

HP XC System Software

Release Notes

October 5, 2005

Product Version: HP XC System Software Version 2.1

This document contains release notes that apply to HP XC System Software Version 2.1 and its accompanying documentation set.

© Copyright 2003–2005 Hewlett-Packard Development Company, L.P.

AMD and AMD Opteron are trademarks or registered trademarks of Advanced Micro Devices, Inc.

FLEXlm is a trademark of Macrovision Corporation.

InfiniBand is a registered trademark and service mark of the InfiniBand Trade Association.

Intel, the Intel logo, Itanium, Xeon, and Pentium are trademarks or registered trademarks of Intel Corporation in the United States and other countries.

Linux is a U.S. registered trademark of Linus Torvalds.

LSF, Platform Computing, and the LSF and Platform Computing logos are trademarks or registered trademarks of Platform Computing Corporation.

Myrinet and Myricom are registered trademarks of Myricom, Inc.

Nagios and the Nagios logo are registered trademarks of Ethan Galstad.

The Portland Group and PGI are trademarks or registered trademarks of The Portland Group Compiler Technology, STMicroelectronics, Inc.

Quadrics and QsNet^{II} are registered trademarks of Quadrics, Ltd.

Red Hat and RPM are registered trademarks of Red Hat, Inc.

syslog-ng is copyrighted by BalaBit IT Security.

SystemImager is a registered trademark of Brian Finley.

TotalView is a registered trademark of Etnus, Inc.

UNIX is a registered trademark of The Open Group.

Confidential computer software. Valid license from HP required for possession, use, or copying. Consistent with FAR 12.211 and 12.212, Commercial Computer Software, Computer Software Documentation, and Technical Data for Commercial Items are licensed to the U.S. Government under vendor's standard commercial license.

The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

Contents

About This Document

1 New Features

1.1	Base Distribution and Kernel	1-1
1.2	Additional Hardware Models Supported	1-1
1.3	The discover Command	1-1
1.4	Cluster Configuration	1-1
1.5	The startsys Command	1-2
1.6	Image Distribution	1-2
1.7	Resource Management	1-2
1.8	Interconnects	1-2
1.9	Serviceability	1-3
1.10	Documentation Changes	1-3
1.10.1	<i>HP XC Hardware Preparation Guide</i>	1-3
1.10.2	<i>HP XC System Software Installation Guide</i>	1-3
1.10.3	<i>HP XC System Software Administration Guide</i>	1-3
1.10.4	<i>HP XC System Software User's Guide</i>	1-4
1.10.5	<i>HP XC System Software Overview</i>	1-4

2 General Notes

2.1	XC System Naming	2-1
2.2	Master Firmware Revision Table	2-1
2.3	Obtaining XC Documentation	2-1

3 Hardware Preparation Notes

3.1	Configuring Disks Into the Smart Array	3-1
3.2	Incorrect Instruction for Preparing HP ProLiant DL145 G2 Servers	3-1
3.3	Preparing HP Integrity rx4640 Servers	3-1
3.4	Preparing HP Integrity rx2620 Servers	3-4

4 Installation Notes

4.1	Notes to Read Before the Kickstart Installation Process	4-1
4.1.1	Prepare Previously Installed Nodes for a Reinstallation	4-1
4.1.2	No Hard Drives Found on Head Nodes with Serial ATA Disks with 4 GB or More Memory	4-1
4.1.3	Installation Supported on SAN on CP6000 Systems But Not on CP4000 Systems	4-1

5 Configuration Notes

5.1	Tasks to Perform After Running the cluster_config Utility	5-1
5.1.1	Required Task: Update Nagios Status	5-1
5.2	Hang May Be Encountered When Logging In To the Head Node	5-1

6	System Administration and Management Notes	
6.1	Running the dgemm Utility	6-1
6.2	Log Files Must Be Rotated and Compressed	6-1
6.3	Recommended NFS Mount Options for External Connections	6-2
7	Programming and User Environment Notes	
7.1	Notes About the HP Math Library	7-1
7.2	Notes About MLIB, HP MPI, and Modulefiles	7-1
7.3	Configuring the Intel Trace Collector and Analyzer with HP MPI on XC	7-2
7.3.1	Installation Notes	7-2
7.3.2	HP MPI and the Intel Trace Collector (OTA)	7-2
8	Load Sharing Facility and Job Management Notes	
8.1	LSF	8-1
8.1.1	LSF Always Runs on the Head Node	8-1
8.2	Job Management	8-1
9	Cluster Platform 3000 Notes	
9.1	Remote Console Logins Do Not Work on HP ProLiant DL140 G2 Nodes	9-1
9.2	The power Command Might Fail on HP ProLiant DL140 G2 Nodes ..	9-1
10	Cluster Platform 4000 Notes	
10.1	Remote Console Logins Do Not Work on HP ProLiant DL145 G2 Nodes	10-1
10.2	The power Command Might Fail on HP ProLiant DL145 G2 Nodes ..	10-1
10.3	Sensor Information for Supermon Not Available	10-1
11	Cluster Platform 6000 Notes	
11.1	Excessive Boot Time with Unzoned SAN Volume Connected Through an A6824A HBA	11-1
11.2	Installing to and Booting from a SAN Volume	11-1
12	Interconnect Notes	
12.1	InfiniBand Interconnect	12-1
12.2	Myrinet Interconnect	12-1
12.2.1	The clear_counters Command Does Not Work on the 256 Port Switch	12-1
12.2.2	New or Changed Myrinet GM Routes May Not Be Found by gm_board_info	12-1
12.3	QsNet ^{II} Interconnect	12-2
12.3.1	ELAN4 Diagnostic Tools Fail on Systems With ELAN3 Interconnect	12-2
12.3.2	OVP Interconnect Tests Fail on ELAN3	12-2
12.3.3	Possible Conflict with Use of SIGUSR2	12-2
12.3.4	ELAN TRAP Queue Error Seen on Some Quadrics MPI Applications	12-3

12.3.5	The qsnet Database May Contain Entries to Nonexistent Switch Modules	12-3
13	Documentation Notes	
13.1	<i>HP XC System Software Administration Guide</i>	13-1
Index		
Figures		
3-1	HP Integrity rx4640 Server Rear View	3-2
Tables		
3-1	IP Addresses for MP Power Management Devices	3-3

About This Document

This document contains release notes for HP XC System Software Version 2.1. This document contains important information about firmware, software, or hardware that may affect your system.

An HP XC system is integrated with several open source software components. Some open source software components are being used for underlying technology, and their deployment is transparent. Some open source software components require HP XC-specific user-level documentation, and that kind of information is included in this document, if required.

HP relies on the documentation provided by the open source developers to supply the information you need to use their product. For links to open source software documentation for products that are integrated with your XC system, see *Supplementary Information*.

Documentation for third-party hardware and software components that are supported on the HP XC system is supplied by the third-party vendor. However, information about the operation of third-party software is included in this document if the functionality of the third-party component differs from standard behavior when used in the XC environment. In this case, HP XC documentation supersedes information supplied by the third-party vendor. For links to related third-party Web sites, see *Supplementary Information*.

Standard Linux[®] administrative tasks or the functions provided by standard Linux tools and commands are documented in commercially available Linux reference documents and on various Web sites. For more information about obtaining documentation for standard Linux administrative tasks and associated topics, see the list of Web sites and additional publications provided in *Related Information*.

Intended Audience

These release notes are intended for anyone who installs and configures an HP XC system, for system administrators who maintain the system, for programmers who write applications to run on the system, and for general users who log in to the system to run jobs.

The information in this document assumes that you have knowledge of the Linux operating system.

Document Organization

This document is organized as follows:

- *Chapter 1* describes the new features delivered in this release.
- *Chapter 2* contains general notes that apply to the overall system.
- *Chapter 3* contains notes about hardware preparation tasks.
- *Chapter 4* contains notes that apply to installing the base operating system and XC software.
- *Chapter 5* contains notes that apply to configuring the system.
- *Chapter 6* contains notes that apply to system administration and system management tasks.
- *Chapter 7* contains notes that apply to programmers and users.

- *Chapter 8* contains notes that apply to the Load Sharing Facility (LSF®) and interactive job management commands.
- *Chapter 9* contains notes that apply only to Xeon® with EMT64-based systems.
- *Chapter 10* contains notes that apply only to AMD Opteron™-based systems.
- *Chapter 11* contains notes that apply only to Intel® Itanium®-based systems.
- *Chapter 12* contains notes that apply to the interconnects.
- *Chapter 13* contains notes that apply to the HP XC System Software Documentation Set and XC manpages.

