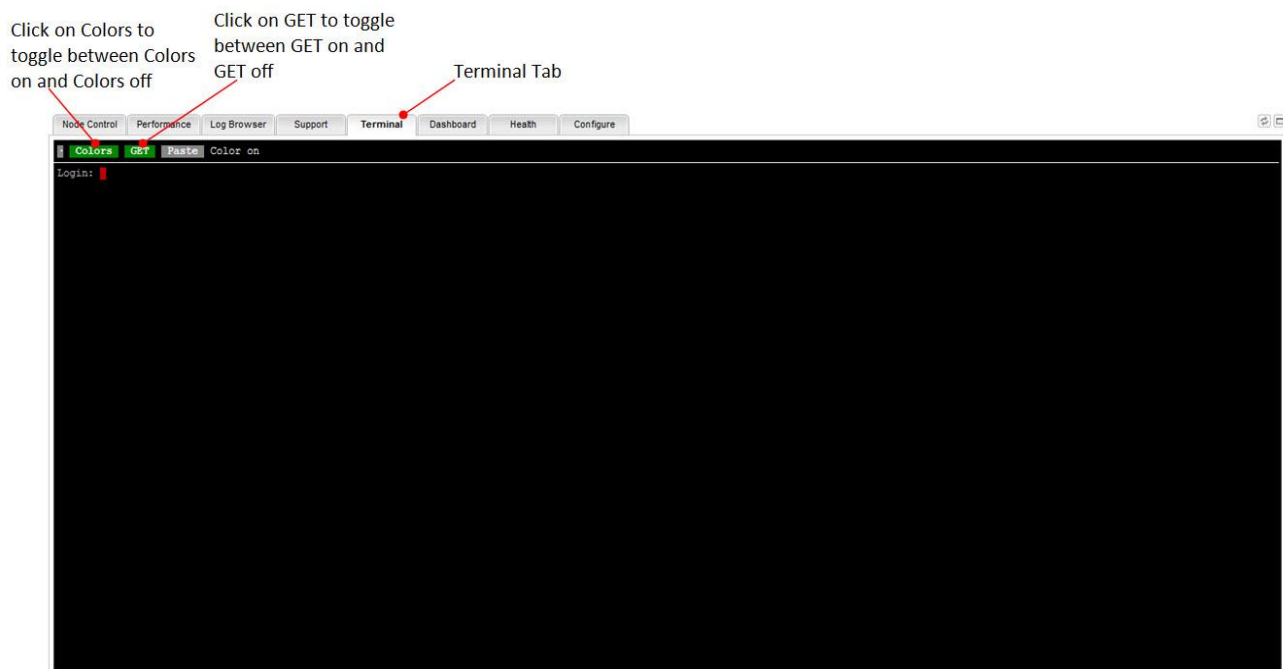
The **Terminal** tab provides a connection to the management node. It is equivalent to logging in as an SSH client. It allows the administrator to access the system via "shell" command line. Login credentials are the same as for the browser-based application. This feature provides a BASH shell session and allows execution of cscli commands on the management node.

Administrators may want or need to perform Linux or Lustre commands out of comfort if the preference is to use the command line over the UI or necessity if the operation needed to be performed is not available via the UI. For example, operations such as:

- Executing MDRAID commands to get status or ID information on arrays or drives
- Executing CRM commands to get status on Cluster processes

Accessing the command line shell within the GUI could be more convenient for the user instead of having to access or physically hook up a separate client terminal to accomplish the same level of access.

The **Terminal** Tab might not be available on all systems.



4.8 Health tab

The **Health** tab monitors the system's health and performance, checking hosts and services and notifying you of problems and recoveries.

Features include:

- Monitoring of network services (SMTP, POP3, HTTP, NNTP, PING, etc.).
- Monitoring of host resources (CPU load, disk usage, etc.).

- Monitoring of components (ArrayDevices, BMC, Batteries, PSUs, Cooling Fans, I/O modules, disk drives, enclosure electronics, etc.).
- Parallelized service checks.
- Ability to define network host hierarchy using "parent" hosts, allowing detection of and distinction between hosts that are down and those that are unreachable.
- Contact notifications when service or host problems occur and get resolved.
- Ability to define event handlers to be run during service or host events for proactive problem resolution.
- Automatic log file rotation.
- Supports redundant monitoring hosts.

The **Health** tab displays monitoring screens. An example is provided below of the Main screen where everything comes together. The default display is the Status group: Tactical monitoring overview (see page 44).

Table 4 on page 41, is a guide to the colors used to indicate health status.

Detailed descriptions are provided below.

4.8.1 Header

The header consists of the summary information Status for Hosts and Services that provides tactical information. Clicking any of the summary components will load the page into the alert summary window. For example, clicking on the Services Warning button will open the Service Status Details for all hosts.

The Status shows host and service counters for their respective states. If the count is zero, the color remains grey. Clicking on a state opens in the main view. On the right hand side, general process information listed as:

- Hosts|Services (active/passive)
- Host|Services execution time (min/avg/max)
- Host|Services latency (min/avg/max)

4.8.2 Link groupings

The left pane has groups of links to screens summarizing various kinds of information:

- **Status** group shows statuses in the categories Tactical Overview, Host Detail, Service Detail, Hostgroup Overview, Servicegroup Overview and Status map
- **Problems** group selects categories of problems: Service Problems, Unhandled Services, Host Problems, Unhandled Hosts, and All Unhandled Problems, and Network Outages.
- **System** group: Comments, Downtime, Process Information, Performance Info, and Scheduling Queue
- **Reporting** group creates report summaries for availability, alert history, alert summary, notifications and event logs. These reports can be exported to a CSV file, JSON, or XML.

4.8.3 Status group: Tactical monitoring overview

The tactical monitoring overview presents a birds-eye view of network monitoring activity. It lets you quickly see network outages, host status, and service status. It distinguishes between problems that have been handled in some way (for example, been acknowledged, had notifications disabled, etc.) and those that have not been handled and thus need attention. This is useful if you are monitoring a large number of hosts and services, and you need a single screen to alert you of problems.

Service and host check retries check

In order to prevent false alarms from transient problems, the monitoring tool allows you to define how many times a service or host should be (re)checked before it is considered to have a “real” problem. This is controlled by the `max_check_attempts` option in the host and service definitions. Understanding how hosts and services are (re)checked in order to determine if a real problem exists is important in understanding how state types work.

Setting the Service and Host Check retry counts

The service and host check retry counts can be done manually by editing Icinga configuration files and then running **puppet**. It cannot be set via the Sonexion System Manager GUI.

The `max_check_attempts` value is set to 3 for all hosts and services in file `/etc/puppet/modules/icinga/files/objects/templates.cfg`. It can also be set to a different value per each service or host. The configuration files which hold the hosts and services in Icinga are in directory: `/etc/icinga/objects/`. They are generated by:

- The puppet module `/etc/puppet/modules/icinga/`
- `/opt/xyratex/python/t0/monitoring/monitoring2puppet.py`

4.8.4 Reporting group: Alert Summary

You can generate an Alert Summary report using the Alert Summary link in the Reporting Group. This lets you view the current status of all hosts and services that are being monitored. The status component can produce two main types of output: a status overview of all host groups (or a particular host group) and a detailed view of all services (or those associated with a particular host).

Below is an example screen shot of an Alert Summary.

The screenshot displays the Icinga Alert Summary interface. At the top, it shows 'Host Alert History' and 'All Hosts and Services'. The 'Log Navigation' section indicates the latest archive is from 'Wed Oct 23 00:00:00 CDT 2013' to 'Present..'. On the right, there are controls for 'State type options' (set to 'All state types') and 'History detail level for all hosts' (set to 'All alerts'). Below these are checkboxes for 'Hide Flapping Alerts', 'Hide Downtime Alerts', 'Hide Process Messages', and 'Older Entries First', along with an 'Update' button. The main content area shows a list of alerts for 'October 23, 2013 08:00'. The first three alerts are 'SERVICE ALERT' entries for 'snx11022n003-Enclosure-rack1-36U' with state types 'OK', 'SOFT', and 'OK' respectively. The 'SOFT' state type is highlighted with a red box. The remaining three alerts are 'SERVICE ALERT' entries with state types 'UNKNOWN', 'SOFT', and 'UNKNOWN'. The 'SOFT' state type is also highlighted with a red box.

4.8.5 State types

The current state of monitored services and hosts is determined by two components (highlighted in the above screen shot):

- The status of the service or host, such as OK, WARNING, UP, DOWN, or UNKNOWN.
- The state type of the service or host, which can be SOFT or HARD, as explained under the following headings. State types are used to determine when event handlers are executed and when notifications are initially sent out.

Soft states

Soft states occur in the following situations:

- When a service or host check results in a non-OK or non-UP state and the service check has not yet been (re)checked the number of times specified by the

`max_check_attempts` directive in the service or host definition. This is called a soft error.

- When a service or host recovers from a soft error. This is considered a soft recovery.

The following events occur when hosts or services experience SOFT state changes:

- Event handlers are executed to handle the SOFT state.

An event handler such as `cscli alerts_threshold` can help you proactively fix a problem before it turns into a HARD state. Event handlers can be tweaked manually in Icinga. For more information on event handlers, see the following site:

<http://docs.icinga.org/latest/en/statetypes.html>

- The SOFT state is logged.

SOFT states are logged only if you enabled the `log_service_retries` or `log_host_retries` options in your main configuration file, which is located in directory `/etc/puppet/modules/icinga/files/icinga.cfg`. You can edit the file on both MGMT nodes and restart puppet on those nodes. The following command can be run with root privileges:

```
/opt/xyratex/bin/beUpdatePuppet -s -g mgmt
```

Hard states

Hard states occur for hosts and services in the following situations:

- When a host or service check results in a non-UP or non-OK state and it has been (re)checked the number of times specified by the `max_check_attempts` option in the host or service definition. This is a hard error state.
- When a host or service transitions from one hard error state to another error state (e.g., WARNING to CRITICAL).
- When a service check results in a non-OK state and its corresponding host is either DOWN or UNREACHABLE.
- When a host or service recovers from a hard error state. This is considered to be a hard recovery.
- When a passive host check is received. Passive host checks are treated as HARD unless the `passive_host_checks_are_soft` option is enabled.

The following things occur when hosts or services experience HARD state changes:

- The HARD state is logged.
- Event handlers are executed to handle the HARD state.
- Contacts who are subscribers are notified of the host or service problem or recovery. Such contacts and notifications are configurable via the sub-command: `cscli alerts_config`.

The following table is an example of how state types are determined, when state changes occur, and when event handlers and notifications are sent out. The table shows consecutive checks of a service over time. The service has a `max_check_attempts` value of 3.

Time	Check #	State	State Type	State Change	Notes
0	1	OK	HARD	No	Initial state of the service
1	1	CRITICAL	SOFT	Yes	First detection of a non-OK state. Event handlers execute.
2	2	WARNING	SOFT	Yes	Service continues to be in a non-OK state. Event handlers execute.
3	3	CRITICAL	HARD	Yes	Max check attempts has been reached, so service goes into a HARD state. Event handlers execute and a problem notification is sent out. Check # is reset to 1 immediately after this happens.
4	1	WARNING	HARD	Yes	Service changes to a HARD WARNING state. Event handlers execute and a problem notification is sent out.
5	1	WARNING	HARD	No	Service stabilizes in a HARD problem state. Depending on what the notification interval for the service is, another notification might be sent out.
6	1	OK	HARD	Yes	Service experiences a HARD recovery. Event handlers execute and a recovery notification is sent out.
7	1	OK	HARD	No	Service is still OK.
8	1	UNKNOWN	SOFT	Yes	Service is detected as changing to a SOFT non-OK state. Event handlers execute.
9	2	OK	SOFT	Yes	Service experiences a SOFT recovery. Event handlers execute, but notification are not sent, as this wasn't a "real" problem. State type is set HARD and check # is reset to 1 immediately after this happens.
10	1	OK	HARD	No	Service stabilizes in an OK state.

Service checks

Services are checked by the monitoring daemon:

- At regular intervals, as defined by the `check_interval` and `retry_interval` options in your service definitions.
- On-demand as needed for predictive service dependency checks.

On-demand checks are performed as part of the predictive service dependency check logic. These checks help ensure that the dependency logic is as accurate as possible. If you don't make use of service dependencies, monitoring won't perform any on-demand service checks.

Parallelization of service checks

Scheduled service checks are run in parallel. When monitoring needs to run a scheduled service check, it will initiate the service check and then return to doing other work (running host checks, etc.). The service check runs in a child process that was forked from the main daemon. When the service check has completed, the child process will inform the main monitoring process (its parent) of the check results. The main monitoring process then handles the check results and takes appropriate action (running event handlers, sending notifications, etc.).

On-demand service checks are also run in parallel if needed. As mentioned earlier, monitoring can forego the actual execution of an on-demand service check if it can use the cached results from a relatively recent service check.

Service States

Services that are checked can be in one of four different states:

- **OK** - if the round trip average (RTA) is less than 200 ms and the packet loss is less than 20%
- **WARNING** - if the round trip average (RTA) is greater than 200 ms or the packet loss is 20% or more.
- **UNKNOWN** - unable to determine the state.
- **CRITICAL** - if the round trip average (RTA) is greater than 600 milliseconds (ms) or the packet loss is 60% or more.

Accessing service properties

On the **Health** tab, you can display the service properties as shown in the following screen shots. In the **Status** group, click the **Service Detail** link.

Host	Service	Status	Last Check	Duration	Attempt	Status Information
snx11022n000	Arrays and Disk Status	OK	2013-10-24 18:09:24	37d 4h 15m 48s	1/3	All arrays are operating normally
	Current Load	OK	2013-10-24 18:09:24	62d 3h 34m 5s	1/3	OK - load average: 0.00, 0.01, 0.01
	Current Users	OK	2013-10-24 18:09:24	62d 3h 34m 5s	1/3	USERS OK - 0 users currently logged in
	Free Space	OK	2013-10-24 18:09:24	62d 3h 34m 5s	1/3	DISK OK - free space: / 161860 MB (87% inode=99%); /mnt/mgmt 757754 MB (98% inode=99%);
	Network	OK	2013-10-24 18:09:24	62d 3h 34m 5s	1/3	NET OK - (Rx/Tx) eth0=(17.9B/14.3B), eth1=(36.3B/78.2B), eth2=

Service State Information

Current Status:	OK (for 1d 2h 22m 45s)
Status Information:	Summary: 4 Fan Sensors available. All Sensors readings are within normal operating levels Fan0-OK:3450RPM, min:2000RPM, max:20480RPM; Fan1-OK:2700RPM, min:2000RPM, max:20480RPM; Fan2-OK:3300RPM, min:2000RPM, max:20480RPM; Fan3-OK:2470RPM, min:2000RPM, max:20480RPM;
Performance Data:	Fan0=3450RPM Fan1=2700RPM Fan2=3300RPM Fan3=2470RPM

Multi Line Status Information

Fan Statistics	OK	04-24-2012 10:39:32	1d 2h 18m 10s	1/3	Summary: 4 Fan Sensors available. All Sensors readings are within normal operating levels
Thermal Statistics	WARNING	04-24-2012 10:39:32	1d 2h 18m 10s	3/3	Summary: 8 Thermal Sensors available. There are sensors showing warnings
Voltage Statistics	OK	04-24-2012 10:39:32	1d 2h 18m 10s	1/3	Summary: 4 Voltage Sensors available. All Sensors readings are within normal operating levels

Click for Details

Online Summary

Service state determination

Service checks are performed by plug-ins, which can return a state of OK, WARNING, UNKNOWN, or CRITICAL. These states directly translate to service states.

Services state changes

When monitoring checks the status of services, it will be able to detect when a service changes between OK, WARNING, UNKNOWN, and CRITICAL states and take appropriate action. These state changes result in different state types (HARD or SOFT), which can trigger event handlers to be run and notifications to be sent out. Service state changes can also trigger on-demand host checks. Detecting and dealing with state changes is what Health monitoring is all about.

When services change state too frequently they are considered to be "flapping." Monitoring can detect when services start flapping, and can suppress notifications until flapping stops and the service's state stabilizes.

4.9 Configure tab

The screen for the **Configure** tab has three or four features depending on the network interconnect technology. The **InfiniBand Settings** feature appears.

You will also have features for configuring network IP Routing for complex routing and network Firewall Settings, as well as configuring the Minimum Rebuild Rates for storage.

4.9.1 Network

The following options are available under the **Configure** tab "Network" feature.

InfiniBand settings

The **InfiniBand Settings** screen lets you enable, disable and set the priority for Subnet Manager on the MGMT, MDS, or MGS.

The settings let you enable or disable this Subnet Manager, and establish a priority. A reset button is provided to clear the current settings and return them to the values set on the server; that is, the default settings.

1. Click **InfiniBand Settings** in the list of options on the left side of the screen.

2. To enable this Subnet Manager, click the **Enable Subnet Manager** button. To disable, deselect the option.
3. Adjust the slider to establish the priority for the manager functions. The values are 0-15, with 15 being the highest priority.
4. Click the **Apply** button to save the changes.

IP routing

This feature provides the option to configure IP routing to multiple networks with separate gateways on the Lustre data network (LNET).

1. Click **IP Routing** in the list of features on the left side of the screen.

2. Click the green "+" symbol to add a route to the routing table.
3. Enter the system's IP address in the **Destination** field, which shows the remote network to be reached through the router.

IP Routing table:

Use 'default' as destination to add default route

Destination : Prefix : /32 Router :

4. Enter the correct prefix value in the **Prefix** field. 'Prefix' stands for 'routing prefix'. The routing prefix is expressed in CIDR notation. The 'Destination' and 'Prefix' define the remote network to be reached through the router. See the example shown below.
5. Enter the correct router IP address in the **Router** field.
6. To add another route to the routing table, click the green "+" symbol again and repeat steps 3-5 for the other routings.
7. If you wish to remove a route, click the red "-" symbol.
8. Click the **Apply** button to save the changes.

Firewall settings

By default, Sonexion configures a firewall on the nodes that allows only Lustre traffic on the Lustre network (LNET). The Firewall Settings feature allows this restriction to be relaxed so that other kinds of traffic can be sent over this network.

1. Click **Firewall Settings** in the list of features on the left side of the screen.

Firewall Settings

This setting applied to the firewalls run on the Lustre nodes only. Firewalls are always active on management nodes.

Firewall protection : On (recommended)
 No filtering (risky)
 Off (dangerous)

2. Select the level of firewall protection you desire. Click on one of the radio buttons.

LEVEL	Description	Network Security Risk
On	Default setting, only Lustre nodes are protected.	Recommended
No filtering	Any traffic is allowed over the Lustre Client Network.	Risky
Off	The firewall is completely stopped on all Lustre nodes, but is active on the management (MGMT) nodes.	Dangerous

3. Click the **Apply** button to save the changes.

4.9.2 Storage

The following two RAID setting options are available under the **Configure** tab "Storage" feature.

Configuring minimum rebuild rates

The **Configuring minimum rebuild rates** feature offers the option to configure the global rebuild rate (rebuild speed) for the arrays with either a single or multiple drive failures.

CSSM offers predefined default settings for the minimum and maximum array rebuild speed. The minimum setting is 1000 KB/s and the maximum rebuild rate is 200,000 KB/s. The minimum rebuild rates setting allows you to override the minimum default setting with a user custom setting for a single drive failure and a multi drive failure.

In the case of a single drive failure, the situation is serious but not as critical since the rebuilding array could sustain another drive failure and still remain operational, therefore the minimum setting for a single drive failure would be much different than for a multiple drive failure. In the case of a multiple drive failure, this would be a very critical condition and you may want to allocate as much speed as possible to the rebuild process so that the exposure of another drive failure during rebuild is minimized. You may wish to set the maximum value for the minimum on multiple drive failures; that is, 200,000 KB/s.

You can also apply rebuild rate settings on a per-node basis from within the **Node Control** tab on the **Selected Nodes** dropdown menu. See page 28. This second option lets you override global settings for an individual node(s).

1. Click **Minimum Rebuild Rates** in the list of options on the left side of the screen.

Minimum Rebuild Rates

This option sets the global rebuild rate (rebuild speed) for the arrays with either a single or multiple drive failures. The global setting can be overridden on a per node basis, see Minimum Rebuild Rates under the Selected Nodes drop-down menu.

Single drive failure : KB/sec

Multiple drive failures : KB/sec

Reset per-node settings : no nodes with per-node settings

2. Enter the rebuild rate expressed in KB/sec allocated to the rebuild for both a single drive failure and multiple drive failures.
3. Although, no value is displayed when the screen is displayed, the default settings for a Single Drive Failure is 50000 KB/sec and for Multiple Drive Failures is 80000 KB/sec. Enter the amounts you wish to use for each situation.

4. Click the check box if you wish to reset the global rebuild speed settings. This resets values the defaults. The interface indicates the number of nodes affected by this global reset. In the example above, two nodes will be affected.
5. Click the **Apply** button to save the changes.

5. Operations

5.1 Managing Nodes

This topic discusses the procedures to start and stop the Lustre file systems using CSSM.

Sonexion manages Lustre distributions, therefore CSSM provides the methods for starting and stopping Lustre, and viewing the current status of the Lustre file systems. Essentially, starting and stopping of Lustre in this context means to mount and un-mount all the relevant Lustre file systems on each server node.

5.1.1 Starting Lustre

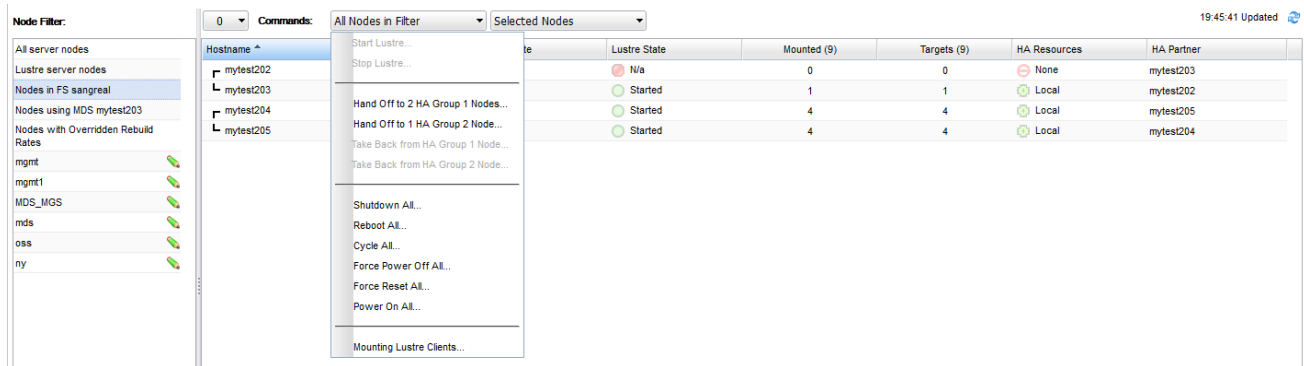
The left pane (filter pane) of the screen contains the “Node Filter” list. There should always be a node or nodes selected in this list.

The right pane (node interaction pane) of the screen dynamically displays details about the node(s) of the file system selected.

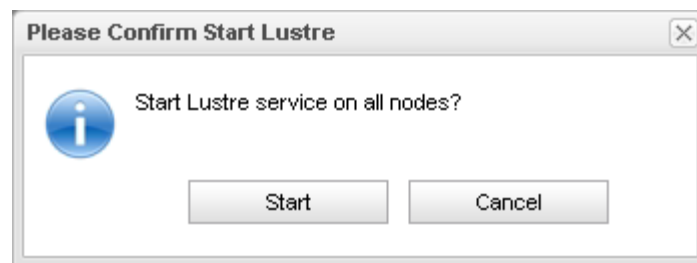
Upon initial startup of the system, the Lustre file system will be in a stopped state.

1. Click the **Node Control** tab, if not already selected.
2. Select a node(s) in the **Node Filter** list.

Using the mouse pointer click to select a node, and continue selecting multiple nodes with additional point and clicks. If you make a mistake, selecting the node a second time will unselect the node.



3. Click the **All Nodes in Filter** button.
4. Select the Start Lustre... menu item.



You will receive a confirmation message that you are starting Lustre on the selected nodes. Click the **Start** button to continue or Cancel to stop the operation.

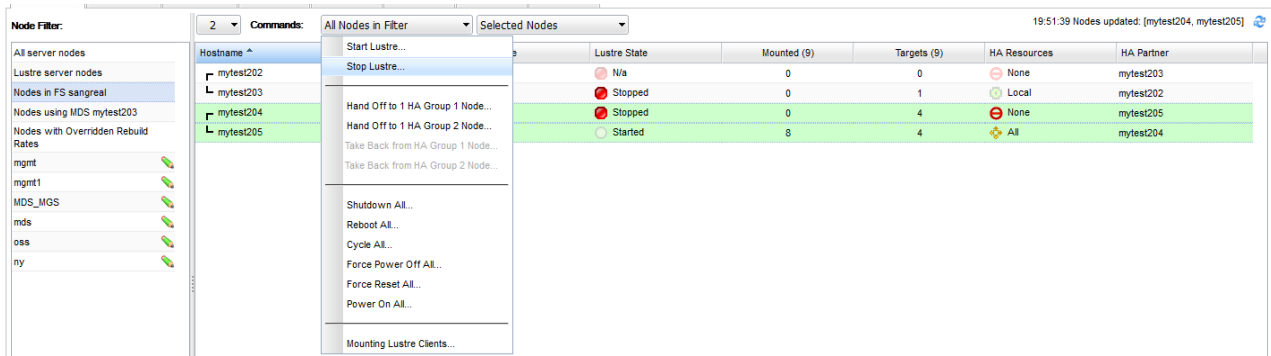
The Lustre file system will begin mounting the targets and the status is automatically updated and displayed in the right pane. It will continue to update approximately every 5 seconds until the start request has been completed.

Should an error occur, there will be no icon displayed for the selected node's Lustre State, instead the word **Error** will appear. Hover the mouse pointer over the word "Error" and a pop-up will display the reason for the error. Access the Log Browser and review the logs for a possible cause.

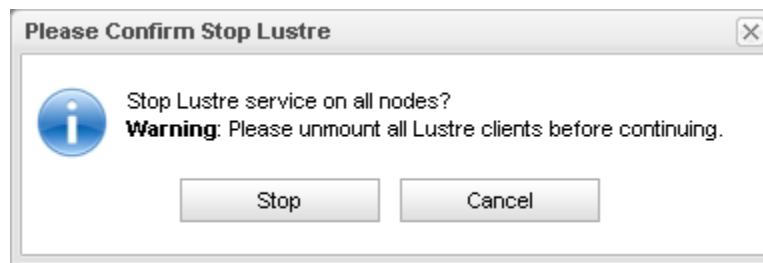
5.1.2 Stopping Lustre

1. Select a node(s) in the "Node Filter" list.

Using the mouse pointer click to select a node, and continue selecting multiple nodes with additional point and clicks. If you make a mistake, selecting the node a second time will unselect the node.



2. Click the **All Nodes in Filter** button.
3. Select the **Stop Lustre...** menu item.



4. You will receive a warning message reminding you to first unmount the Lustre clients, if this was already accomplished, click the **Stop** button. If the Lustre clients were not unmounted, click the **Cancel** button and unmount the Lustre clients. The repeat the steps above.

The Lustre file system will begin stopping Lustre on each selected node, and the display will update and refresh about every 5 seconds until the stop request has been completed.

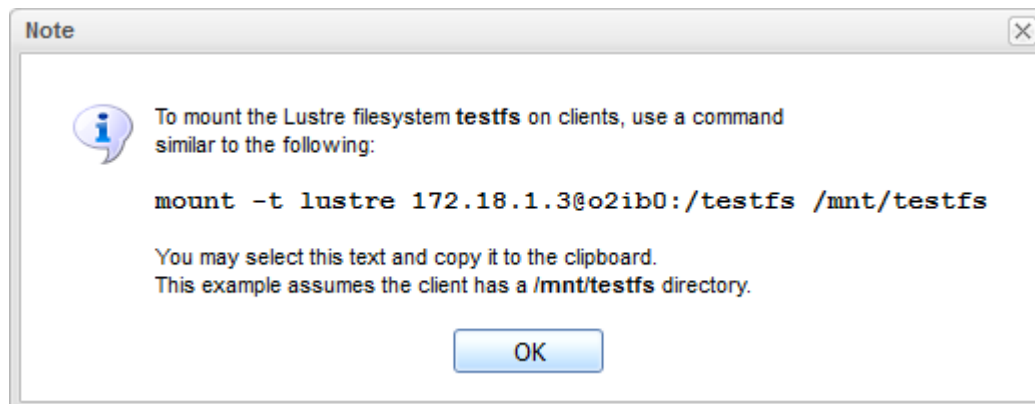
5.1.3 Mounting Lustre

The `cscli mount` command is available at the bottom of the **All Nodes in Filter** menu, option **Mounting Lustre Clients...**

Selecting the menu item causes a dialog window to open, displaying the following command (InfiniBand example):

```
mount -t lustre 172.18.1.3@o2ib0:/testfs /mnt/testfs
```

The text in the dialog window is selectable, so you may highlight and copy the command and then paste it into a terminal window to modify and execute.



5.2 Managing a failover

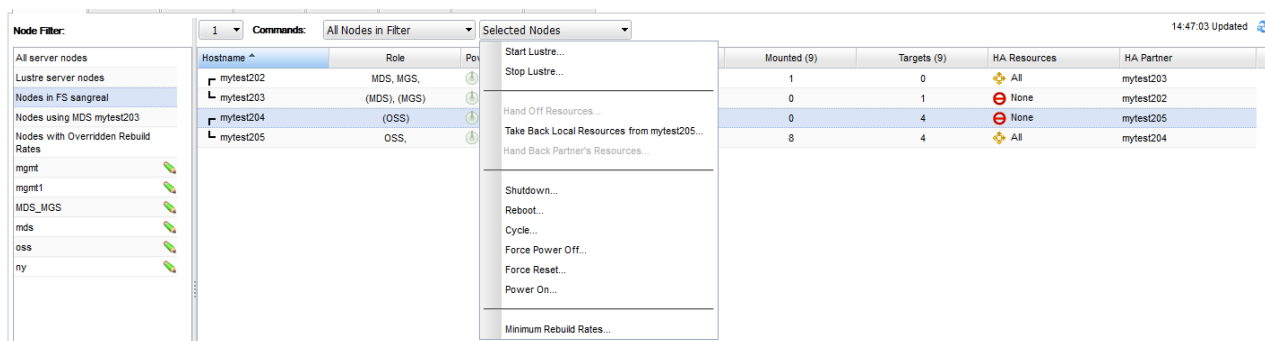
A chassis and two Object Storage Server (OSSs) are bundled in the modular SSU. Each OSS hosts one OSS node; there are two OSS nodes per SSU. Within an SSU, the OSS nodes are organized in an HA pair with sequential numbers (for example, snx11000n004/snx11000n005). If an OSS node goes down because its OSS fails, its resources migrate to the HA partner/OSS node in the other OSS.

Occasionally, there may be a requirement to manually transfer or handoff the resources to the HA partner. In that case, CSSM provides the means to perform this task.

5.2.1 Hand off resources to the HA partner OSS

In the event, you must failover or handoff the resources to the HA partner OSS, do the following:

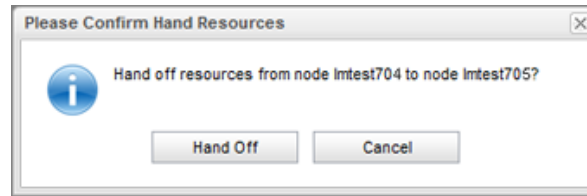
1. Click the **Node Control** tab, if not already selected.
2. Select a node in the **Node Filter** list. Then select the node, in the **Hostname** column.



3. Click the **Selected Nodes** dropdown menu and choose Hand Off Resources to snx11000n004

The menu item is dynamic and will rename to match the HA partner name.

4. You will receive a confirmation dialog, click **Hand Off** to continue or Cancel.



To manage the handing back or taking back of the resources, see [Managing Failback](#).

5.3 Managing failback

The improved High Availability (HA) service integrated with the core OS platform, performs automatic fail over of the resources to ensure continuity of management and data services. This solution provides the means for an enclosure to hand off (fail over) resources from a failed server to a paired backup server.

Two SSU OSSs are configured in an active-active pair, each serving one half of the configured OSTs. If a failure occurs, the remaining OSS automatically takes over the OSTs of the failed OSS.

Similarly, if one of the MGMT nodes fails, the remaining MGMT node takes over all of the required management services. In either case, the Lustre file system continues operation and can be mounted by Lustre clients.

Automatic failback is not currently supported by the HA service. When a failed node is brought back online, the system administrator must go to the **Node Control** tab and "take back" services in order to migrate functionality back to the recovered node. Manual failback only is employed since an automatic feature could possibly interfere with the ability to verify readiness of a recovered node.

There are three possible methods to fail back resources. They are:

- Access the primary MGMT node (typically n000) and execute `cscli failback`.
- Use the CSSM GUI to take back the resources from the recovered node,
- Use the CSSM GUI to hand back the resources from the surviving node,

5.3.1 Command line failback

Now that the issue with the node is corrected and the node is returned to online status, its resources can be failed back to the original node from the partner node using a CLI command.

1. Log into the management node via SSH.
2. `$ ssh -l admin primary_management_ip`
3. Change to root user, by entering:
4. `$ sudo -s`
5. Enter the following command:

```
# /opt/xyratex/bin/cscli failback -c cluster_name -n node_name
```

Specify the cluster name and node name to which to fail back the resources.

Verify that the resources were restored correctly.

5.3.2 Use the CSSM GUI

Take back the resources

Once the issue with the node is corrected and the node is returned to online status, its resources that were handed off to the partner node can now be taken back by the original node.

1. Select the node that had been failed and is now online.
2. Click the **Selected Nodes** drop down menu and choose **Take Back Resources from name....**

The screenshot shows the CSSM GUI interface. On the left, there is a 'Node Filter' sidebar with various node categories. The main area displays a table of nodes with columns for Hostname, Role, and Power status. A context menu is open over the 'mytest205' node, with the option 'Take Back Local Resources from mytest205...' selected. To the right, a table shows resource status for different nodes, including 'Mounted (9)', 'Targets (9)', 'HA Resources', and 'HA Partner'.

Mounted (9)	Targets (9)	HA Resources	HA Partner
1	0	All	mytest203
0	1	None	mytest202
0	4	None	mytest205
8	4	All	mytest204

3. You will receive a confirmation dialog, click **Take Back** or **Cancel**.

Hand back the resources

Once the issue with the node is corrected and the node is returned to online status, its resources that were handed off to the partner node can now be handed back to the original node from the partner node.

4. Select the partner node that had taken over the resources for the failed node.
5. Click the **Selected Nodes** drop down menu and choose **Hand Back name Resources....**

The screenshot shows the CSSM GUI interface. The 'mytest205' node is selected in the main table. A context menu is open over it, with the option 'Hand Back mytest204 Resources...' selected. The right-hand table shows resource status for various nodes, including 'Mounted (9)', 'Targets (9)', 'HA Resources', and 'HA Partner'.

