HP AlphaServer SC

Release Notes

Order Number: AV-RLAYE-TE

November 2003

This document provides updated information on restrictions to the software and documentation for the AlphaServer SC system from the Hewlett-Packard Company ("HP"). This document also provides information on new and changed features in this release.

Revision/Update Information	This version supersedes the *HP AlphaServer SC Release Notes* issued in September 2002 for HP AlphaServer SC Version 2.5 and those issued in March 2003 for HP AlphaServer SC Version 2.5 (Update Kit 1).
Operating System and Version:	HP Tru64 UNIX Version 5.1B, Patch Kit 3
Software Version:	Version 2.6
Maximum Node Count:	1024 nodes
Node Type:	HP AlphaServer ES45 HP AlphaServer ES40 HP AlphaServer DS20L

Legal Notices

The information in this document is subject to change without notice.

Hewlett-Packard makes no warranty of any kind with regard to this manual, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. Hewlett-Packard shall not be held liable for errors contained herein or direct, indirect, special, incidental or consequential damages in connection with the furnishing, performance, or use of this material.

Warranty

A copy of the specific warranty terms applicable to your Hewlett-Packard product and replacement parts can be obtained from your local Sales and Service Office.

Restricted Rights Legend

Use, duplication or disclosure by the U.S. Government is subject to restrictions as set forth in subparagraph (c) (1) (ii) of the Rights in Technical Data and Computer Software clause at DFARS 252.227-7013 for DOD agencies, and subparagraphs (c) (1) and (c) (2) of the Commercial Computer Software Restricted Rights clause at FAR 52.227-19 for other agencies.

HEWLETT-PACKARD COMPANY
3000 Hanover Street
Palo Alto, California 94304 U.S.A.

Use of this manual and media is restricted to this product only. Additional copies of the programs may be made for security and back-up purposes only. Resale of the programs, in their present form or with alterations, is expressly prohibited.

Copyright Notices

© Copyright 2002, 2003 Hewlett-Packard Development Company, L.P.

The Load Sharing Facility (LSF) software product is developed by Platform Computing Corporation ("Platform"). LSF has been licensed by Platform to HP for inclusion in HP AlphaServer SC systems.

The HP MPI software that is included in this HP AlphaServer SC software release is based on the MPICH V1.2.4 implementation of MPI, which includes the following copyright notice:

© 1993 University of Chicago
© 1993 Mississippi State University

Permission is hereby granted to use, reproduce, prepare derivative works, and to redistribute to others. This software was authored by:

Argonne National Laboratory Group
W. Gropp: (630) 252-4318; FAX: (630) 252-7852; e-mail: gropp@mcs.anl.gov
E. Lusk: (630) 252-5986; FAX: (630) 252-7852; e-mail: lusk@mcs.anl.gov
Mathematics and Computer Science Division, Argonne National Laboratory, Argonne IL 60439

Mississippi State Group
N. Doss and A. Skjellum: (601) 325-8435; FAX: (601) 325-8997; e-mail: tony@erc.msstate.edu
Mississippi State University, Computer Science Department & NSF Engineering Research Center for Computational Field Simulation, P.O. Box 6176, Mississippi State MS 39762

GOVERNMENT LICENSE

Portions of this material resulted from work developed under a U.S. Government Contract and are subject to the following license: the Government is granted for itself and others acting on its behalf a paid-up, nonexclusive, irrevocable worldwide license in this computer software to reproduce, prepare derivative works, and perform publicly and display publicly.

DISCLAIMER

This computer code material was prepared, in part, as an account of work sponsored by an agency of the United States Government. Neither the United States, nor the University of Chicago, nor Mississippi State University, nor any of their employees, makes any warranty express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

Trademark Notices

Microsoft®, Windows®, and Windows NT® are trademarks of Microsoft Corporation in the U.S. and /or other countries.

UNIX® is a registered trademark of The Open Group in the U.S. and /or other countries.

The following product names refer to specific versions of products developed by Quadrics. These products combined with technologies from HP form an integral part of the supercomputing systems produced by HP and Quadrics. These products have been licensed by Quadrics to HP for inclusion in HP AlphaServer SC systems.

- *Interconnect hardware developed by Quadrics, including switches and adapter cards*
- *Elan, which describes the PCI host adapter for use with the interconnect technology developed by Quadrics*
- *PFS or Parallel File System*
- *RMS or Resource Management System*

Contents

1 New and Changed Features

1.1 CMF: Name of Console Log Directory Has Been Changed . . . 1–2
1.2 Controller Firmware Changes. . . . 1–3
1.3 DCE/DFS Are Not Supported . . . 1–3
1.4 Diagnostics: Node Diagnostics Using System Event Analyzer and Consolidated Logging . . 1–4
1.5 Documentation Changes . . . 1–5
1.5.1 Terms Used in Documentation . . . 1–5
1.5.2 Changes to Manual Names . . . 1–6
1.6 Elan Kernel Comms: Reduced Impact of Heartbeat Routines . . . 1–6
1.7 File System Management Changes. . . . 1–6
1.7.1 New and Changed File System Management Features Delivered in Version 2.6 . . . 1–6
1.7.2 New and Changed File System Management Features Delivered in Version 2.5 (UK1) 1–9
1.8 LSF Changes . . . 1–10
1.9 Memory Page Size: Big Page Support. . . . 1–12
1.10 MPI Changes . . . 1–13
1.11 MSA1000: Support for Modular SAN Array 1000 . . . 1–15
1.12 Network Cards . . . 1–15
1.13 Operating System . . . 1–15
1.14 RMS Changes . . . 1–15
1.14.1 New and Changed RMS Features Delivered in Version 2.6 . . . 1–15
1.14.2 New and Changed RMS Features Delivered in Version 2.5 (UK1). . . . 1–24
1.15 Routing: Default Route Configuration . . . 1–25
1.16 SC Database: Daemon Upgraded . . . 1–25
1.17 SC Database: Database Integrity Check Utility. . . . 1–25
1.18 SC Viewer: New Overview Tab . . . 1–26
1.19 SCRUN: New Options . . . 1–26
1.20 SCRUN: umask Propagation . . . 1–26
1.21 Shmem Changes . . . 1–27
1.22 SRA: New Commands . . . 1–27
1.23 Switch Management Changes. . . . 1–27
1.23.1 New and Changed Switch Management Features Delivered in Version 2.6 . . . 1–28
1.23.2 New and Changed Switch Management Features Delivered in Version 2.5 (UK1) . . . 1–29
1.24 Switches: Support for ProCurve Switch . . . 1–29
1.25 System Check Tool: New sclogs Plug-In Component . . . 1–29
1.26 Upgrade: Check to Ensure that EVM is Running Correctly . . . 1–29
1.27 Upgrade: External RIS Server No Longer Required . . . 1–30
1.28 Upgrade: RIS Server Does Not Need to be Specified . . . 1–30

1.29 Upgrade: Upgrade Check Utility 1–30
1.30 Upgrade: Upgrade Restart Facility 1–31

2 Problem Solutions Delivered

2.1 Problem Solutions Delivered Since HP AlphaServer SC Version 2.5 (UK1) 2–1
2.1.1 Adapter Drivers 2–1
2.1.2 CFS 2–2
2.1.3 Cluster Alias 2–2
2.1.4 ICS/Elan 2–2
2.1.5 LSM 2–3
2.1.6 MPI 2–3
2.1.7 msqld 2–3
2.1.8 NFS 2–3
2.1.9 Operating System 2–4
2.1.10 PFS 2–4
2.1.11 RMS 2–4
2.1.12 SCFS 2–5
2.1.13 SRA 2–5
2.1.14 Virtual Memory 2–5
2.2 Problem Solutions Delivered In HP AlphaServer SC Version 2.5 (UK1) 2–6
2.2.1 Boot 2–6
2.2.2 CAA 2–6
2.2.3 CFS 2–6
2.2.4 Cluster Alias 2–7
2.2.5 elantest 2–7
2.2.6 ICS/Elan 2–7
2.2.7 Install/Upgrade 2–8
2.2.7.1 Installation 2–8
2.2.7.2 Upgrade 2–8
2.2.8 IO/DRD 2–9
2.2.9 KCH 2–9
2.2.10 LSF 2–10
2.2.11 msqld 2–10
2.2.12 PFS 2–11
2.2.13 RMS 2–11
2.2.14 SC Monitor 2–14
2.2.15 SC Viewer 2–14
2.2.16 SCFS 2–15
2.2.17 scfsmgr 2–15
2.2.18 scload 2–15
2.2.19 scmountd 2–16
2.2.20 scrun 2–16
2.2.21 Shmem/libelan.so 2–16

2.2.22 SRA . . . 2–17
2.2.23 sramon . . . 2–17
2.2.24 Switch Module Firmware — swm-upgrade-02111408 . . . 2–17

3 General Restrictions

3.1 Booting and Shutting Down Nodes — Reminder . . . 3–1
3.2 CAA Daemon Location . . . 3–1
3.3 CAA Placement Policy for srad . . . 3–2
3.4 cfengine Command . . . 3–2
3.5 Disabled Daemons . . . 3–2
3.5.1 AdvFS GUI . . . 3–3
3.5.2 clu_mibs Daemon . . . 3–3
3.5.3 Insight Manager Daemon . . . 3–3
3.5.4 Printing Daemon . . . 3–3
3.5.5 SNMP and Environmental Information Monitoring Daemons . . . 3–3
3.5.6 SysMan Station Daemon . . . 3–3
3.5.7 X Windows Login Daemon . . . 3–3
3.6 Do Not Directly Mount or Unmount SCFS and PFS File Systems . . . 3–4
3.7 Do Not Rebuild the Kernel Before Creating the Domains . . . 3–4
3.8 Do Not Include Blank Lines in the /etc/routes File . . . 3–4
3.9 DRD: Do Not Increase drd_remote_maxphys . . . 3–4
3.10 GPL Code . . . 3–4
3.11 IP Over Elan Not Supported . . . 3–4
3.12 Lazy Swap is Not Supported . . . 3–4
3.13 Local Area Transport Protocol Not Loaded . . . 3–4
3.14 LSF: All LSF Hosts Must be Upgraded to HP AlphaServer SC Version 2.6 . . . 3–5
3.15 LSF: Chunk Job Dispatch Feature Not Supported . . . 3–5
3.16 LSF: LSF Administrator Must Be Known User on Nodes Running LSF Daemons . . . 3–5
3.17 LSF: Preemption Preferences . . . 3–5
3.18 LSF: Restricted Usage . . . 3–5
3.19 LSF: "span[ptile=*n*]" Resource String Not Supported . . . 3–6
3.20 Rack Names Must be Populated and Unique . . . 3–6
3.21 Restriction on Using addvol . . . 3–6
3.22 Readline and History Support Removed from rmsquery . . . 3–6
3.23 RMS: Support for prun -r and prun -dr Commands . . . 3–6
3.24 Routing: routed is Not Supported . . . 3–7
3.25 Upgrade: Do Not Leave the System Partially Upgraded . . . 3–7
3.26 Upgrade: Upgrade Shortcut Not Supported . . . 3–7
3.27 Use /tmp — Not /usr/tmp . . . 3–7
3.28 64-Bit Shmem is Not Supported . . . 3–8

4 Known Problems in HP AlphaServer SC Version 2.6

4.1 Booting: “sed: Function s/-netmask-/255.255.0.0 cannot be parsed” Error 4–1
4.2 Booting: “The boottime cluster check found a potential problem” Error 4–1
4.3 CMF: Daemon May Need to be Manually Restarted 4–1
4.4 Elan Kernel Comms: Reference Pages Not Updated 4–2
4.5 ICS: Error Message: icssvr_daemon_from_pool: invalid IPL 4 4–2
4.6 Installation/Upgrade: Benign Errors 4–2
4.7 Installation/Upgrade: clu_quorum Check 4–2
4.8 Installation/Upgrade: Problem Installing Patch Kit 3 4–2
4.9 Installation/Upgrade: TCP/IP Ephemeral Range 4–3
4.10 Install: Cannot Install Certain Layered Products 4–3
4.11 Install: ifaccess.config File Set Up Only for Head Node in Domain 4–3
4.12 Install: Possible Errors Seen with Tru64 UNIX Patch 1276.00 4–3
4.13 I/O May Hang on HSG80 in Multibus Mode 4–4
4.14 Kernel: Nodes May Crash with “ics_elan: timeout on free tmds” String 4–4
4.15 Kernel: Nodes May Crash with “malloc_internal: invalid link address may occur” String . . 4–4
4.16 LSF: bjobs -l Times Out 4–5
4.17 LSF: Delay Between Resource Allocation and Job Start-Up 4–5
4.18 LSF: Imprecise Pending Reason When the Node Required by a Job is Configured Out . . . 4–5
4.19 LSF: Reservation and User Limits 4–6
4.20 LSF: Reservation for Jobs with [hosts=1] 4–6
4.21 LSF: Slow Update of MXJ=! 4–6
4.22 MPI: Invalid DMA May Be Seen With Onset of Paging 4–6
4.23 MPI: Large Number of MPI Communicators/Not Enough Global Memory 4–6
4.24 MPI: LIBELAN_WAITTYPE Variable Must Be Set to “POLL” 4–7
4.25 MPI: Sending Large Messages Without Buffering 4–7
4.26 Networks: Configuration of Default Gateway on Peripheral Hardware 4–8
4.27 Networks: Quiet Management Networks can be Misinterpreted by niffd 4–8
4.28 NFS: “Stale File Handle” or “Cannot Overwrite Existing File” Errors 4–9
4.29 Performance Visualizer: pvis Command Causes Stack Trace 4–9
4.30 Performance Visualizer: pvis Command Not Removed 4–10
4.31 PFS: Can Create, and Write to, Files on a Full PFS File System 4–10
4.32 PFS: Changing the Default File Mapping May Not Take Effect Globally 4–10
4.33 PFS: Do Not Delete an Active PFS Mount Point 4–10
4.34 PFS: Do Not Mount a PFS on a PFS Directory 4–10
4.35 PFS: File Locking is Not Supported 4–11
4.36 PFS: Link Operations May Fail on a PFS File System 4–11
4.37 PFS: PFS File System Cannot be Deleted if SCFS is Broken 4–11
4.38 PFS: pfsmgr show Displays “new” State 4–12
4.39 PFS: ro (read-only) Access is Not Supported 4–12
4.40 PFS/SCFS: Mounts and Unmounts Rely on scrun Command 4–12
4.41 PFS/SCFS: Mount Status Not Updated After Shutdown 4–13
4.42 RMS: Detection and Reporting of PSU or Fan Failures on HP AlphaServer ES45 Nodes . . 4–13
4.43 RMS: Event Manager Exits Due to Low Memory Availability 4–13

4.44 RMS: “failed to allocate elan capability” Error 4–14
4.45 RMS: prun and Non-Global File Systems 4–14
4.46 RMS: prun Fails with “failed to get capability” Error 4–15
4.47 RMS: prun Fails with “failed to create capability: Invalid argument” Error 4–15
4.48 RMS: rmsd Memory Size 4–16
4.49 scevent: <EvmITEM_CLUSTER_NAME> in NFS Events 4–16
4.50 SCFS: File System Status May Be Set To “Unknown” 4–16
4.51 SCFS: File System May Not Mount If File Server in Single-User Mode 4–17
4.52 SCFS: Fortran Formatted I/O May Perform Poorly for FAST File Access 4–17
4.53 SCFS: May Require Reboot to Release Kernel Resources for Reuse 4–18
4.54 SCFS: scfsmgr export or scfsmgr destroy Print Error Message Twice 4–18
4.55 SCFS: SysMan Menu Refresh Button Behaves Incorrectly 4–18
4.56 SC Monitor: Cannot Monitor HSV110 Cells with Multiword Names 4–18
4.57 SC Monitor: HSG Must Have CCL 4–18
4.58 SC Monitor: HSV RAID Systems May Not be Monitored 4–19
4.59 SC Monitor: May Not Mark Objects As Being Stale 4–19
4.60 SC Viewer: “Can't find class p3_desktop.Desktop” Error 4–19
4.61 SC Viewer: Elan Status May Be Incorrect 4–19
4.62 SC Viewer: “/usr/bin/scviewer: cd: too many arguments” Error 4–20
4.63 Shutting Down: Do Not Use “shutdown -sh now” 4–20
4.64 SRA: Installing on a non-RAID Cluster Disk and Generic Boot Disk 4–20
4.65 SRA: No Sanity Check on the sra shutdown Command 4–21
4.66 SRA: Output Uses the Term “Cluster” 4–21
4.67 SRA: Rerunning the sra setup Command 4–21
4.68 SRA: sra cookie Command Is Only Accurate on rmshost 4–22
4.69 SRA/Sysman: Updating the /etc/hosts File 4–22
4.70 Sysman: Display Error When Configuring NIC On Other Node 4–22
4.71 TotalView: Defining the Path 4–22
4.72 TotalView: Zombie Resources 4–22
4.73 TotalView: TotalView Client May Not Communicate with TotalView Server 4–23
4.74 Upgrade: Benign Error 4–23
4.75 Upgrade: “could not locate Tru64 UNIX in RIS database” Error 4–23
4.76 Upgrade: Member-Specific Crontab Files Not Saved 4–24
4.77 Upgrade: “Unexpected exit while: Building boot disk” Error 4–24
4.78 Upgrade: Upgrade Fails if HP AlphaServer SC-Specific UNIX Accounting is in Place 4–24
4.79 Zombie Resources May Not Be Removed 4–25

About These Release Notes

This document describes new and changed features and problems solved in Version 2.6 of the system software and documentation for an AlphaServer SC system from the Hewlett-Packard Company ("HP"). It also provides information on known restrictions and problems in this release.