HP XC Information

The HP XC System Software Documentation Set includes the following core documents. All XC documents except the *HP XC System Software Release Notes* are shipped on the XC documentation CD. All XC documents, including the *HP XC System Software Release Notes*, are available on line at the following URL:

http://www.hp.com/techservers/clusters/xc_clusters.html

<i>HP XC System Software Release Notes</i>	Contains important, last-minute information about firmware, software, or hardware that might affect your system. This document is available only on line.
<i>HP XC Hardware Preparation Guide</i>	Describes tasks specific to HP XC that are required to prepare each supported cluster platform for installation and configuration, including the specific connections of nodes to switch ports.
<i>HP XC System Software Installation Guide</i>	Provides step-by-step instructions for installing the HP XC System Software on the head node and configuring the system.
<i>HP XC System Software Administration Guide</i>	Provides an overview of the HP XC system administration environment and describes cluster administration tasks, node maintenance tasks, LSF® administration tasks, and troubleshooting procedures.
<i>HP XC System Software User's Guide</i>	Provides an overview of managing the HP XC user environment with modules, managing jobs with LSF, and how to build, run, debug, and troubleshoot serial and parallel applications on an HP XC system.

The following documents are also provided by HP for use with your HP XC system:

Linux Administration Handbook

A third-party Linux reference manual, *Linux Administration Handbook*, is shipped with the HP XC System Software Documentation Set. This manual was authored by Evi Nemeth, Garth Snyder, Trent R. Hein, et al (NJ: Prentice Hall, 2002).

QuickSpecs for HP XC System Software

Provides a product overview, hardware requirements, software requirements, software licensing information, ordering information, and information about commercially available software that has been qualified to interoperate with the HP XC System Software.

The QuickSpecs are located at the following URL:

http://www.hp.com/techservers/clusters/xc_clusters.html

HP XC Program Development Environment

The following URL provides pointers to tools that have been tested in the HP XC program development environment (for example, TotalView® and other debuggers, compilers, and so on):

<ftp://ftp.compaq.com/pub/products/xc/pde/index.html>

HP Message Passing Interface

HP Message Passing Interface (MPI) is an implementation of the MPI standard for HP systems. The home page is located at the following URL:

<http://www.hp.com/go/mpi>

HP Mathematical Library

The HP math libraries (MLIB) support application developers who are looking for ways to speed up development of new applications and shorten the execution time of long-running technical applications. The home page is located at the following URL:

<http://www.hp.com/go/mlib>

HP Cluster Platform Documents

The cluster platform documents describe site requirements, show you how to physically set up the servers and additional devices, and provide procedures to operate and manage the hardware. These documents are shipped with your hardware.

Documentation for the HP Integrity and HP ProLiant servers is available at the following URL:

<http://www.docs.hp.com/>

For More Information

The HP Web site has information on this product. You can access the HP Web site at the following URL:

<http://www.hp.com>

Supplementary Information

This section contains links to third-party and open source components that are integrated into the HP XC System Software core technology. In the XC documentation, except where necessary, references to third-party and open source software components are generic, and the XC adjective is not added to any reference to a third-party or open source command or product name. For example, the SLURM `srun` command is simply referred to as the `srun` command.

The location of each Web site or link to a particular topic listed in this section is subject to change without notice by the site provider.

- <http://www.platform.com>

Home page for Platform Computing, the developer of the **Load Sharing Facility** (LSF). LSF, the batch system resource manager used on an XC system, is tightly integrated with the HP XC and SLURM software.

For your convenience, the following Platform LSF documents are shipped on the HP XC documentation CD in PDF format. The Platform LSF documents are also available on the XC Web site.

- *Administering Platform LSF*
- *Administration Primer*
- *Platform LSF Reference*
- *Quick Reference Card*
- *Running Jobs with Platform LSF*
- <http://www.llnl.gov/LCdocs/slurm/>
Home page for the Simple Linux Utility for Resource Management (SLURM), which is integrated with LSF to manage job and compute resources on an XC system.
- <http://www.nagios.org/>
Home page for Nagios®, a system and network monitoring application. Nagios watches specified hosts and services and issues alerts when problems occur and when problems are resolved. Nagios provides the monitoring capabilities on an XC system.
- <http://supermon.sourceforge.net/>
Home page for Supermon, a high-speed cluster monitoring system that emphasizes low perturbation, high sampling rates, and an extensible data protocol and programming interface. Supermon works in conjunction with Nagios to provide XC system monitoring.
- <http://www.llnl.gov/linux/pdsh/>
Home page for the parallel distributed shell (pdsh), which executes commands across XC client nodes in parallel.
- http://www.balabit.com/products/syslog_ng/
Home page for syslog-ng, a logging tool that replaces the traditional syslog functionality. The syslog-ng tool is a flexible and scalable audit trail processing tool, and it provides a centralized, securely stored log of all devices on your network.
- <http://systemimager.org>
Home page for SystemImager®, which is the underlying technology that is used to install the XC software, distribute the golden image, and distribute configuration changes.
- <http://www.etnus.com>
Home page for Etnus, Inc., maker of the TotalView parallel debugger.
- <http://www.macrovision.com>
Home page for Macrovision®, developer of the FLEXlm™ license management utility, which is used for HP XC license management.
- <http://sourceforge.net/projects/modules/>
Home page for Modules, which provide for easy dynamic modification of a user's environment through modulefiles, which typically instruct the module command to alter or set shell environment variables.
- <http://dev.mysql.com/>
Home page for MySQL AB, developer of the MySQL database. This Web site contains a link to the MySQL documentation, particularly the *MySQL Reference Manual*.

Manpages

Manpages provide online reference and command information from the command line. Manpages are supplied with the HP XC system for standard HP XC components, Linux user commands, LSF commands, and other software components that are distributed with the HP XC system.

Manpages for third-party vendor software components may be provided as a part of the deliverables for that component.

Using the `discover(8)` manpage as an example, you can use either of the following commands to display a manpage:

```
$ man discover
$ man 8 discover
```

If you are not sure about a command you need to use, enter the `man` command with the `-k` option to obtain a list of commands that are related to the keyword. For example:

```
$ man -k keyword
```

Related Information

This section provides pointers to the Web sites for related software products and provides references to useful third-party publications. The location of each Web site or link to a particular topic is subject to change without notice by the site provider.

Related Linux Web Sites

- <http://www.redhat.com>
Home page for Red Hat®, distributors of Red Hat Enterprise Linux Advanced Server, a Linux distribution with which the HP XC operating environment is compatible.
- <http://www.linux.org/docs/index.html>
Home page for the Linux Documentation Project (LDP). This Web site contains guides covering various aspects of working with Linux, from creating your own Linux system from scratch to `bash` script writing. This site also includes links to Linux HowTo documents, frequently asked questions (FAQs), and manpages.
- <http://www.linuxheadquarters.com>
Web site providing documents and tutorials for the Linux user. Documents contain instructions on installing and using applications for Linux, configuring hardware, and a variety of other topics.
- <http://linuxvirtualserver.org>
Home page for the Linux Virtual Server (LVS), the load balancer running on the Linux operating system that distributes login requests on the XC system.
- <http://www.gnu.org>
Home page for the GNU Project. This site provides online software and information for many programs and utilities that are commonly used on GNU/Linux systems. Online information include guides for using the `bash` shell, `emacs`, `make`, `cc`, `gdb`, and more.

Related MPI Web Sites

- <http://www.mpi-forum.org>

Contains the official MPI standards documents, errata, and archives of the MPI Forum. The MPI Forum is an open group with representatives from many organizations that define and maintain the MPI standard.

- <http://www-unix.mcs.anl.gov/mpi/>

A comprehensive site containing general information, such as the specification and FAQs, and pointers to a variety of other resources, including tutorials, implementations, and other MPI-related sites.

Related Compiler Web Sites

- <http://www.intel.com/software/products/compilers/index.htm>

Web site for Intel compilers.

- <http://support.intel.com/support/performance/tools/>

Web site for general Intel software development information.

- <http://www.pgroup.com/>

Home page for The Portland Group™, supplier of the PGI® compiler.

Additional Publications

For more information about standard Linux system administration or other related software topics, refer to the following documents, which must be purchased separately:

- *Linux Administration Unleashed*, by Thomas Schenk, et al.
- *Managing NFS and NIS*, by Hal Stern, Mike Eisler, and Ricardo Labiaga (O'Reilly)
- *MySQL*, by Paul Debois
- *MySQL Cookbook*, by Paul Debois
- *High Performance MySQL*, by Jeremy Zawodny and Derek J. Balling (O'Reilly)
- *Perl Cookbook, Second Edition*, by Tom Christiansen and Nathan Torkington
- *Perl in A Nutshell: A Desktop Quick Reference*, by Ellen Siever, et al.