Mounted (9)	Targets (9)	HA Resources	HA Partner
1	0	All	mytest203
0	1	None	mytest202
0	4	None	mytest205
8	4	All	mytest204

6. You will receive a confirmation dialog, click **Hand Back** or **Cancel**.

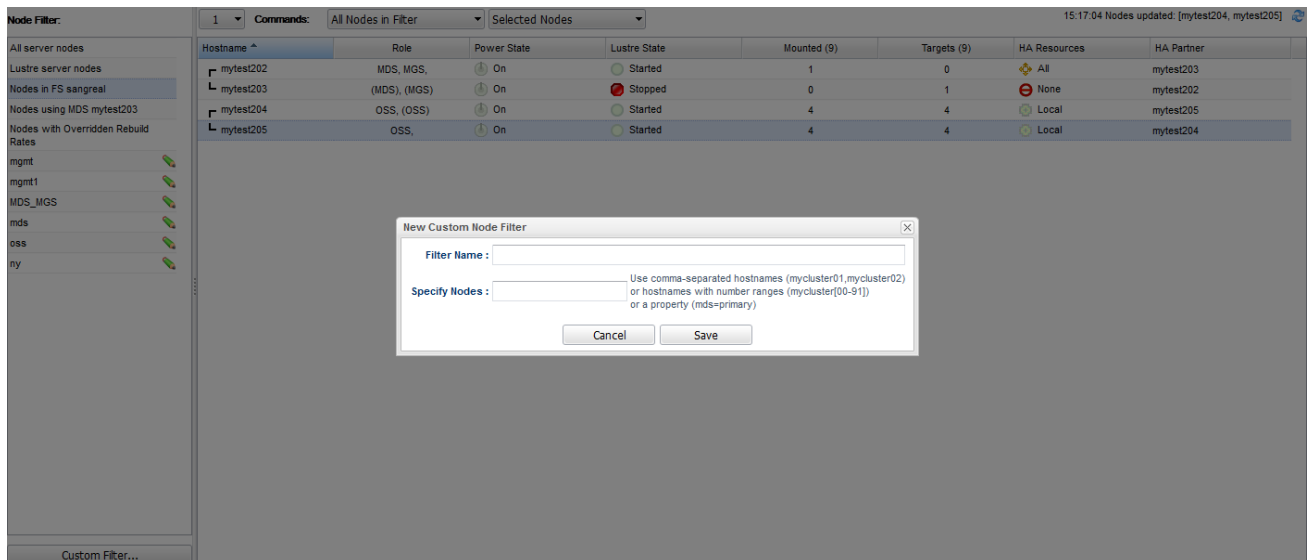
5.4 Using node filters

The left side of the **Node Control** screen is the Node Filter pane. You can select from all server nodes, Lustre nodes, one of the pre-defined filtered nodes, or create your own using the **Custom Filter...** button at the bottom of the list.

The pre-defined sets of filters are: All server nodes, Lustre server nodes, Nodes using MDS *name*, and Nodes in a FS *name*.

5.4.1 Creating a custom filter

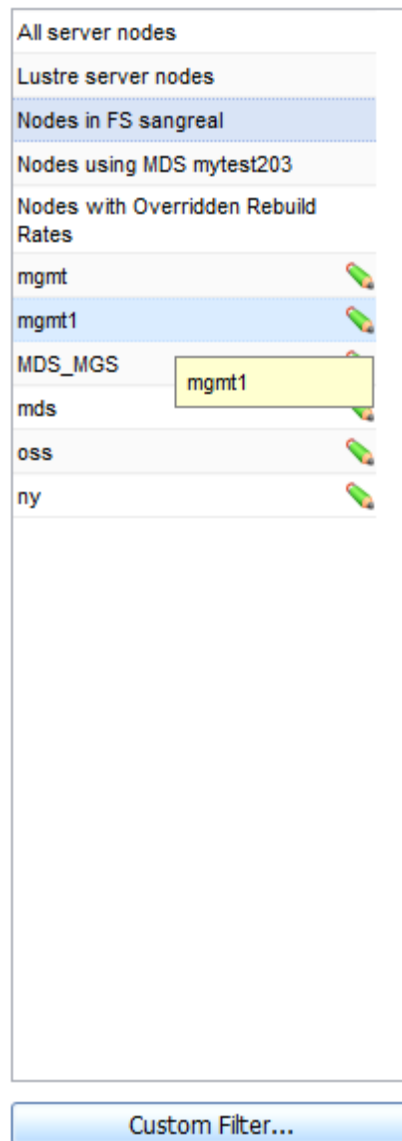
1. To create your own node filter, click the **Custom Filter...** button.



A **New Custom Node Filter** dialog opens to name the custom filter and enter a value or expression for the node(s) you wish to filter upon.

2. Enter a unique name for the filter in the **Filter Name:** field.
The filter name is limited to 64 characters.
3. Specify the nodes to filter upon, such as **mds**, **oss**, etc.
 - Use comma-separated hostnames (**mycluster01,mycluster02**)
 - Hostnames with number ranges (**mycluster[000-091]**)
 - A property (**mds=primary**)
4. Click the **Save** button, and a new filter will appear in the side **Filter** pane.

Custom filters are easy to distinguish from the standard filters, they have a pencil icon adjacent to the item name.



5.4.2 Deleting a custom filter

1. Click the “pencil” icon (Edit - see illustration above) for the custom filter you wish to delete.
2. Click the **Delete** button. A confirmation dialog window will appear, click to confirm the deletion.

5.5 Configuring rebuild rates

This feature provides the option to establish the array minimum rebuild rate (speed) for single or multiple drive failures on a per node basis overriding the global settings from the Configure tab selection.

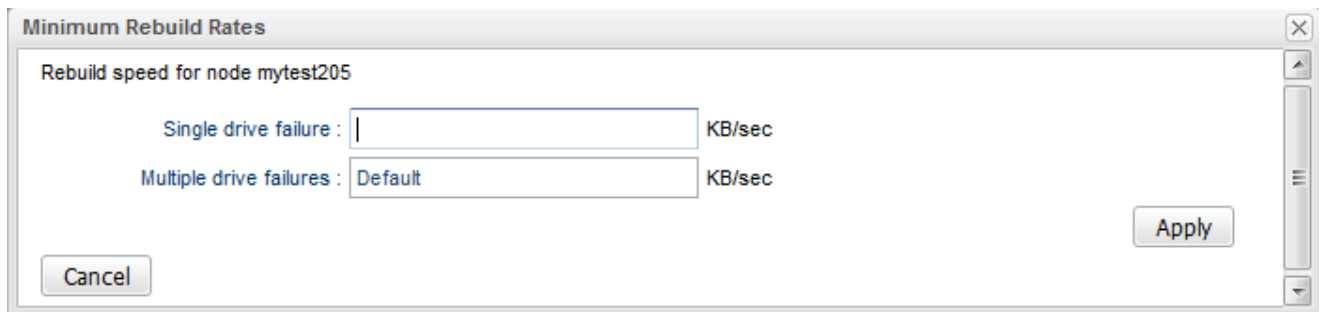
Minimum rebuild rates per node

CSSM has predefined default settings for the minimum array rebuild rates. You can use the minimum rebuild rates setting to override the default settings for a single drive failure and a multi-drive failure.

In the case of a single drive failure, the situation is serious but not as critical since the rebuilding array could sustain another drive failure and still remain operational, therefore the minimum setting for a single drive failure is typically less than that for a multiple drive failure. In the case of a multiple drive failure, you may want to increase the minimum rebuilt rate to minimize exposure to another drive failure.

To set rebuild policy for individual nodes:

1. From the **Node Filter** list, select a custom or default node name.
 Select the node you wish to make rebuild rate policy changes by clicking with the mouse pointer.
2. From the **Selected Nodes** drop down menu, select **Minimum Rebuild Rates...**



3. Enter the rebuild rate expressed in KB/sec allocated to the rebuild for both a single drive failure and multiple drive failures. Although no value is displayed when the screen is displayed, the default setting for a single drive failure is 50000 KB/sec and for multiple drive failures is 80000 KB/sec. Enter the amounts you wish to use for each situation.
4. Click the **Apply** button to continue and save the changes.

NOTE: The GUI does not correctly validate the numbers entered for the Minimum Rebuild Rate field, and may result in an error for very large numbers.

Values in the range of 1000 to 200,000 KB/sec can be entered. Higher values than 200,000 are unlikely to cause shorter overall rebuild time, though I/O performance will be affected. Values under 1000 are unlikely to show an increase in I/O performance during rebuilds. A value of 0 is permitted, but this will halt rebuild activity which increases the risk to stored data.

Most Sonexion operations can be performed using the CSSM GUI or through the Command Line Interface (CLI or cscli) from the **Terminal** tab or a terminal application.

6. Monitoring

6.1 Viewing the host status

On the **Health** tab, use options in the **Status** group at left to display the current network status summary: service status for all host groups, host status details for all host groups, status overview of all host groups and status grid for all host groups.

On the **Host Detail** listings shown below, you can view additional status through the links provided or drill into the host status by clicking the “gear” (perform extra host action) icon.

Looking across the top menu to the right of the **Host Status**, the display shows host counters for their respective states. The general process information is provided next to each status.

Table 4, page 41, is a guide to the colors used to indicate host status.

Host Process

View Links

Commands Unique to Hosts

Host Process (General Process Information)

Export to CSV file

Export to JSON

Link to this page

Service Overview of the Host/Perform Actions

6.1.1 Commands

You may perform several unique commands on each host by clicking the check box on the right hand side and then select a command from the dropdown menu. To locate the host and service for which you need details, scroll down.

Commands for checked services

Select command

Submit

Host	Service	Status	Last Check	Duration	Attempt	Status Information	
mytest202	crmd cpu usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...	<input type="checkbox"/>
	crmd memory usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...	<input type="checkbox"/>
	heartbeat cpu usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...	<input type="checkbox"/>
mytest203	crmd cpu usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...	<input type="checkbox"/>
	crmd memory usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...	<input type="checkbox"/>
	heartbeat cpu usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...	<input type="checkbox"/>

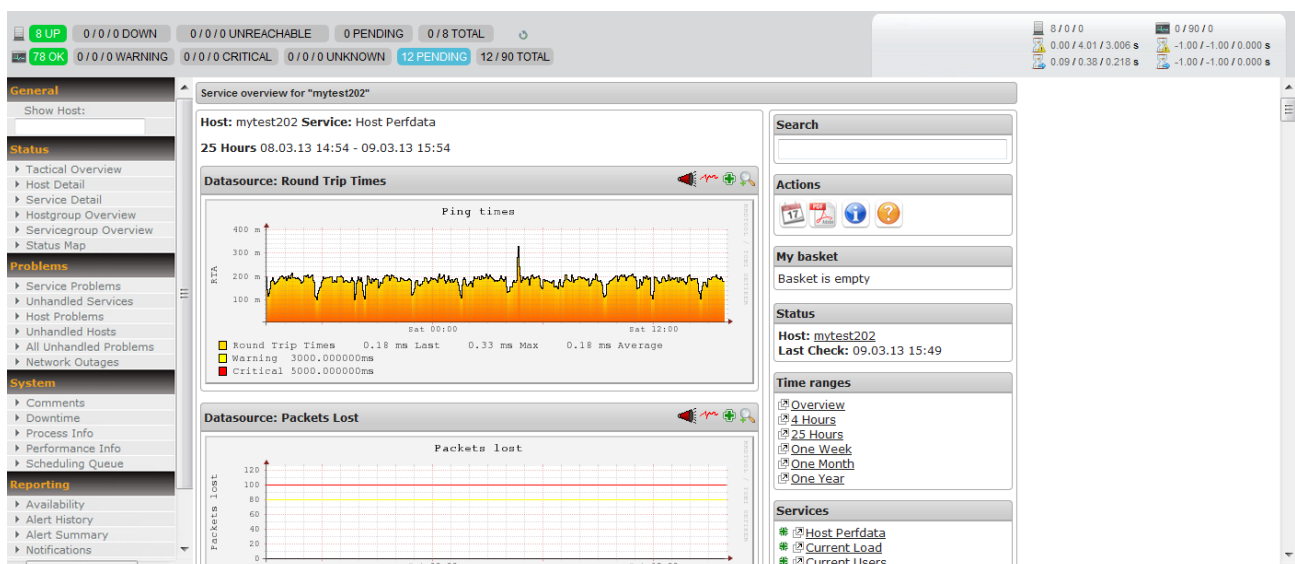
Selecting a node via the check box and clicking the dropdown menu for **Commands for checked host(s)** will provide the following list of commands:



Select a command from the list to act upon the selected host which may be to enable or disable a feature, or start or stop an operation.

6.1.2 Extra host actions

You may also examine the specific host by clicking its “perform extra host action” icon (gear) and the following screen will appear:



From this pane you will find graphs for the different packet states, a search function, general status, time range filters and services. You also have a set of **Action** icons to perform functions on the data such as exporting to a PDF. This feature is provided by PnP4Nagios, a plug-in for Icinga, and provides for charting analysis. **My basket** is located on the right hand side of the window under the **Action** icons.

In the upper right hand corner of each report graph are four icons.



- Megaphone icon - Opens the Most Recent Service Alerts for this host.
- Graphing line icon - Opens the Total Processes report.
- Green Plus symbol icon - Adds to your basket. You can add individual charts to your basket for quick and easy viewing.
- Magnifying glass - Opens the specific graph in a separate window to allow for zooming of details.

6.1.3 Exporting

You may export the results of the checks to several formats such as comma delimited format or JSON, along with establishing a link directly to the panes.

Select a host with the check boxes on the right hand side, then click the JSON icon and a dialog window will open with the option to save the file or open it with an associated application.

Select a host with the check boxes on the right hand side, click the Export to CSV icon (MS Excel icon) and the contents are exported to a new screen. You may copy and paste the contents.

Flapping

Monitoring supports optional detection of hosts that are *flapping*. Flapping occurs when a host changes state too frequently, resulting in a storm of problem and recovery notifications. Flapping can be indicative of configuration problems (i.e., thresholds set too low), troublesome services, or real network problems.

Whenever monitoring checks the status of a host, it will check to see if it has started or stopped flapping. It does this by:

- Storing the results of the last 21 checks of the host.
- Analyzing the historical check results and determine where state changes/transitions occur.
- Using the state transitions to determine a percent state change value (a measure of change) for the host.
- Comparing the percent state change value against low and high flapping thresholds.

A host is determined to have started flapping when its percent state change first exceeds a high flapping threshold.

A host is determined to have stopped flapping when its percent state goes below a low flapping threshold (assuming that it was previously flapping).

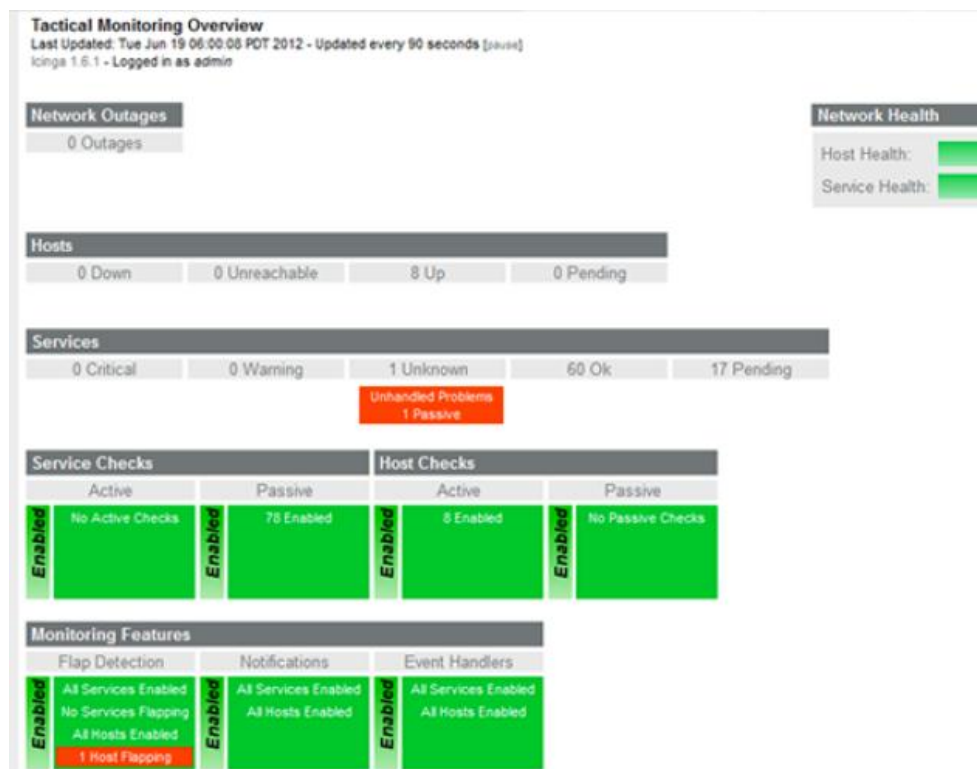
Host flap detection works in a similar manner to service flap detection, with one important difference: Monitoring will attempt to check to see if a host is flapping whenever:

- The host is checked (actively or passively)
- Sometimes when a service associated with that host is checked. More specifically, when at least "x" amount of time has passed since the flap detection was last performed, where "x" is equal to the average check interval of all services associated with the host.

6.1.4 Flap handling

When a host is first detected as flapping, monitoring will:

1. Log a message indicating that the host is flapping.
2. Add a non-persistent comment to the host indicating that it is flapping.
3. Send a “flapping start” notification for the host to appropriate contacts.
4. Suppress other notifications for the host (this is one of the filters in the notification logic).



When a host stops flapping, monitoring will:

1. Log a message indicating that the host has stopped flapping.
2. Delete the comment that was originally added to the host when it started flapping.
3. Send a “flapping stop” notification for the host to appropriate contacts.
4. Remove the block on notifications for the host (notifications will still be bound to the normal notification logic).

6.2 Viewing service status

The Service Status Details link displays the status of all services for each host.

You can view additional status through the links provided or drill into the specific service status by clicking the “perform extra service action” icon (gear).

Looking across the top menu to the right of the **Service Status**, the **Service Status Display** shows host counters for their respective states. The general process information is provided next to each status.

Table 4, page 41, is a guide to the colors used to indicate service status.

Service Status (General Process Information)

Service Status

Status Logistics

View Links

Commands Unique To Services

Commands for checked services

Link to this page

Export to JSON

Export to CSV

Perform Extra Service Action (Gear Icon)

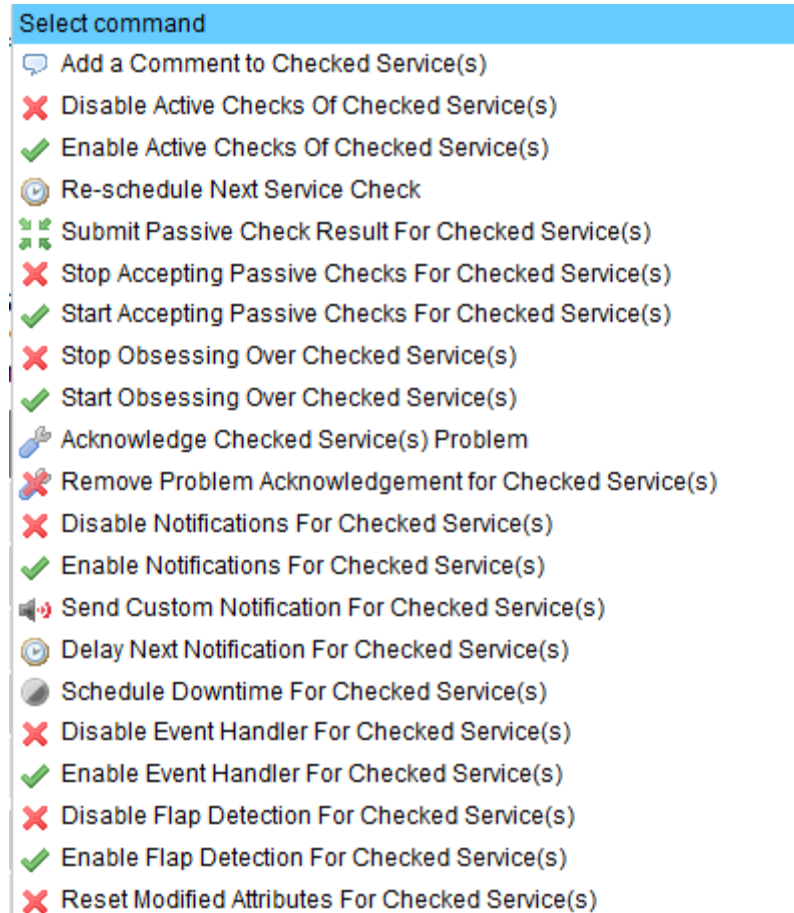
Perform an Active or Passive Check (4-arrow Icon)

Host	Service	Status	Last Check	Duration	Attempt	Status Information
mytes202	cmd cpu usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...
	cmd memory usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...
	heartbeat cpu usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...
mytes203	cmd cpu usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...
	cmd memory usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...
	heartbeat cpu usage	PENDING	N/A	1d 13h 26m 43s+	1/3	Service is not scheduled to be checked...

6.2.1 Commands

You may perform several unique commands on each host service by clicking on the check box on the right hand side for the host to select it, and then choose from the dropdown menu to apply a specific command. To locate the host and service you require details on, scroll down through the list.

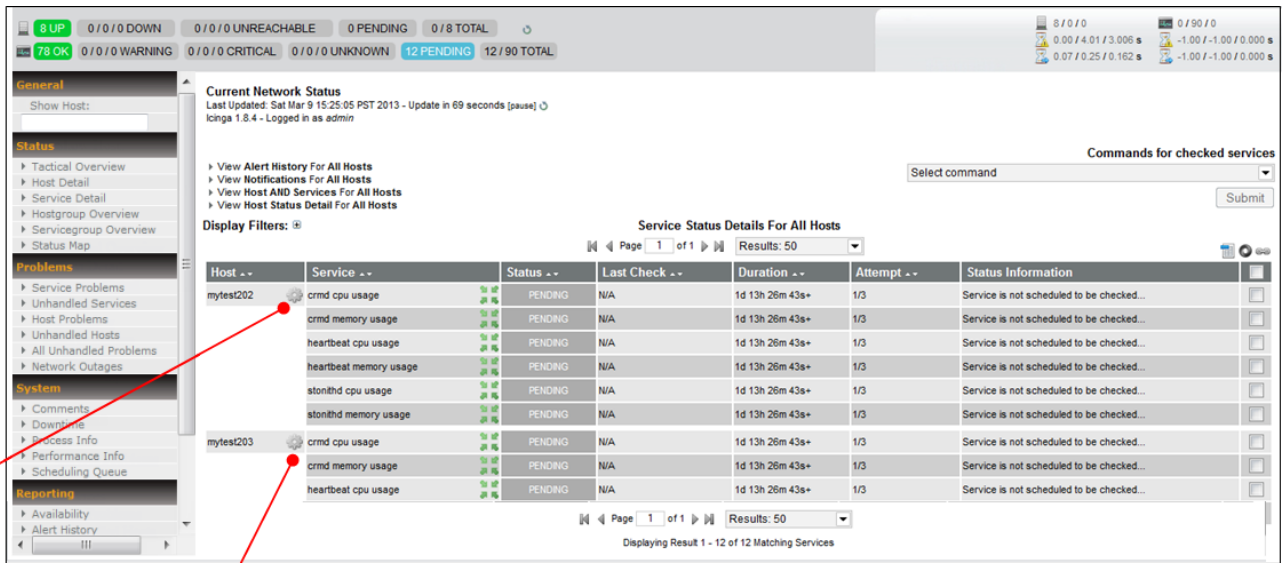
Selecting a node with the check box and clicking the dropdown menu for **Commands for checked services** will provide the following list of commands.



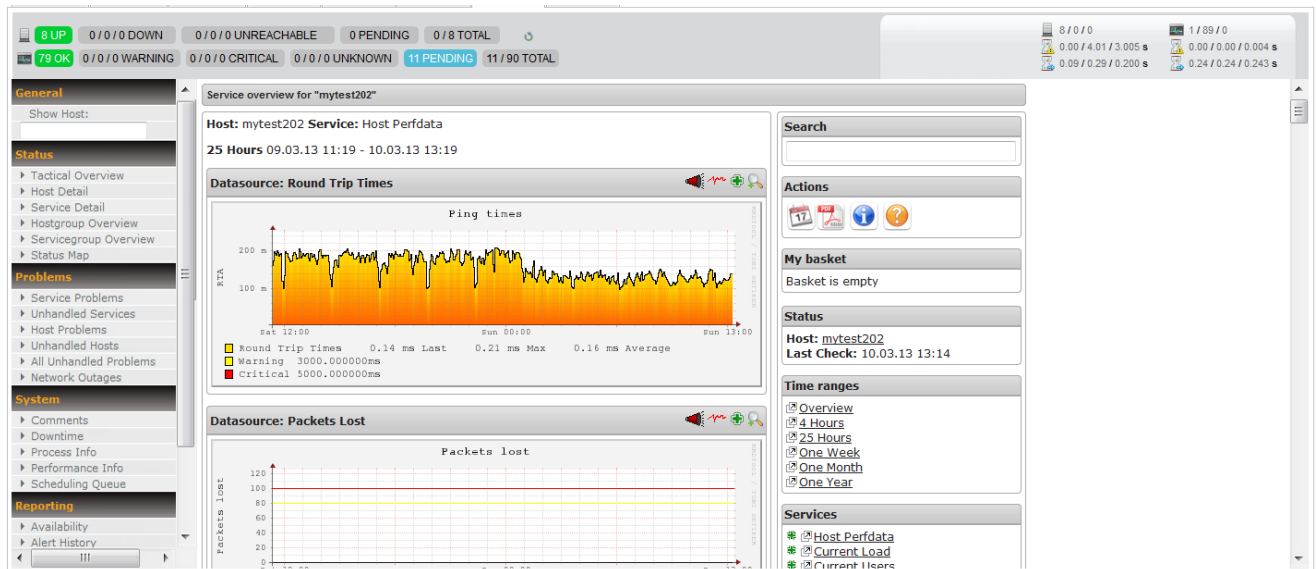
Select a command from the list to act upon the selected host which may be to enable or disable a feature, or start or stop an operation.

6.2.2 Extra service actions

You may also examine the specific host service details by clicking the “gear” icon, “perform extra service action”.






From this pane you will find graphs for the different packet states, a search function, general status, time range filters and services. You also have a set of Action icons to perform functions on the data such as exporting to a PDF.




This feature is provided by PnP4Nagios, a plug-in for Icinga, and provides for charting analysis. **My basket** is located on the right hand side of the window under the Action icons.

In the upper-right corner of each report graph are four icons, shown in the following bar:



-  Megaphone: opens the **Most Recent Service** alerts for this host.
-  Graphing line: opens the **Total Processes** report.
-  Green Plus symbol: adds to your basket. You can add individual charts to your basket for quick and easy viewing.

-  Magnifying glass: Opens the specific graph in a separate window to allow for zooming of details.

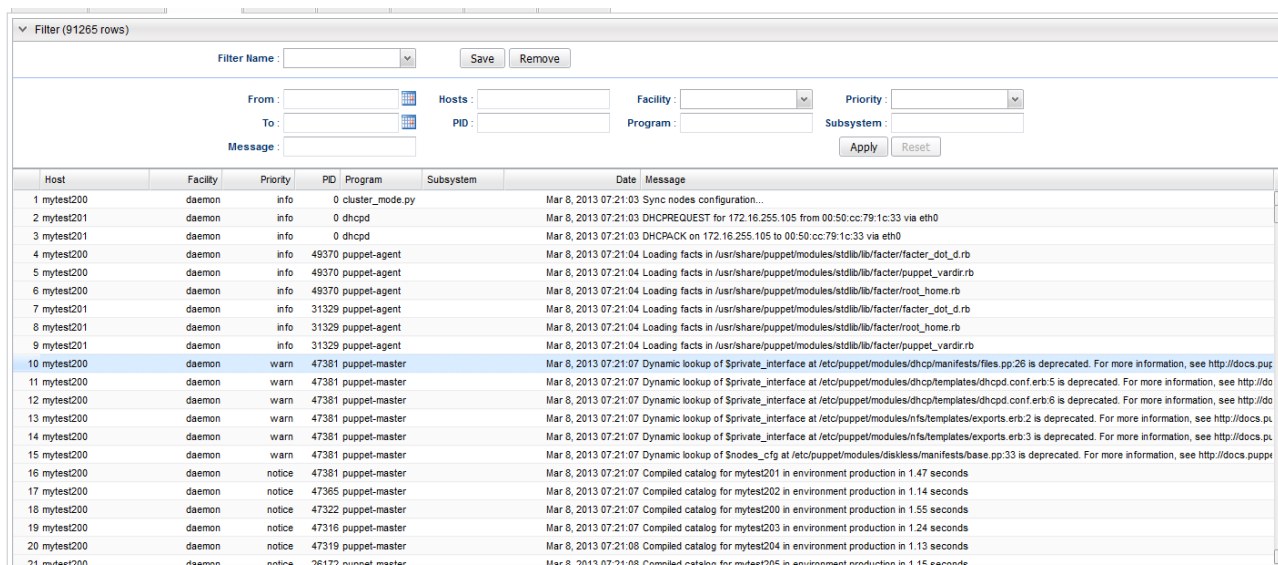
6.3 Viewing log data

Diagnostics and troubleshooting should include viewing the log files. CSSM provides a tab to view the combined logs of all Lustre nodes. To enhance the usability of the logs, a filter is provided. The filter is quite useful to focus on the logs you are searching for as the log files can grow considerably large and it may take time to service them.

Along the header of the table of logs, you may click on the column fields to sort the data. Since these logs may have many thousands of entries, it may take a few moments to sort or display the entries. The “wait” icon displayed in the upper right hand of the screen will appear when the applet is busy.

Clicking the Filter down arrow in the title bar of the Filter section, will collapse the filter options, clicking it again will display them. The value expressed in the parenthesis in the title bar indicates the number of rows of total log messages for the filter applied.

The following is a sample Log Browser screen:



The screenshot shows a web-based Log Browser interface. At the top, there is a title bar that says "Filter (91265 rows)". Below this is a search filter section with a "Filter Name" dropdown, "Save", and "Remove" buttons. Further down are input fields for "From", "To", "Message", "Hosts", "PID", "Facility", "Program", "Priority", and "Subsystem", along with "Apply" and "Reset" buttons. The main part of the screen is a table with the following columns: Host, Facility, Priority, PID, Program, Subsystem, Date, and Message. The table contains 20 rows of log entries, with the 10th row highlighted in blue. The entries include various system messages such as "Sync nodes configuration...", "DHCPREQUEST", "DHCPACK", "Loading facts in /usr/share/puppet/modules/stdlib/lib/facter/facter_dot_d.rb", "Dynamic lookup of \$private_interface at /etc/puppet/modules/dhcp/manifests/files.pp:26 is deprecated.", and "Compiled catalog for mytest202 in environment production in 1.47 seconds".

Host	Facility	Priority	PID	Program	Subsystem	Date	Message
1 mytest200	daemon	info	0	cluster_mode.py		Mar 8, 2013 07:21:03	Sync nodes configuration...
2 mytest201	daemon	info	0	dhcpcd		Mar 8, 2013 07:21:03	DHCPREQUEST for 172.16.255.105 from 00:50:cc:79:1c:33 via eth0
3 mytest201	daemon	info	0	dhcpcd		Mar 8, 2013 07:21:03	DHCPACK on 172.16.255.105 to 00:50:cc:79:1c:33 via eth0
4 mytest200	daemon	info	49370	puppet-agent		Mar 8, 2013 07:21:04	Loading facts in /usr/share/puppet/modules/stdlib/lib/facter/facter_dot_d.rb
5 mytest200	daemon	info	49370	puppet-agent		Mar 8, 2013 07:21:04	Loading facts in /usr/share/puppet/modules/stdlib/lib/facter/puppet_var_dir.rb
6 mytest200	daemon	info	49370	puppet-agent		Mar 8, 2013 07:21:04	Loading facts in /usr/share/puppet/modules/stdlib/lib/facter/root_home.rb
7 mytest201	daemon	info	31329	puppet-agent		Mar 8, 2013 07:21:04	Loading facts in /usr/share/puppet/modules/stdlib/lib/facter/facter_dot_d.rb
8 mytest201	daemon	info	31329	puppet-agent		Mar 8, 2013 07:21:04	Loading facts in /usr/share/puppet/modules/stdlib/lib/facter/root_home.rb
9 mytest201	daemon	info	31329	puppet-agent		Mar 8, 2013 07:21:04	Loading facts in /usr/share/puppet/modules/stdlib/lib/facter/puppet_var_dir.rb
10 mytest200	daemon	warn	47381	puppet-master		Mar 8, 2013 07:21:07	Dynamic lookup of \$private_interface at /etc/puppet/modules/dhcp/manifests/files.pp:26 is deprecated. For more information, see http://docs.pu...
11 mytest200	daemon	warn	47381	puppet-master		Mar 8, 2013 07:21:07	Dynamic lookup of \$private_interface at /etc/puppet/modules/dhcp/templates/dhcpd.conf.erb:5 is deprecated. For more information, see http://do...
12 mytest200	daemon	warn	47381	puppet-master		Mar 8, 2013 07:21:07	Dynamic lookup of \$private_interface at /etc/puppet/modules/dhcp/templates/dhcpd.conf.erb:6 is deprecated. For more information, see http://do...
13 mytest200	daemon	warn	47381	puppet-master		Mar 8, 2013 07:21:07	Dynamic lookup of \$private_interface at /etc/puppet/modules/inf/templates/exports.erb:3 is deprecated. For more information, see http://docs.pu...
14 mytest200	daemon	warn	47381	puppet-master		Mar 8, 2013 07:21:07	Dynamic lookup of \$private_interface at /etc/puppet/modules/inf/templates/exports.erb:3 is deprecated. For more information, see http://docs.pu...
15 mytest200	daemon	warn	47381	puppet-master		Mar 8, 2013 07:21:07	Dynamic lookup of \$nodes_cfg at /etc/puppet/modules/diskless/manifests/base.pp:33 is deprecated. For more information, see http://docs.pu...
16 mytest200	daemon	notice	47381	puppet-master		Mar 8, 2013 07:21:07	Compiled catalog for mytest201 in environment production in 1.47 seconds
17 mytest200	daemon	notice	47365	puppet-master		Mar 8, 2013 07:21:07	Compiled catalog for mytest202 in environment production in 1.14 seconds
18 mytest200	daemon	notice	47322	puppet-master		Mar 8, 2013 07:21:07	Compiled catalog for mytest200 in environment production in 1.55 seconds
19 mytest200	daemon	notice	47316	puppet-master		Mar 8, 2013 07:21:07	Compiled catalog for mytest203 in environment production in 1.24 seconds
20 mytest200	daemon	notice	47319	puppet-master		Mar 8, 2013 07:21:08	Compiled catalog for mytest204 in environment production in 1.13 seconds
21 mytest200	daemon	notice	36172	puppet-master		Mar 8, 2013 07:21:08	Compiled catalog for mytest205 in environment production in 1.15 seconds

See page 37 for more information on the **Log Browser** screen.

6.4 Viewing performance data

The Performance tab enable users to view performance data for any Sonexion node (MDT, OSS or OST). This topic describes how to view performance data and metrics for individual nodes.

6.4.1 Viewing MDT performance data

The Sonexion administrator can monitor compute cluster and node performance from within the Performance tab. For more details on field descriptions, colors and icons, see “Performance tab”, page 33.

Charting of MDS, OST and OSS node performance is provided from within the LMT web applet.

To view MDT performance data:

Click the **Performance** tab, MDT data is presented on the left hand side of the screen.

The screenshot displays the Performance tab interface with three main data panels:

- MDT Performance (Left Panel):** A table showing various operations and their performance metrics.

sangreal-MDT000 2013-03-10 13:24:40.0					
%CPU		%KB		%inodes	
Operation	Samples	Sample /Sec	Avg Value	Std Dev	Units
close	0	0.00	0.00	0.00	reqs
connect	0	0.00	0.00	0.00	reqs
create	0	0.00	0.00	0.00	reqs
destroy	0	0.00	0.00	0.00	reqs
disconnect	0	0.00	0.00	0.00	reqs
getattr	0	0.00	0.00	0.00	reqs
getxattr	0	0.00	0.00	0.00	reqs
link	0	0.00	0.00	0.00	reqs
llog_init	0	0.00	0.00	0.00	reqs
mkdir	0	0.00	0.00	0.00	reqs
mknod	0	0.00	0.00	0.00	reqs
notify	0	0.00	0.00	0.00	reqs
open	0	0.00	0.00	0.00	reqs
process_config	0	0.00	0.00	0.00	reqs
quotactl	0	0.00	0.00	0.00	reqs
reconnect	0	0.00	0.00	0.00	reqs
rename	0	0.00	0.00	0.00	reqs
rmdir	0	0.00	0.00	0.00	reqs
setattr	0	0.00	0.00	0.00	reqs
statfs	0	0.00	0.00	0.00	reqs
unlink	0	0.00	0.00	0.00	reqs
- OST Performance (Middle Panel):** A table showing performance metrics for individual OSTs.