Audience

This document is intended for use by those who install, maintain, and use HP AlphaServer SC systems.

Structure of This Document

These Release Notes provide the following information for HP AlphaServer SC Version 2.6:

- Chapter 1: New and Changed Features
- Chapter 2: Problem Solutions Delivered
- Chapter 3: General Restrictions
- Chapter 4: Known Problems in HP AlphaServer SC Version 2.6

Related Documentation

You should have a hard copy or soft copy of the following documents:

- *HP AlphaServer SC Installation Guide*
- *HP AlphaServer SC System Administration Guide*
- *HP AlphaServer SC Best Practices I/O Guide*
- *HP AlphaServer SC Interconnect Installation and Diagnostics Manual*
- *HP AlphaServer SC RMS Reference Manual*

- *HP AlphaServer SC User Guide*
- *Administering HP AlphaServer SC Platform LSF®*
- *HP AlphaServer SC Platform LSF® Reference*
- *Running Jobs with HP AlphaServer SC Platform LSF®*
- *HP AlphaServer SC Platform LSF® Quick Reference*
- HP *AlphaServer ES45 Owner's Guide*
- HP *AlphaServer ES40 Owner's Guide*
- HP *AlphaServer DS20L User's Guide*
- HP *StorageWorks HSG80 Array Controller CLI Reference Guide*
- HP *StorageWorks HSG80 Array Controller Configuration Guide*
- HP *StorageWorks Fibre Channel Storage Switch User's Guide*
- HP *StorageWorks Enterprise Virtual Array HSV Controller User Guide*
- HP *StorageWorks Enterprise Virtual Array Initial Setup User Guide*
- HP *StorageWorks Modular SAN Array 1000 User Guide*
- HP *SANworks Tru64 UNIX Kit for Enterprise Virtual Array Release Notes*
- HP *SANworks Tru64 UNIX Kit for Enterprise Virtual Array Installation and Configuration Guide*
- HP *SANworks Scripting Utility for Enterprise Virtual Array Reference Guide*
- HP *TruCluster Server Cluster Release Notes*
- HP *TruCluster Server Cluster Technical Overview*
- HP *TruCluster Server Cluster Hardware Configuration*
- HP *TruCluster Server Cluster Administration*
- HP *TruCluster Server Cluster Highly Available Applications*
- HP *Tru64 UNIX Release Notes*
- HP *Tru64 UNIX Installation Guide*
- HP *Tru64 UNIX Network Administration: Connections*
- HP *Tru64 UNIX Network Administration: Services*
- HP *Tru64 UNIX System Administration*

- HP *Tru64 UNIX System Configuration and Tuning*
- HP *Tru64 UNIX AdvFS Administration*
- HP *Tru64 UNIX Logical Storage Manager*
- HP *Tru64 UNIX Advanced Printing Software User Guide*
- HP *Management and Configuration Guide for the HP ProCurve Series 4100GL Switches, Series 2600 Switches, and Switch 6108*
- *Summit Switch Hardware Installation Guide* from Extreme Networks, Inc.
- *ExtremeWare Software User Guide* from Extreme Networks, Inc.

Terms Used in These Notes

Clusters of nodes within an HP AlphaServer SC system are called domains (except in the case of clustered management servers). However, as domain management in HP AlphaServer SC systems is based on cluster management in TruCluster Server, many of the terms used in domain management include the word “cluster” — for example, cluster alias, cluster quorum, cluster application availability (CAA), cluster root, cluster disk.

The term “LSF cluster” is used to describe a group of machines controlled by LSF. An LSF cluster can consist of one or more HP AlphaServer SC systems, or part of an HP AlphaServer SC system; it can also include additional machines of arbitrary architectures.

TruCluster Server Documentation

The HP TruCluster Server documentation set provides a wealth of information about clusters and cluster management; however, there are a number of differences between the management of domains in an HP AlphaServer SC system, and the management of TruCluster Server clusters. These differences are described in the *HP AlphaServer SC System Administration Guide*.

You should use the HP TruCluster Server documentation set to supplement the HP AlphaServer SC documentation set — if there is a conflict of information, use the instructions provided in the HP AlphaServer SC documentation set.

Abbreviations

Table 0–1 lists the abbreviations that are used in this document.

Table 0–1 Abbreviations

Abbreviation	Description
AdvFS	Advanced File System
API	Application Programming Interface
ATM	Asynchronous Transfer Mode
CCL	Command Console LUN
CD-ROM	Compact Disc — Read-Only Memory
CDSL	Context-Dependent Symbolic Link
CFS	Cluster File System
CLI	Command-Line Interface
CMF	Console Management Facility
CPU	Central Processing Unit
CS	Compute-Serving
DCE	Distributed Computing Environment
DFS	Distributed File Service
DMA	Direct Memory Access
DNS	Domain Name System
EVM	Event Manager
FDDI	Fiber-optic Digital Data Interface
FS	File-Serving
GUI	Graphical User Interface
HBA	Host Bus Adapter
HiPPI	High-Performance Parallel Interface
IDCD	InterDomain Connectivity Daemon
IP	Internet Protocol

Table 0–1 Abbreviations

Abbreviation	Description
KMF	Kernel Memory Fault
LIM	Load Information Manager
LSF	Load Sharing Facility
LSM	Logical Storage Manager
LUN	Logical Unit Number
MBD	Master Batch Daemon
MIB	Management Information Base
MPI	Message Passing Interface
NFS	Network File System
NIS	Network Information Service
PFS	Parallel File System
RAID	Redundant Array of Independent Disks
RIS	Remote Installation Services
RMS	Resource Management System
SC	SuperComputer
SCFS	HP AlphaServer SC File System
SCSI	Small Computer System Interface
SMP	Symmetric Multiprocessing
SNMP	Simple Network Management Protocol
SQL	Structured Query Language
SRM	System Resources Manager
SROM	Serial Read-Only Memory
TCL	Tool Command Language
VCS	Virtual Controller Software
WEBES	Web-Based Enterprise Service

Documentation Conventions

Table 0–2 lists the documentation conventions that are used in this document.

Table 0–2 Documentation Conventions

Convention	Description
%	A percent sign represents the C shell system prompt.
$	A dollar sign represents the system prompt for the Bourne and Korn shells.
#	A number sign represents the superuser prompt.
P00>>>	A P00>>> sign represents the System Resources Manager (SRM) console prompt.
`Monospace type`	Monospace type indicates file names, commands, system output, and user input.
Boldface type	Boldface type in interactive examples indicates typed user input.
Italic type	Italic (slanted) type indicates emphasis, variable values, placeholders, function argument names, menu options, and complete titles of documents.
Underlined type	Underlined type emphasizes important information.
[] {|}	In syntax definitions, brackets indicate items that are optional and braces indicate items that are required. Vertical bars separating items inside brackets or braces indicate that you choose one item from among those listed.
...	In syntax definitions, a horizontal ellipsis indicates that the preceding item can be repeated one or more times.
⋮	A vertical ellipsis indicates that a portion of an example that would normally be present is not shown.
cat(1)	A cross-reference to a reference page includes the appropriate section number in parentheses. For example, `cat(1)` indicates that you can find information on the `cat` command in Section 1 of the reference pages.
Ctrl/x	This symbol indicates that you hold down the first named key while pressing the key or mouse button that follows the slash.
Note	A note contains information that is of special importance to the reader.
`atlas`	Throughout this document, `atlas` is an example system name.

Supported Network Adapters

Table 0–3 lists the associated device names for each supported network adapter. The examples in this guide refer to the DE602 network adapter.

Table 0–3 Network Adapters and Device Names

Network Adapter	SRM Device Name	UNIX Device Name
DE60*x*	`eia0`	`ee0`
DE50*x*	`ewa0`	`tu0`
Gigabit Ethernet	SRM cannot use this device	alt0
Gigabit Ethernet	SRM cannot use this device	bcm0
HiPPI[1, 2]	SRM cannot use this device	hip0
ATM[2]	SRM cannot use this device	lis0
FDDI	SRM cannot use this device	fta0

[1]High-Performance Parallel Interface (HiPPI) is only available if you install an additional HiPPI subset — for HP Tru64 UNIX Version 5.1B, the minimum supported version is HiPPI kit 222.

[2]The `sra install` command does not configure HiPPI and Asynchronous Transfer Mode (ATM) interfaces — you must configure such interfaces manually.

Supported Node Types

HP AlphaServer SC Version 2.6 supports the following node types:

- HP AlphaServer ES45
- HP AlphaServer ES40
- HP AlphaServer DS20L

Comments on this Document

HP welcomes any comments and suggestions that you have on this document. Please send all comments and suggestions to your HP Customer Support representative.

1

New and Changed Features

This chapter provides information on new and changed features in HP AlphaServer SC Version 2.6 since Version 2.5.

Note:

While all of the new and changed features listed here are new in HP AlphaServer SC Version 2.6 since Version 2.5, some features were first included in Version 2.5 Update Kit 1 (UK1).

All functionality that is listed as being provided in HP AlphaServer SC Version 2.5 (UK1) is also provided in Version 2.6.

This chapter is organized as follows:

- CMF: Name of Console Log Directory Has Been Changed (see Section 1.1 on page 1–2)
- Controller Firmware Changes (see Section 1.2 on page 1–3)
- DCE/DFS Are Not Supported (see Section 1.3 on page 1–3)
- Diagnostics: Node Diagnostics Using System Event Analyzer and Consolidated Logging (see Section 1.4 on page 1–4)
- Documentation Changes (see Section 1.5 on page 1–5)
- Elan Kernel Comms: Reduced Impact of Heartbeat Routines (see Section 1.6 on page 1–6)
- File System Management Changes (see Section 1.7 on page 1–6)
- LSF Changes (see Section 1.8 on page 1–10)
- Memory Page Size: Big Page Support (see Section 1.9 on page 1–12)
- MPI Changes (see Section 1.10 on page 1–13)

- MSA1000: Support for Modular SAN Array 1000 (see Section 1.11 on page 1–15)
- Network Cards (see Section 1.12 on page 1–15)
- Operating System (see Section 1.13 on page 1–15)
- RMS Changes (see Section 1.14 on page 1–15)
- Routing: Default Route Configuration (see Section 1.15 on page 1–25)
- SC Database: Daemon Upgraded (see Section 1.16 on page 1–25)
- SC Database: Database Integrity Check Utility (see Section 1.17 on page 1–25)
- SC Viewer: New Overview Tab (see Section 1.18 on page 1–26)
- SCRUN: New Options (see Section 1.19 on page 1–26)
- SCRUN: umask Propagation (see Section 1.20 on page 1–26)
- Shmem Changes (see Section 1.21 on page 1–27)
- SRA: New Commands (see Section 1.22 on page 1–27)
- Switch Management Changes (see Section 1.23 on page 1–27)
- Switches: Support for ProCurve Switch (see Section 1.24 on page 1–29)
- System Check Tool: New sclogs Plug-In Component (see Section 1.25 on page 1–29)
- Upgrade: Check to Ensure that EVM is Running Correctly (see Section 1.26 on page 1–29)
- Upgrade: External RIS Server No Longer Required (see Section 1.27 on page 1–30)
- Upgrade: RIS Server Does Not Need to be Specified (see Section 1.28 on page 1–30)
- Upgrade: Upgrade Check Utility (see Section 1.29 on page 1–30)
- Upgrade: Upgrade Restart Facility (see Section 1.30 on page 1–31)

1.1 CMF: Name of Console Log Directory Has Been Changed

The name of the directory where the CMF console logs are placed has been changed to allow logs to be retained for more than one year. The directory name now includes the year.

For example:
`/var/sra/cmf.dated/19-May-2003-13:31`

The `current` link in the directory above, which points to the current logging directory, has also been updated.

1.2 Controller Firmware Changes

The firmware release has been upgraded to 03031311.

Note:

The switch module firmware version number is a date code that identifies the point at which the release was built. The format is yymmddhh; for example, 02082716 equates to 16:00 on August 27th 2002. The value stored in the `switch_modules.build` field of the RMS database is an integer. To see the date code, convert the integer to hexadecimal.

The following changes have been made to the controller firmware functionality:

- Support has been added for the watchdog script that monitors the `swmserver` daemon. If the file `/.data/watchdog.swmserver` exists, the watchdog script will reboot a HP AlphaServer SC Interconnect control processor if there is no `swmserver` daemon running.
- The watchdog script has been made less invasive.
- Support for specifying an IP gateway has been added to the Setup menu. A gateway is required under the following circumstance:
  - When connecting to an HP AlphaServer SC Interconnect control processor from nodes within a domain, where the source address for such a connection will be the cluster alias. In such a case, the HP AlphaServer SC Interconnect control processor will need to have a return route from the management network to the cluster alias.
  - When the `rmshost` node is on a different Ethernet subnet to the subnet the switch module controllers are on.

For information on how to upgrade the switch module firmware, see Section 7.2 of the *HP AlphaServer SC Interconnect Installation and Diagnostics Manual*.

1.3 DCE/DFS Are Not Supported

Since Version 2.5 (UK1), Entegrity Distributed Computing Environment (DCE) and Entegrity Distributed File Service (DFS) are not supported in the HP AlphaServer SC software.

1.4 Diagnostics: Node Diagnostics Using System Event Analyzer and Consolidated Logging

The following changes have been made to the node diagnostics functionality:

- Consolidated log report

  You can now choose how to use the functionality provided by System Event Analyzer (previously called Compaq Analyze). You can use System Event Analyzer on one node (typically the management server) in your HP AlphaServer SC system to produce a single consolidated report for all (or optionally some) nodes in the system. The binary error log events from the nodes are copied to a single consolidated error log on one (host) node, and the report is then generated from the consolidated log by the local copy of System Event Analyzer. Thus System Event Analyzer can provide a single consolidated report for a number of nodes across the system.

  You can also use the `sra diag` command and System Event Analyzer on each node to generate reports for the individual nodes, as in earlier versions of HP AlphaServer SC software.