Typographical Conventions

<i>Italic font</i>	Italic (slanted) font indicates the name of a variable that you can replace in a command example or information in a display that represents several possible values. Document titles are shown in Italic font. For example: <i>Linux Administration Handbook</i> .
Courier font	Courier font represents text that is displayed by the computer. Courier font also represents literal items, such as command names, file names, routines, directory names, path names, signals, messages, and programming language structures.
Bold text	In command and interactive examples, bold text represents the literal text that you enter. For example: <pre># cd /opt/hptc/config/sbin</pre> In text paragraphs, bold text indicates a new term or a term that is defined in the glossary.

\$ and #	In command examples, a dollar sign (\$) represents the system prompt for the <code>bash</code> shell and also shows that a user is in non-root mode. A pound sign (#) indicates that the user is in root or superuser mode.
[]	In command syntax and examples, brackets ([]) indicate that the contents are optional. If the contents are separated by a pipe character (), you must choose one of the items.
{ }	In command syntax and examples, braces ({ }) indicate that the contents are required. If the contents are separated by a pipe character (), you must choose one of the items.
...	In command syntax and examples, horizontal ellipsis points (...) indicate that the preceding element can be repeated as many times as necessary.
:	In programming examples, screen displays, and command output, vertical ellipsis points indicate an omission of information that does not alter the meaning or affect the user if it is not shown.
	In command syntax and examples, a pipe character () separates items in a list of choices.
discover(8)	A cross-reference to a manpage includes the appropriate section number in parentheses. For example, <code>discover(8)</code> indicates that you can find information on the <code>discover</code> command in Section 8 of the manpages.
Ctrl/x	In interactive command examples, this symbol indicates that you hold down the first named key while pressing the key or button that follows the slash (/). When it occurs in the body of text, the action of pressing two or more keys is shown without the box. For example: Press <code>Ctrl/x</code> to exit the application.
Enter	The name of a keyboard key. Enter and Return both refer to the same key.
Note	A note calls attention to information that is important to understand before continuing.
Caution	A caution calls attention to important information that if not understood or followed will result in data loss, data corruption, or a system malfunction.
Warning	A warning calls attention to important information that if not understood or followed will result in personal injury or nonrecoverable system problems.

HP Encourages Your Comments

HP welcomes your comments on this document. Please provide your comments and suggestions at the following URL:

<http://docs.hp.com/en/feedback.html>

New Features

This chapter describes the new features delivered in HP XC System Software Version 2.1.

1.1 Base Distribution and Kernel

The following table lists the changes made to the base distribution and kernel.

XC Version 2.1	XC Version 2.0A
Enterprise Linux 3 Update 4	Enterprise Linux 3 Update 2
Base Red Hat kernel 2.4.21-27.0.2.EL	Base Red Hat kernel 2.4.21-15.0.4.EL
Quadrics driver kit Version 4.30	Quadrics driver kit Version 4.24
QLogic FC SAN driver 7.01.01	QLogic FC SAN driver 7.00.00b17

1.2 Additional Hardware Models Supported

Support for the following hardware models has been added:

- Cluster Platform 3000, Xeon with EM64T-based servers: HP ProLiant DL140 G2
- Cluster Platform 4000, Opteron-based servers: HP ProLiant DL145 G2 and DL385
- Cluster Platform 6000, Itanium-based servers: HP Integrity rx1620, rx2620, and rx4640

1.3 The discover Command

The following enhancements have been made to the `discover` command:

- The new `--addnode` option provides the ability to add nodes to an existing system. Nodes can be added either on the branch switch or the root switch, however they must be added as a separate operation. The new command syntax to add nodes is:


```
discover --addnode {--root | --branch} [--oldmp] [--noconsole]
```

 Refer to `discover(8)` for details.
- Nodes that do not have console ports are classified as workstations and are automatically discovered during a typical discovery process.
- The new `ic=AdminNet` command line keyword configures the interconnect on the administrative network.

1.4 Cluster Configuration

The following changes have been made to the cluster configuration process:

- Configuration questions for the Simple Linux Utility for Resource Management (SLURM) have been modified. You will be prompted to enter a SLURM user name and password. This user will be created if it does not exist.

The compute node configuration in the `slurm.conf` file is updated automatically.

- The `genelanhosts` script has been replaced by the `spconfig` script, which must be run after the `startsys` command regardless of interconnect type. This change is documented in the *HP XC System Software Installation Guide*.

1.5 The startsys Command

The `--max_at_once` option has been added to the `startsys` command. This option allows you to specify the number of nodes to image simultaneously. Specifying this option on the command line improves head node responsiveness on systems with large node counts and relatively small memory size.

1.6 Image Distribution

The ability to dynamically encounter disks on client nodes during an imaging operation, identify their type (for example, `hda`, `sda`, or `cciss`), and create partitions and file systems on them has been added. This is in contrast to the previous static method in which disk type and disk size for each client node were assumed to be known before the imaging operation occurred.

The dynamic disk discovery feature is now the default behavior during imaging operations. However you can revert to the standard, static imaging behavior if you want to assert control over the partitioning and file system creation of disks within your system.

Refer to the *HP XC System Software Installation Guide* for more information.

1.7 Resource Management

The following enhancements were made to SLURM:

- A SLURM job accounting plugin is installed by default on fresh installations and gathers run-time accounting data by job.
- The `sacct` utility was added to report accounting data. Refer to `sacct(1)` for more information.
- The persistent sequencing of SLURM job IDs was added.
- The ability to enable backfill scheduling in `RootOnly` partitions was added.
- The maximum number of characters in the SLURM job name was increased from 16 to 256.

For the Load Sharing Facility (LSF), the `JOB_STARTER` script was added to default queues to launch user jobs on the first allocated node.

1.8 Interconnects

The following updates were made to the interconnect components:

- Infiniband® update to IB Version 3.0.16.5_2 . This version adds support for Itanium-based systems.
- Myricom® update to GM Version 2.1.7, which also adds support for the Myrinet® 256 port switch.
- QsNet^{II}® RPM updates:
 - `qsnet2libs-1.8.13-0.1hptc`
 - `qsnetlibs-1.4.30-1.1hptc`
 - `hptc-qsnet2-diag-1-4.1sp`

- qsnet2diagscommon-1.0.12-2.1hptc
- qsnet2diags-1.0.12-3.1hptc
- qsnetdiags-1.0.2-14.1hptc

1.9 Serviceability

New versions of the following utilities were added:

- hpasm Version hpasm-7.1.1b-95
- collectl Version hp-collectl-1.3.1-1

1.10 Documentation Changes

All manuals in the HP XC System Software Documentation Set were revised to incorporate the new functionality delivered in this release.

As a convenience for users who are familiar with previous versions of the XC software manuals, the following sections describe information in the HP XC System Software Documentation Set that was enhanced, added, condensed, moved, or removed for this release.

1.10.1 *HP XC Hardware Preparation Guide*

Hardware preparation tasks were added for the following new supported hardware models:

- HP ProLiant DL140 G2
- HP ProLiant DL145 G2
- HP ProLiant DL385
- HP Integrity rx1620
- HP Integrity rx2620
- HP Integrity rx4640

1.10.2 *HP XC System Software Installation Guide*

The following changes were made in this manual:

- The appendix containing the sample Kickstart configuration files was removed. If you want to view the Kickstart configuration file `ks.cfg`, mount the HP XC System Software DVD.
- A new appendix provides a list of appropriate Ethernet interfaces to use as the external network connection of the head node.
- A new appendix describes how to customize client node disk partitions.
- A new appendix describes how to enable telnet on Integrated Lights Out (iLO) devices.
- Appendix I was enhanced to provide detailed descriptions of node roles.
- A new appendix describes how to determine the network type on systems using a QsNet^{II} interconnect.
- A new appendix describes how to install and configure the Maui Scheduler.

1.10.3 *HP XC System Software Administration Guide*

The following changes were made in this manual:

- A new table in Chapter 1 summarizes all XC commands. Previously, each XC command was briefly described in Chapter 2.

For this release, refer to the appropriate manpage for XC command descriptions.

- Chapter 1 includes a new table of recommended administrative tasks, which was formerly Appendix A.
- A new chapter describes how to mount a file system using `csys`.
- A new procedure describes how to add a service to an XC system.
- Information about SLURM job accounting was added to the “SLURM Administration” chapter.
- A new procedure describes how to customize Nagios metrics gathering.
- Information from the discontinued *HP XC System Software Overview* has been incorporated into this manual.