OST 2013-03-10 13:24:40.0					
Ost Name	Read Rate	Write Rate	%CPU	%KB	%inodes
sangreal-OST000	0.00	0.00	0.00	0.00	0.00
sangreal-OST001	0.00	0.00	0.00	0.00	0.00
sangreal-OST002	0.00	0.00	0.00	0.00	0.00
sangreal-OST003	0.00	0.00	0.00	0.00	0.00
sangreal-OST004	0.00	0.00	0.00	0.00	0.00
sangreal-OST005	0.00	0.00	0.00	0.00	0.00
sangreal-OST006	0.00	0.00	0.00	0.00	0.00
sangreal-OST007	0.00	0.00	0.00	0.00	0.00
AGGREGATE	0.00	0.00	0.00	0.00	0.00
MAXIMUM	0.00	0.00	0.00	0.00	0.00
MINIMUM	0.00	0.00	0.00	0.00	0.00
AVERAGE	0.00	0.00	0.00	0.00	0.00
- OSS Performance (Right Panel):** A table showing performance metrics for individual OSSs.

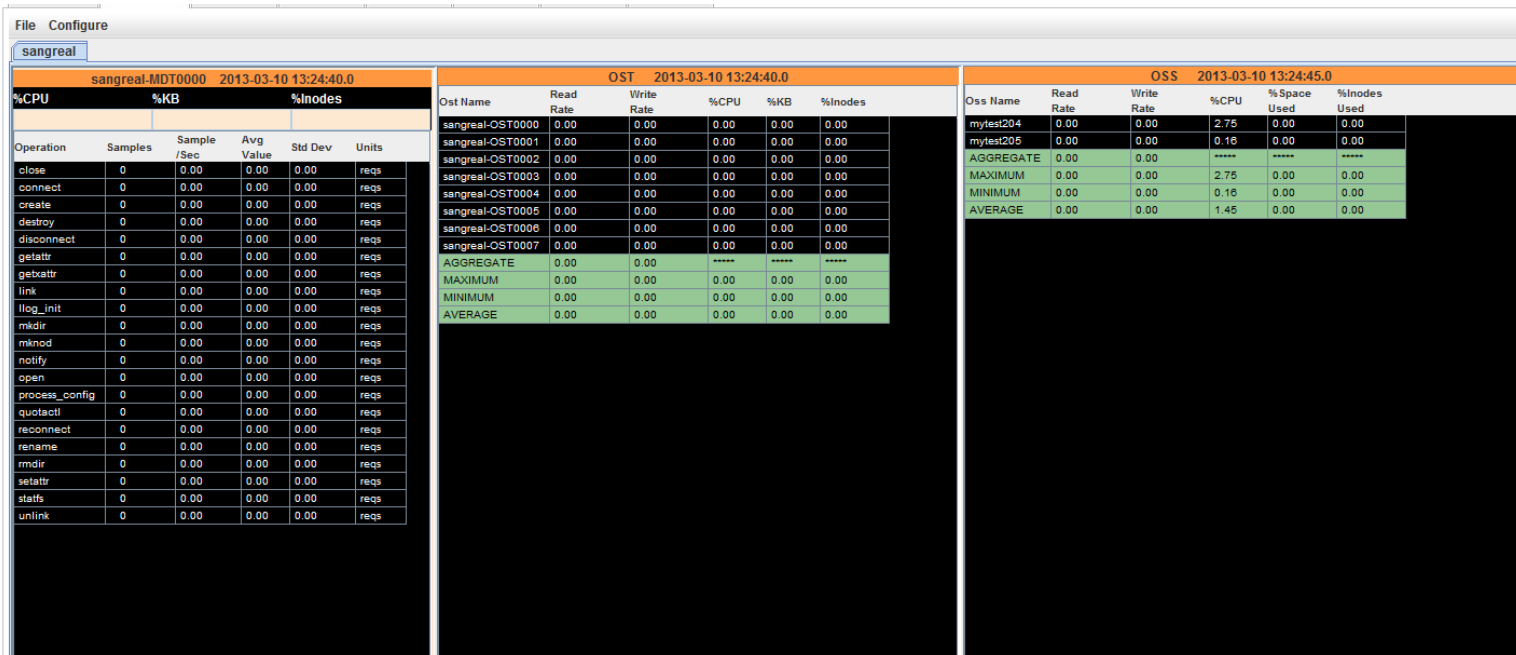
OSS 2013-03-10 13:24:45.0					
Oss Name	Read Rate	Write Rate	%CPU	%Space Used	%inodes Used
mytest204	0.00	0.00	2.75	0.00	0.00
mytest205	0.00	0.00	0.16	0.00	0.00
AGGREGATE	0.00	0.00	0.00	0.00	0.00
MAXIMUM	0.00	0.00	2.75	0.00	0.00
MINIMUM	0.00	0.00	0.16	0.00	0.00
AVERAGE	0.00	0.00	1.45	0.00	0.00

6.4.2 Viewing OST performance data

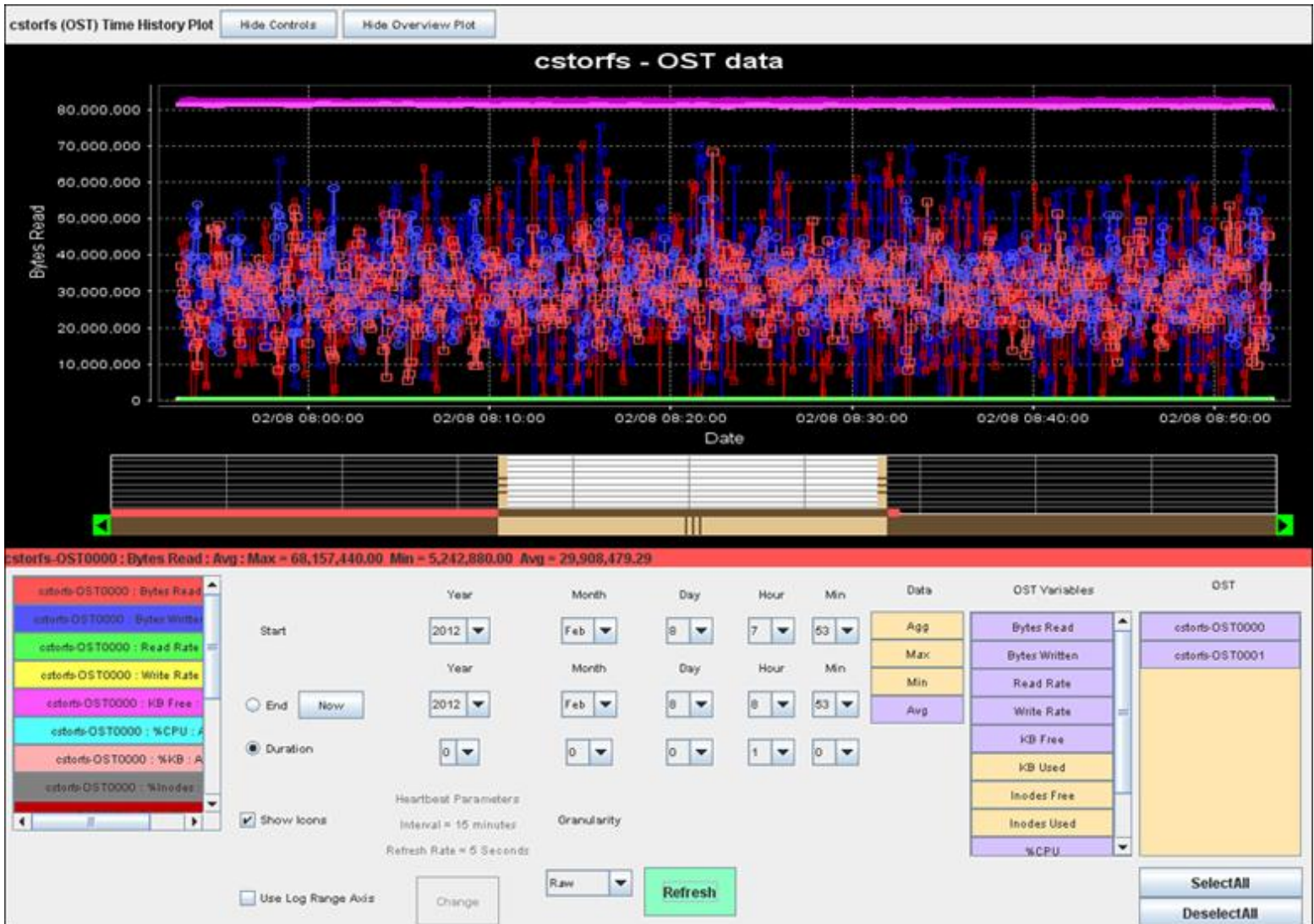
In the OST column (middle pane), performance metrics for individual OSTs are listed (read/write rate, % CPU, and % KB used), along with aggregate performance data (maximum/minimum, aggregate, and average). For more details on field descriptions, colors and icons, see the [Performance](#) topic.

To view OST performance data chart:

From the **Performance** tab screen, double-click any of the OST values. They are **Aggregate**, **Maximum**, **Minimum** and **Average**.



An applet opens and charts the OST performance variables in a plotting graph.



The upper pane displays the OST name and buttons to show/hide the plotting graph (Overview Plot) and performance options (Node Controls).

The lower pane lists performance variables that can be graphed and controls for start/end time or duration time.

6.4.3 Customizing the OST performance view

To customize OST performance views for different Sonexion environments, the plotting graph can be modified to display data for specified OSTs and node variables on a specific timeline. Graphed data appears in the upper pane with variable options shown in the lower pane.

1. See the Performance topic for an explanation of the colors used in the charts.

To customize the view of OST performance data:

2. From the **Performance** tab screen, double-click any of the OST values. They are **Aggregate**, **Maximum**, **Minimum** and **Average**.

The plotting graph and performance variables display.



3. If the performance variables are not visible, click the **Node Controls** button to display the lower pane.
4. Specify the performance data to be graphed.
 - a. Select the **OSTs**.
 - b. Select the **performance variables**.
 - c. Indicate **start/end times** or a **time duration for the time line**.

TIP: To immediately stop graphing data, click the **Now** button.
 - d. Specify **optional** parameters.
 - e. Click the **Refresh** button.

An updated plotting graph displays with data for the selected performance variables and timeline.

6.4.4 File preferences

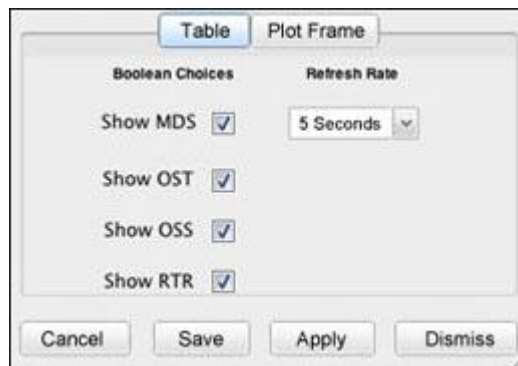
In the upper left corner of the screen you will find two menus, the File and Configure menus.

Within the File menu, the operator has two options **Preferences**" and **Refresh On/Off**.

Preferences has two configurations, **Table** and **Plot Frame**.

Table preferences

Under **Boolean Choices**, you have the option to display the MDS, OST, OSS and RTR (routers) systems in the monitoring table. The standard configuration provides for all to be monitored, however at this release RTR is not fully supported. The three columns of the monitor display as seen in the example are per the selected choices. The first column is the MDS's, the next is the OST's and the third column is the OSS's. If you wish to change the default display place or remove a check mark adjacent your desired options and click the **Apply** button, then click the **Save** button. Close the dialog window.



Set the Refresh rate, the options are 5 seconds, 10 seconds, 15 seconds, and 30 seconds. To change the default select the desired value from the dropdown menu and click the **Apply** button, then click the **Save** button. Close the dialog window.

Plot frame preferences

From the left to right:

- Set the Granularity, from the dropdown menu choose from **Raw**, **Hour**, **Day**, **Week**, **Month**, **Year** and **Heartbeat**.
- Set the Heartbeat Display Interface, from the dropdown menu choose **15 minutes**, **30 minutes**, **1 hour**, **2 hours**, and **4 hours**.
- Set the Heartbeat Refresh Rate, from the dropdown menu choose **5 seconds**, **10 seconds**, **15 seconds**, **30 seconds**, **1 minute**, **2 minutes**, and **5 minutes**.
- Click the **Apply** button, then click the **Save** button. Close the dialog window.



File menu - Refresh

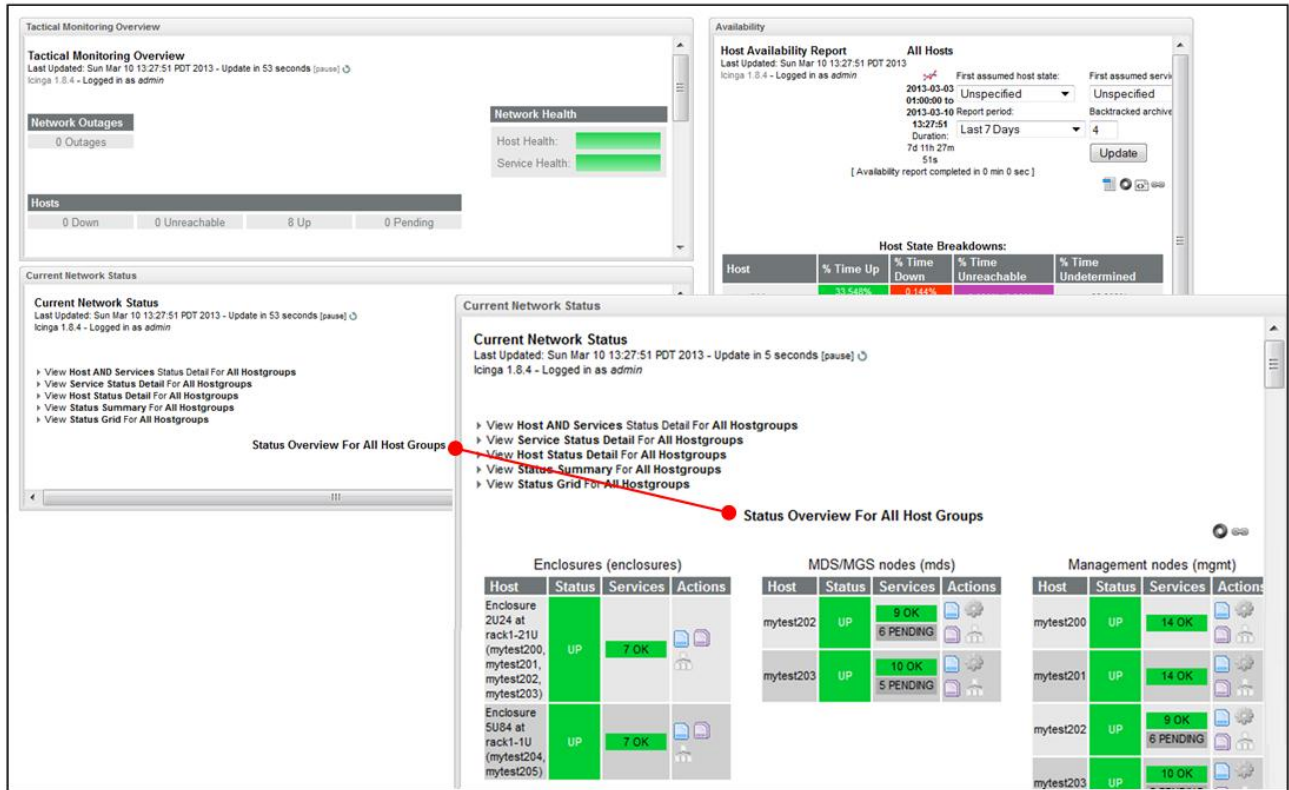
The **File** menu **Refresh** selection is an on/off toggle. Choose it once from the menu and the option is turned On and again the option is turned Off. The default is On.

6.5 Managing a service issue

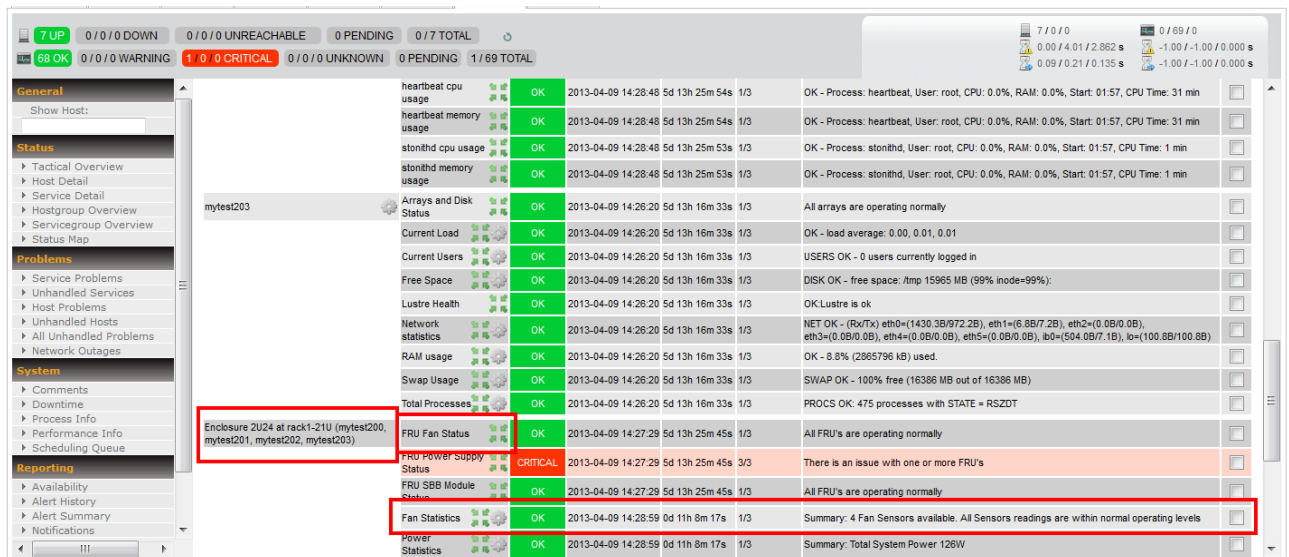
Should one of the nodes experiences an issue, it is reported in the **Service Overview For All Host Groups** pane. When an alert notification is experienced, the operator will acknowledge the alert, and determine the issue and possible cause. Often a RAID array status change would be the issue.

In this example we will use a notification from the Dashboard and look at the details from the Health tab selections to identify the issue and take corrective action.

1. With the Dashboard tab selected as a "daily mode" to monitor the systems, an alert is noticed in the **Current Network Status** panel. From the Current Network Status panel you can do the following:
 - Click the host name to view the **Service Status** details
 - Click the “blue page” icon to access View Extended Information for this Host
 - Click the “gear” icon to “Perform Extra Host Actions”
 - Click the “purple page” icon to access “View Service Details for this Host”



2. Click the **Health** tab, and then from the Tactical Overview (default screen) click the left menu selection for **Hostgroup Overview** under "Status."
3. Depending on which group you will be examining, click the link for the specific group or you may click on the subject node in the Host column.



4. Click on the service that is exhibiting a status issue to drill into the specifics. This information along with the log files should provide sufficient information to diagnose the problem.
5. The following two screens are examples of FRU Fan Status and FAN Statistics:

The screenshot shows the Nagios service status page for 'FRU Fan Status'. The top status bar indicates 7 UP, 0 DOWN, 0 UNREACHABLE, 0 PENDING, 0 TOTAL. The service status is 68 OK, 0 WARNING, 1 CRITICAL, 0 UNKNOWN, 0 PENDING, 1/69 TOTAL. The service is 'Service FRU Fan Status' on host 'mytest203-Enclosure-rack1-21U'. The service state information shows 'Current Status: OK (for 5d 13h 35m 25s)' and 'Status Information: All FRU's are operating normally'. The performance data shows 'Current Attempt: 1/3 (HARD state)' and 'Last Check Time: 2013-04-09 14:37:29'. The service commands section includes options like 'Enable active checks of this service', 'Re-schedule the next check of this service', and 'Submit passive check result for this service'.

The screenshot shows the Nagios service status page for 'Fan Statistics'. The top status bar indicates 7 UP, 0 DOWN, 0 UNREACHABLE, 0 PENDING, 0 TOTAL. The service status is 7 OK, 0 WARNING, 0 CRITICAL, 0 UNKNOWN, 0 PENDING, 1/69 TOTAL. The service is 'Service Fan Statistics' on host 'mytest203-Enclosure-rack1-21U'. The service state information shows 'Current Status: OK (for 0d 11h 20m 0s)' and 'Status Information: Summary: 4 Fan Sensors available. All Sensors readings are within normal operating levels'. The performance data shows 'Current Attempt: 1/3 (HARD state)' and 'Last Check Time: 2013-04-09 14:41:22'. The service commands section includes options like 'Enable active checks of this service', 'Re-schedule the next check of this service', and 'Submit passive check result for this service'.

Should a hardware component require replacement, contact your support representative for servicing the component.

The above example describes only one method to identify a reported issue and steps to diagnose the problem. You will find there are multiple options to access the data presented, and to analyze the health and status of your system.

You may also wish to review the log files that accompany the issues to assist with diagnosing the problem.

6.6 Generating host reports

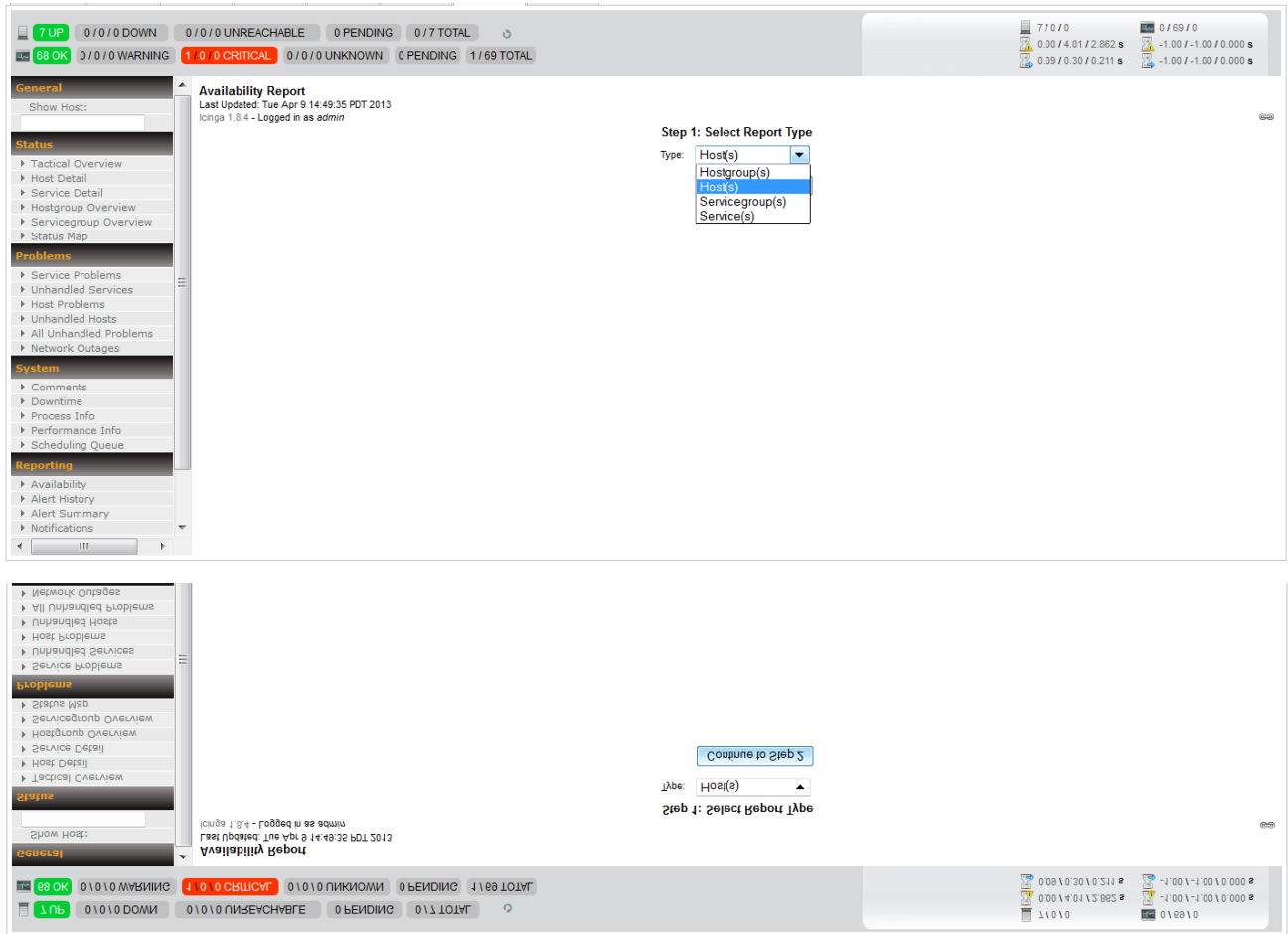
This topic describes the different types of host reports (availability, performance data, and trends) that can be generated from CSSM.

6.6.1 Availability reports

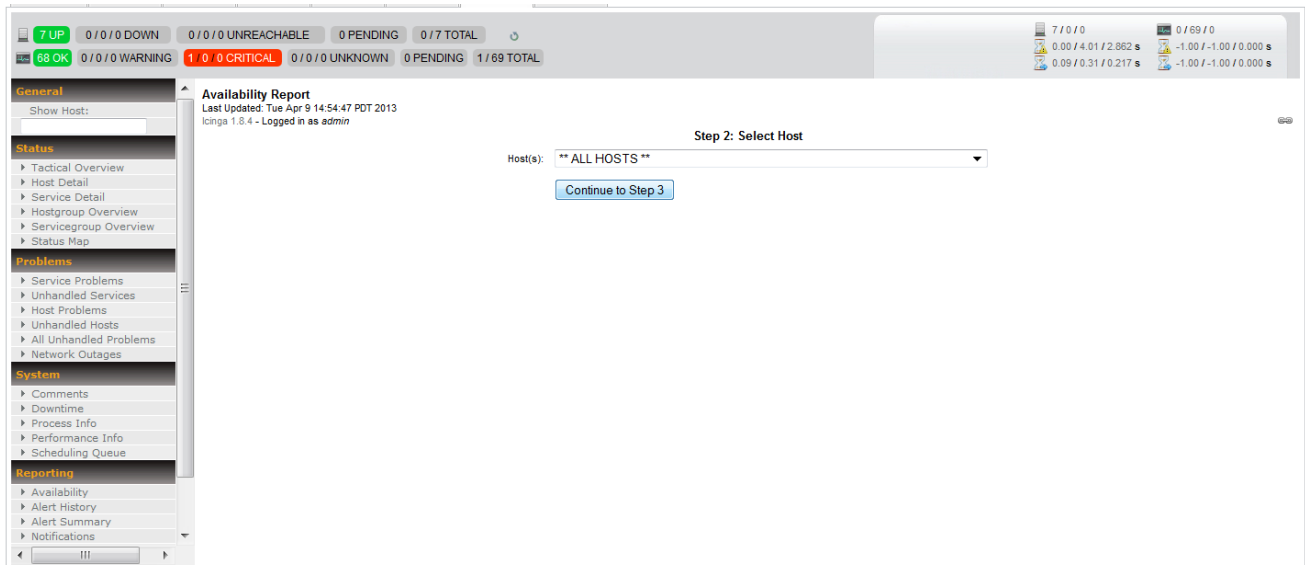
To generate an availability report, do the following:

1. Select the **Health** tab.
2. In the left pane, click **Availability** in the Reporting section.
3. In the Select Report Type area, select **Host(s)** in the dropdown menu and click **Continue to Step 2**.

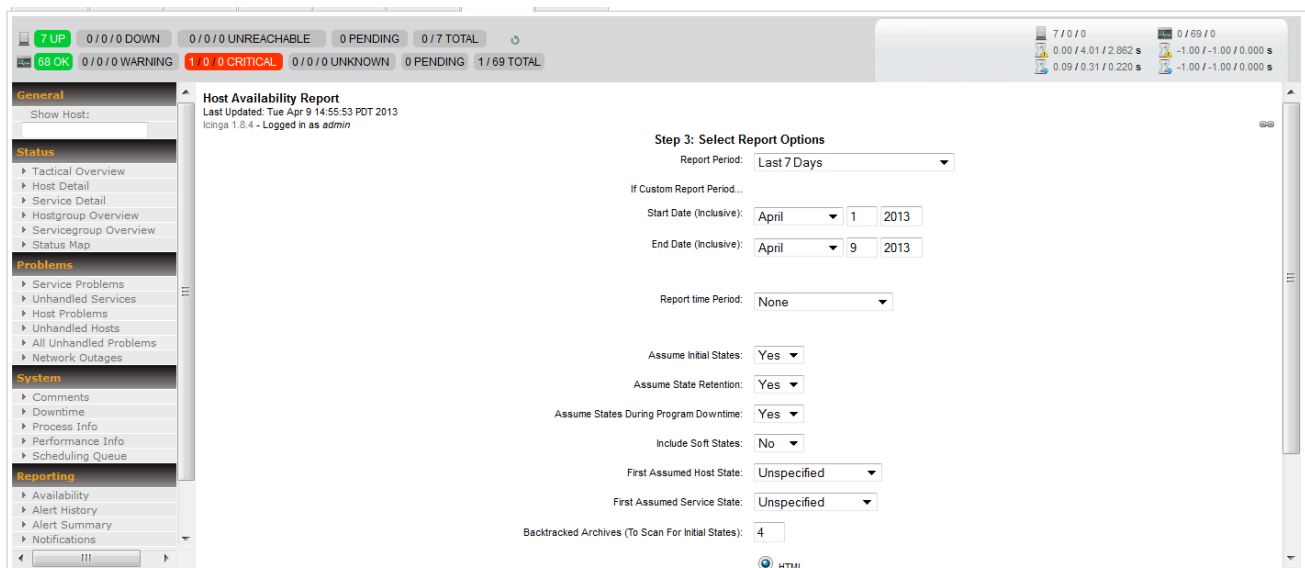
The options are: Hostgroup(s), Host(s), Servicegroup(s), and Service(s)



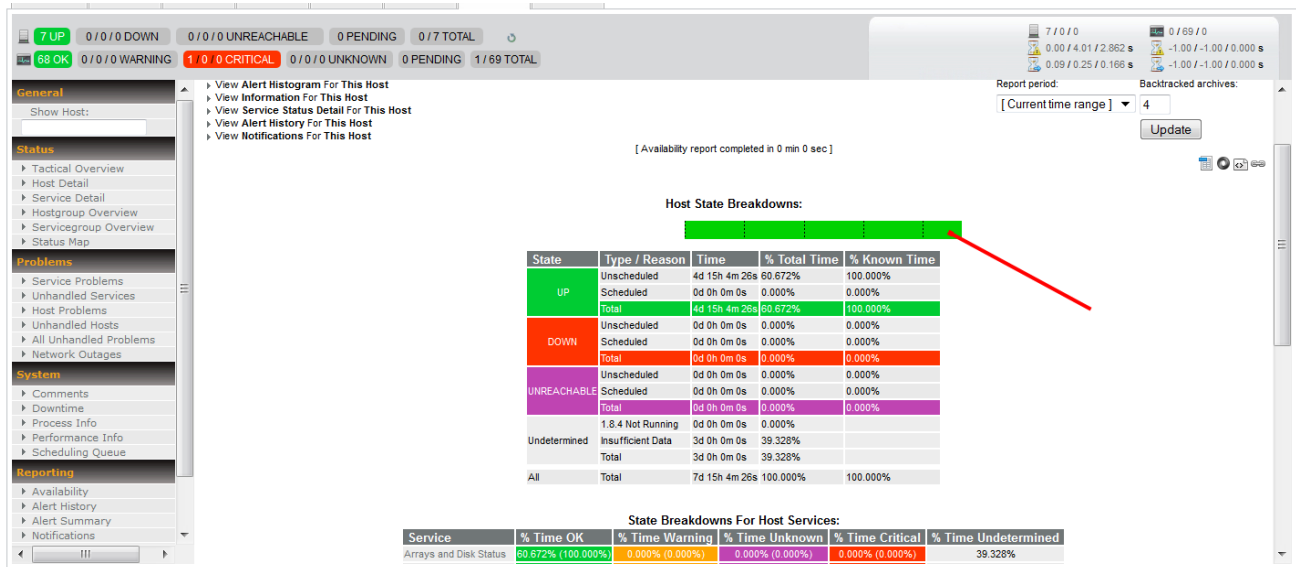
4. In the **Select Host** area, select a Host(s) in the dropdown menu and click **Continue to Step 3**.



5. In the **Select Report Options** screen, select your options and attributes for the report, including the reporting period, various states, and the report format, then click **Create Availability Report**.



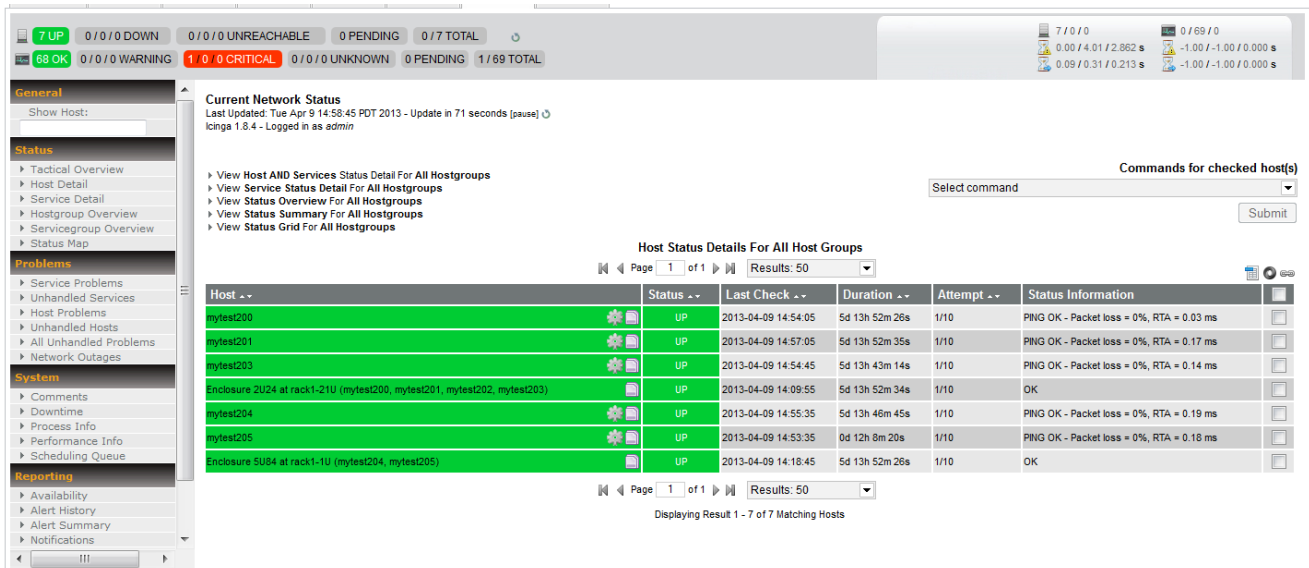
A screen similar to the following displays:



6.6.2 Performance data reports

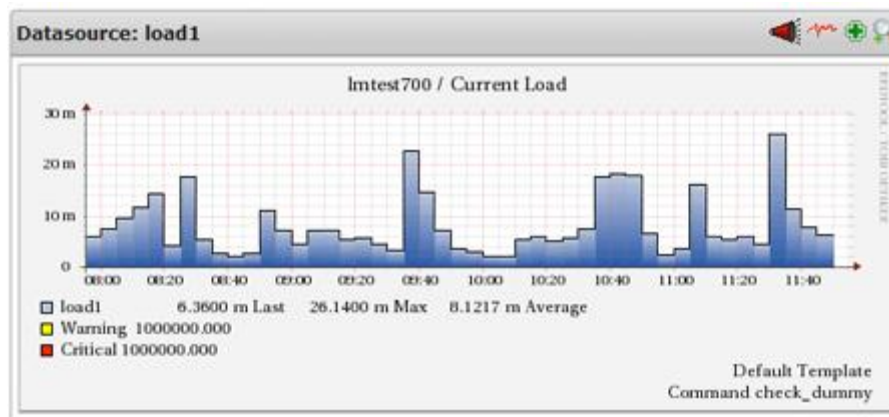
To generate a performance data report, do the following:

1. Select the **Health** tab.
2. In the left pane, click **Host Detail** in the **Status** section.



3. Select the “gear” icon next to the host for which you want to view performance data.

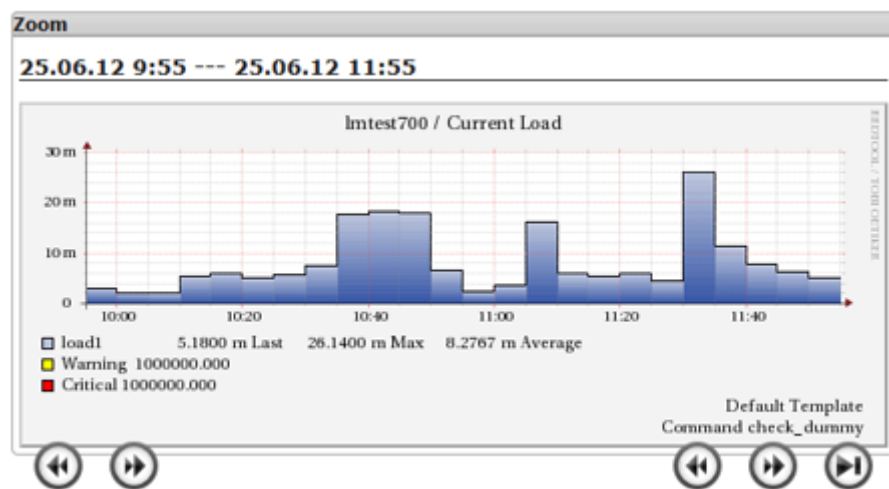
This is a sample screen showing a performance graph for the selected host.