Note:

WEBES Version 4.2 is supported in HP AlphaServer SC Version 2.6, and this is the WEBES version documented in the updated version of the *HP AlphaServer SC System Administration Guide*. However, WEBES Version 4.2 does not support the ability to automatically redirect the default log from the local log (`binary.errlog`) to the consolidated log (`sc.binary.errlog`). Instead, System Event Analyzer can be run manually to produce the report from the consolidated log. The consolidated log, `sc.binary.errlog`, can be specified directly on the System Event Analyzer command line, and the WEBES 4.2 web user interface also allows additional log files to be specified directly in the interface.

WEBES Version 4.1.1 is also supported in HP AlphaServer SC Version 2.6, and this version of WEBES does support the redirection of the default log. The procedure for installing WEBES Version 4.1.1 is similar to that for installing WEBES Version 4.2. If necessary, you can refer to the previous version of the *HP AlphaServer SC System Administration Guide* (Version 2.5) for full instructions for installing WEBES Version 4.1.1.

- Manage the size of error logs

  A new utility, `binex+`, has been provided to allow you to manage the size of the binary error log files. The `binex+` utility is based on the Tru64 UNIX `binex` utility, and operates on the individual log files produced by the binary error log daemons `binlogd`, and the consolidated log produced by the remote binary error log daemon `rbinlogd`. It can be used to summarize the contents of an error log file, to copy a selected range of events to another file for remote analysis, to clean up corrupted areas of the file, or to split the error log file and archive older records.

For more information about node diagnostics, see Chapter 28 of the *HP AlphaServer SC System Administration Guide*.

1.5 Documentation Changes

The following documentation has been updated:

- *HP AlphaServer SC Installation Guide*
- *HP AlphaServer SC System Administration Guide*
- *HP AlphaServer SC Release Notes*
- *HP AlphaServer SC Interconnect Installation and Diagnostics Manual*
- *HP AlphaServer SC RMS Reference Manual*
- *HP AlphaServer SC User Guide*
- *HP AlphaServer SC Best Practices I/O Guide*
- *HP AlphaServer SC Platform LSF® Reference*
- *Administering HP AlphaServer SC Platform LSF®*
- *Running Jobs with HP AlphaServer SC Platform LSF®*
- *HP AlphaServer SC Platform LSF® Quick Reference*

1.5.1 Terms Used in Documentation

In documentation for HP AlphaServer SC Version 2.5, a cluster of nodes in an HP AlphaServer SC system was called a “CFS domain”. In documentation for HP AlphaServer SC Version 2.6, the “CFS” has been dropped, and a cluster of nodes is called a “domain”.

The HP AlphaServer SC documentation set has been updated to use the terms “domain” and “cluster” appropriately and consistently. However, as domain management in an HP AlphaServer SC system is based on cluster management in TruCluster Server, many of the terms used in domain management still include the word “cluster” — for example, cluster alias, cluster quorum, cluster application availability (CAA), cluster root, cluster disk.

The term “cluster” is also used in the case of clustered management servers, and in reference to LSF clusters. An LSF cluster can consist of one or more HP AlphaServer SC systems, or part of an HP AlphaServer SC system; it can also include additional machines of arbitrary architectures.

1.5.2 Changes to Manual Names

The names of the following manuals have been changed:

- The *HP AlphaServer SC Platform LSF® Administrator's Guide* is now called *Administering HP AlphaServer SC Platform LSF®*.
- The *HP Alphaserver SC Platform LSF® Reference Guide* is now called *HP AlphaServer SC Platform LSF® Reference*.
- The *HP AlphaServer SC Platform LSF® User's Guide* is now called the *Running Jobs with HP AlphaServer SC Platform LSF®*.

1.6 Elan Kernel Comms: Reduced Impact of Heartbeat Routines

Elan kernel comms has been optimized to reduce the impact of the kernel comms heartbeat routines on user applications and in particular MPI Barrier operations.

1.7 File System Management Changes

This section describes the new file system management features provided in HP AlphaServer SC Version 2.6 since Version 2.5.

1.7.1 New and Changed File System Management Features Delivered in Version 2.6

The following changes to file system management functionality are new in HP AlphaServer SC Version 2.6:

- Quota support has been added for SCFS file systems. Quota support allows you to limit both the number of files and the total amount of disk space that are associated in SCFS file systems on behalf of a given user or group. For more information, see the `scfsmgr(8)` reference page and Chapter 7 of the *HP AlphaServer SC System Administration Guide*.

  Note:

  HP does not recommend enabling quotas for SCFS file systems that are part of a PFS file system, as the quotas will only be applied to the individual SCFS file systems and not to the entire PFS file system.

If quotas are enabled on an SCFS file system that is part of a PFS file system, messages indicating that the user quota has been exceeded may occur when the quota has been exceeded on the SCFS file system, even when there is space remaining on the PFS file system.

For more information on managing PFS file systems, see Chapter 8 of the *HP AlphaServer SC System Administration Guide*.

- File System Performance — Local File Creation Time

  In HP AlphaServer SC Version 2.6, file creation time is reduced on local SCSI disks when the file systems are mounted as `server_only`.

  File creation on TruCluster Server normally requires synchronous writes to the file system log, in order to enable failover. This results in degraded performance as compared with traditional UNIX file systems, where all file I/O is cached. Local file systems, such as `/tmp`, that are mounted as `server_only` are accessible only by the member that mounts them, and do not benefit from failover in any case.

  As an improvement in HP AlphaServer SC Version 2.6, synchronous writes are not required in the case of file systems mounted as `server_only`. As a result, creation time for small files is improved by a factor of 10 on local SCSI disks when the file systems are mounted as `server_only`. This improvement will be of particular benefit to applications such as compilers that make intense use of small files on `/tmp`. No special actions are required to enable this improvement, as local file systems are automatically mounted as `server_only`. You can check to ensure that file systems are mounted as `server_only` by issuing an `/sbin/mount` command.

- To prepare existing PFS file systems for use with HP AlphaServer SC Version 2.6, the `-o u` option has been provided with the `fsck_pfs` command. Running the `fsck_pfs` command with the `-o u` option also creates two new directories, `.pfsbirth` and `.pfsdeath`. The `.pfsbirth` directory is used to store PFS files while they are being created, and the `.pfsdeath` directory is used to store PFS files while they are being deleted.

  A utility has also been provided to assist in the upgrade of PFS filesystems to HP AlphaServer SC Version 2.6. For more information, see Chapter 4 of the *HP AlphaServer SC Installation Guide*.

  The `fsck_pfs` command with the `-o u` option replaces the `clean_pfs` utility in HP AlphaServer SC Version 2.5 (UK1).

- The O_UBCIO option (defined in <`sys/scfs.h`>) and the CFS subsystem `io_env` attribute have been added to allow for more flexibility in choosing whether a file is opened in FAST or UBC mode. For more information, see Chapter 7 of the *HP AlphaServer SC System Administration Guide*.

- The SCFS kernel subsystem can make use of multiple Elan rails, if they are present on the system, to perform SCFS I/O data transfers. This allows SCFS clients to load-balance their requests over all the rails that they are permitted to use and can see the server nodes on. For more information, see Chapter 7 of the *HP AlphaServer SC System Administration Guide*.
- SCFS file systems exported from an FS domain are now mounted from preferred server cluster aliases. In HP AlphaServer SC Version 2.5, the default cluster alias was used to mount SCFS file systems. In HP AlphaServer SC Version 2.6, each member of an FS domain has a cluster alias associated with it. This new preferred server cluster alias is now used to mount SCFS file systems, thus reducing the number of network message hops inside the FS domain, and providing performance benefits for large sequential read and write operations.

  The new preferred server cluster aliases are in the format:
  `atlasD`*`n`*`-atlas`*`m`*

  where:

  – *n* is the domain number

  – *m* is the member number

  For example, `atlasD0-atlas0`.

  The new preferred server cluster aliases are associated with the management LAN (10.128).

  For more information, see Chapter 19 and Chapter 22 of the *HP AlphaServer SC System Administration Guide*.

  A consequence of the new preferred server cluster alias scheme is that console messages similar to the following are displayed while a node is being booted:
  `arp - local IP address 10.128.106.1 in use by 00-02-A5-60-4D-62`

  These messages indicate that the preferred server cluster alias scheme is operating correctly, and they can safely be ignored if the IP addresses shown are those of the cluster aliases. If the IP addresses shown are not those of the cluster aliases, please contact your HP Customer Service representative for assistance.
- Three new mount-related sub-commands have been added to the `scfsmgr` command, as follows:

  – A new command allows you to synchronize the mount status of all SCFS and PFS file systems on one specified domain. The new `scfsmgr sync_domain` command is similar to the existing `scfsmgr sync` command, but while the `scfsmgr sync` command runs on the entire system, the `scfsmgr sync_domain` command allows

you to limit the operation to a specified domain. The command checks the mount status of all SCFS and PFS file systems in the domain, and either mounts or unmounts the file systems as appropriate.

– The `scfsmgr try_mount` and `scfsmgr try_unmount` commands have been provided for troubleshooting purposes. These commands can be used on FS domains and CS domains to troubleshoot specific client nodes or client domains that are giving problems. The commands trigger the mount or unmount of a file system on a specified node or domain, and information about the operation is displayed while the file system is being mounted or unmounted. The `try_mount` and `try_unmount` commands are intended for troubleshooting purposes only, and should not be used as substitutes for the `online` and `offline` commands.

  The `try_mount` and `try_unmount` sub-commands are also available with the `pfsmgr` command.

For more information, see the `scfsmgr(8)` and `pfsmgr(8)` reference pages and Chapter 7 and Chapter 8 of the *HP AlphaServer SC System Administration Guide*.

- The default value of the `sync_dirty_size` attribute has been changed from 64MB to 256MB.

  The `sync_dirty_size` attribute specifies the maximum amount of data that SCFS allows a file to have before it must be synchronized to disk. The default value of this attribute has been changed to improve performance; measurements have shown that this change increases SCFS throughput by up to 10%.

- The default value of the `smoothsync_age` attribute has been changed from 30 seconds to 0 (zero) seconds on nodes in FS domains.

  The `smoothsync_age` attribute sets the amount of time, in seconds, that modified (dirty) pages are allowed to age before they must be written to disk. The default value of this attribute has been changed on nodes in FS domains in HP AlphaServer SC systems to improve performance; measurements have shown that this change increases SCFS throughput by up to 40%.

1.7.2 New and Changed File System Management Features Delivered in Version 2.5 (UK1)

The following changes to file system management functionality were first delivered in HP AlphaServer SC Version 2.5 (UK1):

- Access Control Lists (ACLs) can be used in SCFS file systems. ACLS allow you to set permissions for specific users and groups on files and directories. For more information, see Chapter 7 of the *HP AlphaServer SC System Administration Guide*.

- Additional mount option arguments can be used with the `scfsmgr create` command. In earlier versions of HP AlphaServer SC software, only the `rw` (read/write) or `ro` (read only) mount option could be specified for SCFS file systems. In HP AlphaServer SC Version 2.5 (UK1), greater flexibility was provided by allowing you to specify a range of AdvFS and NFS mount options for SCFS file systems. For more information, see Chapter 7 of the *HP AlphaServer SC System Administration Guide*.
- The SCFS_SRV_DOMS variable in the `/etc/rc.config.common` file lists the names of all SCFS File Serving domains. This variable is now required.

  In HP AlphaServer SC Version 2.5, although the SCFS_SRV_DOMS variable was used by some elements of the HP AlphaServer SC system software, the SCFS management system did not use it.

  Since HP AlphaServer SC Version 2.5 (UK1), the SCFS management system uses the SCFS_SRV_DOMS variable. Unless the variable is set correctly, the SCFS management system will not operate correctly. Specifically, if a File Server domain is omitted, the mount status on importing domains will not change, even if repeated `sync` commands are issued. For a new file system, the state is `unknown`. For a file system in use prior to an upgrade, the state remains as it was before the upgrade.

  The SCFS_SRV_DOMS variable is a comma-separated list of domain names. For example, if `atlasD0` and `atlasD1` are File Serving domains, the entry in the `/etc/rc.config.common` file on all domains will be as follows:

  ```
  SCFS_SRV_DOMS="atlasD0,atlasD1"
  export SCFS_SRV_DOMS
  ```

  For more information, see Chapter 7 of the *HP AlphaServer SC System Administration Guide*.
- An SCFS client kernel subsystem attribute sets an upper limit for memory usage on all SCFS file systems on a node. The `max_buf_total` attribute specifies the maximum amount of data that a client will allow to be shadow copied in total, before blocking new requests from being issued. By limiting the total amount of memory a client node uses for SCFS file systems, the `max_buf_total` attribute prevents the possibility of the node performance being degraded due to lack of memory. For more information, see Chapter 7 of the *HP AlphaServer SC System Administration Guide*.

1.8 LSF Changes

The following changes have been made to Platform Computing Corporation's LSF® software ("LSF") in HP AlphaServer SC Version 2.6:

- The LSF software provided in HP AlphaServer SC Version 2.6 has been updated to LSF Version 6.0.
- The CAA script takes responsibility for remounting the LSF directory in case of failover.

If the node that NFS-mounts the LSF file system on a domain crashes, the CAA script will automatically remount the LSF file system on the node that takes over the LSF service.

- The preemption, slot reservation and backfill LSF features are now fully supported with all allocation types.
- The syntax for topology scheduler specification and the `bsub -extsched` option has been changed.

  The syntax for specifying topology schedulers at the queue definition level has changed as follows:

  Old syntax:
  parameter=[*allocation_type*][;*topology*][;*flags*]

  New syntax:
  parameter= RMS[[*allocation_type*][;*topology*][;*flags*]]

  where *parameter* is MANDATORY_EXTSCHED or DEFAULT_EXTSCHED.

  The old syntax has been made obsolete and the new syntax must be used for all new features.

  The syntax of the `-extsched` option of the `bsub` command has changed as follows:

  Old syntax:
  -extsched "[*allocation_type*][;*topology*][;*flags*]"

  New syntax:
  -ext "RMS[[*allocation_type*][;*topology*][;*flags*]]"

  or:
  -extsched "RMS[[*allocation_type*][;*topology*][;*flags*]]"

  For topology scheduler options defined in the obsolete syntax, the following conditions apply:

  – Only the RMS topology options are honoured.
  – Options from the queue and from the `bsub` command are parsed separately before they are merged. For example, if the following options are specified in the old syntax when the `bsub` command is run:
    ```
    bsub -extsched "ptile=1" ...
    ```
    and the default topology scheduler for the queue is RMS_SNODE, the `ptile=1` option is ignored because it is not an RMS topology option. The net result is that the RMS_SNODE allocation type is used.

    In HP AlphaServer SC Version 2.5, the same options would have resulted in the following parameters being used:

    ```
    RMS_SNODE;ptile=1
    ```

- RMS spare node functionality is supported in LSF.

  You can define a number or percentage of nodes in a partition as spare nodes. Under normal operating conditions, LSF does not schedule jobs on these nodes and includes a spare node into scheduling only when an operating node become inaccessible.

- RMS drained node functionality is supported in LSF.

  LSF is made aware of nodes that are marked as drained in RMS and will not attempt to schedule jobs on these nodes.

- The output from the `bjobs` and `bhist` commands now shows the RMS resource ID.

- Pre-execution scripts can now be run on RMS resources.

  You can now run pre-execution LSF scripts on a number of nodes of a parallel job. Instructions for running pre-execution scripts are included in Chapter 4 of the *HP AlphaServer SC System Administration Guide* and in the *HP AlphaServer SC User Guide*.

- Support for suspending and resuming jobs in LSF has been added.

- Support for running RMS jobs from non-RMS hosts, and *vice versa* has been added.

  LSF now allows you to submit jobs from an HP AlphaServer SC system to other architectures, and *vice versa*.

For more information on these new features, see Chapter 4 of the *HP AlphaServer SC System Administration Guide*.