1.10.4 *HP XC System Software User’s Guide*

The following changes were made in this manual:

- Information about how SLURM and LSF-HPC interact has been updated.
- Information from the discontinued *HP XC System Software Overview* has been incorporated into this manual.

1.10.5 *HP XC System Software Overview*

This manual was discontinued from the HP XC System Software Documentation Set; its contents was moved to appropriate places in the remaining XC software manuals.

General Notes

This chapter contains general information that applies to the XC system as a whole.

2.1 XC System Naming

Throughout the HP XC System Software Documentation Set, the following terms are used to denote an HP Cluster Platform on which HP XC System Software has been installed:

XC Name	Cluster Platform (CP) Model	Chip Architecture
XC3000	Cluster Platform 3000	Xeon with EM64T
XC4000	Cluster Platform 4000	AMD Opteron
XC6000	Cluster Platform 6000	Intel Itanium 2

2.2 Master Firmware Revision Table

The HP XC System Software is qualified against specific firmware revisions, and you can obtain the table of supported firmware revisions on line at:

http://www.hp.com/techservers/clusters/xc_clusters.html

2.3 Obtaining XC Documentation

In addition to the printed copy of the HP XC System Software Documentation Set you received with your system, all XC manuals except the *HP XC System Software Release Notes* are shipped in PDF and HTML format on the XC documentation CD you received with your software distribution kit.

All XC manuals, including the *HP XC System Software Release Notes*, are available on line from the XC Clusters Web site at:

http://www.hp.com/techservers/clusters/xc_clusters.html

Additional XC-specific topics that are documented after the HP XC System Software Documentation Set is published are issued on line as HowTo documents, which are available at the same URL as the XC manuals.

As an added convenience, the XC manuals and HowTo documents are also available from the official HP documentation Web site, which is located at:

<http://www.docs.hp.com>

The HP XC documentation is located under the Linux category.

Hardware Preparation Notes

Hardware preparation tasks are documented in the *HP XC Hardware Preparation Guide*. This chapter contains information that was not included in the manual at the time of publication.

The following topics are included in this chapter:

- Configuring disks into the smart array (Section 3.1)
- Incorrect instruction for preparing HP ProLiant DL145 G2 Nodes (Section 3.2)
- Preparing HP Integrity rx4640 servers (Section 3.3)
- Preparing HP Integrity rx2620 servers (Section 3.4)

3.1 Configuring Disks Into the Smart Array

On server models such as the HP ProLiant DL385, DL585, DL360 G4, and DL380 G4 with smart array cards, you must add the disk or disks to the smart array before attempting to image the node.

To do so, watch the screen carefully during the power on self-test phase of the node, and press the F8 key when prompted to configure the disks into the smart array.

Specific instructions are outside the scope of the XC documentation; refer to the documentation that came with your model of HP ProLiant server for more information.

3.2 Incorrect Instruction for Preparing HP ProLiant DL145 G2 Servers

The *HP XC Hardware Preparation Guide* has an error in Section 3.4.2, *Preparing HP ProLiant DL145 G2 Nodes*. Step 3 of the hardware preparation tasks instructs you to disable hyperthreading. Hyperthreading is only available on Itanium-based systems, and the HP ProLiant DL145 G2 is an Operton-based system. Therefore, ignore this step.

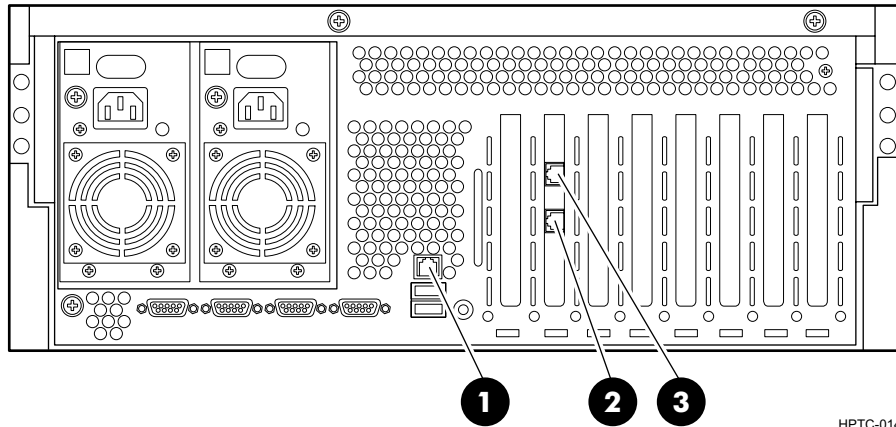
3.3 Preparing HP Integrity rx4640 Servers

HP Integrity rx4640 servers are supported as XC nodes in a Cluster Platform 6000 configuration (CP6000).

Follow the procedures in this section to prepare each HP Integrity rx4640 server before installing and configuring the HP XC System Software.

Figure 3-1 shows a rear view of the HP Integrity rx4640 server and the appropriate port assignments for an XC system.

Figure 3-1: HP Integrity rx4640 Server Rear View



- 1 The port labeled MP LAN is the MP connection to the ProCurve Console Switch.
- 2 The port labeled LAN Gb connects to the Administrative Switch (branch or root).
- 3 This unlabeled port is used for an external connection.

Perform the following hardware preparation tasks on each HP Integrity rx4640 server in your CP6000 system:

1. For each node in the XC system, ensure that the power cord is connected but that the CPU is not turned on.
2. Follow this procedure to connect a personal computer (PC) to the Management Processor:
 - a. Connect a three-way DB9–25 cable to the MP DB-25 port on the back of the HP Integrity rx2600 server.
 - b. Connect the CONSOLE connector to a null modem cable, and connect the null modem cable to the PC COM1 port.
 - c. Use a terminal emulator, such as HyperTerminal, to open a terminal window.
 - d. Press the Enter key to access the MP. If there is no response, press the MP reset pin on the back of the MP and try again.
 - e. Log in to the MP using the default user name and password shown on the screen. The MP Main Menu is displayed.
3. Enter **SL** to clear the error logs (**CLR**).
4. Enter **CM** to display the Command Menu.
5. Ensure that all MP interfaces have had their IP address, subnet mask, and default gateway address preconfigured before you begin the installation. This procedure is described in the documentation that came with your model of HP Integrity server.

Use the **LC** command to set the LAN configuration.

Typically, the MP on the head node is not connected to the internal network; it is connected to the external network. By default, a 16-node system with a single utility cabinet has the following IP addresses set for the MP on the head node:

- The MP IP address on the external network is set by you. Using the sample IP addresses used in this manual, it is set to 192.168.0.1.
- Gateway address 172.21.0.16 (default based on 16 nodes).

- Subnet mask address 255.0.0.0.

In this example, IP addresses for additional nodes are assigned as shown in Table 3-1.

Table 3-1: IP Addresses for MP Power Management Devices

Node	IP Address
First node after the head node is n15	172.21.0.15
Second node after the head node is n14	172.21.0.14
Third node after the head node is n13	172.21.0.13
:	:
n3	172.21.0.3
n2	172.21.0.2
Last node is n1	172.21.0.1

6. Enter **XD** to apply your changes. Enter **R** to restart the MP.
7. Enter **CM** to return to the Command menu.
8. Enter **SO** to set the MP user name and password. The user name must have a minimum of 6 characters, and the password must have a minimum of 8 characters. You must set the same user name and password on every node. The user name and password are required to access the power management device and console, for example, when you issue the console `nodename` command.
9. Enter **PC** (power cycle) to turn on power to the node; then choose the Boot Option Maintenance Menu.
10. Press **Ctrl/b** to return to the Main menu.
11. Enter **CO** to connect to the console.
12. From the Boot Menu screen, which is displayed during the power on of the node, choose the Boot Configuration Menu. Perform this step on all nodes except the head node.
 - a. Choose Add a Boot Entry.
 - b. Choose Load File [Core LAN Gb A] as the network boot choice, which is the Gigabit Ethernet (GigE) port.
 - c. Enter the string **Netboot** as the boot option description. This entry is required and must be set to the string Netboot.
 - d. Enter **N** for No Boot Option when prompted for the Boot Option Data Type.
 - e. Enter **Y** to save the entry to NVRAM.
 - f. Choose Exit to quit.

For more information about how to work with these menus, see the documentation that came with your model of HP Integrity server.

13. From the Boot Configuration screen, choose the option to Edit OS Boot Order. Perform this step on all nodes except the head node:
 - a. Use the navigation instructions on the screen to move the Netboot entry you just defined to the top of the boot order.
 - b. Enter **Y** to save the entry to NVRAM.
 - c. Choose Exit to close the menu.