By default, the graphs present data for the entire day.

- To view performance data for a shorter period of time, click on the “magnifying glass” icon in the upper right corner of the window of the graph to zoom in.

This is a detailed sample screen for the load graph shown above.



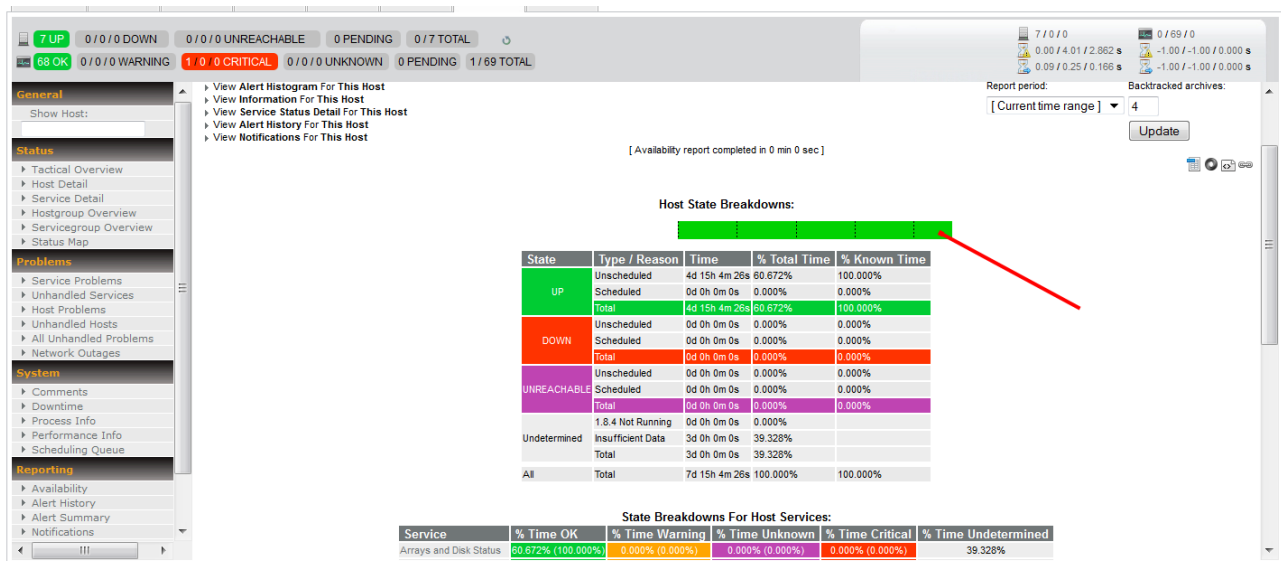
6.6.3 Trend reports

A trend report is a variation of the availability report, and can be accessed in using one of two different methods. To generate a trend report:

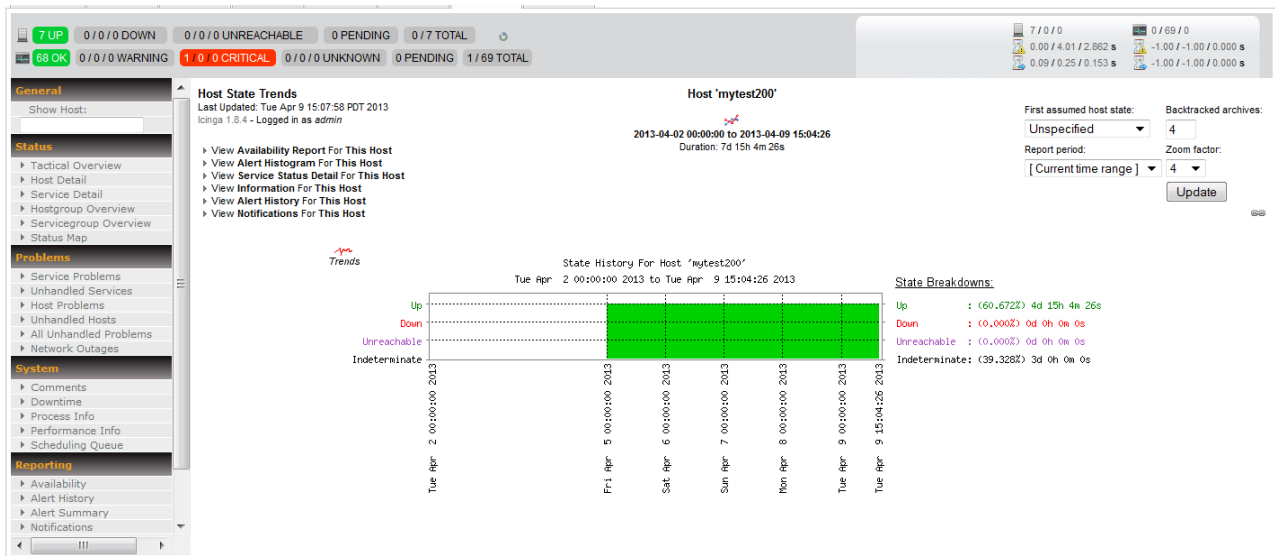
From the Availability Report, create a report as you normally would -

- Select the **Health** tab.
- In the left pane, click **Availability** in the **Reporting** section.
- In the Select Report Type area, select **Host(s)** in the dropdown menu and click **Continue to Step 2**.
- In the Select Host area, select a Host(s) in the dropdown menu and click **Continue to Step 3**.

- In the **Select Report Options** screen, select your options and attributes for the report, including the reporting period, various states, and the report format, then click **Create Availability Report**.



- Click the green block just below the table title, "Host State Breakdowns:" and the **Trend Report** will be created.



OR

- Navigate to https://your_mgs_server_address/icinga/cgi-bin/trends.cgi.
- Select from the dropdown menu either Host or Service, and click **Continue to Step 2** button.

Host and Service State Trends
Last Updated: Mon Jun 25 12:12:30 EDT 2012 - Icinga 1.6.1 - Logged in as
admin

Step 1: Select Report Type

Type:

3. Select a host from the dropdown menu, and click **Continue to Step 3** button.

Host and Service State Trends
Last Updated: Mon Jun 25 12:08:32 EDT 2012 - Icinga 1.6.1 - Logged in as
admin

Step 2: Select Host

Host:

4. Select the options and attributes for the report, including the reporting period and various states and click **Create Report**.

Host Availability Report

Last Updated: Tue Apr 9 15:09:20 PDT 2013
Icinga 1.8.4 - Logged in as admin

Step 3: Select Report Options

Report Period: Last 7 Days

If Custom Report Period...

Start Date (Inclusive): April 1 2013

End Date (Inclusive): April 9 2013

Report time Period: None

Assume Initial States: Yes

Assume State Retention: Yes

Assume States During Program Downtime: Yes

Include Soft States: No

First Assumed Host State: Unspecified

First Assumed Service State: Unspecified

Backtracked Archives (To Scan For Initial States): 4

Output Format:

HTML

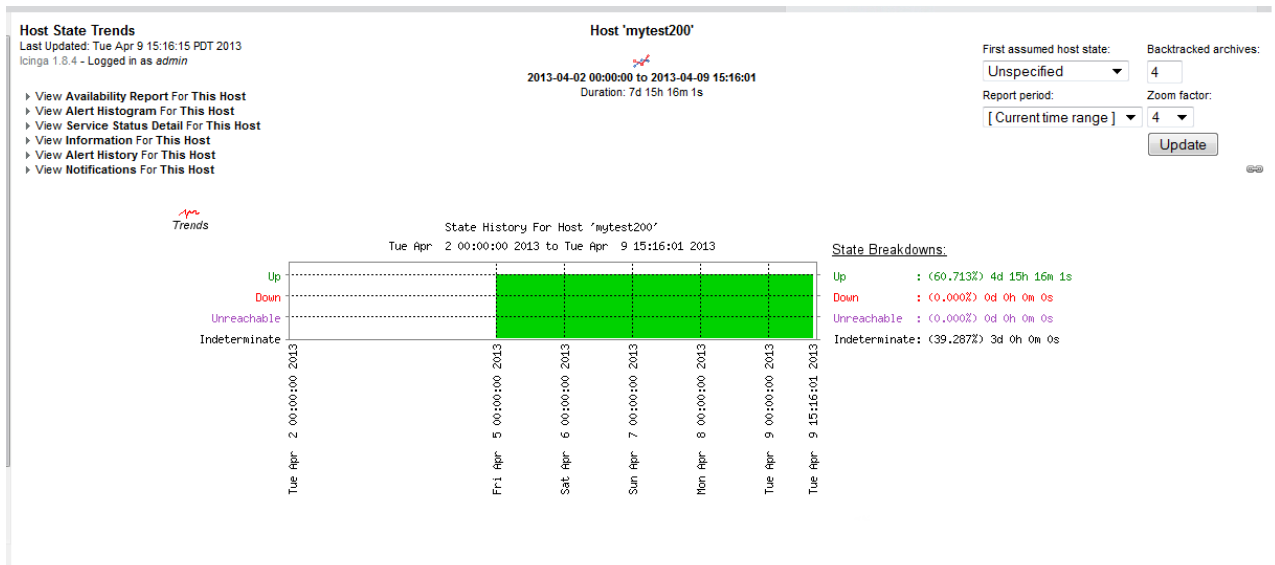
CSV

JSON

XML

Create Availability Report!

There is a slight difference in the way the Trends Report is presented whether you are using the Health tab or a URL.



6.7 Understanding thresholds

The following topic describes the thresholds established for each service. Each description should provide information to help understand the different states of monitoring.

Changes to some thresholds can be accomplished from cscli commands, refer to the [Managing Operations via CLI](#) topic.

6.7.1 Thresholds for general-purpose services

The following are terms used in the discussions related to thresholds:

Term	Description
mgmt	The management node, for Icinga "management" is the localhost.
all	All physical nodes (except virtual "enclosure" nodes).
all-but-mgmt	All physical nodes except management nodes, for Icinga those are remote nodes.
oss	Object Storage Server (OSS).
mgs	Metadata Server (MDS).
enclosure	A virtual node for the enclosures, they are virtual because there is no physical server that can be pinged.

Current load

The service analyzes three load average values for the past 1 minute, 5 minutes, and 15 minutes - load1, load5, load15. These are then compared to warning and critical thresholds.

Example

Thresholds:

- Warning load1, load5, load15 thresholds are: 5.0, 5.0, 5.0
- Critical load1, load5, load15 thresholds are: 10.0, 10.0, 10.0

Example values:

If the load average is...			The status is:	Because:
Load1	Load5	Load15		
0.65	0.43	0.53	OK	All values are below Warning.
0.65	5.43	0.53	WARNING	Load5 > warning and < critical
0.65	5.43	10.53	CRITICAL	Load15 > critical

Status Information example

OK - load average: 0.25, 0.31, 0.68

WARNING - load average: 0.25, 5.31, 0.68

CRITICAL - load average: 0.25, 5.31, 10.68

Metrics

Returns three metrics - load1, load5, and load15.

Node	Status	Thresholds
all	Warning	1000000
	Critical	1000000

The value "1000000" refers to infinity. It is essentially an unreachable threshold to prevent warning and critical states from occurring. This service provides useful metrics for statistics.

Current users

Checks the number of users currently logged into the local system and generates an error if the number exceeds the thresholds specified.

Status Information example

USERS OK - 1 user currently logged in

USERS WARNING - 20 users currently logged in

USERS CRITICAL - 100 users currently logged in

Metrics

Returns the number of users currently logged in.

Node	Status	Thresholds
all	Warning	10

Critical	50
----------	----

Network statistics

Checks the network statistics. Only status is indicated, no thresholds are supported.

Status Information example:

NET OK - (Rx/Tx) eth0=(367.7B/215.4B), eth1=(110.4B/78.2B), lo=(22.2B/22.2B)

where "lo, eth0, and eth1" are interface names

367.7B/215.4B - input and output traffic in bytes B, kilobytes kb, MB, GB, or TB.

Metrics

Returns input and output traffic values for each network interface.

Node	Status	Thresholds
all	OK	The service is always OK, and is used to collect metrics history.

Ping node

This check uses the ping command to check connection statistics for a remote host. Packet loss (expressed in percentage) and round trip average (in milliseconds) are analyzed.

The thresholds are:

- The round trip average travel time (ms) - rta
- Percentage of packet loss - pl

Status Information example

PING OK - Packet loss = 0%, RTA = 0.78 ms

CRITICAL - Could not interpret output from ping command

Metrics

The *rta* and *pl* values are returned.

Node	Status	Thresholds
all	Warning	rta=CS-1300.0, pl=80%
	Critical	rta=5000.0, pl=100%

Root partition

Checks the amount of used disk space of the root partition on each node and generates an alert if free space is less than one of the threshold values. The threshold is the percentage of free disk space remaining.

Status Information example

DISK OK - free space: / 4770 MB (72% inode=86%):

DISK CRITICAL - free space: / 514 MB (7% inode=85%):

Metrics

Used disk space in MB is returned.

Node	Status	Thresholds
all	Warning	20% free space
	Critical	10% free space

RAM usage

Checks RAM usage on each node. Threshold is the percentage of used memory.

Status Information example

OK - 20.1% (386364 kB) used

Metrics

- Total memory
- Used
- Free
- Caches

Node	Status	Thresholds
all	Warning	80%
	Critical	90%

Swap usage

Checks swap space on each node. Threshold is the percentage of free swap space left.

Status Information example

SWAP WARNING - 85% (866 MB out of 1027 MB)

Metrics

Used swap space in MB is returned.

Node	Status	Thresholds
mgmt and mds	Warning	50% free swap space
	Critical	30% free swap space
oss	Warning	99%
	Critical	50%

Total processes

Checks all processes and generates WARNING or CRITICAL states if the number of processes is outside the required threshold ranges. Checks the process that have, in the output of "ps", one or more of the following status flags: RSZDT.

Status Information example

PROCS OK: 316 processes with STATE = RSZDT

Metrics

Total number of processes in RSZDT.

Node	Status	Thresholds
all	Warning	1000000
	Critical	1000000

The value "1000000" refers to infinity. It is essentially an unreachable threshold to prevent warning and critical states from occurring. This service provides useful metrics for statistics.

CPU/memory usage

Checks if a process consumes too much processor time or memory. The check is performed via ps. Threshold is the percentage of overall CPU or memory consumed by a process. There are six services which checks CPU and memory consumption by four different processes:

- Heartbeat memory usage
- Heartbeat cpu usage
- STONITHD memory usage
- STONITHD CPU usage
- CRMD memory usage

- CRMD CPU usage

Status Information example

OK - Process: crmd, User: 498, CPU: 0.0%, RAM: 0.2%, Start: Feb14, CPU Time: 255 min

Metrics

Total number of processes in RSZDT.

Node	Status	Thresholds
mgmt	Warning	5% of CPU or memory consumed
	Critical	10% of CPU or memory consumed

Lustre Health

Checks Lustre file system health.

Status Information example

OK: Lustre OK

CRITICAL: Lustre critical

UNKNOWN: Lustre status is unknown

Metrics

None.

Node	Status	Thresholds
oss, mds	OK	If Lustre reports that it is "Healthy."
	Critical	If Lustre reports that it is "Not Healthy."
	Unknown	Lustre reports neither "Healthy" or "Not Healthy."

Array and disk status

This status is calculated by the results status of this check based on the worst status for the arrays. For example, if a node has 1 array OK, 1 array Degrade and 1 array Critical, then the results of the check will be Critical. If some internal error has occurred, then the status is Unknown and you will probably see some form of python exception in the status message.

What is displayed in the message field?

Array node checks will always display all node array status in the status message field. Only the first line displayed in "Service Detail." If you click on a specific service, you will get information about all arrays (this check has multiple line information).

If the array status differs from OK, then all related disks to this array having a status different from OK are displayed, as they may be causing the array to fail.

If the array status is OK, then no information about the disks is displayed.

Status Information example

- Example 1: (OK)
 - Total number of disk slots available: 24
 - Total number of disks found: 24
 - Array: md1, status: OK
 - Array: md0, status: OK
- Example 2: (WARNING)
 - Total number of disk slots available: 24
 - Total number of disks found: 24
 - Array: md1, status: Degraded
 - Slot: 22, wwn: 5000cca01301b7bc, cap: 100030242304, dev: sdw, parts: 0, status: Failed
 - Array: md0, status: OK
- Example 3: (CRITICAL)
 - Total number of disk slots available: 24
 - Total number of disks found: 24
 - Array: md1, status: Failed
 - Slot: 22, status: Unplugged
 - Slot: 23 status:Unplugged
 - Array: md0, status: OK
- Example 4: (UNKNOWN)
 - Traceback (most recent call last):
 - File "*stdin*", line 1, in *module*
 - IndexError

Metrics

None.

Node	Status	Thresholds
all	Warning	If any array at host is degraded.
	Critical	If any array at host is failed.
	Unknown	No arrays found, received unsupported output of dm_report.

SES sensors

SES (SCSI Enclosure Services) elements are part of the enclosure, and therefore belong to the "Enclosure" hosts, not to the node hosts. The list of sensors are:

- Thermal Statistics
- Power Statistics
- Voltage Statistics
- Fan Statistics

There is one service for each type of sensor. The Status Information of the service is a multi-line output. The first line of the Status Information is overall status of all sensors of this type. The other lines are details for each individual sensor, one per line.

The sensors have internal thresholds, but those thresholds usually must not be changed. For example if the maximum allowed temperature is set for a device, it means that the device will become corrupt at higher temperatures. The minimum and maximum thresholds for each sensor are included in the Status Information.

The status of the service is the worst status among the individual sensors.

Status Information example

Summary: 4 Fan Sensors available. All sensors readings are within normal operating levels

Fan0-OK:3370RPM, min:2000RPM, max:20480RPM;

Fan1-OK:2700RPM, min:2000RPM, max:20480RPM;

Fan2-OK:3450RPM, min:2000RPM, max:20480RPM;

Fan3-OK:2620RPM, min:2000RPM, max:20480RPM;

Power Statistics does not display status and is used for power consumption metrics.

Summary: 4 Fan Sensors available. All sensors readings are within normal operating levels

Power PSU 0 5V 30W

Power PSU 0 12V 53W

Power PSU 1 5V 23W

Power PSU 1 12V 60W

Metrics

The service provides one metric per sensor:

Fan0=3370RPM Fan1=2700RPM Fan2=3450RPM Fan3=2620RPM

Node	Status	Thresholds
enclosure	Warning	If the value of any of the sensors exceeds the threshold.

FRU

FRU is an acronym for Field Replaceable Unit. The list of monitored FRUs are:

- FRU SBB Module Status (formerly Enclosure_Electronics)
- FRU Fan Status (formerly Cooling)
- FRU Power Supply Status (formerly PSU)

Individual FRU devices are grouped by the device type into common services similar to sensors.

Note that fans are both sensors and FRUs. They have two separate services, "Fan Statistics" (sensor) and "FRU Fan Status" (FRU). Those services reflect different aspects of the state of the same physical devices. A pair of fans in an SSU represents a single FRU.

Status Information example

- All FRUs are operating normally
 - Fan0-1: OK
 - Fan2-3: OK
- There is an issue with one or more FRUs.
 - PSU0: CRITICAL: Unrecoverable
 - PSU1: OK

Metrics

None

Node	Status	Thresholds
enclosure	OK	OK
	Warning	NON_CRITICAL NOT_INSTALLED
	Critical	CRITICAL UNRECOVERABLE
	Unknown	UNKNOWN NOT_AVAILABLE

7. Support Files

To successfully debug many Sonexion problems, Cray personnel must collect various event logs from field systems. Sonexion provides a mechanism for collecting sets of these logs, called support bundles. The mechanism to collect support bundles may be initiated manually or may be triggered automatically by certain events (for example, Lustre bugs and failover events). These support bundles should be provided to Cray personnel in the course of requesting technical support.

A support bundle is a standard UNIX-compressed file (**tar-gzip**) that aggregates the following system files:

- System logs for all nodes for a variable length of time before an error occurred.
45 minutes is the default time period for the data collected in a support bundle. This value can be changed for collecting a support bundle manually.
- A list of all cluster nodes and the following information for each node:
 - Software version
 - Linux kernel version and patches
 - Sonexion RPMs
 - OSTs mounted on the node
 - RAID states
 - FRU configuration dump
 - GEM dump
 - IPMI information
 - Power states
 - Resource states
 - Relevant processes
 - Serial connection log
 - HA state

- Syslog messages (/var/log/messages)
- The current Apache/WSGI logs from the MGMT nodes
- Application state data (MySQL database dump)
- Diagnostic and performance test logs

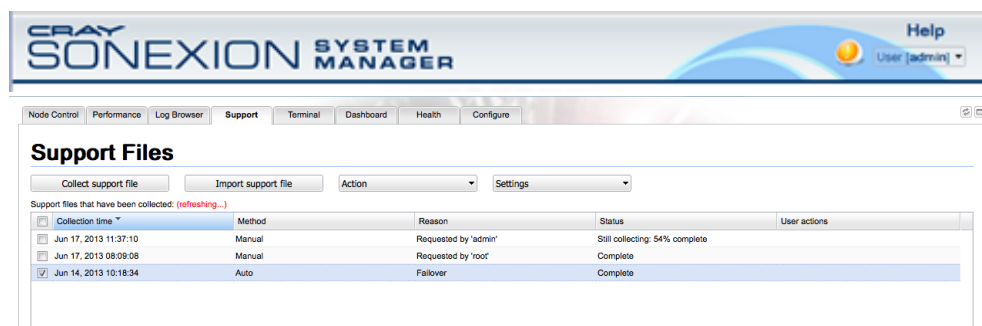
7.1 Obtain a support bundle via the CSSM GUI

When certain types of Sonexion errors occur, support bundle collection is triggered automatically and users cannot terminate or cancel the operation. Alternately, a user can start data collection and create a support bundle manually. Unlike the automatic process, a manual data collection operation can be canceled.

IMPORTANT: Before performing the procedures in this section, verify that all pop-up blockers in the browser are disabled. Pop-up blockers will prevent the CSSM's dialog boxes from displaying correctly.

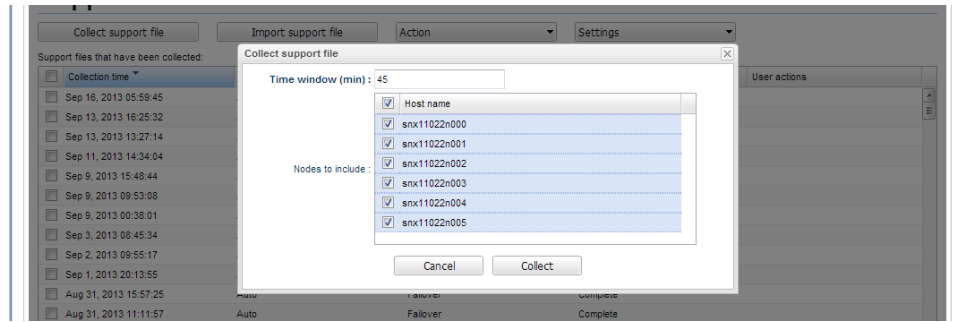
To manually collect a support bundle via the CSSM GUI:

1. Log into the CSSM.
2. Click the **Support** tab to open the **Support Files** screen (see the following figure).



3. Click **Collect support file**.
The **Collect support file** screen opens, displaying a list of all nodes in the cluster.
4. Specify the data collection parameters for the support file:
 - a. Select the **time period** to look back for syslog data to be collected. The default time is 45 minutes.
 - b. Select the **nodes** for which data will be collected.

TIP: Check the box next to **Hostname** to select all nodes.



5. Click **Collect**.

The data collection process starts, using the specified parameters. A status indicator will show the operation's completion percentage. For example, "Still collecting, 64% complete." When data collection is complete, the status will indicate, "Done."

To terminate the data collection operation at any point, click **Cancel**.

When the operation is complete and the support file has been created, the **Support Files** screen will refresh and display an entry for the newly collected support file.

See the section 7 introduction, page 96, for details about the information that is collected in a support bundle.

7.2 Download a support bundle via CSSM GUI

Use the download file feature to save a local copy of a selected support bundle.

1. In CSSM, click the **Support** tab.

The **Support Files** screen displays and lists the available support bundles.

2. Select the support bundle to download.
3. In the row above the support files list, click **Action**.

Available user actions will be displayed as shown below.

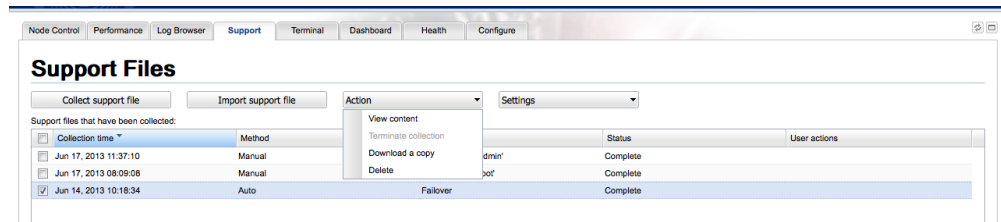


Figure 2 - Support Tab Actions

4. Click **Download a Copy**.

A dialog box opens and displays several file options.

5. Specify a location to save the file on your system.
6. Select **Save File** and click **OK**.

The support bundle downloads to the specified location.

7.3 Get a support bundle via cscli

To collect a support bundle manually using CLI commands:

1. Log into the primary MGMT node via SSH. Run:

```
$ ssh -l admin primary_MGMT_node
```

2. Change to root user. Run:

```
$ sudo su -
```

3. Collect the support bundle.

- To collect the bundle using the default 45 minute time period, run:

```
[root@n000]# cscli support_bundle -c
```

Here is sample output:

```
[root@dvtrack200 ~]# cscli support_bundle -c
Collecting support bundle: id:4, nodes:all, time-window:45
minute(s)
```

- To collect the bundle with a different time period, run:

```
[root@n000]# cscli support_bundle -c -t minutes
```

Here is a sample command and output:

```
[root@dvtrack200 ~]# cscli support_bundle -c -t 90
Collecting support bundle: id:4, nodes:all, time-window:90 minute(s)
```

4. Use the `scp` command to retrieve the collected support bundle, which can be found in the `/var/lib/xyratex/CSSM_cust_debug_bundles/` directory on the active MGMT node.

7.4 Reference for cscli support_bundle command

The following command reference from *Sonexion 1.3.1 CLI User Documentation* is repeated here for convenience.

The `support_bundle` command manages support bundles and support bundle settings.

Synopsis

```
$ cscli support_bundle [-h] [-c] [-n nodes] [-t minutes] [-e bundle_id]
[--disable-trigger trigger] [--get-purge-limit] [--set-purge-limit percents]
```

where

Option	Description
<code>-h</code> <code>--help</code>	Displays the help message and exits.
<code>-c</code> <code>--collect-bundle</code>	Collects the support bundle.
<code>-n NODES</code> <code>--nodes nodes</code>	Shows a comma-separated list of nodes. Default value is "all nodes".
<code>-t MINUTES</code>	Specifies the time window to collect data for the support bundle (in minutes). Default value is 45 minutes.

Option	Description
--time-window <i>minutes</i>	
-e BUNDLE_ID --export-bundle <i>bundle_id</i>	Identifies an export-specified support bundle. This is similar to “Download copy” functionality in CSSM.
--show-triggers	Shows the triggers that initiate automatic collection of support bundles. Current triggers are “LBUG” and “Failover”.
--enable-trigger <i>trigger</i>	Enables a specific trigger for automatic collection.
--disable-trigger <i>trigger</i>	Disables a specific trigger for automatic collection.
--get-purge-limit	Shows the purge limit as a percentage of free file system space. If the purge limit is reached, the Sonexion system purges old support bundle files.
--set-purge-limit <i>percents</i>	Sets the purge limit as a percentage of free file system space.

7.5 Interpret Sonexion support bundles

This section contains an overview of the support bundle contents. Support bundles contain two types of logs: system-wide logs that collect data for the entire system, and node-specific logs that collect data for an individual node.

7.5.1 System-wide logs

- **lbug_syslog.csv**

This file contains syslog messages, in comma-separated value (CSV) format.

NOTE: The following files are not intended for use by Sonexion end users, but they may be valuable to Cray personnel and OEMs to better understand system states and behavior.

- **logs/access.log**

This log contains Apache HTTP access data.

- **logs/data_tables.sql**

This log contains a dump of MySQL database tables. The tables describe internal structures used to manage the cluster, the state of cluster resources, information about hardware, software, firmware, and network configuration, a FRU inventory, etc. The database dump contains all information required to recreate the system state at the time when the support bundle was created.

- **logs/error.log**

This file contains the Apache error log.

- **logs/wsgi_access.log**

This mod_wsgi access log contains records of web service calls made from the CSSM.

7.6 Node-specific Logs:

- **nodes/nodename/conman.log**

This log contains console data captured by CONsole MANager (Conman), a daemon that provides centralized access to node SOL (serial over LAN, IPMI) or real serial consoles. It also provides logging, broadcasting to several consoles or shared console sessions.

- **nodes/nodename/crm.log**

This log contains state data for the RAID and Lustre resources as seen by Pacemaker, an open-source, high-availability resource manager that is suitable for small and large clusters.

- **nodes/nodename/dmesg.log**

This log contains a dump of kernel messages collected from the node.

- **nodes/nodename/fru_dump.yaml**

This file contains an inventory of FRUs for the enclosure hosting the node (DDICs, PSU, fans, power supplies, etc). The dump file includes serial numbers for individual FRU equipment, firmware versions, and states such as **OK** or **Failure**.

- **nodes/nodename/lspci.log**

This log contains a list of PCI devices in a free-form text format generated by the lspci tool. lspci lists PCI devices and their characteristics. lspci can be run in standard or verbose (`-vvv` option) mode.

- **nodes/nodename/mdstat.log**

This log contains state data of the MDRAID arrays, i.e., content of the `/proc/mdstat` file.

- **nodes/nodename/processes.csv**

This file contains a list of processes, a snapshot of 'top', which is a standard monitoring program that reports the top consumers of CPU or memory.

- **nodes/nodename/sgmap.log**

This log contains a list of sg devices and specifies for each device the SCSI address, firmware version, and corresponding block devices.

- **nodes/nodename/software_versions.csv**

This file contains a list of all installed packages with version information (`rpm -qa` output).

- **nodes/nodename/states.csv**

This file contains miscellaneous state data, including power, memory, uptime, CPU load, and Lustre targets.

8. Troubleshooting

This section provides troubleshooting information for the Sonexion system and describes installation and post-installation issues and workarounds. This document also outlines Lustre performance and tuning considerations, CSSM, Networking, RAID, and High Availability (HA).

8.1 Lustre performance considerations and tuning

8.1.1 Prerequisites

- Pre-flight check: Make sure all firmware on tested hardware is at latest stable version and there are no known kernel performance issues related to the hardware.
- Catalog problem areas: Single (slow) disk drives can slow down storage arrays. CPUs can slow down storage arrays. Buggy interconnect drivers can reduce bandwidth and increase roundtrip times.

8.1.2 Hardware performance testing

1. Start from the bottom and work your way up (this is critical).
2. For a full picture of hardware capabilities:
 - Test individual components
 - Test components collectively
3. Single disks:

- Use `dd` and `sgp_dd` (`sgpdd-survey`) - great tools
 - Test various block sizes
 - Test while other disks are being tested at the same time
4. Arrays
 - Use `dd` and `sgpdd-survey`
 - Test various block sizes
 5. OSTs and MDTs
 - OST testing
 - Use `obdfilter-survey` to directly and effectively test OSTs
 - MDT testing
 - Mounting the block device as `ldiskfs` allow you to perform `MDTEST` testing, as well as other methods directly against the MDT -- just **remember to remove the test files once finished.**
 6. Interconnects
 - Lustre LNET self-test is a great tool to test one or more nodes on your network
 - LNET is protocol-neutral and runs at or near full wire speed

8.1.3 Benchmarking interconnect

- Establish a test baseline
 - Are your results consistent with earlier tests?
 - Test different nodes on different switching equipment
 - Test across switching equipment
 - Test multiple nodes? Are the test results expected?
 - Adjust tunable parameters (`max_rpcs_in_flight`, `max_dirty_mb`, etc.)
- How do different parameters affect the test results?

8.1.4 Benchmarking - RAID tuning

- Making Lustre aware of the RAID layout can dramatically improve performance (especially on RAID6 solutions)
- Consider a RAID6 (6+2) configuration consisting of 64kB strides (made of 4kB blocks): $(64\text{kB} / \text{stride}) / 4\text{kB}/\text{blocks} = 16$ blocks per stride $(64\text{kB} * 6 \text{ stripes}) / 4\text{kB}/\text{blocks} = 96$ blocks

Specify:

```
--mkfsoptions="-E stride=16,stripe_width=96"
```

Use the `mkfs.lustre` option when initially formatting the file system (this could be specified in installation YAML file).