1.9 Memory Page Size: Big Page Support

Support for variable memory page size is a new feature in Tru64 UNIX Version 5.1B that is also available in HP AlphaServer SC Version 2.6. By default, this functionality is disabled. You can enable it as follows:

1. Make the following `sysconfigdb` stanza available in a global file system across all domains:

   ```
   vm:
       vm_bigpg_enabled=1
   ```

2. Run the following command from any node:

   ```
   # scrun -n all '/sbin/sysconfigdb -m -f stanza_pointer'
   ```

   where:

   – *`stanza_pointer`* is a pointer to the stanza created in step 1

3. Reboot every node in the system.

For more information about variable memory page size, see section 1.1 of the HP *Tru64 UNIX Release Notes* for Tru64 UNIX Version 5.1B.

For more information about the parameters for variable memory page size, see section 12.7 of the HP *Tru64 UNIX System Configuration and Tuning* manual for Tru64 UNIX Version 5.1B.

1.10 MPI Changes

The following changes have been made in the MPI software used in HP AlphaServer SC Version 2.6:

- The HP MPI software that is included in HP AlphaServer SC Version 2.6 is based on the MPICH V1.2.4 implementation of MPI.
- The `MPI_Cancel` function is no longer supported. According to MPI specifications, `MPI_Cancel` commands are no longer guaranteed to work. This is due to the fact that `eager` sends are now being used for improved performance. This implies that small send operations are likely to have already completed before the user can call the `MPI_Cancel` function.

  For longer messages, the sender will generally have sent some fraction of the data to the destination node, where it is either being buffered or copied directly into a user `recv` buffer. As a result, cancelling sends would also imply cancelling the remote receive and/or cleaning up internal buffers, and this becomes extremely difficult to implement.
- A "compile stamp" or version string has been added to the library. At startup, all processes compare the versions of the `libelan` library across the job. If the versions differ, the process does not run. In the following example, the versions are found to be compatible:

```
atlas0:user> mping --elan-version
MPI: Compile stamp is 'robin@fab0 Wed May 21 15:47:33 BST 2003'
MPI: compiled against libelan version '1.4.10'
MPI: linked against libelan version '1.4.10'
MPI: Compatible libelan versions found
atlas0:user>
```

- Some new `--elan*` command line arguments have been added to MPI programs to allow the behavior of MPI to be changed from the command line. The new arguments are as follows:
  - `--elan-help` shows help on the arguments available
  - `--elan-version` displays the "compile stamp" for the library, the version of the `libelan` library that MPICH was compiled against, and the version number of the `libelan` library the application is linked against; and prints a warning if the two versions are incompatible

- `--elan-no-check-version` specifies not to abort when the application is linked with an incompatible version of the `libelan` library
- `--elan-abort-with-core` specifies whether to dump the core file; using this flag has the same effect as setting the MPI_ABORT_WITH_CORE environment variable
- `--elan-rma` *`bool`* enables/disables the MPI-2 features of the library; the default is set to `on`

The bitmask in the next three arguments can contain the following values:

```
0x002  MPI_Barrier
0x004  MPI_Bcast
0x008  MPI_Gather
0x020  MPI_AllGather
0x080  MPI_AlltoAll
0x200  MPI_Reduce
```

Setting bit `0x001` causes all optimizations to be turned on.

- `--elan-coll-sub` *`bitmask`* controls which functions use the optimized collectives when using MPI sub-group communicators
- `--elan-coll` *`bitmask`* controls which functions use the optimized collectives when using MPI group communicators
- `--elan-coll-simple` *`bitmask`* controls which functions use the optimized collectives when using exact copies of MPI group communicators; this argument will shortly be removed

If you disable optimized collectives for the MPI group communicators, they automatically become disabled for MPI sub-group communicators.

The following five arguments allow you to specify values for MPI to pass into the various `libelan` initialization functions. Setting these arguments can change the behavior of the library. The *`flags`* value is a bitmask.

- `--elan-base-flags` *`flags`* specifies the flags to pass to `elan_baseInit()`
- `--elan-tport-flags` *`flags`* specifies the flags to pass to `elan_tportInit()`
- `--elan-group-flags` *`flags`* specifies the flags to use when creating elan groups
- `--elan-lock-flags` *`flags`* specifies the flags to pass to `elan_lockCtrlInit()`
- `--elan-mqueue-flags` *`flags`* specifies the flags to pass to `elan_mQueueInit()`

1.11 MSA1000: Support for Modular SAN Array 1000

In HP AlphaServer SC Version 2.6, the Modular SAN Array 1000 (MSA1000) is supported for global or external storage. For more information, see Chapter 6 of the *HP AlphaServer SC System Administration Guide*.

1.12 Network Cards

The 3X-DEGXA-SA and 3X-DEGXA-TA Gigabit Ethernet cards (BCM) are supported in HP AlphaServer SC Version 2.6.

1.13 Operating System

In HP AlphaServer SC Version 2.5, the supported operating system was Tru64 UNIX Version 5.1A and associated Patch Kit 2.

In HP AlphaServer SC Version 2.6, the supported operating system is Tru64 UNIX Version 5.1B and associated Patch Kit 3.

(Please see Section 2.1.9 on page 2–4 for information about a fix for a problem with the `find` command as delivered in Tru64 UNIX Version 5.1B Patch Kit 3.)

1.14 RMS Changes

This section describes the changes made to Resource Management System (RMS) functionality in HP AlphaServer SC Version 2.6 since Version 2.5.

1.14.1 New and Changed RMS Features Delivered in Version 2.6

The following changes to RMS functionality are new in HP AlphaServer SC Version 2.6:

- Support has been added for sending a signal to applications when they are disconnected from the Elan rail they are using. If the `elan-disconnected-sig` attribute is set to a signal number, the RMS daemon sends that signal to applications using the Elan if the Elan drops out of the nodeset for its rail. The node is ejected from its domain if the cluster rail is disconnected. This will cause the application to be killed.

  You can create the `elan-disconnected-sig` using the `rcontrol` command, as shown in the following example, where the attribute is set to User Defined Signal 2:

  ```
  # rcontrol create attribute=elan-disconnected-sig val=31
  ```

  You should not set the `elan-disconnected-sig` attribute to SIGKILL (9).

- Memory scheduling has been improved to distinguish the amount of free memory on each node and to split boxes on the free list accordingly. This behavior can be enabled by setting the `memsplit-enabled` attribute to 1 (one). For more information, see Chapter 5 of the *HP AlphaServer SC System Administration Guide*.
- The following new functions have been added to the RMS API. For more information on RMS application programming interface functions, see Appendix D of the *HP AlphaServer SC RMS Reference Manual*:
  - The `rms_allocateHosts()` function allocates CPUs on one or more nodes for a resource that has previously been created by a call to the `rms_createResource()` function.
  - The `rms_version()` function returns the version string of the version an RMS library was built with.
  - The `rms_checkVersion()` function checks that the RMS version an application was compiled with is compatible with the RMS version installed on the machine where the application is run.
  - The `rms_jobIds()` and `rms_killJob()` functions provide support for sending signals to specific jobs rather than to all of the jobs running on an allocated resource.
  - The `rms_closeDatabase()` function closes a socket connection to the RMS database that was previously opened by another RMS API function.
- The following new fields are returned by the `rms_nodeinfo()` function:
  - `fcpumask` is the mask of the free CPUs on the node
  - `domain` is the name of the AdvFS domain
  - `drain` indicates that jobs are being drained from the node
- When called with the name of an active partition, the `rms_nodeinfo()` function returns information on nodes that have been configured out.
- The `rms_allocateCpus()` function supports two new request flags, `baseCpu` and `topCpu`, that allow the function to be used to allocate specific CPUs on each node.
- The `rms_deallocateResource()` function no longer blocks while waiting for CPUs to be deallocated. If the partition manager is down or the request does not complete quickly, `the rms_deallocateResource()` function returns the value RMS_RC_RETRY.
- The RMS partition manager provides support for draining jobs from a set of CPUs, and then configuring out the nodes. For more information, see Chapter 5 of the *HP AlphaServer SC System Administration Guide*.

- RMS now allows you to run user-defined scripts as part of the process of allocating and deallocating CPUs to a user. In the process of allocating CPUs to a user, RMS optionally executes a script; this script must be named `prologue`. In the process of deallocating CPUs, RMS optionally executes a script; this script must be named `epilogue`. The scripts execute as `root`. For more information, see Chapter 5 of the *HP AlphaServer SC System Administration Guide*.
- The `rinfo` command can now be used to view information on the jobs and processes running on the CPUs allocated to a given resource request, using the syntax shown in the following examples:

  ```
  # rinfo -r resource_id
  # rinfo -rl resource_id
  ```

  When run with the `-rl` option, the `-rinfo` command also lists each job associated with a resource, the ID numbers of the job's processes, and the name of the node the processes are running on.

  For more information, see Chapter 5 of the *HP AlphaServer SC System Administration Guide*.
- The `rinfo` command displays the status `zombie` for a resource request that has been killed but has not yet had its CPUs deallocated. If a resource request stays in this state, the `rinfo -rl rid` command (where `rid` is the resource ID) should be used to determine which nodes are still running application processes.
- The `rinfo -cl` command has been extended to list all configurations and partitions, including inactive ones. The output shows the status and the nodes, and for active partitions, the total number of allocated, suspended, queued and blocked resources. For example:

  ```
  # rinfo -cl
  CONFIGURATION  PARTITION      STATUS   ALLC     SUSP QUED BLKD  NODES
  day            parallel       running     0        0    0    0  atlas[0-63]
  day            test           running     0        0    0    0  atlas[64-127]
  night          parallel       -                                 atlas[0-127]
  #
  ```

- The `rinfo` command can now be used to list events by class, name and type. The `-e` (event) option lists events in time order, and can take the following flags:
  - `all` (lists all events)
  - *`class`* (lists all events matching class)
  - *`class`*`:`*`name`*`:`*`type`* (lists all events matching class, name and type).

A banner is printed if the `-l` flag is used. For example, to show all `node` class events, run the following command:

```
# rinfo -e node -l
Date     Time      Class   Name    Type         Description
06/28/02 14:47:00  node    atlas0  status       configured out
07/05/02 11:08:42  node    atlas0  status       not responding
07/18/02 07:45:22  node    atlas1  status       not responding
07/18/02 07:45:22  node    atlas2  status       not responding
07/18/02 08:00:06  node    atlas3  temperature  ambient=30
```

If * (asterisk) is used instead of *`class`*, *`name`*, or *`type`*, events of all classes, names, or types are shown.

For example, to show temperature events for all nodes, run the following command:

```
# rinfo -e node:*:temperature -l
Date     Time      Class   Name        Type             Description
07/18/02 14:59:06  node    atlas3      temperature      ambient=30
07/18/02 14:59:07  node    atlas1      temperature      ambient=31
07/18/02 14:59:17  node    atlas0      temperature      ambient=32
07/18/02 15:35:51  node    atlas1      temperature      ambient=30
```

The addition of the `-H` argument allows events in the last few hours to be selected. For example, to view node status events in the last hour, run the following command:

```
# rinfo -e node:*:status -H 1
Date     Time      Class   Name        Type             Description
07/18/02 07:45:22  node    atlas1      status           not responding
07/18/02 07:45:22  node    atlas2      status           not responding
```

The `-E` option can be used with the `rinfo` command to list event classes and types from the events table, as shown in the following examples:

For example:

```
# rinfo -E class
partition
switch_module
node
server
domain
# rinfo -E type
status
control
psu
fan
error
elan
temperature
```

The equivalent SQL command is:

```
# select distinct class from events
```

- The output of the `rinfo -s [rmsd|rmsmhd]` command is now sorted by hostname, and the `-l` option can be used to provide a summary, as shown in the following examples:

```
atlas0# rinfo -ls rmsd
SERVER               HOSTNAME    STATUS              PID  NODE
rmsd                 atlas0      running           16011    up
rmsd                 atlas1      configured out
rmsd                 atlas2      configured out
rmsd                 atlas3      configured out
rmsd                 atlas4      configured out
rmsd                 atlas5      running            4856    up
rmsd                 atlas6      running           26730    up
rmsd                 atlas8      configured out
rmsd                 atlas9      configured out
rmsd                 atlas10     configured out
rmsd                 atlas11     configured out
rmsd                 atlas12     configured out
rmsd                 atlas13     configured out
rmsd                 atlas14     configured out
rmsd                 atlas15     configured out

STATUS               NODES
running              atlas[0,5-6]
configured out       atlas[1-4,8-15]
```

By default, the `rinfo -s` command does not check the status of RMS servers on nodes that have been configured out. This check is performed if a second `-l` option is used, as shown in the following example:

```
atlas0# rinfo -lls rmsd
SERVER       HOSTNAME    STATUS             PID     NODE
rmsd         atlas0      running            16011   up
rmsd         atlas1      not responding             up (configured out)
rmsd         atlas2      running            4388    up (configured out)
rmsd         atlas3      running            22250   up (configured out)
rmsd         atlas4      running            4176    up (configured out)
rmsd         atlas5      running            4856    up
rmsd         atlas6      running            26730   up
rmsd         atlas8      not responding             down (configured out)
rmsd         atlas9      not responding             down (configured out)
rmsd         atlas10     not responding             down (configured out)
rmsd         atlas11     not responding             down (configured out)
rmsd         atlas12     not responding             down (configured out)
rmsd         atlas13     not responding             down (configured out)
rmsd         atlas14     not responding             down (configured out)
rmsd         atlas15     not responding             down (configured out)

STATUS               NODES
running              atlas[0,2-6]
not responding       atlas[1,8-15]
```

Note the difference between atlas1, which is up but has no RMS running, and atlas8, which is down.

This command can take some time to complete if large numbers of nodes are powered down.

- Signals can now be specified by name or number in the rcontrol command. For example:

```
# rcontrol kill resource 3344 signal SIGTERM
# rcontrol kill resource 3344 signal 9
```

- The prun stdout/stderr handling has been enhanced to label each line of output with a string. The label can be set using the -l option with the prun command, or with the RMS_IOLABEL environment variable. Some characters in the string have special meanings, as follows:
  - %g indicates the global rank of the process
  - %G indicates the number of processes
  - %h indicates the rank of the node
  - %H indicates the number of nodes
  - %l indicates the rank of the process on this node
  - %L indicates the number of processes on the node
  - %m indicates the text base of the hostname
  - %n indicates the numeric part of the hostname
  - %p indicates the pid of the process
  - %t indicates the current time (hh:mm:ss)
  - %@ indicates the hostname of the node (the same as %m%-n)
  - %% indicates a single %

  The existing -t option with the prun command is equivalent to a label of %_g. By default, the prun command pads numeric fields with leading zeros to aid tools that parse this output.

  This behavior can be adjusted by adding a modifier after the % symbol, as follows:

  - (hyphen) indicates do not pad

  _ (underscore) indicates pad with spaces

  Examples:

```
atlas0# prun -N4 -l "%g: " hostname
1: atlas1
3: atlas3
```

```
0: atlas0
2: atlas2

atlas0# prun -N4 -n16 -O -l "%g (%n.%l) " hostname | sort
00 (0.0) atlas0
01 (0.1) atlas0
02 (0.2) atlas0
03 (0.3) atlas0
04 (1.0) atlas1
05 (1.1) atlas1
06 (1.2) atlas1
07 (1.3) atlas1
08 (2.0) atlas2
09 (2.1) atlas2
10 (2.2) atlas2
11 (2.3) atlas2
12 (3.0) atlas3
13 (3.1) atlas3
14 (3.2) atlas3
15 (3.3) atlas3

atlas0# prun -N4 -l "%t " hostname
11:50:42 atlas1
11:50:42 atlas0
11:50:42 atlas2
11:50:42 atlas3
```

- The maximum number of segments in a parallel program has been increased to 150. A segment is a contiguous range of both nodes and CPUs; for example, CPUs 0 and 1 on nodes 4-7.
- The number of concurrent `rsh` processes a `prun -r` command can achieve is limited to 25 by default. This limit can be changed by changing the value of the RMS_RSH_LIMIT environment variable.
- RMS has been integrated with statistics collection in the `libelan` library.