14. Perform this step on all nodes, including the head node:
 - a. Choose the `Select Input Console` option to enable console messages to be displayed on the screen when you turn on the system:
 - i. Enable the `Acpi (HWP0002, 0) / Pci (1 | 1) / Uart (9600 N81) / VenMsg (Vt100+)` option.
 - ii. Enter **Y** to save the entry to NVRAM.
 - iii. Choose `Exit` to return to the menu.
 - b. Choose the `Select Output Console` option to enable console messages to be displayed on the screen when you turn on the system:
 - i. Enable the `Acpi (HWP0002, 0) / Pci (1 | 1) / Uart (9600 N81) / VenMsg (Vt100+)` option.
 - ii. Enable the `Acpi (HWP0002, 0) / Pci (4 | 0)` option.
 - iii. Enter **Y** to save the entry to NVRAM.
 - iv. Choose `Exit` to return to the menu.
 - c. Choose the `Select Active Standard Error Devices` option from the `Boot Option Maintenance Menu` to enable console messages to be displayed on the screen when you turn on the system.
 - i. Enable the `Acpi (HWP0002, 0) / Pci (1 | 1) / Uart (9600 N81) / VenMsg (Vt100+)` option.
 - ii. Enable the `Acpi (HWP0002, 0) / Pci (4 | 0)` option.
 - iii. Enter **Y** to save the entry to NVRAM.
 - iv. Choose `Exit` to return to the menu.
 - d. Press the `Esc` key to return to the `Boot Menu`.
15. Turn off power to the node:
 - a. Press `Ctrl/b` to exit console mode.
 - b. Enter **CM** to display the `Command Menu`.
 - c. Enter **PC** to turn off power to the node.

3.4 Preparing HP Integrity rx2620 Servers

To prepare HP Integrity rx2620 servers, do not use the instructions in the *HP XC Hardware Preparation Guide*, follow the hardware preparation tasks provided for the HP Integrity rx4640 shown in Section 3.3 in this document.

Installation Notes

This chapter contains notes that apply to the XC software installation process.

4.1 Notes to Read Before the Kickstart Installation Process

Read the notes in this section before starting the Kickstart installation process.

4.1.1 Prepare Previously Installed Nodes for a Reinstallation

If you are reinstalling an XC system that is already running an early, advance version of this release, you must first prepare the nodes to network boot before shutting down the system. This procedure is documented in Appendix H in the *HP XC System Software Installation Guide*.

4.1.2 No Hard Drives Found on Head Nodes with Serial ATA Disks with 4 GB or More Memory

This note applies only to head nodes that contain SATA disk drives as their internal storage devices that have 4 GB or more of system RAM.

On a head node with SATA disks with 4 GB or more of memory, no hard drives are found during the installation process. The `ata_piix` driver, which is used for the SATA drives on a number of XC platforms, is sensitive to memory mapping that arises in systems with 4 GB or more of RAM.

To work around this problem, include `mem=2000mb` on the boot command line when starting the installation, as follows:

```
# linux mem=2000mb ks=cdrom:/ks.cfg
```

4.1.3 Installation Supported on SAN on CP6000 Systems But Not on CP4000 Systems

This release can be installed on a SAN device only on a CP6000 system. The procedure is described in Section 11.2.

On CP4000 systems, installation is supported only on disks local to the head node.

If you have a SAN disk connected to the head node on a CP4000 system, examine installation messages closely to determine whether they are related to the local disk on which you are installing. Be sure to select `N` to partition-related questions presented by the installation program.

After the installation Kickstart process has completed and the new kernel has been booted, normal usage of a SAN is supported.

Configuration Notes

This chapter contains information about configuring the system. Notes that describe additional configuration tasks are mandatory and have been organized chronologically. Perform these tasks in the sequence presented in this chapter.

The XC system configuration procedure is documented in Chapter 4 of the *HP XC System Software Installation Guide*.

5.1 Tasks to Perform After Running the `cluster_config` Utility

Perform the tasks described in this section after you run the `cluster_config` utility, before the golden system image is replicated to all nodes.

5.1.1 Required Task: Update Nagios Status

Nagios is started before any client nodes are imaged and started. From time to time, Nagios may fail to update the status of the display. This problem can be identified by pending services displayed on the Nagios Web page or by services that appear to have not updated recently as shown by the time stamp.

To correct this problem, restart Nagios on the head node:

```
# service nagios restart
```

If the Nagios status is not updated, do the following:

```
# service nagios stop
# rm -f /opt/hptc/nagios/var/status.sav
# service nagios start
```

The previous command sequence removes the cached copy of the service states and requires Nagios to acquire all status from the various services.

5.2 Hang May Be Encountered When Logging In To the Head Node

When logging in to the GNOME desktop environment on the head node, you may encounter a blue screen that shows nothing but the mouse pointer. To correct this hang, you must terminate the `gnome-settings-daemon`.

To terminate the `gnome-settings-daemon`, perform the following procedure from a remote login to the head node:

1. Determine the process ID of the `gnome-settings-daemon`:

```
# ps -aef | grep gnome-settings-daemon | grep -v grep
```

Output will be similar to the following; the process ID of the daemon is indicated by `[PID]`:

```
root      [PID]      1  0 11:16 ?          00:00:00
gnome-settings-daemon --oaf-activate-iid=OAFIID:GNOME_SettingsDaemon --
oaf-ior-fd=14
```

2. Terminate the process ID:

```
# kill -9 [PID]
```

Stopping the daemon causes the `gnome-settings-daemon` to restart itself, which enables you to log in to the head node.

System Administration and Management Notes

This chapter contains notes about system administration and management commands and tasks. Perform these tasks only when necessary.

6.1 Running the dgemm Utility

The `dgemm` utility does not run on all supported interconnects. Therefore, it is preferable to run the `dgemm` utility on the Administrative Network instead of the interconnect.

You can do this by invoking `dgemm` with the following options:

```
# mpirun -prot -TCP -srun -v -p lsf -n max /opt/hptc/contrib/bin/dgemm.x
```

6.2 Log Files Must Be Rotated and Compressed

You must rotate and compress the log files in the `/hptc_cluster/adm/logs` directory. There is no generic answer for the size and frequency of the log rotation, but as a rule of thumb, consider sizing the logs based on some percentage of the overall disk that contains them.

For example, you might consider a methodology that takes the size of the `/hptc_cluster` partition and allocates 30 percent of that space to logs. You might then decide you want to keep the last five logs, rotated at some preset size.

The following example demonstrates how to set up the log rotate facility:

1. Add the following lines to the end of the `/etc/logrotate.d/syslog` file on the head node:

```
/var/log/n16 /hptc_cluster/adm/logs/consolidated.log \ [1]
/hptc_cluster/adm/logs/nodenaming_prefix* {
    rotate 5
    create
    compress
    size=max_size_of_file
    postrotate
        /bin/kill -HUP `cat /var/run/syslog-ng.pid 2> /dev/null` 2> \ [2]
        /dev/null || true
    endscript
}
```

- [1] Because this line is too long to fit on a printed page, the backslash character (`\`) indicates line continuation. You must enter the first and second lines of this example on one continuous line.
In this line, `n16` is the name of the head node; replace `n16` with the name of your head node.
- [2] The backslash character indicates line continuation. You must enter this line and the next line on one continuous line.

Follow these guidelines to determine the values required for the entries in the `syslog` file:

- `nodenaming_prefix` represents the node naming prefix defined in the `/opt/hptc/config/discover_data.ini` file.
- `max_size_of_file` is calculated as follows:

$$30\% * (\text{hptc_cluster partition size in MB}) / 5 * (\text{number of nodes in the cluster} + \text{number of syslogng_forward server} + 1)$$

- The `number of syslogng_forward servers` represents the number of management aggregators that have been defined for the system. A `syslogng-forward` service is allocated for each assigned management hub role.

The following command shows that the system has five forwarders:

```
# shownode servers syslogng_forward
<nodenaming_prefix>[252-256]
```

- Use the following command to determine the size of the `/hptc_cluster` partition in megabytes:

```
# df -B M /hptc_cluster | tail -1 | awk '{print $2}'
5424
```

Thus, using the data obtained in this example, use the following calculation to determine the value for `max_size_of_file` on a 256 node cluster (approximately 1,302,436 bytes):

$$0.30 * (5424 \text{ MB}) / 5 * (256 + 5 + 1) = 1.2421 \text{ MB}$$

2. Use the text editor of your choice to modify the template file:

```
/opt/hptc/syslog-ng/etc/global/syslog_ng_global_header_1
```

3. Change the line that looks like this:

```
tcp(ip(0.0.0.0) max-connections(300) port(514));
```

To this:

```
tcp(ip(0.0.0.0) max-connections(300) port(514) keep-alive(yes));
```

4. Save your changes and exit the file.
5. Reconfigure and restart the `syslog-ng` service:

```
# service syslog-ng nconfigure
# service syslog-ng restart
```

6.3 Recommended NFS Mount Options for External Connections

Access to some external NFS mounts may become hung on the head node. The problem is intermittent and is difficult to reproduce. Other symptoms of the problem may include high IOWAIT or load average values and processes stuck in the uninterruptable sleep D state. When the system is in this state, the only method to recover is to reboot the head node.