8.1.5 Benchmarking - direct MDT and OST testing

- MDT testing:
 - MDTEST utility generates lots of small I/O activity against the MDT.
 - To improve metadata performance testing, mount the same Lustre file system multiple times on the clients.
 - Run multiple MDTEST iterations over a specific time period to establish minimum / maximum performance characteristics of the MDT.
- OST testing:
 - `obdfilter-survey` provides detailed data of OST read/write performance through the Lustre block device interface driver (however it may, under some configurations result in severe issues with cluster health, i.e. nodes may get in a panic state).
 - OST pools feature isolates individual OST sets on selected OSSs. OST pools are especially useful when:
 - Only limited clients are available
 - To determine optimal ratios of OSTs to OSSs
 - To locate interconnect bottlenecks between OSSs

8.1.6 Benchmarking - single client testing

- Measure performance characteristics on a single-client basis first
- Test various workloads
- Utilize different striping schemas
- Utilize small and large block size reads / writes to establish where ideal performance numbers can be achieved
- Use tools such as IOR and IOzone to simulate different types of loads

8.1.7 Benchmarking - multi-client testing

- Performance expectations for multi-client testing should be based on the results of earlier single client testing
- Determine the number of clients needed to fully saturate a single OSS
Use OST pools to control the number of clients
- Use MPI IO to collect accurate performance data
- Use IOzone (preferred over IOR) for multi-client testing
- Different tools generate different results which should not be compared to one another
For example, do not compare IOR results against IOzone results (apples to oranges)

8.2 Management software issues

8.2.1 Warning while unmounting Lustre: “Database assertion: created a new connection but pool_size is already reached”

Release 1.2.0

Problem description

The following error message appears after a CCLI unmount:

```
unmount: Database assertion: created a new connection but
pool_size is already reached (4 > 3)!
```

Workaround

This warning indicates the occurrence of a connection leak, meaning that a CSSM instance is using more than three database connections due to a bug in the management software. This warning is benign and can be disregarded.

8.2.2 Invalid puppet certificate on diskless node boot-up

Releases 1.2.0, 1.2.1, 1.2.3, 1.3.1

Problem description

If attempts to login to an OSS node using known-good credentials fail, the node is probably experiencing puppet connection problems. Use this procedure to clear up the puppet configuration on the node:

Workaround

1. SSH into the primary MGMT node.
2. Sudo to root, by entering:


```
[admin@n000]$ sudo su -
```
3. Revoke the certificate for the OSS node and remove the certificate files from the management node, by entering:

```
[root@n000]# puppetca -clean OSS_nodename
```

4. SSH into the OSS node.

If the attempt to SSH into the node succeeds, go to Step 5.

If the attempt to SSH into the node fails, run the command:

1.2.x:

```
[root@n000]# find /var/lib/puppet/ssl_persistent \
    -namehostname.* -delete
```

1.3.1 or 1.4.0:

```
[root@n000]# ssh nfsserv find
    /mnt/nfsdata/var/lib/puppet/ssl_persistent/ -name
    hostname.* -delete
```

Reboot the OSS node (physically or using conman), wait until the node is accessible via ssh, and then go to Step 9.

5. Sudo to root, by entering:

```
[OSS node]$ sudo su -
```

6. Remove the SSL certificate and private key from the OSS node, by entering:

```
[OSS node]# rm -rf /var/lib/puppet/ssl/*
```

7. Run the puppet client. This will regenerate the private key and request a new signed certificate from the management node, by entering:

```
[OSS node]# puppetd -tv
```

8. Exit back out to the management node, by entering:

```
[OSS node]# exit
```

9. Populate the persistent storage with the node's certificate and private key, by entering:

For 1.2.x, run:

```
[root@n000]# rsync -zaHv
    --numeric-idshostname:/var/lib/puppet/ssl/
    /var/lib/puppet/ssl_persistent
```

For 1.3.1 or 1.4.0, run:

```
[root@n000]# ssh nfsserv rsync -zaHv
    --numeric-id hostname:/var/lib/puppet/ssl/
    /mnt/nfsdata/var/lib/puppet/ssl_persistent
```

10. The certificate and associated private key files are regular files, like any other. To verify that the persistent directory has the right files, run:

```
for i in $(nodeattr -s diskless); do
    diff -q <(ssh $i cat /var/lib/puppet/ssl/certs/$i.pem)
    /var/lib/puppet/ssl_persistent/certs/$i.pem 2> /dev/null
    || echo cert for $i is not correct in persistent storage
done
```

The above command checks that the certificate file for each diskless node is the same in that nodes `/var/lib/puppet/ssl/certs` directory and the `/var/lib/puppet/ssl_persistent/certs` directory.

11. Verify that the current puppet certificate is valid. Run:

```
[root@n000]# puppetd -tv
```

12. Using the Legacy HotFix Checker, determine if the HotFix 1.2.0-TRT-2 (the time synchronization hotfix) is installed. If it is not, then install it.

8.2.3 Need to change LDAP settings after GUI/wizard is complete

All releases

Problem description

Provide ability to change the LDAP settings post installation.

Workaround

Run this command from the MGS node:

```
[MGS]# /opt/xyratex/bin/beLDAPConfig.sh -H "host" -b "BaseDN"
-p "UserDN" -g "GroupDN"
```

8.2.4 Unclean shutdown of management node causes database corruption

Release 1.2.0

Problem description

If the management node hosting the MySQL server (usually node 0, the primary MGMT server) is shut down uncleanly, the LMT database (named `filesystem_<filesystem_name>`) can become corrupt. This can manifest in several ways, including out-of-date information in the performance tab and problems accessing the management database `t0db`. There will usually be errors in the file `/var/log/mysql.log` indicating which tables are corrupt:

```
130220 8:20:28 [ERROR] /usr/libexec/mysqld:
Table './filesystem_snx11003/MDS_OPS_DATA' is marked as crashed and
should be repaired
130220 8:20:43 [ERROR] /usr/libexec/mysqld:
Table './filesystem_snx11003/MDS_OPS_DATA' is marked as crashed and
should be repaired
130220 8:20:58 [ERROR] /usr/libexec/mysqld:
Table './filesystem_snx11003/MDS_OPS_DATA' is marked as crashed and
should be repaired
```

Workaround

To repair the corrupted tables, execute the following procedure, saving output using the script command. Note that for very large tables, repair operations can create temporary files that are larger than the available filesystem space. It is recommended that the available space be monitored during this procedure.

1. As root, check all tables in the `filesystem_<filesystem_name>` database:

```
[root@havantfae00 filesystem_fs1]# mysqlcheck filesystem_fs1
filesystem_fs1.EVENT_DATA      OK
filesystem_fs1.EVENT_INFO      OK
filesystem_fs1.FILESYSTEM_AGGREGATE_DAY
```

```

error      : Size of indexfile is: 15360 Should be: 18432
error      : Corrupt
filesystem_fs1.FILESYSTEM_AGGREGATE_HOUR
error      : Size of indexfile is: 224256 Should be: 25CS-1600
error      : Corrupt
filesystem_fs1.FILESYSTEM_AGGREGATE_MONTH    OK
filesystem_fs1.FILESYSTEM_AGGREGATE_WEEK     OK
filesystem_fs1.FILESYSTEM_AGGREGATE_YEAR     OK
filesystem_fs1.FILESYSTEM_INFO               OK

```

Additional output omitted

2. Repair all tables:

```

[root@havantfae00 filesystem_fs1]# mysqlcheck -s -r filesystem_fs1
[root@havantfae00 filesystem_fs1]#

```

3. Verify that repair worked and that all tables are OK:

```

[root@havantfae00 filesystem_fs1]# mysqlcheck filesystem_fs1
filesystem_fs1.EVENT_DATA      OK
filesystem_fs1.EVENT_INFO      OK
filesystem_fs1.FILESYSTEM_AGGREGATE_DAY    OK
filesystem_fs1.FILESYSTEM_AGGREGATE_HOUR  OK
filesystem_fs1.FILESYSTEM_AGGREGATE_MONTH  OK
filesystem_fs1.FILESYSTEM_AGGREGATE_WEEK  OK
filesystem_fs1.FILESYSTEM_AGGREGATE_YEAR  OK
filesystem_fs1.FILESYSTEM_INFO           OK

```

Additional output omitted

If there are still problems after this, contact Cray support.

8.2.5 Many nodes flapping

Release 1.2.0, 1.2.1

This log message indicates that the resources are changing state more often than bebundd expects. The purpose of this threshold is to prevent bebundd from collecting failover-initiated support bundle on each stop-start event. By itself, this error is benign. However, it may suggest other failover-related issues are occurring on the system.

8.3 Networking issues

8.3.1 Recovering from a top-of-rack (TOR) Ethernet switch failure

Release 1.2.0

Problem description

A failure on the top-of-rack switch makes some nodes inaccessible.

Workaround

If the failure occurred on the TOR switch to which the quad node MMU is connected, reboot the entire system. If it affected only expansion racks (and not the MMU), reboot the affected nodes. In either case, refer to the *Sonexion 1.2 Power On / Power Off Procedures*.

8.3.2 Reseating a problematic high-speed network cable

Release 1.2.1

Problem description

On occasion, a node may lose its connection to the InfiniBand fabric.

Loss of connectivity can be caused by an incorrectly seated network cable (leads in the cable/switch not making physical contact), by dust on the leads, or because the cable itself has gone bad. Mellanox cables can only be plugged and unplugged a finite number of times before reaching their lifetime maximum.

A faulty InfiniBand connection can be diagnosed using the `ibcheckerrors` command. This command must return cleanly (no new errors reported) for the high speed network to be considered functional.

Workaround

To reseat a cable, complete the following procedure:

1. SSH into the primary MGMT node.
2. Sudo to root, by entering:

```
[admin@n000]$ sudo su -
```

3. Unmount Lustre.

```
[root@n000]# /opt/xyratex/bin/cscli unmount -c cluster_name -f filesystem_name
```

4. Inspect whether the LED switch for the cable is on.
5. Disable HA's InfiniBand querying, by entering:

```
[root@n000]# ssh nodename stop_ibstat
```

6. Determine the physical location of the cable to be reseated and unplug it.
7. Inspect the cable head for any signs of corrosion or other damage.
8. Blow compressed air over the cable head to remove any dust.
9. Before the cable is replugged, verify failover on the node that was unplugged, by entering:

```
[root@n000]# crm_mon -1
```

10. Replug the cable.

11. If the LED switch for that cable was previously on, verify that it comes back on after the cable has been replugged. Depending on how long is required for discovery, this may take up to a minute.

12. Enable HA's INFINIBAND FABRIC querying, by entering:

```
[root@n000]# ssh nodename start_ibstat
```

13. After reseating the cable, log into the affected node.

14. Replace the cable if it is damaged, or if there are multiple reseats, do not fix the problem.

15. Mount Lustre, by entering:

```
[root@n000]# /opt/xyratex/bin/cscli mount -c cluster_name -f filesystem_name
```

8.4 RAID/HA issues

8.4.1 RAIDs are not assembled correctly on the nodes

Release 1.2.0, 1.2.1, 1.2.2, 1.2.3, 1.3.1

Problem description

When an MDRAID device fails (for example, as a result of a chassis event temporarily removing several disks) the STONITH high-availability (HA) resource detects this change within its monitoring interval (10 minutes) and attempts to reassemble the MDRAID device on its OSS node. If the MDRAID device does not rebuild successfully, then the reassembly attempt times out after three minutes and the STONITH resource records a failed actions message for the OSS node.

The STONITH resource then tries to assemble the MDRAID device on the OSS node's HA partner node. If the rebuild is not successful on the HA partner node, then the reassembly attempt times out after three minutes and the STONITH resource records another failed actions message for the HA partner node.

After these failed attempts, the STONITH resource no longer tries to assemble the -RAID resource but leaves the first three resources in the group assembled.

Workaround

Use the following steps to manually recover the RAID. This procedure assumes that onsite personnel have identified the OSS node(s) that control the failed RAID array(s). Please note that:

- Even numbered OSS nodes natively control even numbered MDRAID devices (md0, md2, md4, and md6).
- Odd numbered OSS nodes natively control odd numbered MDRAID devices (md1, md3, md5, and md7).

- If a native OSS node is in a failover state, control of the MDRAID devices that it natively controls will migrate to its HA partner node. It is possible to recover the MDRAID device using either of these HA partner OSS nodes.

In the Sonexion solution, a chassis and two controllers are bundled in the modular SSU. Each controller hosts one OSS node; there are two OSS nodes per SSU. Within an SSU, the OSS nodes are organized in a High Availability (HA) pair with sequential numbers (for example snx11000n004 / snx11000n005). If an OSS node goes down because its controller fails, its resources migrate to the HA partner/OSS node in the other controller.

The 84 disk drives in a Sonexion SSU are configured as:

- 8 OSTs, each a RAID6 array consisting of 8 data disks and 2 parity disks
- 2 SSDs partitioned to create multiple independent RAID1 slices, used for MDRAID write intent bitmaps and external OST/ldiskfs file system journals
- 2 hot standby spares

The virtual drives defined by the RAID6 arrays are referred to as MDRAID devices, numbered sequentially from 0 through 7, for example, md0. Within the STONITH resource, there are resources defined for each MDRAID device, used by the STONITH resource to control the MDRAID device. For example, snx11000n004_md0-raid is the resource that controls the MDRAID device md0.

When an MDRAID device fails (for example, as a result of a chassis event temporarily removing several disks) the STONITH resource detects this change within its monitoring interval (10 minutes) and attempts to reassemble the MDRAID device on its OSS node. If the MDRAID device does not rebuild successfully, then the reassembly attempt times out after three minutes and the STONITH resource records a failed actions message for the OSS node.

The STONITH resource then tries to assemble the MDRAID device on the OSS node's HA partner node. If the rebuild is not successful on the HA partner node, then the reassembly attempt times out after three minutes and the STONITH resource records another failed actions message for the HA partner node.

After these failed attempts, the STONITH resource no longer tries to assemble the RAID resource but leaves the first three resources in the group assembled.

This procedure describes how to manually recover the RAID array.

Preparing to recover a failed RAID array

1. Log into the primary MGMT node via SSH.
2. Change to root user, by entering:


```
[admin@n000]$ sudo su -
```
3. Determine if either of the OSS nodes that control the failed MDRAID device are offline. If so, power on the downed OSS node(s). On the primary MGMT node, run:

```
[root@n000]# pm -1 OSS_nodename
```

Here is sample output:

```
[root@snx11000n000 ~]# pm -1 snx11000n004
```

Command completed successfully

If both OSS nodes are down, repeat Step 3 on the HA partner node.

4. Wait several minutes, and then log into the previously-downed OSS node via SSH to verify that it is back online, by entering:

```
[root@n000]# ssh OSS_nodename
```

Here is sample output:

```
[root@snx11000n000 ~]# ssh snx11000n004
[root@snx11000n004 ~]#
```

If both OSS nodes were down, repeat Step 4 on the HA partner node.

5. Log into the OSS node that natively controls the MDRAID device, via SSH, by entering:

```
[OSS node]# ssh OSS_nodename
```

6. Use the `crm_mon` utility to verify that a failed actions message was recorded for the failed MDRAID device. Also verify that the first three resources in the failed MDRAID device's resource group have failed over to the HA partner node, by entering:

```
[OSS node]# crm_mon -l
```

IMPORTANT: When reviewing the `crm_mon` output, note the failed MDRAID device's resource group name. You will need this information when performing the procedure to recover the failed RAID array.

Here is sample output showing the resource groups and the failed actions messages:

```
[root@snx11000n004 ~]# crm_mon -l
=====
Last updated: Wed Jan 23 17:30:10 2013
Last change: Wed Jan 23 17:16:30 2013 via cibadmin on snx11000n005
Stack: Heartbeat
Current DC: snx11000n004 (8ab209a5-874a-404d-af1c-1afa84cc18a9) -
partition with quorum
Version: 1.1.6.1-2.el6-0c7312c689715e096b716419e2ebc12b57962052
2 Nodes configured, unknown expected votes
55 Resources configured.
Online: [ snx11000n004 snx11000n005 ]
snx11000n004-stonith (stonith:external/gem_stonith):
Started snx11000n004
snx11000n005-stonith (stonith:external/gem_stonith):
Started snx11000n005
snx11000n004_mdadm_conf_regenerate
(ocf::heartbeat:mdadm_conf_regenerate): Started snx11000n004
snx11000n005_mdadm_conf_regenerate
(ocf::heartbeat:mdadm_conf_regenerate): Started snx11000n005
baton (ocf::heartbeat:baton): Started snx11000n005
snx11000n004_ibstat (ocf::heartbeat:ibstat): Started snx11000n004
snx11000n005_ibstat (ocf::heartbeat:ibstat): Started snx11000n005
Resource Group: snx11000n004_md0-group
snx11000n004_md0-wibr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md0-jnlr (ocf::heartbeat:XYRAID): Started snx11000n004
```



```

snx11000n004_md0-wibs (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md0-raid (ocf::heartbeat:XYRAID): Stopped
snx11000n004_md0-fsys (ocf::heartbeat:XYMNTR): Stopped
snx11000n004_md0-stop (ocf::heartbeat:XYSTOP): Stopped
Resource Group: snx11000n004_md1-group
snx11000n004_md1-wibr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md1-jnlr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md1-wibs (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md1-raid (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md1-fsys (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md1-stop (ocf::heartbeat:XYSTOP): Started snx11000n005
Resource Group: snx11000n004_md2-group
snx11000n004_md2-wibr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md2-jnlr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md2-wibs (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md2-raid (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md2-fsys (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md2-stop (ocf::heartbeat:XYSTOP): Started snx11000n004
Resource Group: snx11000n004_md3-group
snx11000n004_md3-wibr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md3-jnlr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md3-wibs (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md3-raid (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md3-fsys (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md3-stop (ocf::heartbeat:XYSTOP): Started snx11000n005
Resource Group: snx11000n004_md4-group
snx11000n004_md4-wibr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md4-jnlr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md4-wibs (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md4-raid (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md4-fsys (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md4-stop (ocf::heartbeat:XYSTOP): Started snx11000n004
Resource Group: snx11000n004_md5-group
snx11000n004_md5-wibr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md5-jnlr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md5-wibs (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md5-raid (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md5-fsys (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md5-stop (ocf::heartbeat:XYSTOP): Started snx11000n005
Resource Group: snx11000n004_md6-group
snx11000n004_md6-wibr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md6-jnlr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md6-wibs (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md6-raid (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md6-fsys (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md6-stop (ocf::heartbeat:XYSTOP): Started snx11000n004
Resource Group: snx11000n004_md7-group
snx11000n004_md7-wibr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md7-jnlr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md7-wibs (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md7-raid (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md7-fsys (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md7-stop (ocf::heartbeat:XYSTOP): Started snx11000n005
Failed actions:

```

```
snx11000n004_md0-raid_start_0 (node=snx11000n005, call=134, rc=-2,
status=Timed Out): unknown exec error
snx11000n004_md0-raid_start_0 (node=snx11000n004, call=134, rc=-2,
status=Timed Out): unknown exec error
```

7. If the RAID fails to assemble with messages like this:

```
mdadm: ignoring /dev/disk/by-id/wwn-0x5000cca01b3d4224 as it
reports /dev/disk/by-id/wwn-0x5000cca01b3cf13c as failed
mdadm: ignoring /dev/disk/by-id/wwn-0x5000cca01b3d7e24 as it
reports /dev/disk/by-id/wwn-0x5000cca01b3cf13c as failed
mdadm: ignoring /dev/disk/by-id/wwn-0x5000cca01b3d6080 as it
reports /dev/disk/by-id/wwn-0x5000cca01b3cf13c as failed
mdadm: ignoring /dev/disk/by-id/wwn-0x5000cca01c375f04 as it
reports /dev/disk/by-id/wwn-0x5000cca01b3cf13c as failed
```

then the RAID recovery procedure has failed, go to Step 7. If the forceable reassembly does not produce these error messages, go to Step 8.

8. Abort this procedure and contact Cray support.

Cray support will require the mdraid superblock data to be collected in order to debug the problem. Use the **collect_superblock.sh** script to collect this data.

- a. Download the collect_superblock.sh script from Marlin or the XIC to /tmp on the primary MGMT node
- b. Run the script:

```
[root@n000]# /tmp/collect_superblock.sh
/var/lib/mdraidscripts/mdacdm.conf
```

9. If the first three resources in the failed MDRAID device's resource group have failed over to the HA partner node, log into the HA partner node via SSH, by entering:

```
[root@n000]# ssh HA_partner_nodename
```

The procedure to prepare for recovering a failed RAID array is now complete. Proceed to the next section for the procedure to recover the RAID array.

Recovering a failed RAID array

This procedure describes how to force assemble an MDRAID device to recover a failed RAID array.

CAUTION: Assembling a RAID array with the `-force` argument can result in data loss or data corruption. This procedure should only be used as a last resort.

1. Stop the resource group containing the failed MDRAID device, by entering:

```
[OSS node]# stop_xyraid resource_group_name
```

Where *resource_group_name* is the resource group name, which is discussed in the note on page 112 and accompanying output.

Here is sample output:

```
[root@snx11000n005 ~]# stop_xyraid snx11000n004_md0-group
[root@snx11000n005 ~]#
```

2. Unmanage the resource group, which will allow the resources to be started outside of the STONITH high-availability (HA) resource, by entering:

```
[OSS node]# unmanage_xyraid resource_group_name
```

Here is sample output:

```
[root@snx11000n005 ~]# unmanage_xyraid snx11000n004_md0-group
[root@snx11000n005 ~]#
```

3. Clean the resource group to remove the failed actions, by entering:

```
[OSS node]# clean_xyraid resource_group_name
```

Here is sample output:

```
[root@snx11000n005 ~]# clean_xyraid snx11000n004_md0-group
Cleaning up snx11000n004_md0-wibr on snx11000n004
Cleaning up snx11000n004_md0-wibr on snx11000n005
Cleaning up snx11000n004_md0-jnlr on snx11000n004
Cleaning up snx11000n004_md0-jnlr on snx11000n005
Cleaning up snx11000n004_md0-wibs on snx11000n004
Cleaning up snx11000n004_md0-wibs on snx11000n005
Cleaning up snx11000n004_md0-raid on snx11000n004
Cleaning up snx11000n004_md0-raid on snx11000n005
Cleaning up snx11000n004_md0-fsys on snx11000n004
Cleaning up snx11000n004_md0-fsys on snx11000n005
Cleaning up snx11000n004_md0-stop on snx11000n004
Cleaning up snx11000n004_md0-stop on snx11000n005
Waiting for 13 replies from the CRMD..... OK
```

4. If you determined in “Preparing to recover a failed RAID array”, page 111 , that the first three resources in the failed MDRAID device's resource group have failed over to the HA partner node, follow Steps a and b below:

- a. Log into the OSS node that natively controls the MDRAID device, via SSH, by entering:

```
ssh OSS_nodename
```

- b. Fail back resources to the OSS node, by entering:

```
[OSS node]# failback_xyraid
```

Here is sample output:

```
[root@snx11000n004 ~]# failback_xyraid
[root@snx11000n004 ~]#
```

5. Determine if the --force argument is necessary to assemble the MDRAID device, by entering:

```
[OSS node]# mdraid-activate -d resource_group_name
```

Here is sample output showing an unsuccessful attempt to assemble the MDRAID device without the --force argument:

```
[root@snx11000n004 ~]# mdraid-activate -d snx11000n004_md0-group
mdadm: /dev/md/snx11000n004:md128 has been started with 2 drives.
mdadm: /dev/md/snx11000n004:md129 has been started with 2 drives.
mdadm: failed to RUN_ARRAY /dev/md/snx11000n004:md0: Input/output error
```

```
mdadm: Not enough devices to start the array.
mdadm: failed to RUN_ARRAY /dev/md/snx11000n004:md0: Input/output error
mdadm: Not enough devices to start the array.
mdadm: failed to RUN_ARRAY /dev/md/snx11000n004:md0: Input/output error
mdadm: Not enough devices to start the array.
mdadm: failed to RUN_ARRAY /dev/md/snx11000n004:md0: Input/output error
mdadm: Not enough devices to start the array.
mdadm: failed to RUN_ARRAY /dev/md/snx11000n004:md0: Input/output error
mdadm: Not enough devices to start the array.
mdadm: failed to RUN_ARRAY /dev/md/snx11000n004:md0: Input/output error
mdadm: Not enough devices to start the array.
mdadm: failed to RUN_ARRAY /dev/md/snx11000n004:md0: Input/output error
mdadm: Not enough devices to start the array.
mdadm: failed to RUN_ARRAY /dev/md/snx11000n004:md0: Input/output error
mdadm: Not enough devices to start the array.
mdadm: failed to RUN_ARRAY /dev/md/snx11000n004:md0: Input/output error
mdadm: Not enough devices to start the array.
mdadm: failed to RUN_ARRAY /dev/md/snx11000n004:md0: Input/output error
mdadm: Not enough devices to start the array.
mdraid-activate 358: unable to assemble snx11000n004:md0
```

NOTE: If the above assembly attempt was successful, proceed to Step 9.

6. Assemble the MDRAID device using the --force argument. If you are performing this procedure on a system running 1.2.x, run the following command:

```
[OSS node]# mdraid-activate -df resource_group_name
```

On a system running 1.3.1:

```
[OSS node]# mdraid-activate -f
i_am_sure_i_want_to_do_this,exit -d resource_group_name
```

Sample output of a successful 'forced' assembly of an MDRAID device:

```
[root@snx1100005 ~]# mdraid-activate -f i_am_sure_i_want_to_do_this,exit -d
snx1100004_md0-group
mdadm: /dev/md/snx1100004:md129 has been started with 2 drives.
mdadm: failed to RUN_ARRAY /dev/md/snx1100004:md0: Input/output error
mdadm: Not enough devices to start the array.
mdadm: failed to RUN_ARRAY /dev/md/snx1100004:md0: Input/output error
mdadm: Not enough devices to start the array.
/usr/lib/ocf/lib/heartbeat/xrtx-ocf-shellfuncs: line 867: ocf_log: command not found
mdadm: forcing event count in /dev/disk/by-id/wwn-0x5000c500212d3f2b(3) from 19 upto 39
mdadm: clearing FAULTY flag for device 7 in /dev/md/snx1100004:md0 for /dev/disk/by-
id/wwn-0x5000c500212d3f2b
mdadm: Marking array /dev/md/snx1100004:md0 as 'clean'
mdadm: /dev/md/snx1100004:md0 has been started with 8 drives (out of 10).
assembled snx1100004:md0 in 1 tries
[root@snx1100005 ~]
```

On a system running 1.2.x:

```
[OSS node]# mdraid-activate -df resource_group_name
```

Sample output of a successful 'forced' assembly of an MDRAID device:

```
[root@snx11000n004 ~]# mdraid-activate -df snx11000n004_md0-group
```

```
mdadm: /dev/md/snx11000n004:md129 has been started with 2 drives.
mdadm: forcing event count in /dev/disk/by-id/wwn-0x5000cca01c477720(3) from 3 upto 16
mdadm: forcing event count in /dev/disk/by-id/wwn-0x5000cca01c472818(8) from 3 upto 16
mdadm: forcing event count in /dev/disk/by-id/wwn-0x5000cca01b4a0ec4(9) from 3 upto 16
mdadm: clearing FAULTY flag for device 8 in /dev/md/snx11000n004:md0 for /dev/disk/by-
id/wwn-0x5000cca01c477720
mdadm: clearing FAULTY flag for device 7 in /dev/md/snx11000n004:md0 for /dev/disk/by-
id/wwn-0x5000cca01c472818
mdadm: clearing FAULTY flag for device 3 in /dev/md/snx11000n004:md0 for /dev/disk/by-
id/wwn-0x5000cca01b4a0ec4
mdadm: Marking array /dev/md/snx11000n004:md0 as 'clean'
mdadm: /dev/md/snx11000n004:md0 has been started with 10 drives.
assembled snx11000n004:md0 in 1 tries
```

IMPORTANT: If the MDRAID device failed to assemble, stop and contact Cray Support.

7. Run the `e2fsck` command on the MDRAID device.

IMPORTANT: Do not run the `e2fsck` command on an OST larger than 16TB unless an appropriate up-to-date version of the `e2fsck` command is installed at your location. We strongly recommend using the version of the `e2fsck` command that is provided with Sonexion HotFix 1.2.0-MRP-1 or HotFix 1.2.1-MRP-1.

- To check whether the OST size is larger than 16TB, run:

```
[OSS node]# mdadm --misc --detail /dev/md3 | grep Array
```

Example output of the command run on a very small OST:

```
[root@snx11000n204 ~]# mdadm --misc --detail /dev/md3 | grep Array
Array Size : 125829120 (120.00 GiB 128.85 GB)
```

- To check the version of `e2fsck` on your system, run:

```
[OSS node]# e2fsck -V
```

The version to run if the OST is larger than 16TB should be at least 1.42.6.x1.

When you have the correct version of the `e2fsck` command, run:

```
[OSS node]# e2fsck -fp /dev/MDRAID_device
```

Here is sample output:

```
[root@snx11000n004 ~]# e2fsck -fp /dev/md0
testfs-OST0000: recovering journal
testfs-OST0000: 86/7879680 files (2.3% non-contiguous),
509811/31457280 blocks
```

8. Stop the MDRAID device, by entering:

```
[OSS node]# mdraid-deactivate resource_group_name
```

Here is sample output:

```
[root@snx11000n004 ~]# mdraid-deactivate snx11000n004_md0-group
[root@snx11000n004 ~]#
```

9. Manage the MDRAID device's resource group, by entering:

```
[OSS node]# manage_xyraid resource_group_name
```

Here is sample output:

```
[root@snx11000n004 ~]# manage-xyraid snx11000n004_md0-group
[root@snx11000n004 ~]#
```

10. Start the MDRAID device's resource group, by entering:

```
[OSS node]# start_xyraid resource_group_name
```

Here is sample output:

```
[root@snx11000n004 ~]# start-xyraid snx11000n004_md0-group
[root@snx11000n004 ~]#
```

11. Use the `crm_mon` utility to verify that the MDRAID device's resource group started correctly, which can take several minutes, by entering:

```
[OSS node]# crm_mon -1
```

Here is sample output showing a healthy OST group:

```
[root@snx11000n004 ~]# crm_mon -1
=====
Last updated: Wed Jan 23 18:00:58 2013
Last change: Wed Jan 23 18:00:18 2013 via cibadmin on snx11000n004
Stack: Heartbeat
Current DC: snx11000n005 (8ab209a5-874a-404d-af1c-1afa84cc18a9) - partition with
quorum
Version: 1.1.6.1-2.el6-0c7312c689715e096b716419e2ebc12b57962052
2 Nodes configured, unknown expected votes
55 Resources configured.
Online: [ snx11000n004 snx11000n005 ]
snx11000n004-stonith (stonith:external/gem_stonith): Started snx11000n004
snx11000n005-stonith (stonith:external/gem_stonith): Started snx11000n005
snx11000n004_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):Started
snx11000n004
snx11000n005_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):Started
snx11000n005
baton (ocf::heartbeat:baton): Started snx11000n005
snx11000n004_ibstat (ocf::heartbeat:ibstat): Started snx11000n004
snx11000n005_ibstat (ocf::heartbeat:ibstat): Started snx11000n005
Resource Group: snx11000n004_md0-group
snx11000n004_md0-wibr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md0-jnlr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md0-wibs (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md0-raid (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md0-fsys (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md0-stop (ocf::heartbeat:XYSTOP): Started snx11000n004
Resource Group: snx11000n004_md1-group
snx11000n004_md1-wibr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md1-jnlr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md1-wibs (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md1-raid (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md1-fsys (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md1-stop (ocf::heartbeat:XYSTOP): Started snx11000n005
```

```

Resource Group: snx11000n004_md2-group
snx11000n004_md2-wibr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md2-jnlr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md2-wibs (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md2-raid (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md2-fsys (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md2-stop (ocf::heartbeat:XYSTOP): Started snx11000n004
Resource Group: snx11000n004_md3-group
snx11000n004_md3-wibr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md3-jnlr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md3-wibs (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md3-raid (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md3-fsys (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md3-stop (ocf::heartbeat:XYSTOP): Started snx11000n005
Resource Group: snx11000n004_md4-group
snx11000n004_md4-wibr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md4-jnlr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md4-wibs (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md4-raid (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md4-fsys (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md4-stop (ocf::heartbeat:XYSTOP): Started snx11000n004
Resource Group: snx11000n004_md5-group
snx11000n004_md5-wibr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md5-jnlr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md5-wibs (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md5-raid (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md5-fsys (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md5-stop (ocf::heartbeat:XYSTOP): Started snx11000n005
Resource Group: snx11000n004_md6-group
snx11000n004_md6-wibr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md6-jnlr (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md6-wibs (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md6-raid (ocf::heartbeat:XYRAID): Started snx11000n004
snx11000n004_md6-fsys (ocf::heartbeat:XYMNTR): Started snx11000n004
snx11000n004_md6-stop (ocf::heartbeat:XYSTOP): Started snx11000n004
Resource Group: snx11000n004_md7-group
snx11000n004_md7-wibr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md7-jnlr (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md7-wibs (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md7-raid (ocf::heartbeat:XYRAID): Started snx11000n005
snx11000n004_md7-fsys (ocf::heartbeat:XYMNTR): Started snx11000n005
snx11000n004_md7-stop (ocf::heartbeat:XYSTOP): Started snx11000n005

```

8.4.2 Starting Lustre on a given node without mounting the fsys resource

Release 1.2.0, 1.2.1

Problem description

The need may arise to start Lustre on a given node without starting the fsys resource (i.e. mounting the device), or to stop the fsys resource but leave the mdraid assembled.

Workaround

Run one of the following commands, as applicable:

To assemble the RAID arrays without mounting the filesystem, use the following commands:

1. To mount all resources on a node except `fsys`, run:

```
[node]# "crm_mon -lr | awk '/wibr|jnlr|wibs|raid/ {print $1}'
| xargs -I {} start_xyraid {}"
```

2. To mount all resources except `fsys` on all OSSes, run:

```
[root@n000]# pdsh -g oss=primary "crm_mon -lr | awk
'/wibr|jnlr|wibs|raid/ {print \$1}' | xargs -I {}
start_xyraid {}"
```

NOTE: The `\` in `\$1` is required to escape the `$1`.

To stop only the `fsys` resource while leaving the mdraid assembled, use the following commands:

3. To stop only the `fsys` resource on a node, run:

```
[node]# "crm_mon -lr | awk '/fsys/ {print $1}' | xargs -I {}
stop_xyraid {}"
```

4. To stop only the `fsys` resources on all OSS nodes, run:

```
[root@n000]# pdsh -g oss=primary "crm_mon -lr | awk '/fsys/
{print $1}' | xargs -I {} stop_xyraid {}"
```

8.4.3 Lost one OST during forced failover. Release 1.2.0

Problem description

During a system test with a forced failover, an OST was lost.

Workaround

Trigger a `kdump` on the node. In the console log from the trigger, you should be able to see that the node successfully boots into the `kdump` kernel, but shuts down when its DHCP request is not answered. In this case, it will either be a bad cable or NIC. Run the following command to check the connection speed; notice how the connection speed of `eth0` is only 100Mbit/sec, when all the other nodes are at 1000Mbit/sec.

Sample `kdump`:

```
Sending discover...
Unable to get a DHCP address ret[ 26.547454] md: stopping all md devices.
ry...
No lease, failing
eth0 failed to come up
[ 27.665615] sd 10:0:72:0: [sdbt] Synchronizing SCSI cache
[ 27.671577] sd 10:0:13:0: [sdp] Synchronizing SCSI cache
[ 27.677282] sd 3:0:0:0: [sdc] Synchronizing SCSI cache
```



```
[ 27.682684] sd 3:0:0:0: [sdc] Stopping disk
[^@ 27.691845] e
[^@ 27.767758] e
[ 27.772799] [ 27.787961] mpt2sas0: sending message unit reset !!
[ 27.794746] mpt2sas0: message unit reset: SUCCESS
[ 27.799677] mpt2sas 0000:12:00.0: PCI INT A disabled
[^@ 31.344980] A[ 31.458896] Disabling non-boot CPUs ...
[ 31.462916] Power down.
```

1. Log in to the affected node using an admin account with the password set during the first-run (customer wizard) configuration, by entering:

```
[MGMT]$ ssh admin@oss_nodename
```

2. Change to root user.

```
[OSS]$ sudo su -
```

3. Check the connection speed, by entering:

```
[OSS]# pdsh -g oss "ethtool eth0 | grep Speed"
```

4. The following is sample output:

```
[root@snx11000n000 ~]# pdsh -g oss "ethtool eth0 | grep Speed"
snx11000n006: Speed: 1000Mb/s
snx11000n007: Speed: 1000Mb/s
snx11000n004: Speed: 1000Mb/s
snx11000n005: Speed: 100Mb/s
```

Replace the cable and run the check again. If the speed comes back, then it was the cable. If the speed did not return to normal, then replace the NIC card. Verify that the speed has now returned to normal.