  The Elan libraries can generate statistics on the interprocess communication calls in a parallel program. Statistics are written to a file in the current working directory on delivery of a signal and as a process exits. The RMS_STATINTERVAL environment variable controls the rate at which RMS sends the LIBELAN_STATSIG signal to the application. For example, if the RMS_STATINTERVAL variable is set to 10, an update is sent every 10 seconds. Alternatively, you can set the LIBELAN_STATSIG variable and use the `rcontrol` command to send the signal as and when required. For example:

```
# setenv LIBELAN_STATSIG 31
# prun -N8 program
...

# rinfo -jl
```

```
JOB               CPUS     STATUS    TIME     USERNAME     NODES
parallel.758         2    running   00:20       duncan     atlas[0-1]
# rcontrol kill job 758 signal 31
```

- Two new queueing interfaces, `elan_queueRx*` and `elan_queueTx*`, have been added. If you require more details on these interfaces, please contact your HP Customer Support representative.
- The behavior of the `rmstbladm` backup function (affecting `rmsbackup`) has been changed to print and store in the database the compressed backup file name (that is, the file name with the `.gz` suffix) rather than the name of the uncompressed file.
- Support has been added to the `rcontrol` command for checking whether a node that is being configured out is in use or not. This check is only used when the `rcontrol` command is run interactively. For example:

  ```
  # rcontrol configure out atlas2
  rcontrol: will kill user jobs. Continue (y/n): n
  ```

- You can now control whether individual partitions use HP AlphaServer SC Elan adapter cards. If the `use_elan` field is set to 0 (zero) for a partition, RMS does not attempt to initialize HP AlphaServer SC Elan adapter cards on the nodes in the partition, nor does it create Elan adapter card capabilities for parallel applications. This control replaces the use of the `ignore-elan` attribute in earlier releases. For more information, see Chapter 5 of the *HP AlphaServer SC System Administration Guide*.
- It is no longer necessary for all nodes in an RMS system to have HP AlphaServer SC Elan adapter cards. For more information, see Chapter 5 of the *HP AlphaServer SC System Administration Guide*.
- Support for the Tru64 UNIX processor set has been added to RMS. For more information, see Chapter 5 of the *HP AlphaServer SC System Administration Guide*.
- The mechanism for creating new log files has been improved. If a SIGHUP signal is sent to one of the RMS daemons, it moves the existing log file to `/var/rms/adm/log.dated/dd-mm-ccyy/` (creating the directory if necessary) and opens a new log file.
- The `-e` (exit) option has been added to the `rmsquery` command. This forces the command to exit immediately if an error occurs on connecting to the database server. The default remains (as in earlier versions) to retry the connect for a period of approximately a minute.
- Information on the hostname and process ID has been added to the output generated when an application is killed by a signal, as shown in the following example output:
  `prun: ./segv (host atlas2 process 1 pid 1621806) killed by signal 11`

- A new attribute (`partition-blocked-timeout`) instructs the partition manager to be more proactive in searching for non-responsive nodes when that partition is blocked. The attribute defines the length of time (default 15 minutes or 900 seconds) the partition manager waits when a partition is blocked before polling each `rmsd` daemon in turn. Any nodes that do not respond will be configured out, with the comment "no response" in the partition manager log file.

  This process will continue until a working set of nodes is reached.

  The value of the `partition-blocked-timeout` attribute must be at least twice the value set in the `inactivity-timeout` attribute.

  To create and set the `partition-blocked-timeout` attribute, run the following command:

  ```
  # rcontrol create attribute=partition-blocked-timeout val=900
  ```

  You must restart RMS for this change to take effect.

- Table 1–1 lists the fields that have been added to the RMS database:

Table 1–1 New Fields in RMS Database

Table	Field	Type	Default	Description
partitions	use_elan	int	1	Controls whether a partition uses the HP AlphaServer SC Elan adapter card (1) or not (0).
nodes	drain	int	0	Set to 1 when jobs are being drained from a node. Cleared to 0 when the drain is completed and the node configured out or if the drain is cancelled.
nodes	pagesize	int	8192	Page size in use on this node.
elites	routeUp	int	4-7	Link used by an elite to route to top of tree in broadcast DMA algorithm
elites	bcastTop	int	3,5 or 7	Defines which links an elite uses to broadcast DMAs down the tree. For example, a value of 3 implies broadcast DMA over links 0, 1, 2, and 3.

In addition, a definition for the use_elan field in the partitions table has been added to the fields table, allowing you to set the values 0 or 1 using the rcontrol command. An index has been added to the names field of the link_errors table to reduce the amount of time spent processing the link error data in the neterror script.

1.14.2 New and Changed RMS Features Delivered in Version 2.5 (UK1)

The following changes to RMS functionality were first delivered in HP AlphaServer SC Version 2.5 (UK1):

- RMS preserves resource numbers when a partition is restarted.

  In HP AlphaServer SC Version 2.5, if you stopped a partition and then restarted it, resources that already had assigned CPUs retained the same resource numbers when the partition restarted; however, resources that had not been assigned any CPUs (that is, resources that were waiting in a queued or blocked state when the partition was stopped) were assigned new resource numbers. This behavior has been changed, so that when a partition is restarted, all resource numbers are now preserved.

- Support for the set-vm-limit and inherit-user-limits attributes has been added. For more information, see Chapter 5 of the *HP AlphaServer SC System Administration Guide*.

- Node Statistics

  You can now reduce the amount of time that RMS spends on collecting the per-node statistics. By default, RMS will collect statistics every 30 seconds. You can increase this interval by setting the cpu-stats-poll-interval attribute to (for example) 120 seconds. Making this change will mean that data in the node_stats table is older, but this will not normally be a problem.

  If you set a higher value in the cpu-stats-poll-interval attribute (for example, 180 or 240 seconds), you should also increase the value of the inactivity-timeout attribute, which is used by the RMS Machine Manager (mmanager) to determine how long a node can be out of contact before assuming that the node is inactive. The value of the inactivity-timeout attribute should be at least 5 minutes, and at least 3 times the value of the cpu-stats-poll-interval attribute.

  HP does not recommend setting the cpu-stats-poll-interval to values in excess of 5 minutes.

  If you wish, you can also increase the value of the rms-poll-interval attribute. This will reduce the amount of time RMS spends on node housekeeping operations.

- File System Statistics

  Collecting the local file system statistics (local_usage, tmp_usage and freetmp) makes up a significant part of the periodic work done by the RMS daemon on each node. In general, these statistics vary slowly and it is acceptable to collect them less frequently.

A new attribute (`fsys-stats-multiplier`) has been added to control how frequently this information is collected. With the default value of `fsys-stats-multiplier` set to 10, and the default value of `rms-poll-interval` set to 30, the local file system statistics are collected every 300 seconds, once every 10 samples.

- Support has been added for synchronizing RMS work against the UNIX clock. If the UNIX clocks are kept synchronized (with NTP for example), then this new functionality will ensure that the periodic operations performed by RMS occur at the same time on each node. Synchronizing RMS work in this way reduces the impact of system "noise", particularly for larger systems.
- The `rms_whichPartition` function has been added to the RMS API. This function, which returns the name of the active partition containing a given node, replaces the old `rms_partitionName()` function, which was only good until the next call to `rms_partitionName`.

  The syntax of the new function is as follows:
  `int rms_whichPartition(const char *nodename, char *partition, size_t maxlen)`

1.15 Routing: Default Route Configuration

The default route mechanism has changed for HP AlphaServer SC Version 2.6. All nodes that have an external interface will have their default routes set to the system gateway. All nodes that do not have an external interface will obtain a default route setting *via* the `gated` daemon. The distribution and subsequent learning of the default route *via* the `gated daemon` is controlled by the `aliasd` daemon, and the `/etc/gated.conf` and `/etc/clua_default` files.

For more information, see Chapter 19 and Chapter 22 of the *HP AlphaServer SC System Administration Guide*.

1.16 SC Database: Daemon Upgraded

The SC database server daemon has been upgraded to `msql3d`.

1.17 SC Database: Database Integrity Check Utility

A new database internal integrity check utility, `sra dbcheck`, has been added. The new utility identifies internal inconsistencies in the SC database, and can be used to help to identify the cause of problems that might be occurring in your HP AlphaServer SC system, and prevent potential problems arising from database inconsistency.

For more information on the database integrity check utility, see Chapter 3 of the *HP AlphaServer SC System Administration Guide*.

1.18 SC Viewer: New Overview Tab

A new Overview tab has been added to SC Viewer. The Overview tab shows the status of every node in the system.

The main pane on the Overview tab shows a row of nodes for each domain. The overview status, which is a combination of the operational status and the environmental status of the node, is shown for each node.

The properties pane on the Overview tab displays the attributes of a domain selected in the main pane.

For more information on SC Viewer, see Chapter 10 of the *HP AlphaServer SC System Administration Guide*.

1.19 SCRUN: New Options

Two new options have been added to the `scrun` command:

- The `-t` option allows you to specify a timeout period. If the command issued by `scrun` has not completed on all nodes within the timeout period, `scrun` terminates the command on any nodes where the command has not completed. If the `-t` option is used with the `-w` option, the timeout period is reset when the next series of commands is issued.
- The `-w` option forces `scrun` to execute commands sequentially, rather than in the default parallel execution mode. The `-w` option takes an argument that specifies the maximum number of nodes or domain members the command is to be run on in parallel.

For more information about these options, and for examples of their use, see the `scrun(8)` reference page and Chapter 12 of the *HP AlphaServer SC System Administration Guide*.

1.20 SCRUN: umask Propagation

The `umask` setting for commands invoked by the `scrun` command has been changed to provide increased security. In HP AlphaServer SC Version 2.5, a default `umask` setting of 0 (zero) was used for commands invoked by the `scrun` command; this resulted in files created by these commands having read/write permissions for all users. In HP AlphaServer SC Version 2.6, the default `umask` setting has been removed, and the setting of the process in which the `scrun` command is invoked is propagated to the executing nodes.

For example, running the following commands:

```
# umask 077
# scrun -n all touch /tmp/f
```

creates files that have read/write permissions for the `root` user only.

1.21 Shmem Changes

The following changes have been made to Shmem functionality:

- The `-f putnb`, `-f getnb`, and `-b batch` options have been added to the `gping` command to show the performance of Shmem non-blocking puts and gets when running the `gping` command across nodes in the system
- Support for non-blocking puts and gets has been added to the `libshmem` library.
- The Shmem `fn globalexit()` function initiates a clean exit of the whole program when run under RMS. This functionality was already present, but was not documented.
- Support has been added for striping `elan_put()`, `elan_get()`, `elan_putv()` and `elan_getv()` calls over multiple rails, where available. This facility is used by the `libshmem` library. The `shmem_put()` and `shmem_get()` functions, and their type-specific variants, split a single transfer over multiple rails if the transfer size is 1024 bytes or greater. Transfers smaller than 1024 bytes are allocated rails on a "round robin" basis. The thresholds for rail striping can be changed by setting the LIBSHMEM_SPLITPUTSIZE and LIBSHMEM_SPLITGETSIZE environment variables.

1.22 SRA: New Commands

The following changes have been made to the `sra` command:

- Two new `sra` commands have been added:
  - The `sra command_info` command displays information on `sra` commands. The `-states` option allows you to specify which command states to display. The states are: `Allocated`, `Unallocated`, `Success`, `Error`, and `Abort`.
  - The `sra upgrade_info` command displays information on the upgrade status of each targeted domain.
- A new option has been added to the `sra shutdown` command. The `-force` option allows you to specify whether the node is to be shut down if it will result in the domain losing quorum.

For more information on the `sra` command, see the `sra(8)` reference page, and Chapter 16 of the *HP AlphaServer SC System Administration Guide*.

1.23 Switch Management Changes

This section describes the changes made to switch management functionality in HP AlphaServer SC Version 2.6 since Version 2.5.

1.23.1 New and Changed Switch Management Features Delivered in Version 2.6

The following changes to switch management functionality are new in HP AlphaServer SC Version 2.6:

- A new utility, `swctrl`, has been provided to perform the following functions:
  - The `swctrl check` command can be used to check the internal link state of each module. It also reports stale module update times to indicate that the switch network manager is no longer receiving module heartbeats.
  - The `swctrl configure` command can be used to configure a link, or an entire broadcast tree, in or out of the network
  - The `swctrl set` command can be used to set the clock source, network-type, and broadcast-tree attributes.
  - The `swctrl info` command can be used to view information on the entire network.

  For more information on using the new `swctrl` utility, see Chapter 7 of the *HP AlphaServer SC Interconnect Installation and Diagnostics Manual*.
- The following changes have been made to the switch network management daemon `swmgr`:
  - The `swmgr` daemon now creates database entries for Elites.
  - Entries for the broadcast top and route up have been added to the `elites` table. These parameters define the behavior of broadcast DMAs in the interconnect network.
  - The `swmgr` daemon now updates the `started` field in the `switch_modules` table each time the `swmserver` daemon restarts.
  - The `swmgr` daemon now posts a `switch_module` event of type `clock` if the frequency of the Elite clock in a switch module departs from the expected value of 400MHz.
  - A new event handler, `rmsevent_qsnet`, has been added. Its default behavior is to log all network events and send e-mail to administrator users for the more serious problems (for example, clock failure, module becoming unresponsive). Site or vendor specific behavior can be enabled using the `event-handler-path` mechanism.

1.23.2 New and Changed Switch Management Features Delivered in Version 2.5 (UK1)

The following change to switch management functionality was first delivered in HP AlphaServer SC Version 2.5 (UK1):

- An `updateModule` function has been added to the `jtest info` command. This allows you to update environment details while within an interactive client session.

1.24 Switches: Support for ProCurve Switch

The HP ProCurve 2650 (48-Port) switch is now supported. For more information, see Chapter 1 of the *HP AlphaServer SC System Administration Guide*.

1.25 System Check Tool: New sclogs Plug-In Component

An `sclogs` plug-in component has been created to support the use of the `sys_check` tool in HP AlphaServer SC systems.

HP provides a system check tool (`sys_check`) that creates a configuration report and, when invoked with the `-escalate` option, collects system log and crash files.

The task of collecting the system log and crash files takes only a few minutes, whereas the task of producing the configuration report can take several hours on an HP AlphaServer SC system.

By enabling only the `sclogs` component, you can collect the system log files and crash files (including the `sra` and `rms` logs files), without producing the configuration report, by running the following command:

```
# sys_check -escalate sclogs
```

The output from the system check can be sent to your HP Customer Support representative to assist in diagnosing system problems.

For information on how to install and use the `sys_check` tool and the `sclogs` plug-in component, see Chapter 29 of the *HP AlphaServer SC System Administration Guide*.

1.26 Upgrade: Check to Ensure that EVM is Running Correctly

In earlier versions of HP AlphaServer SC software, if the Event Manager (EVM) was not working correctly, the upgrade would hang during patch installation, without providing information on the source of the problem.

The check phase of the upgrade process now checks to ensure that EVM is running correctly; if it is not, the upgrade stops, and the following message is displayed:

```
Event Management System (EVM) reported errors
```

```
Such errors are usually caused by bad configuration files such as /etc/evm.auth
being empty. EVM is required by the patch process.
```

As the `/etc/evm.auth` file is generic, the corrupt file can be replaced with a copy of a valid version of the same file, possibly from the management server or from another HP AlphaServer SC system. If you need assistance in locating a valid copy of the file, please contact your HP Customer Support representative.

1.27 Upgrade: External RIS Server No Longer Required

In earlier versions of HP AlphaServer SC software, an external RIS server was required when upgrading the Tru64 UNIX software on the first domain in a system that does not have a management server. This requirement has been removed: the first domain can now be upgraded directly from the Tru64 UNIX operating system CD-ROM.