HP has not been able to reproduce the problem when external NFS mounts use the `noac` and `noacl` NFS mount options. Therefore, until this problem is better understood and fixed in a future update of the XC system software, HP recommends using the `noac` and `noacl` NFS mount options on all external NFS mounts.

Programming and User Environment Notes

This chapter contains information that applies to the programming and user environment.

7.1 Notes About the HP Math Library

The following notes apply to Intel compilers and the HP Math Library (MLIB):

- After installation, MLIB directory information is located in the `/opt/mlib/README` file.
- MLIB requires the Intel Fortran Compiler.
- When using `/opt/mlib/intel_7.1/hpmpi_2.0`, use the Intel Version 7 compilers.
- When using `/opt/mlib/intel_8.0/hpmpi_2.0`, use the Intel Version 8 compilers.
- When using `/opt/mlib/pgi_5.2/hpmpi_2.0`, use the Portland Group (PGI) Version 5.2 compilers.
- Use the following specific Fortran compilers:
 - Version 8.0 Fortran Compiler
 - RPM: `intel-ifort8-8.0-57.ia64.rpm`
 - tar file: `1_fc_pc_8[1].0.046.tar.gz`
 - Version 7.1 Fortran compiler
 - RPM: `intel-efc7-7.1-41.ia64.rpm`
 - tar file: `1_fc_pc_7[1].1.040.tar`
 - Version 5.2 Portland Group (PGI) Fortran compiler
 - tar file: `5.2.4-linux86-64[1].tar`

7.2 Notes About MLIB, HP MPI, and Modulefiles

When building and running an application built against MLIB, it is crucial that the environment is consistent. Modulefiles can make it easier to access a package, therefore, if modulefiles are used, it is necessary to use a consistent set of modulefiles.

In particular, modulefiles can be used to select a compiler, both making its command available in `$PATH` and making its shared objects available in `$LD_LIBRARY_PATH`. MLIB has a modulefile corresponding to each supported compiler, making its shared objects available in `$LD_LIBRARY_PATH`. If modulefiles are used to facilitate the user environment, failure to use companion modulefiles will result in build and run time errors or both.

If the HP Message Passing Interface (MPI) is used as well, it is important to make sure the `mpi**` compiler scripts use the intended compiler, for example, by setting the `MPI_CC` or `MPI_F90` environment variables (or both). Failure to do so may cause the compiler scripts to discover a compiler that is not the intended compiler, and thus introduce an unintended inconsistency.

For more information, refer to the *MLIB User's Guide*, which is located at the following URL and on the XC documentation CD:

<http://www.hp.com/go/mlib>

7.3 Configuring the Intel Trace Collector and Analyzer with HP MPI on XC

The Intel Trace Collector was formerly known as VampirTrace. The Intel Trace Analyzer was formerly known as Vampir.

7.3.1 Installation Notes

The following are installation-related notes:

- Installation kits:
 - ITC-IA64-LIN-MPICH-PRODUCT.4.0.2.1.tar.gz
 - ITA-IA64-LIN-AS21-PRODUCT.4.0.2.1.tar.gz
- Installation locations:
 - The Intel Trace Collector is installed in the `/opt/IntelTrace/ITC` directory.
 - The Intel Trace Analyzer is installed in the `/opt/IntelTrace/ITA` directory.
 - The license file is located in the `/opt/IntelTrace/` directory so both tools can find it.

7.3.2 HP MPI and the Intel Trace Collector (OTA)

The following are notes regarding building a program with OTA.

HP MPI is MPICH compatible if you use the following HP MPI MPICH scripts, which are located in the `/opt/mpi/bin` directory:

- `mpicc` is replaced by `mpicc.mpich`
- `mpif77` is replaced by `mpif77.mpich`
- `mpirun` is replaced by `mpirun.mpich`

In summary: `mpiXX` becomes `mpiXX.mpich`

As an example, the `examples` directory under `/opt/IntelTrace/ITC` was copied to a home directory and renamed to `ITC_examples_xc6000`. The GNU makefile now looks as follows:

```
CC          = mpicc.mpich
F77         = mpif77.mpich
CLINKER     = mpicc.mpich
FLINKER     = mpif77.mpich
IFLAGS      = -I$(VT_ROOT)/include
CFLAGS      = -g
FFLAGS      = -g
LIBS        = -lvtunwind -ldwarf -lnsl -lm -lelf -lpthread
CLDFLAGS    = -static-libcxa -L$(VT_ROOT)/lib $(TLIB) -lvtunwind -ldwarf -lnsl -lm -lelf -lpthread
FLDFLAGS    = -static-libcxa -L$(VT_ROOT)/lib $(TLIB) -lvtunwind -ldwarf -lnsl -lm -lelf -lpthread
```

When the Intel compilers are used, add `-static-libcxa` to the link line; otherwise, the following errors are generated at run-time:

```
[n1]/nis.home/sballe/xc_PDE_work/ITC_examples_xc6000 >mpirun.mpich
-np 2 ~/xc_PDE_work/ITC_examples_xc6000/vtjacobic warning: this is a
development version of HP MPI for internal R&D use only
/nis.home/sballe/xc_PDE_work/ITC_examples_xc6000/vtjacobic:
error while loading shared libraries: libcprts.so.6:
```

cannot open shared object file: No such file or directory MPI Application
rank 0 exited before MPI_Init() with status 127 mpirun exits with
status: 127 [n1]/nis.home/sballe/xc_PDE_work/ITC_examples_xc6000 >

For more information, go to the following URL:

<http://support.intel.com/support/performance/c/linux/sb/CS-010097.htm>

Running Your Program

Both the C and Fortran runs were successful when the `-static-libcxa` flag was added. This will only work if you use `mpirun.mpich` to launch your program.

The following is a C example called `vtjacobic`:

```
# mpirun.mpich -np 2 ~/xc_PDE_work/ITC_examples_xc6000/vtjacobic
warning: this is a development version of HP MPI for internal R&D use only
/nis.home/user_name/xc_PDE_work/ITC_examples_xc6000/vtjacobic: 100 iterations in
0.228252 secs (28.712103 MFlops), m=130 n=130 np=2
[0] Intel Trace Collector INFO: Writing tracefile vtjacobic.stf in
/nis.home/user_name/xc_PDE_work/ITC_examples_xc6000
mpirun exits with status: 0
```

The following is a Fortran example called `vtjacobif`:

```
# mpirun.mpich -np 2 ~/xc_PDE_work/ITC_examples_xc6000/vtjacobif
warning: this is a development version of HP MPI for internal R&D use only
 2 Difference is 0.390625000000000
 4 Difference is 0.123413085937500
 6 Difference is 6.341743469238281E-002
 8 Difference is 3.945139702409506E-002
10 Difference is 2.718273504797253E-002
12 Difference is 1.992697400677912E-002
14 Difference is 1.520584760276139E-002
16 Difference is 1.192225932086809E-002
18 Difference is 9.527200632430437E-003
20 Difference is 7.718816241067778E-003
22 Difference is 6.318016878920021E-003
24 Difference is 5.211741863535576E-003
26 Difference is 4.324933536667125E-003
28 Difference is 3.605700997191797E-003
30 Difference is 3.016967266492488E-003
32 Difference is 2.531507385360385E-003
34 Difference is 2.128864474525351E-003
36 Difference is 1.793360334698915E-003
38 Difference is 1.512772311960036E-003
40 Difference is 1.277430422527046E-003
42 Difference is 1.079587281504879E-003
44 Difference is 9.129693228356968E-004
46 Difference is 7.724509294615510E-004
48 Difference is 6.538134083283944E-004
50 Difference is 5.535635456297787E-004
52 Difference is 4.687947140887371E-004
54 Difference is 3.970788908277634E-004
56 Difference is 3.363815146174385E-004
58 Difference is 2.849935053418584E-004
60 Difference is 2.414763917353628E-004
62 Difference is 2.046176064805586E-004
64 Difference is 1.733937801556939E-004
66 Difference is 1.469404085386082E-004
68 Difference is 1.245266549586746E-004
70 Difference is 1.055343296682637E-004
72 Difference is 8.944029434752290E-005
74 Difference is 7.580169395426893E-005
76 Difference is 6.424353519703476E-005
78 Difference is 5.444822123484475E-005
80 Difference is 4.614672291984789E-005
82 Difference is 3.911112299221254E-005
84 Difference is 3.314831465581266E-005
86 Difference is 2.809467246160129E-005
```

```
88 Difference is 2.381154327036583E-005
90 Difference is 2.018142964565221E-005
92 Difference is 1.710475838933507E-005
94 Difference is 1.449714388058985E-005
96 Difference is 1.228707004052045E-005
98 Difference is 1.041392661369357E-005
[0] Intel Trace Collector INFO: Writing tracefile vtjacobif.stf in
/nis.home/user_name/xc_PDE_work/ITC_examples_xc6000
mpirun exits with status: 0
```

Across Nodes (using LSF)

```
# bsub -n4 -I mpirun.mpich -np 2 ./vtjacobic
```

The license file and the OTC directory need to be distributed across the nodes.