8.4.4 Multiple HDDs spontaneously drop out of RAID arrays (heap overflows)

Release 1.2.0

Problem description

Errors in the SAS subsystem can trigger a reset of the SAS firmware or hardware. Even if the hardware/firmware recovers, the effect of this reset is seen at the Sonexion level as a large number of HDDs dropping out of RAID arrays. The most common cause of these resets is a heap overflow in the GEM firmware.

Workaround

First, determine if the drive drop-out were caused by a heap overflow. To determine if a drive drop-off was caused by a heap overflow, collect the GEM logs using `ddump`:

```
[MGMT0] conman nodename-gem
[gem] ddump
[gem] -ddump
```

The `ddump` command collects the local canister's GEM logs, and `-ddump` collects the partner node's GEM logs. The output from these commands is written to:

```
/var/log/conman/nodename-gem.log
```

Search the GEM log for the phrase “heap overflow detected”.

If these entries are not present, a heap overflow may not be responsible for the drive drop-offs. Please contact Cray support.

If these entries are present, then the problem was caused by a heap overflow. Now, determine if the software automatically recovered. Examine the kernel messages (which can be obtained using the `dmesg` command) for the following entries:

```
Restoring state sled x element x
Sled x element x version x.x.x.xx
```

If these kernel messages are not present, the problem can manifest itself in `dm_report` as empty drive slots. For example:

```
[root@n000]# cat dm_report.txt
Diskmonitor Inventory Report: Version: 1.0-2020.xrtx.2206 Host:
  snx110001 Time: Mon Mar 18 08:33:12 2013
encl:  0, wwn: 50050cc10c40036d, dev:    /dev/sg0, slots:  84,
  vendor: CRAY , product_id: UD-8435-CS-1600
slot:  0, status: Empty
slot:  1, status: Empty
slot:  2, status: Empty
slot:  3, status: Empty
slot:  4, status: Empty
slot:  5, status: Empty
slot:  6, status: Empty
slot:  7, status: Empty
slot:  8, status: Empty
slot:  9, status: Empty
slot: 10, status: Empty
slot: 11, status: Empty
slot: 12, status: Empty
slot: 13, status: Empty
slot: 14, wwn: 5000cca01c3f18f0, cap: 2000398933504, dev:
  sdce, parts: 0, status: Foreign Arrays
slot: 15, wwn: 5000cca01c3f1130, cap: 2000398933504, dev:
  sdcc, parts: 0, status: Foreign Arrays
slot: 16, wwn: 5000cca01c3f13b4, cap: 2000398933504, dev:
  sdcd, parts: 0, status: Foreign Arrays
slot: 17, wwn: 5000cca01c3e159c, cap: 2000398933504, dev:
  sdcf, parts: 0, status: Foreign Arrays
```

The expander is either hung or has become defective. Issue a reset to the expander and reboot GEM using the following commands:

1. SSH into the problematic OSS node's partner
2. Sudo to root
3. Unmanage the HA resources:

```
[OSS]# unmanage xyraid all
```

- From the MGMT node, conman into GEM on the problematic OSS node:

```
[root@n000]# conman nodename-gem
```

NOTE: Steps 5 and 6 must be performed within one second of each other:

- Reboot all the expanders on this node:

```
[GEM] gncli exp:local reboot
```

- Reboot GEM:

```
[GEM] reboot
```

- Exit conman.

- Wait approximately one minute, then run the following command on the problematic node:

```
[OSS node]# sg_map -i
```

This will check the status of the previously empty drive slots. If HDDs are present, then the reset cleared the expander. If drives are still not present, contact Cray support.

8.4.5 MD device fails to assemble with the message: mdadm: cannot reread metadata from /dev/disk/by-id/WWN - aborting

Release 1.2.0

Problem Description

When attempting to mount Lustre, an MD device fails to assemble with the error: mdadm: cannot reread metadata from /dev/disk/by-id/WWN - aborting.

This problem is caused by a faulty hard drive. The following procedures will guide you to restore the MD device.

Workaround

Power down the bad drive and re-activate the raid array. At the earliest opportunity, replace the hard drive using Field Replacement of 5U84 DDIC (in the SSU).

- SSH into the primary management node. Run

```
[Client] ssh -l admin primary_MGMT_node
```

- Determine which OSS node controls the failed hard drive.

- SSH into that OSS node. Run:

```
[MGMT0] ssh OSS_node
```

- Determine the slot of the drive that is faulty with the drive's WWN (which is in the error message). Run:

```
[OSS node]# dm_report | grep WWN
```

It might be necessary to remove the last character from the WWN string.

5. Use the `poweroffdrive` command to power down the failed drive. Run:

```
[OSS node]# echo "poweroffdrive Slot" | wbccli /dev/device
```

For example, to power off `/dev/sg0` located in slot 45, run:

```
[root@cstor012 ~]# echo "poweroffdrive 45" | wbccli /dev/sg0
Error: I/O timeout. ***
GEMRITE>
GEMRITE>[root@cstor012 ~]#
```

6. Verify that the drive slot is now empty with the command `dm_report`. Run:

```
[OSS node]# dm_report
```

Sample output

```
[root@cstor012 ~]# dm_report
Diskmonitor Inventory Report: Version: 1.0-2020.xrtx.2206 Host: cstor012 Time: Sat Aug
 24 21:34:07 2013
encl: 0, wwn: 50050cc10c400107, dev: /dev/sg0, slots: 84, vendor: CRAY , product_id: UD-
 8435-CS-1600
...
slot: 42, wwn: 5000c50034ce95ef, cap: 2000398933504, dev: sdm, parts:
0, status: Ok
slot: 43, wwn: 5000c500348c8c83, cap: 2000398933504, dev: sdk, parts:
0, status: Foreign Arrays
slot: 44, wwn: 5000c500348cf957, cap: 2000398933504, dev: sdl, parts:
0, status: Foreign Arrays
slot: 45, status: Empty
slot: 46, wwn: 5000c50034a1632f, cap: 2000398933504, dev: sdi, parts:
0, status: Foreign Arrays
slot: 47, wwn: 5000c5003488751f, cap: 2000398933504, dev: sdj, parts:
0, status: Ok
...
```

If the drive slot is not 'empty', contact Cray support at MyCray.com.

7. Clean the fail-counts on this HA resource. Run:

```
[OSS node]# clean_xyraid HA_OST_resource_group
```

After the fail counts have been cleared, the HA resource should start the problematic array. If HA fails again while starting this HA group, contact Cray support at MyCray.com.

8. If the OST starts on the node that doesn't normally host it, fail it back with this command. Run:

```
[OST node]# failback_xyraid
```

on the OST's normal host.

9. From the MGMT node, mount Lustre: Run:

```
[root@n000]# cscli mount -f filesystem_name
```

10. At the earliest opportunity, replace the failed hard drive using Field Replacement of 5U84 DDIC (in the SSU).

8.4.6 MD device fails to assemble with the message: “mdadm: cannot reread metadata from /dev/disk/by-id/WWN - aborting.”

Release 1.2.0

Problem Description

When attempting to mount Lustre, an MD device fails to assemble with the error: mdadm: cannot reread metadata from /dev/disk/by-id/WWN - aborting.

This problem is caused by a faulty hard drive. The following procedures will guide you to restore the MD device.

Workaround

Power down the bad drive and re-activate the raid array. At the earliest opportunity, replace the hard drive. (See publication HR5-6098, *Maintenance and Replacement Procedures for Cray Sonexion Storage Systems*, “5U84 Disk”.)

1. SSH into the primary management node. Run

```
[Client] ssh -l admin primary_MGMT_node
```

2. Determine which OSS node controls the failed hard drive.

3. SSH into that OSS node, by entering:

```
[MGMT0] ssh OSS_node
```

4. Determine the slot of the drive that is faulty with the drive's WWN (which is in the error message), by entering:

```
[OSS node]# dm_report | grep WWN
```

NOTE: It might be necessary to remove the last character from the WWN string.

5. Use the poweroffdrive command to power down the failed drive, by entering:

```
[OSS node]# echo "poweroffdrive Slot" | wbcli /dev/device
```

For example, to power off /dev/sg0 located in slot 45, run:

```
[root@snx11000n012 ~]# echo "poweroffdrive 45" | wbcli /dev/sg0
```

```
Error: I/O timeout. ***
```

```
GEMLITE>
```

```
GEMLITE>[root@snx11000n012 ~]#
```

6. Verify that the drive slot is now empty with the command dm_report, by entering:

```
[OSS node]# dm_report
```

Sample output:

```
[root@snx11000n012 ~]# dm_report
```

```
Diskmonitor Inventory Report: Version: 1.0-2020.xrtx.2206
```

```
Host: snx11000n012 Time: Sat Aug 24 21:34:07 2013
```

```

encl: 0, wwn: 50050cc10c400107, dev: /dev/sg0, slots: 84,
  vendor: CRAY , product_id: UD-8435-CS-1600
...
slot: 42, wwn: 5000c50034ce95ef, cap: 2000398933504, dev:
  sdm, parts: 0, status: Ok
slot: 43, wwn: 5000c500348c8c83, cap: 2000398933504, dev:
  sdk, parts: 0, status: Foreign Arrays
slot: 44, wwn: 5000c500348cf957, cap: 2000398933504, dev:
  sdl, parts: 0, status: Foreign Arrays
slot: 45, status: Empty
slot: 46, wwn: 5000c50034a1632f, cap: 2000398933504, dev:
  sdi, parts: 0, status: Foreign Arrays
slot: 47, wwn: 5000c5003488751f, cap: 2000398933504, dev:
  sdj, parts: 0, status: Ok
...

```

If the drive slot is not 'empty', contact Cray support at MyCray.com.

- Clean the fail-counts on this HA resource, by entering:

```
[OSS node]# clean_xyraid HA_OST_resource_group
```

After the fail counts have been cleared, the HA resource should start the problematic array. If HA fails again while starting this HA group, contact Cray support at MyCray.com.

- If the OST starts on the node that doesn't normally host it, fail it back with this command, by entering:

```
[OST node]# failback_xyraid
```

on the OST's normal host.

- From the MGMT node, mount Lustre, by entering:

```
[root@n000]# cscli mount -f filesystem_name
```

- At the earliest opportunity, replace the failed hard drive, (See publication HR5-6098, *Maintenance and Replacement Procedures for Cray Sonexion Storage Systems*, “5U84 Disk”.)

8.5 Other Issues

8.5.1 Nodes are shown in “unknown” state in GUI

Release 1.2.0

Problem Description

If a node shows up as 'unknown' in the GUI, this indicates that the management nodes are unable to communicate with that node's IPMI interface.

Workaround

This problem is caused by an unresponsive BMC on the OSS node.

NOTE: This problem occurs less frequently in more recent USM releases. A USM upgrade may be advisable.

1. If possible, log into the node that has state 'unknown' as user root. Run :

```
[Node]# ipmitool bmc reset cold
```

2. Wait approximately 2 minutes for the BMC to reboot, after which the node's status should no longer be 'unknown'.
3. If the node is still in the unknown state after resetting the BMC, check that this node's BMC network configuration is correct. On the node that's showing up as unknown, list the BMC's network parameters with the command:

```
ipmitool lan print 1
```

Example output:

```
[root@csstor01 ~]# ipmitool lan print 1
Set in Progress           : Set Complete
Auth Type Support         : NONE MD5 PASSWORD
Auth Type Enable          : Callback : NONE MD5 PASSWORD
                          : User      : NONE MD5 PASSWORD
                          : Operator : NONE MD5 PASSWORD
                          : Admin   : NONE MD5 PASSWORD
                          : OEM     :
IP Address Source         : Static Address
IP Address                 : 172.16.0.101
Subnet Mask                : 255.255.0.0
MAC Address                : 00:1e:67:66:db:32
SNMP Community String     : public
IP Header                  : TTL=0x00 Flags=0x00 Precedence=0x00
                          TOS=0x00
BMC ARP Control           : ARP Responses Enabled, Gratuitous
                          ARP Disabled
Gratuitous ARP Intrvl     : 0.0 seconds
Default Gateway IP        : 172.16.0.101
Default Gateway MAC       : 00:00:00:00:00:00
Backup Gateway IP         : 0.0.0.0
Backup Gateway MAC        : 00:00:00:00:00:00
802.1q VLAN ID            : Disabled
802.1q VLAN Priority       : 0
RMCP+ Cipher Suites      :
                          1,2,3,4,6,7,8,9,11,12,13,15,16,17,18,0
Cipher Suite Priv Max     : caaaaXaaaaXaaaX
                          : X=Cipher
Suite Unused              :
                          : c=CALLBACK
                          : u=USER
                          : o=OPERATOR
                          : a=ADMIN
                          : O=OEM
[root@snx11000n000 ~]
```

4. Verify that the field 'IP Address Source' is set to 'Static Address', not 'dhcp'. If 'IP Address Source' is set to 'dhcp', fix this with the command, run:

```
[root@n000]# ipmitool lan set 1 ipsrc static
```

5. Also verify that the field 'IP Address' is correct. The correct address can be obtained with the command, run:

```
[root@n000]# host hostname-ipmi
```

Example output:

```
[root@snx11000n004 ~]# host snx11000n004-ipmi
snx11000n004-ipmi has address 172.16.0.110
[root@snx11000n004 ~]#
```

6. If this isn't set correctly in the BMC, fix it using a command in the following form:

```
ipmitool lan 1 set ipaddr 172.16.0.101
```

Example output:

```
[root@snx11000n004 ~]# ipmitool lan set 1 ipaddr 172.16.0.110
Setting LAN IP Address to 172.16.0.110
[root@snx11000n004 ~]#
```

7. If the node still shows up as 'unknown' after correcting the BMC network settings, contact Cray support at my.Cray.com.

8.5.2 SSUs failed after AC power loss

Release 1.2.0

Problem Description

Several SSU's failed on an 18 SSU file system as a result of an AC power loss.

Workaround

When a power loss occurs, the Sonexion system will automatically fail over the HA components ensuring continuous operation. In the event of a total power loss, the entire system will shutdown. To understand what may have caused the situation, review the GEM logs for messages similar to the following:

```
2012-10-23 10:19:16.005; ENC_MGT; batt_manager; 01; Power Loss (AC Fail) detected
2012-10-23 10:19:16.005; ENC_MGT; drive_manager; 01; Enclosure power loss detected
2012-10-23 10:19:16.005; ENC_MGT; power_manager; 01; Enclosure Power Loss (AC Fail)
detected
```

Once power is restored and the system booted, manually re-power (re-boot) the failed nodes from the primary MGMT node, by entering:

```
[root@n000]# pm -1 nodename
```


8.5.3 CS-1600 OSS will not power up, BMC out of memory

Release 1.2.1. The CS-1600 OSS will not power up because the BMC is out of memory.

Problem Description

It was observed in the field that some CS-1600 OSS units were failing to power up. Investigation determined that the BMC IPMI system event logs (sel logs) on those nodes had grown so large that they had consumed all the BMC memory.

An out-of-memory BMC is known to cause problematic behavior.

Workaround

To clear the sel log on an affected node, run this command:

```
[root@n000]# ipmitool -H nodename-ipmi -U admin -P admin sel clear
```

To clear all the sel logs on a machine, run this command:

```
[root@n000]# for addr in $(awk '/ipmi/ {print $1}' /etc/hosts);
do echo $addr; ipmitool -H $addr -U admin -P admin sel clear;
done
```

8.5.4 No response when attempting to physically connect to the serial port

Release 1.2.1

Problem Description

There is no response when physically connecting to the serial port and starting a hyperterminal session when using the following settings:

- Baud Rate: 115200
- Data bits: 8
- Parity: none
- Stop bits: 1
- Flow control: none
- Function Keys are set to VT100+

Workaround

Only one serial connection at a time is possible. This includes virtual serial connections. If the serial port is not responding to a physical connection, it is very likely that the controller is connected somewhere else using a Serial-On-LAN (SOL) connection.

1. First, forcibly disconnect any serial connections (SOL sessions), use the following command:

```
[root@n000]# ipmitool -H nodename-ipmi -U admin -P admin bmc reset cold
```

2. Then, attempt to connect to the physical serial port using the above hyperterminal settings.

8.5.5 OSS/MDS nodes go down during FS testing

All releases

Problem Description

This problem is likely caused by a Lustre crash and a kernel panic. To verify this problem, connect to the OSS node using a serial cable (refer to section 7.6 for the hyperterminal settings) or conman and press "&L". If there is a stacktrace with an LBUG error, this is the issue.

Workaround

Power-cycle the node.

8.5.6 Non-responsive server

Release 1.2.0

Problem Description

A server node failed and is not responding to power manage commands to reboot.

Workaround

To revive the node, using conman, run the ipmi command to start the node.

1. Log in via console manager, by entering:

```
[node]# conman nodename-gem
```

2. Issue the following command, by entering:

```
[gem]# -ipmi_power 4
```

9. Upgrading Lustre RPMs

This section presents two procedures for upgrading Lustre RPMs: one that involves first deactivating Lustre, and one during which Lustre continues to function throughout the procedure.

9.1 RPM upgrade with Lustre inactive

Use the following procedure to apply Lustre system updates on the Cray Sonexion system. The RPMs needed vary by Sonexion release, as listed in the Sonexion Release Notes.

1. Log in via SSH into the primary MGMT node.
2. Change to root by entering:

```
[admin@n000]$ sudo su -
```

3. Unmount Lustre by entering:

```
[root@n000]# /opt/xyratex/bin/cscli unmount -c cluster_name -f filesystem_name
```

4. Copy the Lustre RPMs to **/tmp** on the primary MGMT server.

NOTE: Newer Lustre system updates may be distributed as a zip file of RPMs. Please unzip these files before copying the RPMs over to the MGMT node. Newer Lustre system updates may be distributed as a zip file of RPMs. Please unzip these files before copying the RPMs over.

5. Update the RPMs on the primary MGMT node, by entering:

```
[root@n000]# yum update -y /tmp/lustre-*.rpm
```

- Repeat Steps 1, 2, 4, and 5 on the secondary MGMT node.

NOTE: When run on the secondary MGMT node, Step 5 must use the

`--disablerepo="*"` parameter or the command will fail.

NOTE: In order to copy the Lustre RPMs from MGMT0 to MGMT1, use this command:

```
[root@n000]# scp /tmp/lustre* root@MGMT1:tmp
```

- Update the Lustre RPMs on the diskless image, by entering:

```
[root@n000]# yum update -y --installroot=
/mnt/mouse/images/pristine/$(nodeattr -VU ver)/appliance.x86_64
/tmp/lustre-*.rpm
```

- From the primary MGMT node, power down the secondary MGMT, MGS, MDS, and OSS nodes, by entering:

```
[root@n000]# pm -0 system_name[001-007]
```

Example:

```
[root@snx11000n000 ~]# pm -0 snx11000n[001-007]
```

Wait until these nodes have powered down before proceeding.

- Reboot the primary MGMT node by entering:

```
[root@n000]# reboot
```

Wait several minutes, until the primary MGMT node fully reboots.

- Log in via SSH to the primary MGMT node.

- Change to root by entering:

```
[root@n000]# sudo su
```

- Power on the secondary MGMT, MDS, and MGS nodes, by entering:

```
[root@n000]# pm -1 system_name[001-003]
```

Examples:

```
[root@snx11000n000 ~]# pm -1 snx11000n[001-003]
Command completed successfully
[root@snx11000n000 ~]# pm -q
on: snx11000n[000-003]
off: snx11000n[004-007]
unknown:
[root@snx11000n000 ~]# pm -1 snx11000n[004-007]
Command completed successfully
[root@snx11000n000 ~]# pm -q
on: snx11000n[000-007]
off:
unknown:
```

Wait until these nodes have finished fully booting.

13. Power on the OSS nodes (in groups of no more than 50 at a time), by entering:

```
[root@n000]# pm -1 OSS_nodes_#1-50
```

Wait until all nodes have fully booted and check all nodes have been updated, using the following command:

```
[root@n000]# pdsh -a rpm -qa | grep HOTFIX | dshbak -c
```

If any nodes show that some system updates failed to be installed, repeat the update for that node only. Example :

```
[root@localhost-mgmt ~]# pdsh -a "rpm -qa | grep HOTFIX" | dshbak -c
```

```
-----
snx11000n[000,002-007]
-----
lustre-tests-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-ldiskfs-3.3.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-modules-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
-----
snx11000n001
-----
lustre-ldiskfs-3.3.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
```

In the preceding example, node MGMT1 (n001) did not get fully updated, as shown by its omission from the first output grouping and by the separate output that has only one line. In this case you would need to rerun the update for this node as shown in step 7. If the `yum update` option fails to update RPM, rerun using the `install` option.

Following is an example output after a successful update:

```
[root@snx11000n000 ~]# pdsh -a "rpm -qa | grep HOTFIX36" | dshbak -c
```

```
-----
snx11000n[000-001]
-----
lustre-tests-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-debuginfo-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-ldiskfs-3.3.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-source-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-headers-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-modules-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-ldiskfs-debuginfo-3.3.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
-----
snx11000n[002-007]
-----
lustre-tests-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-debuginfo-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-source-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-ldiskfs-3.3.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-headers-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-modules-2.1.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
lustre-ldiskfs-debuginfo-3.3.0.x2-74_2.99.P87_HOTFIX36_2.6.32_131.21.1.e16.lustre.3025.x86_64_g4e0816a.x86_64
```

14. Mount Lustre by entering:

```
[root@n000]# /opt/xyratex/bin/cscli mount -c cluster_name -f filesystem_name
```

NOTE: If MGMT1 fails to update, clean up by executing the following commands:

```
[root@n001]# yum clean all
[root@n001]# rpm -e lustre -ldiskfs
```

This completes the process to update Lustre RPMs.

9.2 Lustre live upgrade

This section describes procedures for applying system patches (HotFixes) without taking the Lustre file system offline. (Previously administrators were required to take Lustre offline before applying any Lustre updates.)

This document includes the following procedures:

- Installing HotFix RPMs. See page 134
- Using HA to Upgrade MDS, MGS. See page 135

9.2.1 Requirements

- **Prerequisites:** HotFix OSG-2 is required for this procedure on 1.2.0 systems only. This fix is incorporated in release 1.2.1 systems and higher.
- **System access requirements:** Certain Sonexion procedures require you to have root (super user) access. Root access is required to perform these procedures on a Sonexion system.
- **Service Interruption:** The procedure can be applied to a live system with no service interruption.
- **Required files:** To obtain the RPM files needed for this procedure, contact your Cray service provider.

9.2.2 Installing HotFix RPMs

1. Log in to the primary MGMT node via SSH by entering:

```
$ ssh -l admin primary_MGMT_node
```

2. Change to root user by entering:

```
[admin@n000]$ sudo su -
```

3. Copy the Lustre RPMs to /tmp on the primary MGMT node. (If these are in a Zip file, it should be unzipped first.) copy the Lustre RPMs to a directory in /tmp, for example /tmp/lustre_rpms.fix_xx.

4. Install the RPMs on the primary MGMT node by entering:

```
[root@n000]# yum install
-y /tmp/lustre_rpms.fix_xx/lustre-*.rpm
```

5. Copy the RPMs to /tmp on the secondary MGMT node, and install.

NOTE: The `disablerepo` argument must be included on the `yum` command for the secondary MGMT node:

```
n001# yum install --disablerepo="*"
-y /tmp/lustre_rpms.fix_xx/lustre-*.rpm
```

6. Update the Lustre RPMs on the diskless image by entering:

```
[root@n000]# yum install -y
--installroot=/mnt/mouse/images/pristine/$(nodeattr -VU ver)/appliance.x86_64
/tmp/lustre_rpms.fix_xx/lustre-*.rpm
```

9.2.3 Using HA to Upgrade MDS, MGS, and OSS

Prerequisite

The following procedure requires all resources to be running on their primary nodes. Failback should be used where necessary.

- All OSS nodes must be up with the appropriate OSTs.
- Verify that the MDT (**md66-group**) begins on node 03 and that the MGS (**md65-group**) begins on node 02, by entering:

```
[MDS]# crm_mon -1
```

Following is an example output:

```
[root@snx11000n003 ~]# crm_mon -1
=====
Last updated: Fri Oct 19 12:22:34 2012
Last change: Fri Oct 19 10:03:56 2012 via crm_resource on snx11000n003
Stack: Heartbeat
Current DC: snx11000n003 (acb50dfd-e3de-4623-afef-bc68cfc51848) - partition with quorum
Version: 1.1.6.1-2.e16-0c7312c689715e096b716419e2ebc12b57962052
2 Nodes configured, unknown expected votes
12 Resources configured.
=====
Online: [ snx11000n003 snx11000n002 ]
snx11000n003-stonith (stonith:external/ipmi): Started snx11000n003
snx11000n002-stonith (stonith:external/ipmi): Started snx11000n002
snx11000n003_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
Started snx11000n003
snx11000n002_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
Started snx11000n002
baton (ocf::heartbeat:baton): Started snx11000n003
snx11000n002_ibstat (ocf::heartbeat:ibstat): Started snx11000n002
Resource Group: snx11000n003_md66-group
snx11000n003_md0-raid (ocf::heartbeat:XYRAID): Started snx11000n003
snx11000n003_md66-raid (ocf::heartbeat:XYRAID): Started snx11000n003
snx11000n003_md66-fsfs (ocf::heartbeat:XYMNTR): Started snx11000n003
Resource Group: snx11000n003_md65-group
snx11000n003_md65-raid (ocf::heartbeat:XYRAID): Started snx11000n002
snx11000n003_md65-fsfs (ocf::heartbeat:XYMNTR): Started snx11000n002
```

In the command output, the highlighted phrases indicate the following:

- The **snx11000n003_md66-group** resources are started on **n003**.
- The **snx11000n003_md65-group** resources are started on **n002**.

If MDT and MGS are not running on their primary servers, fail back the entity that is not on its primary server.

Upgrade MDS node

1. Change to root user by entering:

```
[admin@n000]$ sudo su -
```

2. Log in to the MDS node via SSH, by entering the following on the primary MGMT node:

```
[root@n000]# ssh MDS_node
```

For example:

```
[snx11000n000 ~]$ ssh snx11000n003
```

3. Fail over resources from the MDS node to the MGS node, by entering:

```
[MDS]# failover_xyraid
```

4. Verify that MDT (**md66**) is mounted on the MGS node, by entering:

```
[MDS]# crm_mon -1
```

Following is an example output:

```
[root@snx11000n003 ~]# crm_mon -1
=====
Last updated: Fri Oct 19 12:22:34 2012
Last change: Fri Oct 19 10:03:56 2012 via crm_resource on snx11000n003
Stack: Heartbeat
Current DC: snx11000n003 (acb50dfd-e3de-4623-afef-bc68cfc51848) - partition with quorum
Version: 1.1.6.1-2.el6-0c7312c689715e096b716419e2ebc12b57962052
2 Nodes configured, unknown expected votes
12 Resources configured.
=====
Online: [ snx11000n003 snx11000n002 ]
snx11000n003-stonith (stonith:external/ipmi): Started snx11000n003
snx11000n002-stonith (stonith:external/ipmi): Started snx11000n002
snx11000n003_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
Started snx11000n003
snx11000n002_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
Started snx11000n002
baton (ocf::heartbeat:baton): Started snx11000n003
snx11000n002_ibstat (ocf::heartbeat:ibstat): Started snx11000n002
Resource Group: snx11000n003_md66-group
snx11000n003_md0-raid (ocf::heartbeat:XYRAID): Started snx11000n002
snx11000n003_md66-raid (ocf::heartbeat:XYRAID): Started snx11000n002
snx11000n003_md66-fsys (ocf::heartbeat:XYMNTR): Started snx11000n002
Resource Group: snx11000n003_md65-group
snx11000n003_md65-raid (ocf::heartbeat:XYRAID): Started snx11000n002
snx11000n003_md65-fsys (ocf::heartbeat:XYMNTR): Started snx11000n002
```


In the above output, note that, following the failover, both **mdraid** resources (**snx11000n003_md66-group** and **snx11000n003_md65-group**) resources are now started on **n002**.

- Exit from the MDS back to the primary MGMT node, by entering:

```
[MDS]# exit
```

- Reboot the MDS node by entering the following on the primary MGMT node:

```
[root@n000]# pm -0 MDS_node
```

Wait for one minute, and then run:

```
[root@n000]# pm -1 MDS_node
```

For example:

```
[root@snx11000n000 ~]# pm -0 snx11000n003
```

```
[root@snx11000n000 ~]# pm -1 snx11000n003
```

- It takes 5 or more minutes for the MDS node to reboot and be HA-ready. To confirm that the node is ready, enter `crm_mon -1`, which should have output similar to that shown in step 4. (The node should be listed on the line beginning `Online:` heading.)
- Verify that the MDS node rebooted with the new HotFix, by entering the following on the primary MGMT node:

```
[root@n000]# pdsh -w MDS_node rpm -qa | awk -F. '/lustre-2.1/{print $1"."$5"."$6}' | sort | dshbak -c
```

Following is an example output:

```
[root@snx11000n000 ~]# pdsh -w snx11000n003 rpm -qa | awk -F. '/lustre-2.1/{print $1"."$5"."$6}' | sort | dshbak -c
```

```
-----  
snx11000n003
```

```
-----  
lustre-2.99.P68_HOTFIX28_2
```

- Log in to the MDS node via SSH by entering the following on the primary MGMT node:

```
[root@n000]# ssh MDS_node
```

For example:

```
[root@snx11000n000 ~]# ssh snx11000n003
```

- Fail back MDT (**md66**) by entering:

```
[MDS]# failback_xyraid
```

- Verify that MDT (**md66**) is remounted on the MDS node, by entering:

```
[MDS]# crm_mon -1
```

Following is an example output:

```
[snx11000n003 ~]# crm_mon -1
```

```
=====
```

```
Last updated: Fri Oct 19 12:22:34 2012
```

```
Last change: Fri Oct 19 10:03:56 2012 via crm_resource on snx11000n003
```

```
Stack: Heartbeat
Current DC: snx11000n003 (acb50dfd-e3de-4623-afef-bc68cfc51848) - partition with quorum
Version: 1.1.6.1-2.e16-0c7312c689715e096b716419e2ebc12b57962052
2 Nodes configured, unknown expected votes
12 Resources configured.
=====
Online: [ snx11000n003 snx11000n002 ]
snx11000n003-stonith (stonith:external/ipmi): Started snx11000n003
snx11000n002-stonith (stonith:external/ipmi): Started snx11000n002
snx11000n003_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
  Started snx11000n003
snx11000n002_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
  Started snx11000n002
baton (ocf::heartbeat:baton): Started snx11000n003
snx11000n002_ibstat (ocf::heartbeat:ibstat): Started snx11000n002
Resource Group: snx11000n003_md66-group
  snx11000n003_md0-raid (ocf::heartbeat:XYRAID): Started snx11000n003
  snx11000n003_md66-raid (ocf::heartbeat:XYRAID): Started snx11000n003
  snx11000n003_md66-fsfs (ocf::heartbeat:XYMNTR): Started snx11000n003
Resource Group: snx11000n003_md65-group
  snx11000n003_md65-raid (ocf::heartbeat:XYRAID): Started snx11000n002
  snx11000n003_md65-fsfs (ocf::heartbeat:XYMNTR): Started snx11000n002
```

In the above output, note that MDT (**md66**) is back on the MDS node, n003.

- Verify that Lustre recovery has completed by entering:

```
n003# lctl get_param mdt.*.recovery_status
```

You can proceed with the next steps when the output includes this indication:

```
status: COMPLETE
```

Upgrade the MGS node

- Change to root user by entering:

```
[MGMT0]$ sudo su -
```

- Log in to the MGS node via SSH, by entering the following on the primary MGMT node:

```
[root@n000]# ssh MGS_node
```

For example:

```
[snx11000n000 ~]$ ssh snx11000n002
```

- Verify that MDT (**md66**) is mounted on the MDS node and that MGS (**md65**) is mounted on the MGS node, by entering:

```
[MGS]# crm_mon -1
```

Following is an example output:

```
[root@snx11000n002 ~]# crm_mon -1
=====
Last updated: Fri Oct 19 12:22:34 2012
Last change: Fri Oct 19 10:03:56 2012 via crm_resource on snx11000n002
Stack: Heartbeat
```

```

Current DC: snx11000n002 (acb50dfd-e3de-4623-afef-bc68cfc51848) - partition with quorum
Version: 1.1.6.1-2.el6-0c7312c689715e096b716419e2ebc12b57962052
2 Nodes configured, unknown expected votes
12 Resources configured.
=====
Online: [ snx11000n003 snx11000n002 ]
  snx11000n003-stonith (stonith:external/ipmi): Started snx11000n003
  snx11000n002-stonith (stonith:external/ipmi): Started snx11000n002
  snx11000n003_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
Started snx11000n003
  snx11000n002_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
Started snx11000n002
  baton (ocf::heartbeat:baton): Started snx11000n003
  snx11000n002_ibstat (ocf::heartbeat:ibstat): Started snx11000n002
Resource Group: snx11000n003_md66-group
  snx11000n003_md0-raid (ocf::heartbeat:XYRAID): Started snx11000n003
  snx11000n003_md66-raid (ocf::heartbeat:XYRAID): Started snx11000n003
  snx11000n003_md66-fsfs (ocf::heartbeat:XYMNTTR): Started snx11000n003
Resource Group: snx11000n003_md65-group
  snx11000n003_md65-raid (ocf::heartbeat:XYRAID): Started snx11000n002
  snx11000n003_md65-fsfs (ocf::heartbeat:XYMNTTR): Started snx11000n002

```

In the above output, note that MGS and MDS are on their primary servers (**snx11000n003_md66-group** resources on **n003** and **snx11000n003_md65-group** on **n002**).

16. Fail over resources from the MGS node to the MDS node, by entering:

```
[MGS]# failover_xyraid
```

17. Verify that MGS (**md65**) is mounted on the MDS node (**n003** rather than **n002** as shown in the previous output), by entering:

```
[MGS]# crm_mon -l
```

Following is an example output:

```

[root@snx11000n002 ~]# crm_mon -l
=====
Last updated: Fri Oct 19 12:22:34 2012
Last change: Fri Oct 19 10:03:56 2012 via crm_resource on snx11000n002
Stack: Heartbeat
Current DC: snx11000n002 (acb50dfd-e3de-4623-afef-bc68cfc51848) - partition with quorum
Version: 1.1.6.1-2.el6-0c7312c689715e096b716419e2ebc12b57962052
2 Nodes configured, unknown expected votes
12 Resources configured.
=====
Online: [ snx11000n003 snx11000n002 ]
  snx11000n003-stonith (stonith:external/ipmi): Started snx11000n003
  snx11000n002-stonith (stonith:external/ipmi): Started snx11000n002
  snx11000n003_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
Started snx11000n003
  snx11000n002_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
Started snx11000n002
  baton (ocf::heartbeat:baton): Started snx11000n003
  snx11000n002_ibstat (ocf::heartbeat:ibstat): Started snx11000n002
Resource Group: snx11000n003_md66-group

```

```

snx11000n003_md0-raid (ocf::heartbeat:XYRAID): Started snx11000n003
snx11000n003_md66-raid (ocf::heartbeat:XYRAID): Started snx11000n003
snx11000n003_md66-fsfs (ocf::heartbeat:XYMNTTR): Started snx11000n003
Resource Group: snx11000n003_md65-group
snx11000n003_md65-raid (ocf::heartbeat:XYRAID): Started snx11000n003
snx11000n003_md65-fsfs (ocf::heartbeat:XYMNTTR): Started snx11000n003

```

In the above output, note that, following the failover, both **mdraid** resources (**snx11000n003_md66-group** and **snx11000n003_md65-group**) are now started on node **n003**.