To upgrade the Tru64 UNIX software on the first domain directly from the Tru64 UNIX operating system CD-ROM, use the `-cd` option with the `sc_upgrade install` command to specify the directory where the CD-ROM has been mounted. When the `-cd` option is specified, the `-rishost` option is not needed; if both are specified, the `-cd` option overrides the `-rishost` option.

The `-cd` option can also be used with the `sc_upgrade check` and `sc_upgrade setup` commands.

For more information, see Chapter 4 of the *HP AlphaServer SC Installation Guide*.

1.28 Upgrade: RIS Server Does Not Need to be Specified

In earlier versions of HP AlphaServer SC software, when upgrading the domains, the RIS server had to be specified using the `-rishost` option with the `sra upgrade` command. In HP AlphaServer SC Version 2.6, if the `-rishost` option is not specified, the default RIS server is the management server if one is used; if no management server is used, the default RIS server is the cluster alias of the first domain. For more information, see Chapter 4 of the *HP AlphaServer SC Installation Guide*.

1.29 Upgrade: Upgrade Check Utility

A new upgrade check utility has been provided that allows you to check if your system is ready to be upgraded to HP AlphaServer SC System Version 2.6, and helps you to identify problems that might cause an upgrade to fail. The new utility runs a series of tests that examine a wide range of system configuration and status conditions that could cause problems during the upgrade process. Failed tests are reported to standard output (`stdout`) and a more complete log is written to the `/var/sra/adm/log/InstallSC/upgrade_check.log` file.

The upgrade check utility is provided on the HP AlphaServer SC System Version 2.6 Software CD-ROM; it must be run by the `root` user from the management server (if used) or from Node 0 (if not using a management server). The utility can be run directly from the CD-ROM; no installation is required. For more information on running the new utility, and a list of the tests it performs, see Chapter 4 of the *HP AlphaServer SC Installation Guide*.

1.30 Upgrade: Upgrade Restart Facility

In HP AlphaServer SC Version 2.5, if an upgrade failed at any point, you could not restart the process. The operating system version had to be restored from backup, and the upgrade process had to be launched again from the beginning. In HP AlphaServer SC Version 2.5 (UK1), an upgrade restart facility was introduced that allows you to restart the upgrade process if it fails at any point. A state-based mechanism controls the upgrade. For more information, see Chapter 4 of the *HP AlphaServer SC Installation Guide*.

Note:

In some circumstances, an interrupted upgrade can leave the system in an unrecoverable state, and the upgrade restart facility cannot be used. In such cases, the upgrade process must be launched again from the beginning.

2
Problem Solutions Delivered

This chapter lists the problems for which solutions have been delivered since HP AlphaServer SC Version 2.5, and is organized as follows:

- Problem Solutions Delivered Since HP AlphaServer SC Version 2.5 (UK1) (see Section 2.1 on page 2–1)
- Problem Solutions Delivered In HP AlphaServer SC Version 2.5 (UK1) (see Section 2.2 on page 2–6)

2.1 Problem Solutions Delivered Since HP AlphaServer SC Version 2.5 (UK1)

This section lists the problems for which solutions have been delivered since HP AlphaServer SC Version 2.5 (UK1). These solutions were delivered as patches for either Version 2.5 (UK1) or Version 2.5 and are included in HP AlphaServer SC Version 2.6, unless they have been superseded by new or changed functionality. Where available, the associated problem codes are shown.

2.1.1 Adapter Drivers

The following adapter driver problems were fixed:

- Intermittent system crashes occur when halting or rebooting a system that is using either DAPBA or DAPCA ATM adapters. (95814)
- Packets can be sent from a node with an incorrect MAC address. (101332)
- The Ethernet driver for DE60x Ethernet adapters allows a buffer to be re-used by the adapter after the Ethernet driver has given it to the stack, thus causing timeout errors and/or a network hang. (101735, 102137)

2.1.2 CFS

The following CFS problems were fixed:

- Assertion in `cfs_vnops.c` during boot. (99458)
- Assertions at `cfs_server.c:2084` and `svrcfs.c:1353` in CFS module. (99539, 99655)
- Deadlock between CFS client and NFS client: associated with ESTALE I/O errors. (100574)
- Multiple cluster members crash with Kernel Memory Fault in `rfs_find_fsid()`. (100700, 100835)
- The `setld` command sometimes fails because of running out of inodes; this results from an accounting error in AdvFS. (101387)
- A cluster member can panic with a kernel memory fault in `kdm_isenabled()`. The panic can happen when a write to a file fails due to a "file system full" condition, if that file has an improperly initialized recycled `vnode`. (102450)

2.1.3 Cluster Alias

The following cluster alias problem was fixed:

- Threads hang in `clua_lock_port` and `clua_portcache_wait` with the `clua_lp_thread` hung in `dlm_wait_on_event`. (100529)

2.1.4 ICS/Elan

The following ICS/ELAN problems were fixed:

- A kernel memory fault occurs in `simple_lock`, `fuser`. (96014)
- `icssvr_daemon_from_pool: invalid IPL 4` messages occur. (96956, 97965)
- PANIC: `ics_elan: sequence out of range`. (99716)
- Problem with freeing large routes on systems with more than 1024 nodes. (100051)
- Panic occurs during boot, caused by a problem in ICS initialization. (100863, 101038)
- Process with threads in `icscli_wait` can hang after being sent an `exit` signal in certain circumstances. (100950, 101399)
- One node being shut down can cause other nodes to temporarily leave the elan nodeset and permanently leave a domain. (102437)

2.1.5 LSM

The following LSM problem was fixed:

- Bad configuration results in LSM problem and `vold` core dumping. The fix allows LSM to recover from configurations where disk groups are comprised of local-only disks. (94133)

2.1.6 MPI

The following MPI problems were fixed:

- Memory exhaustion problems with repeated calls to `MPI_Gather()` function. (99656)
- MPI applications hang in `MPI_Reduce` and occasionally cause a kernel memory fault in `FlushDisconnect`. (100377, 100890, 100930)
- Problem with `libelan` global memory fragmentation when a block of memory in the middle of the heap is allocated, freed and then reallocated. (101374)

2.1.7 msqld

The following `msqld` problem was fixed:

- Problem in the `msqld` daemon causes certain updates to be excluded during cache `msync` operations, leading to updates not being written to disk and consequent data loss in the event of failover on a clustered management server. (99877)

2.1.8 NFS

The following NFS problems were fixed:

- Ill-formed TCP connections can cause RPC-based services to hang in an `accept(2)` state, waiting for the connection to be completed. The ill-formed TCP connections typically come as a result of portscanning activity on a network. (95128)
- A hard-mounted NFS file system on an NFS client incurs ETIMEDOUT errors. This situation can occur when the NFS client is accessing an NFS file system on a server that is down (either when the server is actually down or temporarily appears to be down because it is busy). (97673)
- Problem with uninterruptible processes when the MVFS ClearCase file system is in use. (98662, 99220)
- Problem with the way the NFS client handles full NFS V3 64-bit file IDs. When traversing an NFS V3 server file system that generated file IDs greater than 32 bits, messages similar to the following are displayed on the console:
  `NFS3 LOOKUP: directory attributes from <servername> are of the wrong dir`

The fix allows the full 64-bit file ID to be stored and the actual value is used for comparisons. (101488, 101585)

2.1.9 Operating System

The following operating system problems were fixed:

- Client (`login`, `su`, `rshd`, `edauth`, and `sshd2`) hangs and long delays are experienced when booting cluster under Enhanced Security; also some intermittent errors or failures are seen with `prpasswdd` or `rpc.yppasswdd`. (96980)
- Cluster members can crash when running `dupatch`. (99451)
- Shutdown hangs on some nodes if the HP AlphaServer SC-specific UNIX accounting is configured. (100875)
- The `/usr/sbin/dirclean` utility attempts to remove the AdvFS `.tags` directory or the `quota.group` and `quota.user` files. (102508)
- The `find` command as delivered in the Tru64 UNIX Version 5.1B Patch Kit 3 aggregate patch kit contains a regression that affects the traversing of very large directory structures.

  Early Release Patch T64KIT0020545-V51BB24-E-20031104 was made available to remedy this problem. The files made available in that Early Release Patch have been integrated with the HP AlphaServer SC Version 2.6 release and there is no additional requirement to install Early Release Patch T64KIT0020545-V51BB24-E-20031104.

2.1.10 PFS

The following PFS problems were fixed:

- Panic because PFS releases a CFS `vnode` that is still open. This occurs when PFS encounters a problem (for example, a network error) opening either the `.pfsbirth` or `.pfsdeath` directory while mounting a file system. (99673, 99701)
- Problem writing to PFS file systems after boot. (100266)

2.1.11 RMS

The following RMS problems were fixed:

- Following RMS failover on a clustered management server, the `rmsd` daemon on the nodes does not retrieve the new value of `rmshost` from the database. (99155)
- The `rmsd` daemon running on the node responsible for updating the `mmanager` daemon with the cluster status can crash. (100243)
- Starting the `rmsd` daemon fails on an XP1000 server. (100382)

- A problem causes the `prun` command to intermittently fail to start on all nodes when the `-Rone-process-per-node` parameter is used. (100728)
- Starting a number of jobs in an allocated resource can result in an incorrect value for the `acctstats.cpus` field. (102451)
- Calculation of `acctstats.cpus` and `acctstats.utime` is incorrect when a node is configured out. (103133).
- Negative CPU usage counts reported after repeated resource suspend/resume actions.
- An RMS partition appears to block indefinitely if, for example, a node's `rmsd` daemon is caught up in a system call.

2.1.12 SCFS

The following SCFS problems were fixed:

- Threads hang indefinitely on the head node of a domain in `tok_wait`. The problem was seen on SCFS file systems running jobs that were concurrently accessing the same file. (99196, 100430, 101838)
- SCFS file systems sometimes mount with UBC access mode, even though the database indicates that these file systems were to be mounted with FAST mode. (99480)
- Assertion in `scfs_failover.c:357` in SCFS module. (101019)
- Writes to an SCFS file system in FAST mode can stall after approximately 2 billion writes. (101019, 100266)
- SCFS server miscalculates the number of pages in a request. This problem can result in the I/O or request manager threads on an SCFS file-serving node accessing invalid pages, resulting in a panic. (102652)

2.1.13 SRA

The following SRA problem was fixed:

- On a large system, when a boot of all nodes is performed, some nodes occasionally do not boot. (100121)

2.1.14 Virtual Memory

The following virtual memory problems were fixed:

- Node panics with `u_anon_unlock_page: anon already unlocked` message. (95802, 102744)

- A `delete_pv_entry` panic occurs when kernel virtual address space has high usage. (97119)
- Deadlocks in `vm_fault` and `u_stack_grow` lead to uninterruptible RMS orphans. (95052)
- A memory leak causes an out-of-swap situation and therefore a performance issue. (102738)

2.2 Problem Solutions Delivered In HP AlphaServer SC Version 2.5 (UK1)

This section describes the problem solutions that were delivered in HP AlphaServer SC Version 2.5 (UK1); it also describes some changes that were introduced as a result of problem solutions.

The solutions and changes described here are also included in HP AlphaServer SC Version 2.6, unless they have been superseded by new or changed functionality.

2.2.1 Boot

The following boot problem was fixed:

- Failing to connect to the Console Management Facility (CMF)

2.2.2 CAA

The following CAA problem was fixed:

- Missing CAA `.cap` files, which caused CAA to be incorrectly installed

2.2.3 CFS

The following CFS problems were fixed:

- `PANIC (ASSERT FAILED: HP->SCT_VP_HOLDS TNC_CFE/SVRCFS.C:1365`
- Kernel memory fault (KMF) while printing message after CFS token timeout
- `ASSERT FAILED: CFSTOK_HOLDCNT(VP) == 0 OR TOK_DESTROY: TOKENS ARE PRESENT`
- `PANIC: CFS_FO_ISSUE_BID_SERVER: KGS_SEND_PROPOSAL FAILED`
- Minimum SPL violation in `CFS_REISSUE_KGS_JOIN_THREAD`
- CFS panics if you set `KMEM_DEBUG` to `0X40`

- Token deadlocks (`token_wait`) were causing processes associated with RMS jobs to become orphans. The CFS token code was changed to use complex locks instead of simple ones.

2.2.4 Cluster Alias

The following cluster alias problems were fixed:

- Cluster alias ICS performance problems in HP AlphaServer SC clusters.
- Cluster alias lockaside list causes port starvation on HP AlphaServer SC clusters.
- EADDRINUSE errors for ports in cluster alias lockaside list.
- Cluster alias ICS command thread does not re-initialize the `tries` variable.
- Lock hierarchy panic in cluster alias.
- Cluster alias not handling rolling upgrade properly.
- When an application tries to use a port that exists in a lockaside list, the application fails with EADDRINUSE even if the port is not in use.
- Loss of port registrations.
- Problem that prevented the `nessus` security scan from running. The problem caused multiple nodes to panic due to kernel memory faults occurring in `clua_udp_input_tunnel`.

2.2.5 elantest

The following elantest problems were fixed:

- Missing reference pages
- Problem with `-RD` option (corrected to get `devinfo` before generating route)
- Network error handling problem in `cabletest` script
- Statistics displaying incorrectly
- DMA not stopping `check_data` checking beyond data size

2.2.6 ICS/Elan

The following ICS/Elan problem was fixed:

- Multiple nodes were crashing with the panic string:

  `cmn_err: CE_PANIC: ics_elan: sequence delay greater than 1000`

  with simultaneous crashes with the panic string:

```
elan0: remote panic call from elan node
```

In addition, the following ICS/Elan change was made:

- Elan ICS was modified to improve tunneling performance. This change improves the performance of ICS tunneling for small data sizes, and improves NFS/SCFS server performance.

2.2.7 Install/Upgrade

2.2.7.1 Installation

The following installation problems were fixed:

- Errors in the following scripts:
  - `configureUNIX.exp`
  - `haltcluster.exp`
  - `member_delete.exp`
  - `update_firmware.exp`
- A problem that existed during setup or upgrade when setting the value of the `cmf.subnet` SC database field. This value is derived from the management LAN or external LAN IP/subnet mask settings for node 0 (zero). The external IP address for node 0 was not being set under some circumstances.

In addition, the following changes were made to the installation process:

- Improved console connection retry mechanism.
- Renamed CAA services: `srad` is now `SC15srad`, `cmf` is now `SC10cmf`.
- License units are now calculated from the node types (for example, HP AlphaServer DS20L nodes only require 200 units as they only have 2 processors).
- Adding entries to the `/etc/hosts` file is now faster.
- Databases can now start at a node or domain other than 0 (zero).

2.2.7.2 Upgrade

The following upgrade problems were fixed:

- Error in the following script:
  - `upgrade.exp`
- Space in a password causing login of the `sra` command to fail
- Problem when appending to the `/etc/exports` file

- Problem of upgrading from HP AlphaServer SC Version 2.0 in an HP AlphaServer SC system that does not have a management server
- Upgrade failing to re-add members
- `sra delete_member` command failing if a previous add aborted, leaving the boot domain pointing at the Tru64 disk swap partition

In addition, the following changes were made to the upgrade process:

- In the boot to single-user mode procedure, the `sra` command now waits for the shell prompt before continuing in single-user mode.
- The SRA database translator (the `sc_upgrade_db` command) now sets the `current_state` and `bootable` fields in the `sc_nodes` table appropriately when translating the SRA database from the text file to the SQL database.
- The SRA database translator now ensures that the values in the `.entries` fields of the text file are valid.
- During upgrade, the check phase now fails if the `lockmode` kernel attribute has a value greater than 2, as any value greater than 2 in this attribute can cause problems for the upgrade.
- A check to ensure that the cluster alias IP address is correct has been added.
- The `dupatch` script has been modified to ignore point patches.
- The `-noris` option has been added to the `InstallSC` command, to be used only by the upgrade process to avoid a problem with not finding the RIS area when run from within an upgrade.
- Added a check to ensure that the backup disk is big enough before proceeding with backup.