Load Sharing Facility and Job Management Notes

This chapter contains notes about the following topics

- Load Sharing Facility (LSF) (Section 8.1)
- Job management with SLURM (Section 8.2)

8.1 LSF

The Load Sharing Facility (LSF), developed by Platform Computing, is available for use in this release. This version of LSF HPC has been integrated with the SLURM functionality in order to couple LSF's extensive scheduling tools with SLURM's parallel job-launching facilities into a single comprehensive job management environment on the XC system.

8.1.1 LSF Always Runs on the Head Node

LSF always runs on the head node, by default, even if the resource management role is assigned to a different node.

Use the following command to relocate LSF to a different node if the node has been assigned the resource management role:

```
# controllsf set primary node_name
```

8.2 Job Management

At the time of publication, there are no notes that apply to the Simple Linux Utility for Resource Management (SLURM). SLURM provides commands for launching, monitoring, and controlling jobs.

Refer to the *HP XC System Software User's Guide* for more information about using SLURM.

Cluster Platform 3000 Notes

This chapter contains information that applies only to Cluster Platform 3000 systems.

9.1 Remote Console Logins Do Not Work on HP ProLiant DL140 G2 Nodes

Logging in remotely to a console, either through the `XC console` command or by a `telnet` session to the lights-out 100i (LO-100i) remote management processor, does not work on HP ProLiant DL140 G2 nodes.

Secure shell (`ssh`) logins can be opened to the nodes through the administrative network, interconnect, or external connections. `telnet` connections to the management processor can be established and other operations can be performed, but Linux logins to the nodes are not possible through that interface.

The following message might be observed on the console or in message logs:

```
init: Id "co" respawning too fast: disabled for 5 minutes
```

This problem will be resolved in a future revision of the HP ProLiant DL140 G2 firmware.

9.2 The power Command Might Fail on HP ProLiant DL140 G2 Nodes

Under certain conditions, HP ProLiant DL140 G2 nodes might fail to turn off power with the `XC power` command, with other commands that invoke the `power` command, or with the power control commands of the LO-100i remote management firmware.

This problem may occur if the node is halted or if the kernel panics or is hung. This problem may also happen on an initial installation when an operating system is not running on the node. If this happens, turn off power to the node by manually holding down the power button or removing power.

This problem will be resolved in a future revision of LO-100i remote management firmware.

Cluster Platform 4000 Notes

This chapter contains information that applies only to Cluster Platform 4000 systems.

10.1 Remote Console Logins Do Not Work on HP ProLiant DL145 G2 Nodes

Logging in remotely to a console, either through the XC console command or by a telnet session to the lights-out 100i (LO-100i) remote management processor, does not work on HP ProLiant DL145 G2 nodes.

Secure shell (`ssh`) logins can be opened to the nodes through the administrative network, interconnect, or external connections. `telnet` connections to the management processor can be established and other operations can be performed, but Linux logins to the nodes are not possible through that interface.

The following message might be observed on the console or in message logs:

```
init: Id "co" respawning too fast: disabled for 5 minutes
```

This problem will be resolved in a future revision of the HP ProLiant DL145 G2 firmware.

10.2 The power Command Might Fail on HP ProLiant DL145 G2 Nodes

Under certain conditions, HP ProLiant DL145 G2 nodes might fail to turn off power with the XC `power` command, with other commands that invoke the `power` command, or with the power control commands of the LO-100i remote management firmware.

This problem may occur if the node is halted or if the kernel panics or is hung. This problem may also happen on an initial installation when an operating system is not running on the node. If this happens, turn off power to the node by manually holding down the power button or removing power.

This problem will be resolved in a future revision of LO-100i remote management firmware.

10.3 Sensor Information for Supermon Not Available

Sensor information for Supermon is not currently supported for Opteron-based systems. Work is in progress to determine how to enable this feature for Opteron-based systems in a future release.

Cluster Platform 6000 Notes

This chapter contains information that applies only to CP6000 systems.

11.1 Excessive Boot Time with Unzoned SAN Volume Connected Through an A6824A HBA

When a SAN volume is connected to an HP Integrity rx2600 system that has an A6824A (dual-channel fibre) host bus adapter (HBA) installed, excessive boot times (in the 2-4 hour range) have been observed. The solution to this problem is to implement zoning on the SAN switch.

This does not actually represent a problem, but rather a misconfiguration of the SAN; it is likely that an administrator experienced with SAN volumes will have already implemented zoning on the SAN switch. This note is targeted primarily to administrators who are configuring a SAN volume on an HP Integrity rx2600 for the first time.

For detailed information on implementing zoning, refer to the latest version of the user guide for HP Storage Works Zoning, which is located at the following URL (search for the string `zoning guide` in the Search box on this page):

<http://www.hp.com/country/us/eng/prodserv/storage.html>

At a minimum, the following items must be included in the pertinent zone:

- The World Wide Name (WWN) of the adapter nodes and ports on the A6828A HBA, which is located in the `/proc/scsi/qla2300/#` file, where `#` represents a numeric file name (these vary between systems).
- The numbers of the ports (expressed as `blade#,port#`) on the SAN switch that are connected to the HBA.
- The WWN of the controllers for the physical storage on the SAN.

After zoning is successfully implemented, the boot time is reduced to a few minutes.

This problem was not observed with the FCA2214 HBA, but this adapter is intended for 32-bit environments and is not supported for the HP Integrity rx2600 (although it does seem to work).

11.2 Installing to and Booting from a SAN Volume

You can install the XC System Software to a SAN volume and use it as the boot device for an HP Integrity rx2600 head node. The HP Integrity rx2600 must be connected to the SAN volume through an A6826A HBA, and certain procedures must be followed in order to ensure successful booting.

1. Perform the installation using only one of the two channels on the HBA connected to the SAN switch. This means that either the second port on the HBA should be physically disconnected from the SAN switch during installation, or the SAN volume should be presented (using the GUI on the SAN Appliance) to only one of the two ports on the HBA. If each of the ports are both physically and virtually connected, the system will be unable to remount the root file system in read/write mode during Linux boot; the remount attempt will fail with the following message:

duplicate LABELS

2. It is not currently possible to determine the device to select at installation time. The SAN volume is shown twice in the list of devices; it may appear as `/dev/sdb` and `/dev/sdc` (the exact device names depend on the number of SCSI hard disks in the system). During the installation, select a device, and if the installation process is unable to use it, restart the installation and select the other device.
3. When the installation is complete, in order to connect the second HBA port to the SAN switch (either physically or virtually, that is, presented through the SAN Appliance GUI), you must edit the `/etc/fstab` file.

To do so, determine the partitions that correspond to the labels in the `/etc/fstab` file (the first column in the entries in that file will contain either a partition name, such as `/dev/sdb1`, or a label identifier, such as `LABEL=/1`).

To determine the partition names corresponding to the label identifiers, use the `findfs` command for each label identifier in the `/etc/fstab` file. For example, to determine the partition that corresponds to `LABEL=/1`, enter the following command:

```
# findfs LABEL=/1
```

4. When you have identified the label identifiers and partitions, edit the `/etc/fstab` file (make sure to save a backup copy) and replace all of the label identifiers with the corresponding partition names. Save your changes to the file. Then, connect the second HBA port, and reboot the system.

Interconnect Notes

This chapter contains generic information that applies to the supported interconnect types:

- InfiniBand® interconnect (Section 12.1)
- Myrinet® interconnect (Section 12.2)
- QsNet^{II}® interconnect (Section 12.3)

12.1 InfiniBand Interconnect

At the time of publication, there are no release notes specific to the InfiniBand interconnect.

12.2 Myrinet Interconnect

The following release notes are specific to the Myrinet interconnect.