18. Exit from the MDS back to the primary management node by entering:

```
[MDS]# exit
```

19. Reboot the MGS node, by entering the following on the primary MGMT node:

```
[root@n000]# pm -0 MGS_node
```

Wait for one minute, and then run:

```
[root@n000]# pm -1 MGS_node
```

For example:

```
[root@snx11000n000 ~]# pm -0 snx11000n002
[root@snx11000n000 ~]# pm -1 snx11000n002
```

20. Verify that the MGS node rebooted with the new HotFix, by entering the following on the primary MGMT node:

```
[root@n000]# pdsh -w MGS_node rpm -qa | awk -F. '/lustre-2.1/
{print $1"."$5"."$6}' | sort | dshbak -c
```

Following is an example output:

```
[root@snx11000n000 ~]# pdsh -w snx11000n002 rpm -qa | awk -F.
'/lustre-2.1/ {print $1"."$5"."$6}' | sort | dshbak -c
-----
snx11000n002
-----
lustre-2.99.P68_HOTFIX28_2
```

21. Verify that the MGT node has rejoined the cluster by running `crm_mon -1` and seeing the same output as in step 15, page 138.
22. Log in to the MGS node via SSH, by entering the following on the primary MGMT node:

```
[root@n000]# ssh MGS_node
```

For example:

```
[root@snx11000n000 ~]# ssh snx11000n002
```

23. Fail back MGS (**md65**) by entering:

```
[MGS]# failback_xyraid
```

24. Verify that MGS (**md65**) is remounted on the MGS node, by entering:

```
[MGS]# crm_mon -1
```

Following is an example output:

```
[snx11000n002 ~]# crm_mon -1
=====
Last updated: Fri Oct 19 12:22:34 2012
Last change: Fri Oct 19 10:03:56 2012 via crm_resource on snx11000n002
Stack: Heartbeat
Current DC: snx11000n002 (acb50dfd-e3de-4623-afef-bc68cfc51848) - partition with quorum
Version: 1.1.6.1-2.el6-0c7312c689715e096b716419e2ebc12b57962052
2 Nodes configured, unknown expected votes
12 Resources configured.
=====
Online: [ snx11000n003 snx11000n002 ]
snx11000n003-stonith (stonith:external/ipmi): Started snx11000n003
snx11000n002-stonith (stonith:external/ipmi): Started snx11000n002
snx11000n003_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate): Started
snx11000n003
snx11000n002_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate): Started
snx11000n002
baton (ocf::heartbeat:baton): Started snx11000n003
snx11000n002_ibstat (ocf::heartbeat:ibstat): Started snx11000n002
Resource Group: snx11000n003_md66-group
snx11000n003_md0-raid (ocf::heartbeat:XYRAID): Started snx11000n003
snx11000n003_md66-raid (ocf::heartbeat:XYRAID): Started snx11000n003
snx11000n003_md66-fsys (ocf::heartbeat:XYMNTR): Started snx11000n003
Resource Group: snx11000n003_md65-group
snx11000n003_md65-raid (ocf::heartbeat:XYRAID): Started snx11000n002
snx11000n003_md65-fsys (ocf::heartbeat:XYMNTR): Started snx11000n002
```

This completes the steps for using HA to upgrade MDS and MGS.

Verify that all OSS nodes are up and accessible

25. Verify that all OSS nodes are up and accessible, by entering:

```
[root@n000]# pdsh -g all date
```

26. Verify that all OSS nodes have 4 OSTs mounted, by entering:

```
[root@n000]# /opt/xyratex/bin/cscli fs_info
```

NOTE: If any OSTs are failed over, they should be failed back prior to starting the next steps.

Upgrade secondary OSS nodes

The following steps describe how to upgrade the secondary OSS nodes. Secondary nodes are those with odd-numbered OSS node names.

27. Log in to the primary MGMT node via SSH by entering:

```
$ ssh -l admin primary_MGMT_node
```

28. Change to root user by entering:

```
[admin@n000]$ sudo su -
```

29. Fail over the OSTs from the secondary OSS nodes to their partner nodes, which are identified as primary OSS nodes, by entering:

```
[root@n000]# pdsh -g oss=secondary failover_xyraid
```

30. From a new terminal, you can watch the failover complete. On the primary MGMT node, as root user, run:

```
[root@n000]# watch "pdsh -g oss 'mount -t lustre | wc -l' | dshbak -c"
```

31. Once all of the OSTs have failed over, enter the following:

```
[root@snx11000n000 ~]# cscli -c snx11000n show_nodes
```

In the output of the above command, confirm that 8 targets are mounted on the primary OSS servers, 0 on the secondary servers.

32. Get a list of all the secondary OSS nodes by entering:

```
[root@n000]# pm -q $(nodeattr -s oss=secondary)
```

33. Turn off all secondary OSS nodes by entering:

```
[root@n000]# pm -0 $(nodeattr -s oss=secondary)
```

34. Turn on the secondary OSS nodes. Perform one of the following:

- For clusters with 100 or fewer secondary OSS nodes, run:

```
[root@n000]# pm -1 $(nodeattr -s oss=secondary)
```

- For clusters with more than 100 secondary OSS nodes, run the following command for the first hundred secondary OSS nodes:

```
[root@n000]# pm -1 snx11000n[first_100_secondary_OSS_nodes]
```

Wait 5 minutes. Then run the command for each remaining hundred secondary OSS nodes:

```
[root@n000]# pm -1 snx11000n[additional_100_secondary_OSS_nodes]
```

NOTE: After the nodes boot up, it takes about 10 minutes before they become accessible via SSH.

35. Check the version of the HotFix on the secondary OSS nodes, by entering:

```
[root@n000]# pdsh -g oss=secondary rpm -qa | awk -F.
'/lustre-2.1/ {print $1"."$5"."$6}' | sort | dshbak -c
```

Following is an example output showing the latest HotFix applied.

```
[root@snx11000n000 ~]# pdsh -g oss=secondary rpm -qa | awk -F.
'/lustre-2.1/ {print $1"."$5"."$6}' | sort | dshbak -c
-----
snx11000n[005,007,009,011]
-----
lustre-2.99.P68_HOTFIX28_2
```

36. Wait 10 minutes. Verify that the nodes have joined the cluster by running `crm_mon -1` and seeing output similar to that in step 15.

37. Fail back the OSTs to the secondary OSS nodes, by entering:

```
[root@n000]# pdsh -g oss=secondary failback_xyraid
```

38. From a new terminal, you can watch the failback complete. On the primary MGMT node, as root user, run:

```
[root@n000]# watch "pdsh -g oss 'mount -t lustre | wc -l' | dshbak -c"
```

39. Once all of the OSTs have failed back, enter the following:

```
[root@snx11000n000 ~]# cscli -c snx11000n show_nodes
```

In the output of the above command, confirm that 4 targets are mounted on each primary OSS server, 4 on each secondary server.

40. Confirm that Lustre recovery has completed by entering:

```
[root@n000]# pdsh -g oss=secondary "lctl get_param
obdfilter.*.recovery_status | grep status"
```

Make sure the status is `COMPLETE` for all OSTs before proceeding.

Upgrade primary OSS nodes

The following steps describe how to upgrade the primary OSS nodes. Primary nodes are those with even numbered OSS node names.

41. Log in to the primary MGMT node via SSH, by entering:

```
$ ssh -l admin primary_MGMT_node
```

42. Change to root user by entering:

```
[admin@n000]$ sudo su -
```

43. Fail over the OSTs from the primary OSS nodes to their partner nodes, which are identified as secondary OSS nodes, by entering:

```
[root@n000]# pdsh -g oss=primary failover_xyraid
```

44. From a new terminal, you can watch the failover complete. On the primary MGMT node, as root user, run:

```
[root@n000]# watch "pdsh -g oss 'mount -t lustre | wc -l' | dshbak -c"
```

45. Once all of the OSTs have failed over, enter the following:

```
[root@snx11000n000 ~]# cscli -c snx11000n show_nodes
```

In the output of the above command, confirm that 0 targets are mounted on the primary OSS servers, 8 on the secondary servers.

46. Get a list of all the primary OSS nodes by entering:

```
[root@n000]# pm -q $(nodeattr -s oss=primary)
```

47. Turn off all of the primary OSS nodes by entering:

```
[root@n000]# pm -0 $(nodeattr -s oss=primary)
```

48. Turn on the primary OSS nodes. Perform one of the following:

- For clusters with 100 or less primary OSS nodes, run:

```
[root@n000]# pm -c $(nodeattr -s oss=primary)
```

- For clusters with more than 100 primary OSS nodes, run the following command for the first half of the primary OSS nodes:

```
[root@n000]# pm -c snx11000n[first_half_of_primary_OSS_nodes]
```

Wait 5 minutes, and then run the command for the remaining half of the primary OSS nodes:

```
[root@n000]# pm -c snx11000n[second_half_of_primary_OSS_nodes]
```

NOTE: After the nodes boot up, it takes about 10 minutes before they become accessible via SSH.

49. Check the version of the HotFix on the primary OSS nodes, by entering:

```
[root@n000]# pdsh -g oss=primary rpm -qa | awk -F.
'/lustre-2.1/ {print $1"."$5"."$6}' | sort | dshbak -c
```

Following is an example output showing the latest HotFix applied:

```
[root@snx11000n000 ~]# pdsh -g oss=primary rpm -qa | awk -F.
'/lustre-2.1/ {print $1"."$5"."$6}' | sort | dshbak -c
-----
snx11000n[004,006,008,010]
-----
lustre-2.99.P68_HOTFIX28_2
```

50. Wait 10 minutes. Verify that the nodes have joined the cluster by running `crm_mon -1` and seeing output similar to that in step 15.

51. Verify again that all of the OSS nodes are up and accessible via SSH, by entering:

```
[root@n000]# pdsh -g oss 'uname -r' | dshbak -c
```

52. Fail back the OSTs to the primary OSS nodes, by entering:

```
[root@n000]# pdsh -g oss=primary failback_xyraid
```

53. From a new terminal, you can watch the failback complete. On the primary MGMT node, as root user, run:

```
[root@n000]# watch "pdsh -g oss 'mount -t lustre | wc -1' | dshbak -c"
```

54. Once all of the OSTs have failed back, enter the following:

```
[root@snx11000n000 ~]# cscli -c snx11000n show_nodes
```

55. In the output of the above command, confirm that 4 targets are mounted on each primary OSS server, 4 on each secondary server.

56. As a final check, verify that all OSS, MDS, and MGS nodes are all at the latest HotFix level, by entering:

```
[root@n000]# pdsh -a rpm -qa | awk -F. '/lustre-2.1/ {print
$1"."$5"."$6}' | sort | dshbak -c
```

Following is an example output:

```
[root@snx11000n000 ~]# pdsh -a rpm -qa | awk -F.
'/lustre-2.1/ {print $1"."$5"."$6}' | sort | dshbak -c
-----
snx11000n[002-011]
```



```
-----  
lustre-2.99.P68_HOTFIX28_2
```

This completes the steps to upgrade the OSS nodes.

Upgrade the management nodes

The new Lustre version will be picked up on the management nodes the next time the servers are rebooted, or the next time the Lustre modules are loaded.

10. CSSM CLI User Documentation

This appendix provides reference information for Sonexion's CLI command interface.

CLI commands are organized by mode; that is, certain commands are available according to the mode (state) of the Sonexion system. Two modes are relevant to customers – Customer Wizard Mode and Daily Mode. A third mode, OEM Mode, is relevant only to Manufacturing and factory personnel. OEM Mode commands are not included in this document.

- Customer Wizard Mode
- Daily Mode

10.1.1 Customer Wizard mode

Use Wizard mode to configure the Sonexion system for customer use (after factory provisioning and before daily operations mode). Customer Wizard (custWizard) Mode commands are available after the Sonexion system has been fully provisioned and before the system runs in Daily Mode. These commands enable users to specify customer configuration settings, apply or reset network cluster settings, obtain FRU information, upgrade Sonexion software on Lustre nodes, and toggle between Customer Wizard and Daily Modes.

10.1.2 Daily mode

Use Daily Mode mode when the Sonexion system is fully operational and available to manage the Lustre file system and cluster nodes.

Daily Mode commands are available after the Sonexion system has been fully provisioned and configured for customer use. These commands enable users to fully manage the Lustre

file system and cluster nodes, including mount/unmount, power-cycle, failover/failback, and control node filters and exports. Daily Mode commands also enable users to obtain FRU information and upgrade Sonexion software on Lustre nodes.

10.1.3 CLI command summary

Table 5 summarizes the CLI commands, with columns indicating the mode or modes that include each command.

Table 5. CLI Command Summary

Wizard Mode	Daily Mode	Command	Description
<i>Network Setup Commands</i>			
x		set_network	Specifies a Sonexion network setup.
x		show_network_setup	Shows a Sonexion network setup.
x		reset_network_setup	Resets the network setup of an existing Sonexion system.
x		apply_network_setup	Applies a network setup to a Sonexion system.
<i>User setup commands</i>			
x		get_lustre_users_ad	Shows the Lustre file system's AD settings.
x		get_lustre_users_ldap	Shows the Lustre file system's LDAP settings.
x		get_lustre_users_nis	Shows configured NIS settings.
x		set_lustre_users_ad	Sets the Lustre file system's AD configuration.
x		set_lustre_users_ldap	Sets the Lustre file system's LDAP configuration.
x		clear_lustre_users_ad	Clears the Lustre file system's AD settings.
x		set_lustre_users_nis	Configures Filesystem NIS settings.
x		clear_lustre_users_ldap	Clears the Lustre file system's LDAP settings.
x		clear_lustre_users_nis	Clears the Lustre file system's NIS settings.
<i>System alert commands</i>			
x	x	alerts	Displays current and historical system health alerts.
x	x	alerts_config	Shows and updates the alerts configuration.
x	x	alerts_notify	Enables or disables alert notifications.
<i>Node Control Commands</i>			
x	x	autodiscovery_mode	Enables or disables auto-discovery mode on system nodes.
	x	failback	Fails back resources for the specified node.

Wizard Mode	Daily Mode	Command	Description
	x	failover	Fails over resources to the specified node.
x	x	mount	Mounts the Lustre file system in the cluster.
x	x	unmount	Unmounts Lustre clients or targets on the file system.
x	x	power_manage	Specifies node power management options.
	x	show_nodes	Displays node information.

Administrative Commands

x	x	fs_info	Retrieves file system information.
x	x	cluster_mode	Toggles the system among 'daily mode', 'custWizard' and 'pre-shipment' modes.
x	x	fru	Retrieves FRU (replacement) information.
x	x	list	Lists all supported commands.
x	x	syslog	Retrieves syslog entries.
x	x	batch	Runs a sequence of CSCLI commands in a batch file.
x	x	ip_routing	Manages IP routing.
x	x	set_admin_passwd	Changes administrator user password on an existing Sonexion system.

Configuration Commands

	x	configure_hosts	Configures host names for discovered nodes.
	x	configure_oss	Configures a new OSS node.
	x	show_new_nodes	Displays a table with new OSS nodes and their resources.

Filter Commands

	x	create_filter	Creates customer filters for nodes.
	x	delete_filter	Deletes customer filters for nodes.
	x	show_filters	Shows customized and predefined node filters.

Updating System Software

	x	prepare_update	Updates the specified node.
	x	split_ha_partners	Splits a set of nodes into two sets with each set containing no HA pairs.
	x	update_node	Updates the software version on the specified node.
	x	show_node_versions	Shows the current software version on the specified nodes.

Wizard Mode	Daily Mode	Command	Description
	x	show_version_nodes	Shows all nodes at the specified software version.
	x	show_update_versions	Shows available software versions in the Sonexion Management Server repository.

Managing node position in a Sonexion rack

	x	get_rack_position	Indicates the specified node's position in the Sonexion rack.
	x	set_rack_position	Changes a given node position in the Sonexion rack.

Monitoring System Health

x	x	monitor	Monitors the current health of the cluster nodes and elements.
x	x	netfilter_level	Manages the netfilter level.

Enabling RAID Checks

	x	raid_check	Enables RAID checks on RAID devices.
x	x	rebuild_rate	Manages the RAID rebuild rate.
	x	set_date	Manages the date setting on the Sonexion system
	x	set_timezone	Manages the timezone setting on the Sonexion system.
x	x	sm	Manages the InfiniBand Subnet Manager.
x	x	support_bundle	Manages support bundles and support bundle settings.

10.2 Summary of changes in release CS 1.3.1

Table 6 and Table 7 show commands that were added to this release:

Table 6. New CLI Commands, Customer Wizard Mode

No.	Command	Description	Component
1	<code>clear_lustre_users_nis</code>	Clear Filesystem NIS settings	NIS Support
2	<code>get_lustre_users_nis</code>	Show configured NIS settings	NIS Support
3	<code>set_lustre_users_nis</code>	<code>set_lustre_users_nis</code>	NIS Support
4	<code>support_bundle</code>	Manage support bundles and support bundle settings	Support Bundles

Table 7. New CLI Commands, Daily Mode

No.	CLI Command	Description	Component
1	<code>configure_hosts</code>	Configure hostname for discovered node	SSU Addition
2	<code>raid_check</code>	Enable RAID checks on RAID devices	XYRAID
3	<code>support_bundle</code>	Manage support bundles and support bundle settings	Support Bundles

The following options were added to the `alerts_config` command. See “Manage the alerts configuration”, page 160:

- `email_off`
- `thresholds`
- `email_update`
- `email_server_update`
- `email_delete`
- `email_add`
- `email_on`
- `email_server`
- `emails`

10.3 Network setup commands

The `network_setup` command manages network parameters for the Lustre file system. This command includes functions to show, set, apply, and reset Lustre network parameters.

10.3.1 Show network parameters

The `show_network_setup` command displays the Lustre network configuration. If the Lustre network is not yet configured, no parameters are shown.

Synopsis

```
$ cscli show_network_setup [-h] [-c cluster_name]
```

where:

Optional Arguments	Description
<code>-h</code> <code>--help</code>	Shows the help message and exits.
<code>-c</code> <i>cluster_name</i> <code>--cluster</code> <i>cluster_name</i>	Specifies the cluster name.

10.3.2 Set network parameters

The `set_network` command specifies new Lustre network parameters and adds them to the database.

Synopsis

```
$ cscli set_network [-h] -k netmask -r ipranges [-d dns] [-t ntp] [-c cluster_name]
```

where:

Optional Arguments	Description
<code>-h --help</code>	
<code>-k <i>netmask</i></code> <code> --netmask <i>netmask</i></code>	Specifies the network mask value of the <i>ip</i> address.
<code>-r <i>ipranges</i></code> <code> --range <i>ipranges</i></code>	Specifies the <i>IP</i> address range.
<code>-d <i>dns</i> --dns <i>dns</i></code>	Specifies the <i>DNS</i> server <i>IP</i> address (optional).
<code>-t <i>ntp</i> --ntp <i>ntp</i></code>	Specifies the <i>ntp</i> server's <i>IP</i> address (optional).
<code>-c <i>cluster_name</i></code> <code> --cluster <i>cluster_name</i></code>	Specifies the cluster name.

10.3.3 Reset network parameters

The `reset_network_setup` command resets the Lustre network parameters by removing old values from the database and replacing them with default values.

Synopsis

```
$ cscli reset_network_setup [-h] [-y] [-c cluster_name]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-y --yes</code>	Confirms the action to reset the network parameters.
<code>-c <i>cluster_name</i></code> <code> --cluster <i>cluster_name</i></code>	Specifies the cluster name.

10.3.4 Apply network parameters

The `apply_network_setup` command applies new Lustre network parameters to the database.

Synopsis

```
$ cscli apply_network_setup [-h] [--yes] [-c cluster_name]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code> --yes</code>	Confirms the action that network setup parameters were applied.
<code>-c <i>cluster_name</i></code> <code> --cluster <i>cluster_name</i></code>	Specifies the cluster name.

10.4 User setup commands

User setup commands include functions to configure the system's AD and LDAP settings and change the administrative user's password (used for CSSM login).

10.4.1 Get the file system's AD settings

The `get_lustre_users_ad` command retrieves the file system's AD settings.

Synopsis

```
$ cscli get_lustre_users_ad [-h] [-f fs_name] [--yaml-format]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-f <i>fs_name</i></code> <code> --fs <i>fs_name</i></code>	Specifies the file system name.
<code>--yaml-format</code>	Shows the <i>ad</i> configuration in YAML file format.

10.4.2 Get the file system's LDAP settings

The `get_lustre_users_ldap` command retrieves the file system's *LDAP* settings.

Synopsis

```
$ cscli get_lustre_users_ldap [-h] [--yaml-format]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
--yaml-format	Shows the LDAP configuration in YAML file format.

10.4.3 Get the file system's NIS settings

The `get_lustre_users_nis` command retrieves and displays the file system's NIS settings.

Synopsis

```
$ cscli get_lustre_users_nis [-h] [-f fs_name] [--yaml-format]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-f <i>fs_name</i> --f <i>fs_name</i>	Shows the <i>nis</i> file system name.
--yaml-format	Shows the <i>nis</i> configuration in YAML file format.

10.4.4 Set the file system's AD settings

The `set_lustre_users_ad` command specifies the file system's AD settings.

Synopsis

```
$ cscli set_lustre_users_ad [-h] -f fs_name [-l ldap_uri]  
[-b base_dn] [-i bind_dn] [-p password]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-f <i>fs_name</i> --fs <i>fs_name</i>	Specifies the file system name.
-l <i>ldap_uri</i>	Specifies the LDAP URI. For example:

<code>--ldap-uri</code> <i>ldap_uri</i>	LDAP://127.0.0.1:389
<code>-b base_dn</code> <code>--base-dn base_dn</code>	Specifies the LDAP base DN.
<code>-i bind_dn</code> <code>--bind-dn bind_dn</code>	Specifies the LDAP bind DN.
<code>-p password</code> <code>--password password</code>	Specifies the LDAP bind password.

10.4.5 Set the file system's LDAP settings

The `set_lustre_users_ldap` command specifies the file system's LDAP settings.

Synopsis

```
$ cscli set_lustre_users_ldap [-h] [-N] [-l ldap_uri]
  [-b base_dn] [-u user_dn] [-G group_dn] [-s hosts_dn] [-i bind_dn]
  [-p password]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-N --noauth</code>	Disables the LDAP configuration on the system.
<code>-l <i>ldap_uri</i></code> <code>--ldap-uri <i>ldap_uri</i></code>	Specifies the LDAP URI. For example: LDAP://127.0.0.1:389
<code>-b <i>base_dn</i></code> <code>--base-dn <i>base_dn</i></code>	Specifies the LDAP base DN.
<code>-u <i>user_dn</i></code> <code>--user-dn <i>user_dn</i></code>	Specifies the LDAP user DN.
<code>-G <i>group_on</i></code> <code>--group-dn <i>group_dn</i></code>	Specifies the LDAP group DN.
<code>-s <i>hosts_dn</i></code> <code>--hosts-dn <i>hosts_dn</i></code>	Specifies the LDAP hosts DN.
<code>-i <i>bind_dn</i></code> <code>--bind-dn <i>bind_dn</i></code>	Specifies the LDAP bind DN.
<code>-p <i>password</i></code> <code>--password <i>password</i></code>	Specifies the LDAP bind password.

10.4.6 Configure the file system's NIS Settings

The `set_lustre_users_nis` command configures the file system's NIS settings.

Synopsis

```
$ cscli set_lustre_users_nis [-h] -f fs_name [-s nis_server]
[-d nis_domain]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-f <i>fs_name</i></code> <code> --fs <i>fs_name</i></code>	Shows the file system's name.
<code>-s <i>nis_server</i></code> <code> --nis_server <i>nis_server</i></code>	Shows the NIS server. For example: "10.0.0.10 10.0.0.11" or "server1 server2" .
<code>-d <i>nis_domain</i></code> <code> --nis_domain <i>nis_domain</i></code>	Shows the <i>NIS</i> domain. For example: nisdomain.

10.4.7 Clear the file system's AD settings

The `clear_lustre_users_ad` command clears the file system's AD settings.

Synopsis

```
$ cscli clear_lustre_users_ad [-h] [-f fs_name] [--all]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-f <i>fs_name</i> --fs <i>fs_name</i></code>	Specifies the file system name.
<code>--all</code>	Cleans all file systems' configurations.

10.4.8 Clear the file system's LDAP settings

The `clear_lustre_users_ldap` command clears the file system's LDAP settings.

Synopsis

```
$ cscli clear_lustre_users_ldap [--yes] [-h]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
--yes	Confirms the action to clear the file system's <i>ldap</i> settings.

10.4.9 Clear the file system's NIS settings

The `clear_lustre_users_nis` command clears the file system's NIS settings.

Synopsis

```
$ ccli clear_lustre_users_nis [-h] [-f fs_name] [--all]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-f <i>fs_name</i> --fs <i>fs_name</i>	Confirms the file system's name.
--all	Clears all the file system's configuration

10.5 System alert commands

Alert commands include functions to view and update the alerts configuration, turn on/off alert notifications, and display current and historical system alerts.

10.5.1 Display current and historic system alerts

The `alerts` command displays current and historic health alerts for system nodes and elements, and thresholds for system alerts.

Synopsis

```
$ ccli alerts [-h]
{elements_active,nodes,elements,nodes_active,thresholds}
```

where:

Positional Arguments	Description
nodes	Shows alert history for nodes.

elements	Shows alert history for elements.
nodes_active	Shows current alerts for nodes.
elements_active	Shows current alerts for elements.
thresholds	Shows editable alert thresholds and their current settings.

Optional Arguments	Description
-h --help	Shows the help message and exits.

Subcommand (alerts elements_active)

Synopsis

```
$ cscli alerts elements_active [-h] [-y] [-v] [-x]
[-n node_spec | -g genders_query] [-S element_filter]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-y --yaml	Outputs data in YAML format.
-v --verbose	Outputs extra data.
-x --unhandled	Shows alerts for notifications that have not been turned off. (Default value is all alerts are shown.)
-n <i>node_spec</i> --nodes <i>node_spec</i>	Specifies pdsh-style node hostnames (for example, <code>node[100-110,120]</code>).
-g <i>genders_query</i> --genders <i>genders_query</i>	Specifies node genders attributes query (for example, <code>mds=primary</code>).
-S <i>element_filter</i> --search <i>element_filter</i>	Specifies node genders attributes query (for example, <code>mds=primary</code>).

Subcommand (alerts nodes)

Synopsis

```
$ cscli alerts nodes [-h] [-y] [-s start_time] [-e end_time]
[-m limit] [-n node_name] [-N {down,unreachable,up}]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-y --yaml	Outputs data in YAML format.
-s <i>start_time</i> --start-time <i>start_time</i>	Specifies the alert start time in ISO-8601 format. If --start-time is not specified, then --end-time is ignored and the "last 7 days" period is used.
-e <i>end_time</i> --end-time <i>end_time</i>	Specifies the alert end time in ISO-8601 format. (Default value is "now".)
-m <i>limit</i> --max <i>limit</i>	Specifies the maximum number (limit) of alerts to display.
-n <i>node_name</i> --node <i>node_name</i>	Specifies the node for which to display alerts. Pdsh-style node masks are <u>not</u> allowed here.
-N {down,unreachable,up} --node status	Specifies node status.

Subcommand (alerts elements)

```
$ cscli alerts elements [-h] [-y] [-s start_time] [-e end_time]
[-m limit] [-n node_name] [-U {unknown,warning,ok,critical}]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-y --yaml	Outputs data in YAML format.
-s <i>start_time</i> --start-time <i>start_time</i>	Specifies the start time filter in ISO-8601 format. If --start-time is not specified, then --end-time is ignored and the "last 7 days" period is used.
-e <i>end_time</i> --end-time <i>end_time</i>	Specifies the end time filter in ISO-8601 format (default value is "now").
-m <i>limit</i> --max <i>limit</i>	Specifies the maximum number (limit) of items to display.
-n <i>node_name</i> --node <i>node_name</i>	Specifies the node for which to display items. Pdsh-style node masks are <u>not</u> allowed here.
-U {unknown,warning,ok,critical} --element status	Specifies the element's status.

Subcommand (alerts nodes_active)

```
$ cscli alerts nodes_active [-h] [-y] [-v] [-x] [-n node_spec |
-g genders_query]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-y --yaml	Outputs data in YAML format.
-v --verbose	Outputs extra data.
-x --unhandled	Shows alerts for notifications that have not been turned off (default is all alerts are shown).
-n <i>node_spec</i> --nodes <i>node_spec</i>	Specifies pdsh-style node hostnames. For example: node[100-110,120])
-g <i>genders_query</i> --genders <i>genders_query</i>	Specifies node genders attributes query (e.g. mds=primary).

Subcommand (alerts threshold)

```
$ cscli alerts thresholds [-h] [-y]
```

Threshold fields are:

```
name           Short identifier of the threshold
description    Describes the threshold and gives tips on how to modify it
gender         Type of nodes to which the threshold is applied
warning        Value of the warning threshold
critical       Value of the critical threshold
```

Possible gender values:

```
all           All nodes; general node type that can be overwritten by more specific node types
mgmt          Management nodes (primary and secondary)
mds           Metadata Servers
oss           Object Storage Servers
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.

<code>-y --yaml</code>	Outputs data in YAML format.
-------------------------	------------------------------

10.5.2 Manage the alerts configuration

The `alerts_config` command enables you to view and update the alerts configuration.

Synopsis

```
$ cscli alerts_config [-h]
    {email_off,thresholds,email_update,email_server_update,
    email_delete,email_add,email_on,email_server,emails}
```

where:

Positional Arguments	Description
<code>email_off</code>	Turns off notifications for notification subscribers.
<code>thresholds</code>	Sets the current value of an threshold. This value can be edited
<code>email_update</code>	Sends an email alert with an update.
<code>email_server_update</code>	Sends an email alert with a server update.
<code>email_delete</code>	Deletes the email.
<code>email_add</code>	Adds a new notification subscriber.
<code>email_on</code>	Turns on notifications for notification subscribers.
<code>email_server</code>	Displays the relay <i>SMTP</i> server configuration.
<code>emails</code>	Lists the alert notification subscribers.

Optional Arguments	Description
<code>-h --help</code>	Displays the help message and exits.

Subcommand (`email_off`)

The `email_off` command turns off notifications for subscribers.

Synopsis

```
$ cscli alerts_config email_off [-h] -u email
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-u email --user email</code>	Displays subscriber email. Notifies subscribers have new mail in <code>/var/spool/mail</code> or <code>admin</code> .

Subcommand (`thresholds`)

Current thresholds are applied to the monitoring configuration only if the

`--apply-config` option is used. It may take about 15 seconds to apply the configuration threshold changes.

If a group of changes needs to be made to the thresholds, edit a few threshold values and then add the `--apply-config` option to the last edit to set all the changes at once.

The new thresholds applied to monitoring configuration take effect a few minutes after they are applied when the next scheduled node check is performed.

The only editable thresholds are those listed in the output of the `cscli alerts thresholds` command.

Synopsis

```
$ cscli alerts_config thresholds [-h] -t threshold_name
-g gender_name [-W warning_threshold_value] [-C critical_threshold_value] [-A]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-t <i>threshold_name</i></code> <code> --threshold <i>threshold_name</i></code>	Displays the name of the threshold.
<code>-g <i>gender_name</i></code> <code> --gender <i>gender_name</i></code>	Displays the gender name of the threshold.
<code>-W <i>warning_threshold_value</i></code> <code> --warning <i>warning_threshold_value</i></code>	Displays the warning threshold value.
<code>-C <i>critical_threshold_value</i></code> <code> --critical <i>critical_threshold_value</i></code>	Displays the critical threshold value.
<code>-A --apply-config</code>	Applies the threshold configuration.

Subcommand (`email_update`)

The `email_update` command updates the existing subscriber's notification.

Notification Levels

The `level` option sets the alerts trigger for an email to be sent to a subscriber. The possible level option values are:

- Critical - Notify elements critical or node down statuses
- Warning - Notify elements warning statuses
- Unknown - Notify elements unknown statuses
- Ok - Notify when elements and nodes recover from problems
- Any combination of the above (comma-separated)
- None - No notifications (similar to `cscli alerts_config email_off`)
- All - Send all notifications, including notifications
- When a node/element is flapping between statuses
- When a node/element is in scheduled downtime

Notification Periods

The Notification period are:

- 24x7 - Notify always
- Workhours - Notify only during working days and hours (in the timezone of the server).

Synopsis

```
$ cscli alerts_config email_update [-h] -u email [-M email]
[-N user_full_name] [-P {24x7,workhours}] [-L level]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-u <i>email</i> --user <i>email</i></code>	Displays subscriber email. Notifies you have new mail in <code>/var/spool/mail</code> or <code>admin</code> .
<code>-M <i>email</i></code> <code> --email <i>email</i></code>	Displays the email address.
<code>-N <i>user_full_name</i></code> <code> --name <i>user_full_name</i></code>	Displays a longer name or description for the subscriber.
<code>-P {24x7,workhours}</code> <code> --period</code> <code>{24x7,workhours}</code>	Displays the time periods at which the subscriber is notified. possible values: <code>{24x7,workhours}</code>
<code>-L <i>level</i>,</code> <code> --level <i>level</i></code>	Displays notification level; possible values: any comma-separated combination of <code>{critical,ok,unknown,</code>

	warning}, or "all", or "none".
--	--------------------------------

Subcommand (email_server_update)

The `email_server_update` command configures the SMTP server to send alerts to external email addresses.

Synopsis

```
$ cscli alerts_config email_server_update [-h]
-s smtp_server_address [--port port] [-S email_from] [-d domain]
[-u smtp_user] [-p smtp_password]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-s smtp_server_address</code> <code> --server smtp_server_address</code>	Displays an IP address or hostname of the (relay) SMTP server.
<code>--port port</code>	SMTP server port (default: 25)
<code>-S email_from</code> <code> --sender email_from</code>	Displays the senders email address. If the <code>--domain</code> is set, the default value for the sender is <code>cluster_name@domain</code> . If the <code>--domain</code> is not set, the sender's email address is required.
<code>-d domain</code> <code> --domain domain</code>	Displays the internet hostname of the mail system to be used with email addresses that have no "@".
<code>-u smtp_user, --user smtp_user</code>	Specifies the username if the SMTP server requires authentication.
<code>-p smtp_password</code> <code> --password smtp_password</code>	The password if the SMTP server requires authentication.

Subcommand (email_delete)

The `email_delete` deletes notifications to subscribers.