2.2.8 IO/DRD

The following IO/DRD problem was fixed:

- Domain panic on cluster root: `BS_OSF_COMPLETE METADATA WRITE FAILED`

2.2.9 KCH

The following KCH problem was fixed:

- `PANIC: KCH SET NOT LOCKED BY INITIATOR`

2.2.10 LSF

The following LSF problems were fixed:

- CAA script failed to recognize paged-out daemons
- Non-AlphaServer SC hosts could not be licensed as client hosts
- Missing HP AlphaServer ES45 models
- LSF master did not fail over
- Host Preference not working in multiple partitions
- MBD/RLA internal errors in multi-machine environment
- Interactive jobs being suspended when the RMS partition was down
- `bsub` command not allowing railmask specification
- Allocation left over after a job was killed
- Job's `rusage` not available if PIM was removed

In addition, the following changes were made to LSF:

- Added support for height allocation. By default, LSF now allocates from the left-most node within a domain, and tries to make the allocation as compact as possible.
- When `brun` is being used to force an RMS job to run on specified hosts, the job will run immediately if LSF finds enough resources to satisfy the topological requirement of the job. Otherwise the job will exit with an exit code of `123`.
- LSF no longer requires that RMS partitions on different HP AlphaServer SC systems in the same LSF cluster have unique names. For example, if you have two HP AlphaServer SC systems, one called `athens`, and a second called `apollo`, each system can have a partition called `parallel`. In HP AlphaServer SC Version 2.5 or earlier, one of the partitions would have had to be renamed.

2.2.11 msqld

The following `msqld` problems were fixed:

- `msqld` performance issues.
- High `msql` load when `scfsmgr` was processing a large number of file systems. This could have caused nodes to be non-responsive to RMS.
- Running out of connections for a large number of nodes. (The connection limit has been increased to 4096.)

2.2.12 PFS

The following PFS problems were fixed:

- `vrestore` of `vdump` results in corrupted files; restored files are the correct size, but have many NULLS
- File corruption problem on PFS file systems causing zero-length files to be seen under high I/O load
- Locking problems occurring when multiple nodes opened a file concurrently
- PFS file systems hanging with RM memory usage and `vm_wait` memory allocation
- SysMan menu — `delete` option for PFS not working

In addition, the following change was made to PFS:

- Modified `ioctl` code in PFS to return a message of `not supported` when asked for a file extent map. PFS has never supported extent maps; however, previous versions of PFS would return a success status when asked for one. This was causing problems to programs (such as `vdump`) that use extent maps.

2.2.13 RMS

The following RMS problems were fixed:

- Passing USR1 and USR2 signals to the `rcontrol kill` command caused the process to hang or die.
- `rcontrol` command failing to configure nodes in and out.
- If run in interactive mode, the `rcontrol` command was not updating its internal machine description, so that changes made during the interactive session (for example, configuring nodes in and out, or stopping/starting partitions) were not showing.
- Problem with the event log path in the handler for escalation events (`rmsevent_escalate`).
- `prun -T -1` command failing to select the user's resource when the password files are not consistent across the nodes in an HP AlphaServer SC system. This was causing the TotalView bulk startup mechanism to fail.
- Problem in the `rmsd` daemon prevented updates to some of the environment fields in the `nodes` table (`psus`, `fans` and `environment`) except on restart of the daemon.
- Corrupt account statistics caused by reissuing a `stats` collection command when a blocked partition restarts.
- `prun -T` should have the cyclic flag set by default.

- Memory leak in the partition manager provoked by configuring nodes in or out (or anything that caused a scheduler reset); the size of the leak is dependent on the number of users and projects defined.
- Database server connection leak caused a failure when constructing a machine description.
- `prun -r` signal handling left `rsh` processes defunct.
- `prun` output redirection code caused the output file to be opened once per process rather than once per program.
- Signals not being forwarded to the processes in a parallel program if they arrived while `stdio` was pending.
- `prun` command generating an assert error and associated core dump when the user's environment exceeded the size of a `LoaderCmd?` buffer.
- Partition manager trying to reallocate (as for `prun` after allocate) `procfile` allocations and doing it incorrectly.
- `procfile` command trying to run parallel programs because the `pfloader` code was using `elan_baseInit` to find its local ID. `pfloader` now uses `RMS_RANK` instead.
- Multi-segment allocations could cause corrupt resource records in the database.
- Problem with the status of resource requests killed when the partition is shut down.
- Partition manager not rejecting job start requests after a resource has been killed.
- Problem with responsiveness of `mmanager` under Tru64 UNIX when voting members of a cluster are down and configured out.
- Problem with responsiveness of the `rmsd` daemon under Tru64 UNIX when the high numbered nodes in a cluster are down and configured out.
- `running` and `project` fields not being set on return from `rms_resourceStats()`.
- Problem with output buffering in parallel applications.

  If a user writes 8KB of data including newline characters on `stdout`, only the data up to and including the last newline character is sent to the `prun` command. The remaining data is buffered in the RMS application loader (`rmsloader`) until a newline character is sent or the process exits.
- `rmsloader` exiting when `rcontrol` is used to send SIGUSR1 or SIGUSR2 to the application. These signals are now ignored.
- `prun -r` command returning early on large numbers of nodes.
- `rmsd` daemon dropping the database server connection as soon as possible.

- Problem with the behavior of `RMS_MEMLIMIT`.

  Two new attributes were added for controlling memory limits. For more information, see Section 1.14.2 on page 1–24.

- Jobs being queued rather than an error being returned when `RMS_MEMLIMIT` is set to a size larger than the amount of memory available to the machine.

  The scheduler now determines whether the request can be met and reports if the `RMS_MEMLIMIT` is too high.

- Switch network manager (`swmgr`) core dumping/accessing the `elites` table.

- `rmsd` daemon core dumps following the removal of the `rmshost` alias from the `/etc/hosts` file.

- Jobs being queued indefinitely (forever).

- Resource numbers were being lost because `prun` generated a new request.

  RMS now preserves resource numbers when the partition is restarted. For more information, see Section 1.14.2 on page 1–24.

- The `rinfo` command sometimes shows resources and jobs that have actually finished as still running.

  When a partition stops, jobs can be left running. These jobs may finish during the period that the partition is stopped. When the partition is restarted, `rinfo` was temporarily showing that these jobs were still running, when in fact they had finished. This problem is now fixed.

In addition, the following changes were made to RMS:

- Added fix to `swctrl` to prevent environment checking of `fan` and `psu` status for standalone QMS16 modules.

- Added error message if the `rcontrol` command is used to attempt to stop/start/restart a server on a node that has been configured out. This fails silently in previous versions of RMS.

- Added PRUN_MAX_ENV (12KB) and a check for overrun.

- Added code to the `prun` command to ensure that the `arg` list for the processes in a parallel job will fit in a command buffer. Maximum size is 8KB.

- Improved `prun` debug when `rmsloader` processes fail to sign on, for example:

```
prun: Error: loaders exited without completing initialisation
prun: no signon from atlas[0-31,128-159]
```

- Added support to `rcontrol`, `pmanager` and partition startup script for a new partition type `restricted`. Only the `root` user can run jobs on a partition of this type. The partition startup script disables logins on a `restricted` partition.
- In the RMS API, the following command:
  `rms_nodeInfo(char *partition, rms_nodeinfo_t *info int maxnodes, int options)`

  now returns the value `-1` if the node is configured out. It was previously returning the value `-2`.
- Included a workaround to `core_analysis.OSF1` for a problem where Tru64 truncates the filename to 16 characters.

  The workaround assumes that if the filename extracted from the core file (`corefile_exe`) is exactly 16 characters, and the executable file run by `prun` (`prun_exe`) has a file name longer than 16 characters, with the first 16 characters matching the `corefile_exe` file name, then the two are the same. This fixes the most common scenario.

2.2.14 SC Monitor

The following SC Monitor problems were fixed:

- Problem in `sc_monitor_dev` table usage
- Non-existent `/usr/bin/sssu`

2.2.15 SC Viewer

The following SC Viewer problems were fixed:

- Default switch addresses assigned when creating new network view. The creation of a new Interconnect view was resulting in the incorrect TCP/IP address range being configured in the `switch_modules` table for the HP AlphaServer SC Interconnect switches.
- Improved error handling for unknown `switch_module` table contents. Fixed a problem in SC Viewer where `unknown` module types in the `switch_module` table were not handled gracefully. The treatment now applied is to report the problem to the user and suggest the potential root causes.

In addition, the following changes were made to SC Viewer:

- The `network_layers` attribute is no longer used. Due to a problem in upgrades from HP AlphaServer SC V2.0 and HP AlphaServer SC V2.4, SC Viewer was incorrectly referencing the deprecated `network-layers` field in the `attributes` table in preference to the correct `network-rails` field. The problem prevented SC Viewer from being able to display the second rail in a dual rail system.
- Node numbering that does not start at 0 (zero) is now allowed.
- Added check for DISPLAY variable or `-display` given on the command line. If both are missing, then an error message is issued and the procedure exits. In addition, a usage message is printed out if there are unrecognized command-line arguments.

2.2.16 SCFS

The following SCFS problems were fixed:

- Processes hanging in `scfs_buf_wait` and requiring a reboot to clear.
- Processes were seen to hang in `scfs_buf_wait` when the File Serving node failed over.
- SCFS could potentially consume large amounts of memory in the RM bucket on a client node during a server failover, in extreme cases even causing the client node to hang.
- SCFS client simple lock timeout and panic could occur if a file system was unmounted while a failover was in progress.
- Problem with use of the O_SCFSIO flag on PFS file systems.

2.2.17 scfsmgr

The following `scfsmgr` problems were fixed:

- `scfsmgr scan` failing during upgrade
- SCFS manager performance issue, occurring when SQL was busy, caused nodes to be configured out

2.2.18 scload

The following `scload` change was made:

- The `scload` command no longer clears the screen before displaying information.

2.2.19 scmountd

The following `scmountd` change was made:

- Improved PFS mount time. Mounts that were taking more than 110 minutes now take only 5-15 minutes.

2.2.20 scrun

The following `scrun` problems were fixed:

- `gxclusterd`/`gxmgmtd` deadlock problems caused `scrun` to appear to work on some nodes of the domain but not on others.
- `gxnoded` daemon hanging when a node returns from quorum.
- Problem of running the management station daemon and domain management daemons only on member 1 and member 2 if the daemons are being started on a domain. (This fix was necessary to prevent potential routing issues with the cluster daemon.)

In addition, the following changes were made to `scrun`:

- Use of – (dash) in system names now allowed

2.2.21 Shmem/libelan.so

The following Shmem/libelan.so problems were fixed:

- Intra-node dual memory issue
- Bad mappings
- Dual rail intra-node use of elan
- Corruption issues
- Error in intra-node barrier
- Startup race in the `psync` initialization code that caused Shmem collectives to intermittently block on the first call
- `test_event()` incorrect (reversed) return value

In addition, the following change was made to Shmem:

- Modified the Shmem collectives to use the Elan library barrier and broadcast code when called for all processes in a parallel program

2.2.22 SRA

The following SRA problems were fixed:

- `sra` failing to boot some nodes.
- `sra boot` command failing to boot a number of nodes on a large system with the following message on the console log of the nodes that failed to boot:
  `Timeout while talking to CMF`
- `sra delete_member` command failing if a previous add aborted leaving the boot domain pointing at the Tru64 disk swap partition.
- `sra delete_member` command timing out.

  Deleting a member can sometimes take longer than 4 minutes, therefore this timeout has been increased to 8 minutes.
- Problem when deleting a member.

  The fix removes the member's boot file domain before deleting the member.

In addition, the following change was made to SRA:

- Added a confirmation message to the `sra power_off` command.

2.2.23 sramon

The following `sramon` problem was fixed:

- `msqld` unavailable for a brief time during upgrade

2.2.24 Switch Module Firmware — swm-upgrade-02111408

The following switch module firmware problems were fixed:

- `swmserver` logging incorrect clock failures
- Problem with the test used to determine whether a board has been reset

In addition, the following changes were made to the switch module firmware:

- Added support for the watchdog script monitoring the `swmserver`.
- The `cabletest` script now reports cases where a link is shown as being operational (not in `reset`) and yet the link is not transferring data.
- `elanlinktest`: Added signal handler to report DMA progress when aborted.
- `elantest` changed to use word events rather than block events.

3
General Restrictions

This chapter describes the general restrictions that apply to HP AlphaServer SC Version 2.6.

3.1 Booting and Shutting Down Nodes — Reminder

Use the `sra [boot|shutdown]` command to boot or shutdown nodes of an HP AlphaServer SC system, and do not extend the default width. See Chapter 2 of the *HP AlphaServer SC System Administration Guide* for more information.

3.2 CAA Daemon Location

CAA monitors the status of domain members and when necessary will relocate applications or daemons to another member of the domain based on the placement policy within the CAA profiles. Over the lifetime of a domain, the actual location of these applications may change; for example, if member 1 was rebooted at some time, the daemons may now be running on member 2 or elsewhere. This can lead to unpredictable performance, which may adversely affect barrier operations within parallel programs.

To ensure a predictable performance characteristic, you can manually relocate the CAA applications to a defined member (for example, member 0) as shown in the following example:

1. On each domain, run the following command to determine which member is currently running each resource:
   ```
   atlas0# caa_stat -t
   ```
2. On each domain, run the following command to view the placement policy for all registered applications:
   ```
   atlas0# caa_profile -print
   ```
3. To manually relocate a CAA application, run a command as shown in the following example (where `SC15srad` is the application to be relocated on member `atlas0`):
   ```
   atlas0# caa_relocate SC15srad -c atlas0
   ```

3.3 CAA Placement Policy for srad

The CAA placement policy for the `srad` daemon is based on what nodes can "see" the generic boot disk, typically a disk on a RAID system that is connected to both member 1 and member 2. Each time a node boots, it runs the `/usr/opt/sra/bin/srad.caa` script. This script adds or removes the node from the HOSTING_MEMBERS field of the `srad` daemon resources file (`SC15srad`).

To determine the current placement policy, run the following command:

```
# caa_profile -print SC15srad
```

The PLACEMENT entry should be set to `restricted` and the HOSTING_MEMBERS field should contain the first two members of the cluster (assuming these members have direct access to the generic boot disk). If this is not the case, run the `srad.caa` script on members 1 and 2. If this does not update the placement policy, check the `sc_system_devices` table in the SC database. There should be a cluster-specific entry for the `gen_boot` disk, as shown in the following example:

```
sql> select name,role,image,location from sc_system_devices where
sql> name='atlasD1'
name     role      image  location
-------------------------------------
atlasD1  cluster   first  IDENTIFIER=4
atlasD1  gen_boot  first  IDENTIFIER=5
```

Use the `hwmgr -v d` command to verify that the `location` field is correct for the `gen_boot` disk on the specific domain. If you are unsure which disk was used as the generic boot disk, check the `clu_add_member.log` file in `/cluster/admin`, as follows:

```
atlas64# grep "Checking the member boot disk" clu_add_member.log
Checking the member boot disk: dsk6
```

3.4 cfengine Command

The `cfengine` command is present but not supported in HP AlphaServer SC Version 2.6.

3.5 Disabled Daemons

A number of daemons have been disabled in HP AlphaServer SC Version 2.6 as a general performance-enhancement measure. These daemons can be enabled if required. The disabled daemons are listed in this section.

As a result of these daemons being disabled, the following message is displayed after a cluster configuration check is run during boot time:

```
The boottime cluster check found a potential problem.
```

This is followed by another message pointing to a log file. These messages are benign, and may be ignored.

3.5.1 AdvFS GUI

The AdvFS graphical user interface (`dtadvfs`) cannot manage storage on a domain. The AdvFS daemon, `advfsd`, has been disabled.

3.5.2 clu_mibs Daemon

The data items in the Common Cluster Management Information Base (MIB) and the TruCluster Server MIB are not available. The SNMP subagent daemon for TruCluster Server, `clu_mibs`, has been disabled.