12.2.1 The `clear_counters` Command Does Not Work on the 256 Port Switch

The `/opt/gm/sbin/clear_counters` command does not clear the counters on the Myrinet 256 port switch. The web interface to the Myrinet 256 port switch has changed from the earlier, smaller switches.

To clear the switch counters, you must open an interactive Web connection to the switch and clear the counters using the menu commands. The `gm_prodmode_mon` script, which uses the `clear_counters` command, will not clear the counters periodically, as it does on the smaller switches.

This problem will be resolved in a future software update from Myricom.

12.2.2 New or Changed Myrinet GM Routes May Not Be Found by `gm_board_info`

It is possible for a problem to occur on the Myrinet GM network that causes some nodes to be unable to update their GM map files. This is most noticeable when a node reboots and sees no other nodes on the GM network even though one or more nodes are connected and working on the network.

If this problem occurs, run the following command on all nodes (including SFS servers and clients) connected to the GM network:

```
# /opt/gm/bin/gm_board_info | grep -e "Mapper is" -e "Map version is"
```

Command output will look like this if a node is affected by this problem:

```
Mapper is 00:00:00:00:00:00.
```

```
Map version is 4120027.
```

The `Map version` is a numeric identifier of the current network topology that the node sees on the GM network. If a node is exhibiting the problem, the map version will be different from the other systems because its map of the network will not

have been updated as nodes are booted or shut down. It is possible for the mapper to be correctly set and the map version to be incorrect, so check both.

There is one mapper per port, so there will be two lines of output for Myrinet XP and four lines of output for Myrinet 2XP.

This problem is caused by a node that has either crashed or hung or has become unresponsive in some way but is still powered on. To correct this problem, reboot the unresponsive node.

12.3 QsNet^{II} Interconnect

The following release notes are specific to the QsNet^{II} interconnect.

12.3.1 ELAN4 Diagnostic Tools Fail on Systems With ELAN3 Interconnect

This release note applies only to systems with an ELAN3 interconnect.

The `/opt/hptc/etc/gconfig.d/C30swmlogger gconfigure` script, which is run by the `cluster_config` utility, fails on an ELAN3 system because the `swmlogger` service and the `qsnet` diagnostic database are supported only on ELAN4 systems.

On ELAN3 systems, enter the following commands on the head node to stop the `swmlogger` service from running and to disable the `swmlogger` service from starting at boot time. You must run these two commands any time you rerun the `cluster_config` utility:

```
# service swm stop
# chkconfig --del swm
```

12.3.2 OVP Interconnect Tests Fail on ELAN3

This release note applies only to systems using the ELAN3 interconnect.

The operation verification procedure (OVP) interconnect tests are valid for ELAN4 only and will fail on an ELAN3 system with the following errors; these errors are benign and can be safely ignored:

Verify interconnect:

```
Testing quadrics/qsnet_database ...
    --- FAILED ---
Testing quadrics/swmlogger ...
    +++ PASSED +++
Testing quadrics/network ...
    --- FAILED ---
```

12.3.3 Possible Conflict with Use of SIGUSR2

The Quadrics QsNet^{II} software internally uses SIGUSR2 to manage the interconnect. This can conflict with any user applications that use SIGUSR2, including for debugger use.

To work around this conflict, set the environment variable `LIBELAN4_TRAPSIG` for the application to a different signal number other than the default value 12 that corresponds to SIGUSR2. Doing this instructs the Quadrics software to use the new signal number, and SIGUSR2 can be once again used by the application. Signal numbers are define in the `/usr/include/asm/signal.h` file.

12.3.4 ELAN TRAP Queue Error Seen on Some Quadrics MPI Applications

Some QsNet^{II} MPI applications that generate many concurrent DMA operations might encounter the following error:

```
ELAN TRAP -0- Unknown - Queue Error
```

This error terminates the program, which is believed to be caused by high rates of ELAN PutGet operations.

It is possible to work around this problem by setting the `LIBELAN_PUTGET_THROTTLE` environment variable to a value lower than its default value of 32.

12.3.5 The qsnet Database May Contain Entries to Nonexistent Switch Modules

Depending on your system topology, the `qsnet` diagnostics database may contain entries to nonexistent switches.

This issue is manifested as errors reported by the `/usr/bin/qsctrl` utility similar to the following:

```
# qsctrl
qsctrl: failed to initialise module QR0N03: no such module (-7)
...
```

In the previous example, the `switch_modules` table in the `qsnet` database is populated with `QR0N03` even though the `QR0N03` module is not physically present. This problem has been reported to Quadrics, Ltd.

To work around this problem, delete the `QR0N03` entry (and any other nonexistent switch entries) from the `switch_modules` table, and restart the `swmlogger` service:

```
# mysql -u root -p qsnet
mysql> delete from switch_modules where name="QR0N03";
mysql>quit
# service swm restart
```

In addition to the previous problem, the IP address of a switch module may be incorrectly populated in the `switch_modules` table, and you might see the following message:

```
# qsctrl
qsctrl: failed to parse module name 172.20.66.2
...
```

Resolve this issue by deleting the IP address from the `switch_modules` table and restarting the `swmlogger` service:

```
# mysql -u root -p qsnet
mysql> delete from switch_modules where name="172.20.66.2";
mysql>quit
# service swm restart
```

Note

You must repeat the previous procedure if you rerun the `cluster_config` utility because the `qsnet` database is recreated during a `cluster_config` operation.

This chapter contains notes that apply to the HP XC System Software Documentation Set.

13.1 *HP XC System Software Administration Guide*

There are two omissions in the procedure in Section 17.2 that describes how to replace a node.

- In step 8, the `cluster_config` utility prompts you to regenerate ssh keys. Do not regenerate the ssh keys; answer `n` (no) to this prompt.

If you answer `y` to regenerate ssh keys, XC commands such as `stopsys` will not work and you must reimage all nodes by using the `setnode --resync --all;power --cycle` command before continuing.

- After completing step 9 to turn on power to the replaced node, run the SLURM post-configuration utility:

```
# /opt/hptc/sbin/spconfig
```

This information will be included in the next version of the *HP XC System Software Administration Guide*.

A

A6824A host bus adapter, 11-1

C

clear_counters command, 12-1
CP3000 system, 9-1
CP4000 system, 10-1
 Supermon sensor information, 10-1
CP6000 system, 11-1
 installing to SAN device, 11-1
 SIGUSR2 signal, 12-2

D

dgemm utility, 6-1
documentation
 how-to documents, 2-1
 Web site, 2-1
documentation nodes, 13-1

G

GNOME desktop hang, 5-1
gnome-settings-daemon, 5-1

H

hardware preparation, 3-1
 HP Integrity rx2620, 3-4
 HP Integrity rx4640, 3-1
head node
 hang when logging in, 5-1
 running LSF, 8-1
how-to documents, 2-1
HP MPI, 7-2
 Intel trace collector, 7-2
HP ProLiant DL140 G2, 9-1
HP ProLiant DL145 G2, 3-1, 10-1
hyperthreading, 3-1

I

InfiniBand interconnect, 12-1
installation
 on SAN device, 4-1, 11-1
Intel compiler
 for MLIB, 7-1
Intel trace analyzer, 7-2

interconnect, 12-1
IP addresses
 management processors, 3-3t

J

job management, 8-1

L

log file
 rotating, 6-1
login hang, 5-1
LSF, 8-1
 HP MPI, 7-4
 on head node, 8-1

M

MLIB math library, 7-1
 modulefile, 7-1
modulefile
 MLIB, 7-1
MP
 accessing, 3-2
Myrinet interconnect, 12-1

N

Nagios
 update status, 5-1
new features, 1-1
NFS mount options, 6-2
no hard disks found, 4-1

O

OVP failure, 12-2

P

password
 MP, 3-3

Q

qsnet diagnostics database, 12-3
QsNet^{II}queue error, 12-3
Quadrics OVP test failure, 12-2

Quadrics QsNet interconnect, 12-2

R

reinstalling system, 4-1
remote console login, 9-1, 10-1
replace node procedure, 13-1

S

SAN device, 4-1
 booting from, 11-1
 installing to, 11-1
 unzoned volume, 11-1
SATA disks, 4-1
signal
 Quadrics QsNet, 12-2
SLURM, 8-1
smart array, 3-1
Supermon service
 sensor information, 10-1

U

unzoned SAN volume, 11-1
URL
 XC documentation, 2-1

V

Vampir, 7-2

W

Web site
 XC documentation, 2-1

X

XC3000 system, 9-1
XC4000 system, 10-1
 defined, 2-1
XC6000 system, 11-1
 defined, 2-1