Synopsis

```
$ cscli alerts_config email_delete [-h] -u email
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-u email --user email</code>	Displays subscriber email. Notifies you have new mail in /var/spool/mail or admin.

Subcommand (email_add)

The `email_add` command adds a new notification subscriber.

Synopsis

```
cscli alerts_config email_add [-h] -M email [-N user_full_name]
[-P {24x7,workhours}] [-L level]
```

Notification levels

The `level` option sets the alerts trigger for email to be sent to a subscriber. Possible level option values are:

The possible levels are:

- **Critical** - Notify elements critical or node down statuses
- **Warning** - Notify elements warning statuses
- **Unknown** - Notify elements unknown statuses
- **Ok** - Notify when elements and nodes recover from problems
- Any combination of the above (comma-separated)
- **None** - No notifications (similar to "`cscli alerts_config email_off`")
- **All** - Send all notifications, including notifications when a node/element is flapping between statuses, or when a node/element is in scheduled downtime

Notification periods

Possible Notification Periods:

- **24x7** - Notify always
- **Workhours** - Notify only during working days and hours (in the timezone of the server)

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-M email</code> <code> --email email</code>	Displays subscriber email. Notifies you have new mail in /var/spool/mail or admin. email address.

<code>-N <i>user_full_name</i></code> <code> --name <i>user_full_name</i></code>	Displays a longer name or description for the subscriber.
<code>-P {24x7,workhours}</code> <code> --period</code> <code> {24x7,workhours}</code>	The time periods at which the subscriber is notified. Possible values: {24x7,workhours} (default: 24x7).
<code>-L <i>level</i></code> <code> --level <i>level</i></code>	The notification level. Possible values: any comma-separated combination of: {critical,ok,unknown,warning}, or "all", or "none" (default: all).

Subcommand (email_on)

The `email_on` command turns on notifications for subscribers.

Synopsis

```
$ cscli alerts_config email_on [-h] -u email
```

where:

Optional Arguments	Description
<code>-h</code> <code> --help</code>	Shows the help message and exits.
<code>-u <i>email</i></code> <code> --user <i>email</i></code>	Displays subscriber email. Notifies you have new mail in /var/spool/mail or admin.

Subcommand (email_server)

The `email_server` command displays the relay `smtp` server configuration.

Synopsis

```
$ cscli alerts_config email_server [-h]
```

where:

Optional Arguments	Description
<code>-h</code> <code> --help</code>	Shows the help message and exits.

Subcommand (emails)

The `emails` command displays a list of alert notifications to the subscribers.

Notification levels

The `level` option sets the alerts trigger for email to be sent to a subscriber. Possible level option values are:

The possible Levels are:

- Critical - Notify elements critical or node down statuses
- Warning - Notify elements warning statuses
- Unknown - Notify elements unknown statuses
- Ok - Notify when elements and nodes recover from problems
- Any combination of the above (comma-separated)
- None - No notifications (similar to “`cscli alerts_config email_off`”)
- All - Send all notifications, including notifications
 - When a node or element is flapping between statuses
 - When a node or element is in scheduled downtime

Synopsis

```
$ cscli alerts_config emails [-h] [-y] [-v] [-u email]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-y --yaml</code>	Outputs data in YAML format.
<code>-v --verbose</code>	Outputs extra data in verbose mode.
<code>-u <i>email</i> --user <i>email</i></code>	Displays subscriber email. Notifies you have new mail in <code>/var/spool/mail</code> or <code>admin</code> .

10.5.3 Manage the alerts notification

The `alerts_notify` command turns alert notifications on or off.

Synopsis

```
$ cscli alerts_notify [-h] {on,off} ...
```

where:

Positional Arguments	Description
<code>on</code>	Sets the alert notification on.
<code>off</code>	Sets the alert notification off.

Optional Arguments	Description
-h --help	Displays the help message and exits.

The `alerts_notify on` command turns alert notifications on.

Synopsis

```
$ cscli alerts_notify on [-h] (-n node_spec | -g genders_query)
    [-S element_filter | -E element_name]
```

where:

Positional Arguments	Description
-h --help	Displays the help message and exits.
-n <i>node_spec</i> --node --node_spec --nodes <i>node_spec</i>	Looks through passed hostname elements. Looks for pdsh style nodes host names (e.g. node[100-110,120]).
-g <i>genders_query</i> --genders <i>genders_query</i>	Displays the node genders attributes query (e.g. mds=primary).
-S <i>element_filter</i> --search <i>element_filter</i>	This command searches by element name. The pattern is case sensitive. Regular expressions allowed.
-E <i>element_name</i> --element <i>element_name</i>	Displays the element name.

The `alerts_notify off` command turns alert notifications off.

Synopsis

```
$ cscli alerts_notify off [-h] (-n node_spec | -g genders_query)
    [-S element_filter | -E element_name] [-C comment]
```

where:

Positional Arguments	Description
-h --help	Displays the help message and exits.
-n <i>node_spec</i> --node --node_spec --nodes <i>node_spec</i>	Looks through passed hostname elements. Looks for pdsh style nodes host names (e.g. node[100-110,120]).
-g <i>genders_query</i> --genders <i>genders_query</i>	Displays the node genders attributes query (e.g. mds=primary).
-S <i>element_filter</i> --search <i>element_filter</i>	This command searches by element name. The pattern is case sensitive. Regular expressions allowed.
-E <i>element_name</i>	Displays element name.

--element <i>element_name</i>	
-C <i>comment</i> --comment <i>comment</i>	Displays a brief description of what you are doing.

10.6 Node control commands

The node control commands are used to control individual Lustre nodes (MDS/MGS and OSSs) in a clustered file system. The commands include functions to mount and unmount the Lustre nodes, show nodes in the file system. Additional functions include powering nodes on and off, managing node failover and failback, managing node auto-discovery and controlling exporter nodes.

10.6.1 Manage node auto-discovery

This command manages node auto-discovery in the Sonexion system.

Synopsis

```
$ cscli autodiscovery_mode [-h] [-s] [--mode {enabled,disabled}]
```

where:

Optional Arguments	Description
-h --help	Displays the help message and exits.
-s --status	Indicates the status of the auto-discovery mode.
--mode {enabled,disabled}	Switches to the specified mode. Enables or disables the auto-discovery mode.

10.6.2 Manage node failback and failover

These commands manage node failback and failover in the Sonexion system.

Synopsis

```
$ cscli failback [-h] (-F filter_sid | -n node_spec) |  
-c cluster_name |--cluster cluster_name  
$ cscli failover [-h] (-F filter_sid | -n node_spec) |  
-c cluster_name |--cluster cluster_name
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.

<code>-f filter_sid</code> <code> --filter filter_sid</code>	The filter identifier for the specified node. Failover/failback actions run on the nodes by filtering this filter.
<code>-n node_spec</code> <code> --nodes node_spec</code>	Specifies the nodes on which the failover/failback operations are performed. Node hostnames should be passed in pdsh style. If this parameter is passed, the <code>--filter</code> parameter is ignored.
<code>-c cluster_name</code> <code> --cluster cluster_name</code>	This parameter is deprecated. It is supported only for backward compatibility.

10.6.3 Mount and unmount Lustre targets

The `mount` and `unmount` commands control file system access to the Lustre targets (MDS/MGS and OSSs). The `mount` action enables file system access to the node. The `unmount` action disables file system access to the node.

- If one or more nodes are specified, then the `mount/unmount` action is performed only on the selected nodes in the file system.
- If no server nodes are specified, then the `mount/unmount` action is performed on all server nodes in the file system.

Synopsis

```
$ cscli mount [-h] -f fs_name [-n node_spec] |-c cluster_name |
--cluster cluster_name
$ cscli unmount [-h] -f fs_name [-n node_spec] |-c cluster_name
|--cluster cluster_name
```

where:

Optional Arguments	Description
<code>-h</code> <code> --help</code>	Shows the help message and exits.
<code>-f fs_name</code> <code> --fs-name=fs_name</code>	Specifies the name of the file system.
<code>-n node_spec</code> <code> --nodes=node_spec</code>	Specifies the node(s) on which the mount/unmount action is performed. Node hostnames should be passed in pdsh style.
<code>-c cluster_name</code> <code> --cluster cluster_name</code>	This parameter is deprecated. It is supported only for backward compatibility.

10.6.4 Manage node power

The `power_manage` command manages the power on the Sonexion system. These commands power-cycle nodes on and off and also control HA resource hand-offs.

Synopsis

```
$ cscli power_manage [-h] (--filter filter_sid | -n node_spec)
(--power-on|--power-off|--reboot|--cycle|--reset|--hand-over) [-
-force] -c cluster_name, --cluster cluster_name
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-f <i>filter_sid</i> --filter <i>filter_sid</i>	The filter identifier for the specified node. Failover/failback actions run on the nodes by filtering this filter. If --filter is specified, then --nodes is ignored.
-n <i>node_spec</i>	Specifies the nodes on which failover/failback operations are performed. Node hostnames should be passed in pdsh style.
--power-on	Powers on the specified nodes.
--power-off	Powers off the specified nodes.
--reboot	Reboots the specified nodes.
--cycle	Power-cycles the specified nodes.
--reset	Resets the specified nodes.
--hand-over	Hands over resources.
--force	An optional flag that indicates the node operation should be performed in <i>force</i> mode; should only be used with --power-off.
-c <i>cluster_name</i> --cluster <i>cluster_name</i>	This parameter is deprecated. It is supported only for backward compatibility.

10.6.5 Show node information

This command displays information about specified system nodes.

Synopsis

```
$ cscli show_nodes [-h] [-F filter_sid] [-r] | --refresh
-c cluster_name |--cluster cluster_name
```

where:

Option	Description
-h --help	Shows the help message and exits.

<code>-F filter_sid --filter filter_sid</code>	Specifies the node filter.
<code>-r --refresh</code>	Specifies the refresh mode (press 'q' for quit).
<code>-c cluster_name --cluster cluster_name</code>	This parameter is deprecated. It is supported only for backward compatibility.

10.7 Administrative commands

Administrative commands include functions to get file system and cluster node information, retrieve syslog entries, show FRU information and list available commands.

10.7.1 Show file system information

The `fs_info` command shows all file system information.

Synopsis

```
$ cscli fs_info [-h] [-f fs_name] [-c cluster_name | --cluster cluster_name]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-f fs_name --fs fs_name</code>	Shows the file system name.
<code>-c cluster_name --cluster cluster_name</code>	This parameter is deprecated. It is supported only for backward compatibility.

10.7.2 Retrieve FRU information

The `fru` command lists the defined Field Replaceable Units (FRUs) in the Sonexion system. FRUs are grouped into the following element 'types': ArrayDevice, BMC, Cooling, Enclosure, Enclosure_Electronics, PSU and Battery. FRU information can be retrieved per element type, on a per node basis, or for all nodes in the system.

Synopsis

```
$ cscli fru [-h] (-a | -n node_spec)
[-t ArrayDevice,BMC,Cooling,Enclosure,Enclosure_Electronics,
PSU,Battery}] [-i index] [-l [history]]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-a --all	Shows FRUs (including status) grouped by type, for all nodes in the system.
-n <i>node_spec</i> --nodes <i>node_spec</i>	Shows FRUs (including status) grouped by element type, for a specified node(s) in the system.
-t {ArrayDevice, BMC, Cooling, Enclosure, PSU, Battery, Enclosure_Electronics }	Shows frus (including status) for the specified element type. Examples of element types: array device, BMC, PSU, battery.
-i <i>index</i> --index <i>index</i>	Shows FRUs (including status) for specified elements within a list of elements of the same type.
-l [<i>history</i>] --history [<i>history</i>]	Shows FRU history (default is 10 lines of history).

10.7.3 Change the Sonexion mode

The `cluster_mode` command toggles the Sonexion system among multiple system modes: `daily`, `custWizard` or `pre-shipment`.

Synopsis

```
$ cscli cluster_mode [-h] [-s]
  [--mode {daily,custwiz,pre-shipment}] [--db-only]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-s --status	Shows the status of the cluster.
--mode {daily,custwiz,pre-shipment}	Switches to the specified mode. Switches to either daily mode, customer wizard mode or pre-shipment mode. CAUTION: Use of the pre-shipment option will delete any current configuration settings.
--db-only	Update only the database. Does not sync nodes via puppet. Valid only with the '--mode' argument.

10.7.4 List commands

The `list` command shows a list of available commands in the current Sonexion mode.

Synopsis

```
$ cscli list [-h]
```

where:

Optional Arguments	Description
<code>-h</code> <code>--help</code>	Shows the help message and exits.

10.7.5 Display log information

The `syslog` command displays Lustre log entries.

Synopsis

```
$ cscli syslog [-h] [-m max] [-F] [-d duration] [-s start_time]
[-e end_time] [-r]
```

where:

Optional Arguments	Description
<code>-h</code> <code>--help</code>	Shows the help message and exits.
<code>-m <i>max</i></code> <code>--max=<i>max</i></code>	Specifies the maximum number of entries to return.
<code>-F</code> <code>--follow</code>	Polls for future messages. Only valid without <code>-e</code> , <code>-r</code> arguments.
<code>-d <i>duration</i></code> <code>--duration=<i>duration</i></code>	Specifies duration (in seconds) for which to follow output. Only valid with <code>-F</code> argument.
<code>-s <i>start_time</i></code> <code>--start_time=<i>start_time</i></code>	Specifies the earliest time for which messages should be received.
<code>-e <i>end_time</i></code> <code>--end_time <i>end_time</i></code>	Specifies the latest time for which messages should be received.
<code>-r</code> <code>--reverse</code>	Sorts entries in descending order (by time).

10.7.6 Set administrator password

The `set_admin_passwd` command changes and sets an administrator password.

Synopsis

```
$ cscli set_admin_passwd [-h] [-p password]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-p -- <i>password</i>	Specify the new administrator password string.

10.7.7 Batching commands

The `batch` command runs a sequence of CCLI commands in a batch file.

Synopsis

```
$ cscli batch [-h] -b batch_file
```

where:

Optional Arguments	Description
-h --help	Displays the help message and exits.
-b <i>batch_file</i> --batch-file <i>batch_file</i>	Specifies the command batch file.

10.7.8 Manage IP routing

The `ip_routing` command manages IP routing to and from the system database.

Synopsis

```
$ cscli ip_routing [arguments]
```

where [*arguments*] are:

```
--show|-s [--loadable]
```

or

```
--load path_to_file
```

or

```
--add | -a --dest destination_ip --prefix prefix_len --router router_ip
```

or

```
--update | -u --route-id route_id [--destdestination_ip]  
[--prefixprefix_len] [--routerrouter_ip]
```

or

```
--delete | -d --route-id route_id
```

or

```
--clear | -c
```

or

```
--apply | -a
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-s --show	Shows the current <i>IP</i> routing table in the database.
--loadable	Prints the routing table in loadable format (use with the <i>-show</i> argument).
-c --clear	Clears the routing table in the database.
--apply	Applies IP routing.
--load <i>load</i>	Loads the IP routing table from a file to the database.
-a --add	Inserts IP routing in the database.
-u --update	Updates IP routing in the database.
-d --delete	Deletes IP routing from the database.
--dest <i>dest</i>	Specifies the destination IP address.
--prefix <i>prefix</i>	Specifies the prefix length (0-32).
--router <i>router</i>	Specifies the router IP address.
--route-id <i>route_id</i>	Specifies the route identifier (see <i>ip_routing -show</i>).

10.8 Configuration commands

The configuration commands specifies the *mac* address and hostname for a given node and configures *oss* nodes

10.8.1 Configure hosts

The `configure_hosts` command configures the MAC address and host names for the discovered node.

Synopsis

```
$ cscli configure_hosts [-h] -m mac_address --hostname hostname [-f]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-m <i>mac_address</i> --mac <i>mac_address</i>	Shows the <i>mac_address</i> and node <i>mac</i> address.
--hostname <i>hostname</i>	Shows the new node hostname.
-f --force	Forces the mode (to skip hostname validation).

10.8.2 Configure new OSS nodes

This command configures new OSS nodes in the Sonexion system.

Synopsis

```
$ cscli configure_oss [-h] -n node_spec (-A | -b bind_arrays) |-c  
cluster_name | --cluster cluster_name
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-n <i>node_spec</i> --nodes <i>node_spec</i>	Specifies the hostname of the new OSS node (in genders style).
-A --apply-config	Applies the configuration to the new <i>oss</i> node.
-b <i>bind_arrays</i> --bind-arrays <i>bind_arrays</i>	Specifies comma-separated pairs of array-file system bindings. Each binding should be in this format: <i>array:file_system_name</i> . The <i>array</i> variable can be a genders-style string. For example: <i>md[0-3]</i> .
-c <i>cluster_name</i> --cluster <i>cluster_name</i>	This parameter is deprecated. It is supported only for backward compatibility.

10.8.3 Show information about new OSS nodes

This command displays a table of new OSS nodes and their resources.

Synopsis

```
$ cscli show_new_nodes [-h] [-v]  
|-c cluster_name | --cluster cluster_name
```


where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-v, --verbose	Specifies the verbose mode.
-c <i>cluster_name</i> --cluster <i>cluster_name</i>	This parameter is deprecated. It is supported only for backward compatibility.

10.9 Filter commands

The filter commands create and delete a filter.

10.9.1 Create a filter

The `create_filter` command creates a customer nodes filter.

Synopsis

```
$ cscli create_filter [-h] -i filter_sid -F filter_name -e filter_expr
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-i <i>filter_sid</i> --id <i>filter_sid</i>	Shows the symbol identifier of the filter.
-F <i>filter_name</i> --name <i>filter_name</i>	Shows the filter name.
-e <i>filter_expr</i> --expression <i>filter_expr</i>	Shows the filter expression. Examples: "host1,host2", "host[1-3]", "mds=primary".

10.9.2 Show filters

The `show_filters` command shows all filters.

Synopsis

```
$ cscli show_filters [-h] [-P] [-C]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-P --predefined	Shows only predefined filters.
-C --custom	Shows only custom filters.

10.9.3 Delete a filter

The `delete_filter` command deletes a customer nodes filter.

Synopsis

```
$ cscli delete_filter [-h] -i filter_sid
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-i <i>filter_sid</i> --id <i>filter_sid</i>	Shows the symbol identifier of the filter.

10.10 Updating system software

These commands prepare a software upgrade package for installation and apply it to system nodes.

10.10.1 Prepare a software update

The `prepare_update` command runs the software update preparation process.

Synopsis

```
$ cscli prepare_update [-h] [--run]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
--run	Prepares the software upgrade package for installation.

10.10.2 Update software on a system node

The `update_node` command updates software on the specified node(s).

Synopsis

```
$ cscli update_node [-h] -n node_spec
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-n <i>node_spec</i></code> <code> --node-spec <i>node_spec</i></code>	Specifies hostnames of the nodes on which to update software.

10.10.3 Split HA partners

The `split_ha_partners` command manages support bundles and support bundle settings.

Synopsis

```
$ cscli split_ha_partners [-h] -g genders_query
```

where:

Option	Description
<code>-h --help</code>	Displays the help message and exits.
<code>-g <i>genders_query</i></code> <code> --genders <i>genders_query</i></code>	Specifies a genders style when splitting <i>ha</i> pairs of <i>oss</i> nodes.

Showing nodes at specified software version

The `show_version_nodes` command lists all system nodes at the specified software version.

Synopsis

```
$ cscli show_version_nodes [-h] [-q] -v sw_version
```

where:

Option	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-q --query</code>	Controls the format of the command output. If this flag is

	specified, nodes in output should display in genders style. Example: <code>mycluster[02-05,97-98]</code> .
<code>-v <i>sw_version</i></code> <code> --version <i>sw_version</i></code>	Specifies the Sonexion software version.

10.10.4 Show node versions

The `show_node_versions` command displays the Sonexion software version running on specified nodes.

Synopsis

```
$ cscli show_node_versions [-h] [-q] [-n node_spec]
[-g genders_query] [-c cluster_name]
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-q, --query</code>	Controls output format. If this flag is specified, nodes in the output should be in genders style. Example: <code>mycluster[02-05,97-98]</code>
<code>-n <i>node_spec</i>,</code> <code>--nodes</code> <code><i>node_spec</i></code>	Specifies nodes to indicate the Sonexion software version.
<code>-g <i>genders_query</i></code>	Specifies a gender's style query.
<code>-c <i>cluster_name</i></code>	This parameter is deprecated. It is supported only for backward compatibility.

10.10.5 Showing available software versions

The `show_update_versions` command lists software versions available in the Sonexion Management (MGMT) Server repository.

Synopsis

```
$ cscli show_update_versions [-h]
```

where:

Option	Description
<code>-h --help</code>	Shows the help message and exits.

10.11 Managing node position in a Sonexion rack

The rack position commands manage the location of components (hosting system nodes) in a Sonexion rack. The MMU hosts the primary and secondary MGMT, MGS and MDS nodes. Each SSU hosts OSS nodes (two OSSs per SSU).

10.11.1 Get node position in a Sonexion rack

The `get_rack_position` command indicates the location of server nodes in a Sonexion rack.

Synopsis

```
$ cscli get_rack_position [-h] -r rack_name [--yaml]
```

where:

Option	Description
<code>-h --help</code>	Displays the help message and exits.
<code>-r <i>rack_name</i>, --rack <i>rack_name</i></code> <code> --rack <i>rack_name</i></code>	Specifies the rack containing the node(s).
<code>--yaml</code>	Prints node rack position information in <i>YAML</i> file format.

10.11.2 Set node position in a Sonexion rack

The `set_rack_position` command sets the location of server nodes in the Sonexion rack or moves a node to another rack.

Synopsis

```
$ cscli set_rack_position [-h] -r rack_name [--yaml] -n node_spec  
-p position
```

where:

Optional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-r <i>rack_name</i></code> <code> --rack <i>rack_name</i></code>	Specifies the rack containing the node(s).
<code>-y <i>yaml path</i></code> <code> --yaml <i>yaml_path</i></code>	Loads rack position information in <i>yaml</i> file format.

<code>-n <i>node_spec</i></code> <code> --node <i>node_spec</i></code>	Specifies the node(s) hostname.
<code>-p <i>position</i></code> <code> --position=<i>position</i></code>	Specifies the node position in rack units (Us).

10.11.3 Monitor system health

The `$ cscli monitor` command monitors and displays current health and status information for the cluster nodes and elements.

Synopsis

```
$ cscli monitor [-h] {nodes,elements,health} ...
```

Positional Arguments	Description
health	Current overall health information - status summary.
nodes	Current status for nodes.
elements	Current status for elements.

Optional Arguments	Description
<code>-h</code> <code> --help</code>	Shows the help message and exits.

The `$ cscli monitor nodes` command monitors individual nodes.

Synopsis

```
$ cscli monitor nodes [-h] [-y] [-v] [-n node_spec | -g genders_query]  
[-N {down,unreachable,up,pending}]
```

where:

Positional Arguments	Description
<code>-h</code> <code> --help</code>	Shows the help message and exits.
<code>-y</code> <code> --yaml</code>	Displays output data in YAML format.
<code>-v</code> <code> --verbose</code>	Outputs extra data.
<code>-n <i>node_spec</i></code> <code> --node <i>node_spec</i></code> <code> --nodes <i>node_spec</i></code>	Looks through passed hostname elements. Looks for pdsh style nodes host names. Example: <code>node[100-110,120]</code> .

<code>-g <i>genders_query</i></code>	Displays the node genders attributes query (for example, <code>mds=primary</code>).
<code>-N {down,unreachable,up,pending}</code> <code> --nodestatus</code> <code>{down,unreachable,up,pending}</code> <code>node status</code>	Displays node status.

The `$ cscli monitor elements` command monitors individual nodes.

Synopsis

```
$ cscli monitor elements [-h] [-y] [-v]
[-n node_spec | -g genders_query]
[-N {down,unreachable,up,pending}]
[-U {unknown,warning,ok,critical,pending}] [-S element_filter]
```

where:

Positional Arguments	Description
<code>-h --help</code>	Shows the help message and exits.
<code>-y --yaml</code>	Displays output data in YAML format.
<code>-v --verbose</code>	Outputs extra data.
<code>-n <i>node_spec</i> --node <i>node_spec</i></code> <code> --nodes <i>node_spec</i></code>	Looks through passed hostname elements. Looks for pdsh style nodes host names (for example, <code>node[100-110,120]</code>).
<code>-g <i>genders_query</i></code>	Displays the node genders attributes query (e.g. <code>mds=primary</code>).
<code>-N {down,unreachable,up,pending}</code> <code> --nodestatus {down, unreachable,</code> <code>up, pending} node status.</code>	Displays node status.
<code>-U {unknown, warning, ok,</code> <code>critical, pending}</code> <code> --elementstatus {unknown,</code> <code>warning, ok, critical, pending}</code>	Displays element status.
<code>-S <i>element_filter</i></code> <code> --search <i>element_filter</i></code>	Searches by element name. The pattern is case sensitive. Regular expressions are allowed.

NOTE: If you call this command without any options, you may get thousands of elements on a large system.

10.11.4 Manage the netfilter level

The `netfilter_level` command manages the netfilter level on the Sonexion system.

Synopsis

```
$ cscli netfilter_level [-h] [-s] [-l level] [--force]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-s --show	Shows the current netfilter level.
-l <i>level</i> --level <i>level</i>	Sets the netfilter level (<i>off</i> , <i>lustre</i> , <i>on</i>).
--force	Forces the netfilter level to be set to <i>off</i> .

10.11.5 Enable RAID checks

The `raid_check` command enables RAID check on RAID devices.

Synopsis

```
$ cscli raid_check -h (-a | -n node_list) [-i] [-c {on,off}]
  [--now] [-s a_time]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-a --all	Looks through all nodes elements.
-n <i>node_list</i> --node <i>node_list</i>	Looks through passed hostname elements. Looks for pdsh style nodes host names.
-i --info	Prints the current RAID check status for selected nodes.
-c {on,off} --cron {on,off}	Enables/Disables the cron job for the RAID check.
--now	Performs the raid check now.
-s <i>a_time</i> --set <i>a_time</i>	Specifies a string to set a time to run the RAID check.

10.11.6 Manage the RAID rebuild rate

The `rebuild_rate` command manages the RAID rebuild rate on the Sonexion system.

Synopsis

```
$ cscli rebuild_rate [-h] [-n nodes] [--reset] [-l single_rate]
[-m multiple_rate]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-n <i>nodes</i> --node <i>nodes</i>	Specifies pdsh-style node hostnames. For example, node[100-110,120. Global RAID rebuild rates are installed without this argument.
--reset	Resets the RAID rebuild rate.
-l <i>single_rate</i> --after-first-failure <i>single_rate</i>	Specifies the RAID rebuild rate for a single drive failure.
-m <i>multiple_rate</i> --after-multiple-failures <i>multiple_rate</i>	Specifies the RAID rebuild rate for multiple drive failures.

10.11.7 Manage the administrative password

The `set_admin_passwd` command sets the Sonexion system administrator's user password.

Synopsis

```
$ cscli set_admin_passwd [-h] -p password
```

where:

Option	Description
-h --help	Prints the help message and exits.
-p --password	Sets the system administrator's password.

10.11.8 Manage the system date

The `set_date` command manages the date on the Sonexion system.

Synopsis

```
$ cscli set_date [-h] [-s new_date] [--force-ntp]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-s <i>new_date</i> --set <i>new_date</i>	Specifies the new date in this format: <i>mmddhhmmccyy.ss</i> .
--force-ntp	Forces NTP configuration.

10.11.9 Manage the system timezone

The `set_timezone` command manages the timezone on the Sonexion system.

Synopsis

```
$ cscli set_timezone [-h] [-s new_timezone] [-l]
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-s <i>new_timezone</i> --set <i>new_timezone</i>	Specifies the new time zone location name. For example, "America/Los_Angeles".
-l --list	Lists the available timezones.

10.11.10 Manage the InfiniBand Subnet Manager

The `sm` command manages (enables, disables or prioritizes) the InfiniBand Subnet Manager (SM) integrated with the Sonexion system. The local SM ensures that InfiniBand is properly configured and enabled for use. In situations in which Sonexion is connected to a larger InfiniBand network that already uses a subnet manager, the local SM should be disabled. The `sm` command can also be used to modify subnet manager priorities.

Synopsis

```
$ cscli sm [-h] (-e | -d)
[-P {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}]
|-c cluster_name |--cluster cluster_name
```

where:

Optional Arguments	Description
-h --help	Shows the help message and exits.
-e --enable	Enables the IB storage manager used with the Sonexion system.
-d --disable	Disables the IB storage manager used with the system.
-P {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15} --priority {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}	Sets the priority [0..15] of the <i>IB</i> storage manager used with the system.
-c <i>cluster_name</i> --cluster <i>cluster_name</i>	This parameter is deprecated. It is supported only for backward compatibility.

10.11.11 Manage support bundles

The `support_bundle` command manages support bundles and support bundle settings.

Synopsis

```
$ cscli support_bundle [-h] [-c] [-n nodes] [-t minutes] [-e bundle_id] [--disable-trigger trigger] [--get-purge-limit] [--set-purge-limit percents]
```

where:

Option	Description
-h --help	Displays the help message and exits.
-c --collect-bundle	Collects the support bundle.
-n <i>nodes</i> --nodes <i>nodes</i>	Shows a comma-separated list of nodes. Default value is all nodes.
-t <i>minutes</i> --time-window <i>minutes</i>	Specifies the time window to collect data for the support bundle (in minutes). Default value is 45 minutes.
-e <i>bundle_id</i> --export-bundle <i>bundle_id</i>	Identifies an export-specified support bundle.
--show-triggers	Shows the triggers that initiate automatic collection of support bundles.
--enable-trigger <i>trigger</i>	Enables a specific trigger.
--disable-trigger <i>trigger</i>	Disables a specific trigger.

<code>--get-purge-limit</code>	Shows the purge limit as a percentage of free file system space. If the purge limit is reached, the Sonexion system purges old support bundle files.
<code>--set-purge-limit</code> <i>percents</i>	<i>Sets</i> the purge limit as a percentage of free file system space.

11. GEM CLI Commands

This appendix details the supported command line interface (CLI) provided by the GEM software.

11.1 Serial port settings

Use the following settings for using HyperTerminal or other serial communications GUI to work with the CLI:

Baud rate (bits/sec):	115200
Data bits:	8
Parity:	None
Stop bits:	1
Flow control:	None

The above settings apply to manually typed commands. If multiple commands are sent via a text file, then the baud rate needs to be reduced for all characters to be processed.

Set the baud rate in the running firmware by issuing:

```
rmon baud 0
```

Change the serial communications GUI settings to:

Baud rate (bits/sec):	9600
Data bits:	8
Parity:	None
Stop bits:	1

Flow control: None

NOTE: To return to the higher baud rate, issue: `rmon baud 4`. The complete set of supported values is:

0 = 9600

1 = 19200

2 = 38400

3 = 57600

4 = 115200

11.2 Supported number bases

Numeric parameters passed into CLIs can be in different bases. Decimal is the default. Octal or hexadecimal can be supplied by using a leading code:

Decimal	– Plain number
Octal	– Leading '0'
Hexadecimal	– Leading '0x'

For example, the decimal number 14 would be represented in the following ways:

Decimal	– 14
Octal	– 016
Hexadecimal	– 0xE

11.3 Supported commands

The following CLI commands are supported by the GEM software.

11.3.1 ddump

Command name:	ddump
Command synopsis:	Returns a system-wide diagnostic dump
Command description:	Calls all commands of the command type 'diagnostic' that do not demand an argument, i.e. a simple single-shot diagnostic dump.
Command arguments:	None
Command type:	Diagnostic
Access level:	General

11.3.2 getboardid

Command name:	getboardid
Command synopsis:	Reports the local board slot ID and HA mode
Command description:	Reports the local board slot ID and HA mode in human-readable and machine-readable form.
Command arguments:	hex: Returns the slot ID (Byte 1) and HA mode (Byte 2) in hexadecimal form. If the canister is the master, then the HA mode is set to 0x0. If the canister is the slave, then the mode is 0x00.
Command type:	Debug
Access level:	General

11.3.3 getmetisstatus

Command name:	getmetisstatus
Command synopsis:	Reports Metis status for the enclosure. (Supplies reserve power to protect in-flight storage data, enabling it to be securely stored on persistent media).
Command description:	Invoking this command returns Metis status in human-readable or machine-readable form.
Command arguments:	Argument 1 [hex]: If the "hex" argument is present, the Metis status is reported in machine-readable form. If "hex" is not specified, the status is reported in human-readable form.
Command type:	Diagnostic
Access level:	Engineering

11.3.4 getvpd

Command name:	getvpd
Command synopsis:	Retrieves VPD information from all enclosure FRUs
Command description:	The getvpd command displays the following enclosure VPD data: <ul style="list-style-type: none"> • Enclosure Vendor • Enclosure Product ID • Enclosure WWN • Enclosure Serial Number • Enclosure Part Number • Canister VPD Version • Canister Vendor • Canister Product ID • Canister SAS Address

- Canister Serial Number
- Canister Part Number
- Midplane VPD Version
- Midplane Product ID
- Midplane Serial Number
- Midplane Part Number
- PCM VPD Version
- PCM Vendor
- PCM Product ID
- PCM Serial Number
- PCM Part Number

Command arguments: getvpd – No additional arguments
 Command type: Debug
 Access level: General

11.3.5 help

Command name: help
 Command synopsis: Displays helpful information about the GEM commands
 Command description: Provides a mechanism to discover the available commands and display the command usage information. By default (i.e. no argument supplied), the command only lists the synopsis for those commands with the access level 'general'. The argument `all` lists the synopsis for all commands, regardless of access level. The argument `testing` lists the synopsis for all commands that have the 'testing' access level. If the argument matches a command (for example `help ddump`) then detailed help for the specified command displays instead.
 Command arguments: 1 optional argument - see description above.
 Command type: Control
 Access level: General

11.3.6 ipmi_power

Command name: ipmi_power
 Command synopsis: Performs safe canister-level power control using chassis commands to the BMC
 Command description: This command allows the user to request a canister-level shutdown through the BMC. The benefit of using this command is to cleanly shut down the x86 subsystem using ACPI.
 Command arguments: ipmi_power [type]

Type:	2 "soft" – Orchestrated shutdown of x86 complex. 3 "off" – Immediate shutdown of x86 complex. 4 "cycle" – Canister power cycle. 5 "reset" – Canister reset. 6 "on" – Wake x86 complex from standby/soft-off.
Command type:	Control
Access level:	General Access

11.3.7 ipmi_setosboot

Command name:	ipmi_setosboot
Command synopsis:	Sets a value in the IPMI OS boot sensor indicating that the x86 subsystem has successfully booted. The OS boot sensor value is cleared to zero (0) on x86 resets and BMC firmware upgrades / reboots.
Command description:	<p>This command is intended for use by an application on the local x86 subsystem to set the OS boot sensor to confirm that the system has finished booting and the OS is in full control.</p> <p>This command MUST be invoked by the customer OS on startup. If it is not set and GEM detects an AC loss event, then the module is automatically shut down. This shutdown ensures that the system batteries are not flattened by a module booting at full power.</p> <p>Without a parameter, the command reads the current sensor value. With a parameter of 1, the command sets the sensor to indicate that the system has booted (0x40) and then reads back the sensor for confirmation.</p>
Command arguments:	ipmi_setosboot [setting]
Command type:	Control
Access level:	Engineering

11.3.8 logdump

Command name:	logdump
Command synopsis:	Displays logged messages
Command description:	Provides a mechanism to output logging information.
Command arguments:	<p>6 optional arguments:</p> <p>Argument 1 specifies the area of memory from which to retrieve log messages from. 'r' = RAM, 'n' = non-volatile.</p> <p>Argument 2 specifies the order of the log messages. "old" = oldest first, "new" = newest first.</p> <p>Argument 3 limits the number of logged messages</p>

displayed to *n*. Set to zero (0) or omit the argument to display all logged messages.

Argument 4 controls the generation of a *timestamp* field in the log dump messages. Set to 1 for enable; 0 for disable.

Argument 5 controls the generation of a *subsystem name* field in the log dump messages. Set to 1 for enable; 0 for disable.

Argument 6 controls the generation of a *service name* field in the log dump messages. Set to 1 for enable; 0 for disable.

The default (for omitted command arguments) displays all logged messages from RAM, newest first, with all message fields enabled.

Command type: Diagnostic
 Access level: General

11.3.9 report_faults

Command name: `report_faults`
 Command synopsis: Reports all system-wide faults
 Command description: Outputs all known faults, collected from each GEM service.
 Command arguments: None
 Command type: Diagnostic
 Access level: General

11.3.10 settime

Command name: `settime`
 Command synopsis: Sets GEM logging time in days, hours, minutes and seconds
 Command description: "settime days hh mm ss", for example: "settime 10 9 8 7" sets the logging time to 10 days, 9 hours, 8 minutes and 7 seconds. The new logging time appears in the log timestamps as: 10+09:08:07.123 M0 > Using the "settime" command on its own, without any arguments, prints the current logging time to the CLI.
 Command arguments: days hh mm ss
 Command type: Control
 Access level: General

11.3.11 ver

Command name:	ver
Command synopsis:	Displays version information
Command description:	Displays version numbers and information for the components in the local canister, midplane and PCMs.
Command arguments:	None
Command type:	Diagnostic
Access level:	General