3.5.3 Insight Manager Daemon

The Insight Manager is not needed if the system is not to be remotely configured *via* the web. The Insight Manager daemon, `insightd`, has been disabled.

3.5.4 Printing Daemon

Printing functionality is not normally used in HP AlphaServer SC systems. The printing daemon, `lpd`, has been disabled.

3.5.5 SNMP and Environmental Information Monitoring Daemons

SNMP and environmental information monitoring are not widely used in HP AlphaServer SC systems. The `snmpd`, `envmond`, `cpq_mibs`, and `os_mibs` daemons have been disabled.

Note:

When the `envmond` daemon is disabled, an overheated system will still shut itself down by powering off.

3.5.6 SysMan Station Daemon

Sysman Station, `sysman_station(8)`, cannot be used to monitor a domain. The SysMan Station daemon, `smsd`, and the SysMan authentication daemon, `smauth`, have been disabled.

3.5.7 X Windows Login Daemon

HP AlphaServer SC systems do not contain graphics cards, and do not need to run X Windows login. The `dtlogin` daemon has been disabled.

3.6 Do Not Directly Mount or Unmount SCFS and PFS File Systems

The SCFS and PFS file system management system tries to keep the file systems in a consistent state at all times. In addition, this state is maintained across the system — the `ONLINE`/`OFFLINE` state applies to all domains. It is not possible to mark the file systems on a specific domain as `ONLINE` or `OFFLINE`.

3.7 Do Not Rebuild the Kernel Before Creating the Domains

Do not rebuild the kernel after installing the HP AlphaServer SC system software and before creating the domains. This restriction does not apply to the management server.

3.8 Do Not Include Blank Lines in the /etc/routes File

Do not include blank lines in the `/etc/routes` file — this will cause spurious, but benign, errors about `botched keyword`s in the boot sequence.

3.9 DRD: Do Not Increase drd_remote_maxphys

Do not increase the value of the `drd sysconfig` variable `drd_remote_maxphys` — the maximum supported value is the default value (524288).

3.10 GPL Code

The RMS functionality in HP AlphaServer SC Version 2.6 does not have any dependence on General Purpose License (GPL) code. However, if you want the extra capability that would come from using that code, you can access the relevant code at www.quadrics.com.

3.11 IP Over Elan Not Supported

HP AlphaServer SC Version 2.6 software does not support IP over Elan for inter-process communication.

3.12 Lazy Swap is Not Supported

Lazy swap is not supported in HP AlphaServer SC Version 2.6 — use eager swap only.

3.13 Local Area Transport Protocol Not Loaded

Local Area Transport (LAT) is an older protocol that is seldom used. The LAT kernel subsystem is prevented from loading on boot.

3.14 LSF: All LSF Hosts Must be Upgraded to HP AlphaServer SC Version 2.6

HP AlphaServer SC Version 2.6 hosts and HP AlphaServer SC Version 2.5 hosts cannot coexist in the same LSF cluster. You must upgrade all HP AlphaServer SC hosts in the LSF cluster to HP AlphaServer SC Version 2.6.

3.15 LSF: Chunk Job Dispatch Feature Not Supported

The chunk job dispatch feature is not supported in HP AlphaServer SC Version 2.6.

3.16 LSF: LSF Administrator Must Be Known User on Nodes Running LSF Daemons

The LSF administrator must be a known user on the nodes that are running the LSF daemons. This can be accomplished either by configuring Network Information Service (NIS) on these nodes, or by manually adding an entry to the `/etc/passwd` file. If the LSF administrator is not a known user, error messages similar to the following are logged in the relevant `sbatchd.log.`*`nodename`* file:

```
Aug 21 07:54:37 2002 525902 3 4.2 getManagerId: No LSF administrator defined in
sbdPackage
Aug 21 07:54:37 2002 525902 3 4.2 Daemon on host <atlas224> exiting; cause code
<-1> unknown
```

Do not use the `mount(8)`, `umount(8)`, `mount_pfs(8)`, or `umount_pfs(8)` commands to directly mount or unmount SCFS and PFS file systems. If you do so, the state of the system will be inconsistent and the `scfsmgr show` and `pfsmgr show` commands will not reflect the true state of the system. In addition, the system may undo your mount or unmount action if that action does not agree with the `ONLINE`/`OFFLINE` status.

3.17 LSF: Preemption Preferences

The topological preemption preference is the only preemption preference supported in HP AlphaServer SC Version 2.6.

3.18 LSF: Restricted Usage

LSF usage is restricted to batch operations. Specifically, the `lsrun` command is not supported.

The LSF software provided is specifically designed to work with the HP AlphaServer SC system software. Do not use other versions of the LSF software on an HP AlphaServer SC system.

3.19 LSF: "span[ptile=*n*]" Resource String Not Supported

The "span[ptile=*n*]" resource string is not supported for use in the -R option of the bsub command in HP AlphaServer SC Version 2.6.

3.20 Rack Names Must be Populated and Unique

When you are defining racks in the SC database (using the sc_cabinet command), ensure that each rack has a unique label. Do not leave rack labels blank.

3.21 Restriction on Using addvol

The addvol command can only be issued from the node serving the file system in question. This is because the addvol command creates a temporary file in /tmp and then expects this file to be readable by both the node the command is being run on and the server node of the file system. This restriction only applies if /tmp is mounted using the server_only option.

3.22 Readline and History Support Removed from rmsquery

Readline and history support have been removed from the rmsquery command. This functionality can be obtained by using the Quadrics QSQUERY package; this package can be downloaded from the Quadrics website at www.quadrics.com.

Alternatively, you can create a new version of the rmsquery command, as follows:

1. Back up the original rmsquery command, which is located in the /usr/bin directory.
2. Download the Gnu Readline library from the Open Source Software Collection for Tru64 UNIX website at www3.hp.com.
3. Place the libraries for readline and history in the /usr/opt/rms/lib directory, and place the header files for readline in the /usr/opt/rms/include directory.
4. Recompile the rmsquery command by running the following command:
```
cc -o rmsquery /usr/opt/rms/src/rmsquery.c -I/usr/opt/rms/include \
   -L/usr/opt/rms/lib -lmsql -lreadline -lhistory -lcurses -DUSE_READLINE
```
5. Copy the new rmsquery command created in step 4 above to the original rmsquery command in the /usr/bin directory.

3.23 RMS: Support for prun -r and prun -dr Commands

The prun -r and prun -dr commands are not supported in HP AlphaServer SC Version 2.6 when RMS is only installed on the management server.

3.24 Routing: routed is Not Supported

In HP AlphaServer SC Version 2.6, the only supported router is `gated`. You should not disable `gated` on a domain unless every member has an external network interface (and an accompanying static default route).

3.25 Upgrade: Do Not Leave the System Partially Upgraded

When upgrading to HP AlphaServer SC Version 2.6, you must not leave your system partially upgraded, or attempt to complete the upgrade over a long period, domain by domain. A partial upgrade, or upgrade over a long period, will cause PFS in HP AlphaServer SC Version 2.6 to fail, and will also result in problems with RMS.

For these reasons, you must take your PFS file systems offline, and stop RMS, before starting an upgrade. You must also rename the RMS startup script so that non-upgraded domains will not cause any difficulties for the newly upgraded management server during the course of the upgrade. These tasks are explained in detail in Section 4.2 of the *HP AlphaServer SC Installation Guide*.

3.26 Upgrade: Upgrade Shortcut Not Supported

The upgrade shortcut mechanism that was added in HP AlphaServer SC Version 2.5 (UK1) is not supported for upgrading to HP AlphaServer SC Version 2.6.

If you are upgrading to HP AlphaServer SC Version 2.6 from HP AlphaServer SC Version 2.5, you can perform either a full install or a long upgrade.

If you have already installed HP AlphaServer SC Version 2.6 EFT1 or HP AlphaServer SC Version 2.6 EFT2 and are now upgrading to HP AlphaServer SC Version 2.6, you can either perform a complete install or revert back to HP AlphaServer SC Version 2.5 and perform a long upgrade.

3.27 Use /tmp — Not /usr/tmp

`/usr/tmp` is a symbolic link to the `/var/tmp` directory. The `/var/tmp` file system is served by member 1 or member 2 in each domain. If you use `/usr/tmp` or `/var/tmp` in user applications, you may cause the following problems:

- The file system may fill up.
- You will place a large load on the serving node.

Instead of using `/usr/tmp` or `/var/tmp`, use `/tmp`. The `/tmp` file system is located on the internal disk of each node.

3.28 64-Bit Shmem is Not Supported

HP AlphaServer SC Version 2.6 does not support 64-bit Shmem.

For certain Shmem applications, you can overcome the 64-bit addressing issues by changing the LIBELAN_ALLOC_SIZE environment variable from its default value (200 MB). For some Shmem applications, you may need to increase this environment variable, and for others you may need to decrease it — you should investigate both options.

4

Known Problems in HP AlphaServer SC Version 2.6

This chapter lists the problems that have been identified in HP AlphaServer SC Version 2.6.

4.1 Booting: "sed: Function s/-netmask-/255.255.0.0 cannot be parsed" Error

Messages similar to the following may appear on a node's console while it is being booted:

```
sed: Function s/-netmask-/255.255.0.0 cannot be parsed.
sed: Function s/-interface-/ee0 cannot be parsed.
```

These messages are related to the `cfengine` command. As the `cfengine` command is not supported in HP AlphaServer SC Version 2.6, you can ignore these errors.

4.2 Booting: "The boottime cluster check found a potential problem" Error

A number of daemons have been disabled in HP AlphaServer SC Version 2.6 as a general performance-enhancement measure. The disabled daemons are listed in Section 3.5 on page 3–2.

As a result of these daemons being disabled, the following message is displayed after a cluster configuration check is run during boot time:

```
The boottime cluster check found a potential problem.
```

This is followed by another message pointing to a log file. These messages are benign, and may be ignored.

4.3 CMF: Daemon May Need to be Manually Restarted

If mSQL is configured as a CAA service, CMF must also be configured as a CAA service. If mSQL is CAA-enabled and CMF is not CAA-enabled, the CMF service must be manually started after CAA has started mSQL — for example, after a reboot. This is only relevant for an HP AlphaServer SC system with no management server, or for an HP AlphaServer SC system with a clustered management server.

If mSQL is CAA-enabled, ignore the following console message on the clustered management server or Node 0 during boot:

```
CMF: console logging daemon does not run on this host
```

This message appears because the mSQL service is started later in the boot sequence and the CMF startup script cannot correctly determine the CMF host.

4.4 Elan Kernel Comms: Reference Pages Not Updated

The reference pages for Elan Kernel Comms have not been updated for HP AlphaServer SC Version 2.6.

4.5 ICS: Error Message: icssvr_daemon_from_pool: invalid IPL 4 ...

In some cases, the following error message may be displayed in the console log on the importing node:

```
icssvr_daemon_from_pool: invalid IPL 4...
```

If this error is displayed, please contact your HP Customer Support representative.

4.6 Installation/Upgrade: Benign Errors

During installation and upgrade, the Patch Kit 3 patches for Tru64 UNIX Version 5.1B are configured on the non-lead nodes when they are booting to join their domains for the first time. On some of these non-lead nodes, you may see messages similar to the following:

```
Configuring "Patch: Removes compiler warnings" (OSFPAT00068700540) on member4
cp: /listpatches.tmp2: No such file or directory
cp: /listpatches.tmp2: can't propagate property list
```

These messages are benign and may be ignored.

4.7 Installation/Upgrade: clu_quorum Check

When upgrading a domain, the `clu_quorum` check is issued and must complete successfully before proceeding with the actual upgrade. If `clu_quorum` does not complete, please contact your HP Customer Support representative for assistance.

4.8 Installation/Upgrade: Problem Installing Patch Kit 3

When you are installing Patch Kit 3 for Tru64 UNIX Version 5.1B, you may see the following sequence of error messages and behavior:

```
/usr/sbin/setld: /sbin/tar: arg list too long
tar: No input
ls: *.ctrl not found
.
.
```

.
```
awk: Input line  OSFPAT00113300540 O cannot be longer than 3,000 bytes.
```

The `dupatch` script then exits without installing any patches.

To resolve this problem, disable argument limits by running the following command:

```
atlasms# sysconfig -r proc exec_disable_arg_limit=1
```

After you have run this command, reinstall Patch Kit 3.

For more information, see the *HP AlphaServer SC Installation Guide*.

4.9 Installation/Upgrade: TCP/IP Ephemeral Range

The TCP/IP ephemeral range may need to be moved on a management server. Many of the HP AlphaServer SC daemons use well-known ports, and these ports are in the ephemeral range. The ephemeral range on a management server should be moved to avoid the potential risk of a socket already being in use when a system daemon is attempting to start up.

4.10 Install: Cannot Install Certain Layered Products

Installing certain layered products prior to domain creation can adversely affect the automated installation process, because the product configuration phases interact with the user. Legato NetWorker is an example of such a product. You should install such layered products after you have completed the post-installation tasks described in Chapter 10 of the *HP AlphaServer SC Installation Guide*.

4.11 Install: ifaccess.config File Set Up Only for Head Node in Domain

By default, the `ifaccess.conf` file is set up only for the head node of a domain (that is, member 1) at installation time. If other domain members have external network connections, the `ifaccess.conf` files for these members must be set up manually, as described in the "Updating ifaccess.conf When Changing Network Interfaces" section of Chapter 21 of the *HP AlphaServer SC System Administration Guide*.

4.12 Install: Possible Errors Seen with Tru64 UNIX Patch 1276.00

After installing Tru64 UNIX Patch 1276.00, the following problems have been known to occur:

- The Common Data Security Architecture (CDSA), IP Security Protocol (IPsec), or Single Sign-On (SSO) do not work.

- The following error message is displayed during boot time:
  `CSSM_ModuleLoad: CSSM error 4107`

If you experience these problems, make sure that the following command has been executed:

```
# /usr/sbin/cdsa/mod_install -f -i -s /usr/lib/cdsa/libt64csp.so -d /usr/lib/cdsa/
```

4.13 I/O May Hang on HSG80 in Multibus Mode

When configuring an HSG80 RAID controller, you can set parameters for the connection (each route to a controller from a Host Bus Adapter (HBA)). In multiple-bus failover mode, the default host type is WINNT. When a controller connection is thus configured, the controller ports are seen as 4 active and 0 standby, instead of 2 active and 2 standby.

You can see this by entering the following command:

```
# hwmgr –get attr –id HWID
```

If all four ports are active, a hang may occur. To avoid this, you MUST set each connection type to TRU64_UNIX. You must set this for EACH connection.

If you do not have a serial connection, use the `/usr/lbin/hsxterm5` command to provide CLI access from UNIX. You can use this command to script the setting of connections.

Note:

New connections are created if you unplug and replug any fibre.

4.14 Kernel: Nodes May Crash with “ics_elan: timeout on free tmds” String

Nodes may crash with the following panic string:

```
cmn_err: CE_PANIC: ics_elan: timeout on free tmds 0x0
```

If this occurs, please file a problem report with pointers to this crash file and any other crashes that occur on the domain at the same time.

4.15 Kernel: Nodes May Crash with “malloc_internal: invalid link address may occur” String

Nodes may crash with the following panic string:

```
malloc_internal: invalid link address may occur.
```

The most likely reason for this panic is that the number of cluster aliases has reached the limit defined by the `max_aliasid` attribute in the `clua` subsystem. If, for example, the cluster aliases are moved from one subnet to another, the number of cluster aliases effectively

doubles, as both the original IP addresses and the new IP addresses are stored (and will be reported by the `cluamgr -s all` command). In such a case, the limit defined by the `max_aliasid` attribute may be reached.

To avoid this panic, increase the value of the `max_aliasid` attribute, so that the limit will not be reached.

To recover from this panic, reboot the node. You should also file a problem report with pointers to this crash file and any other crashes that occur on the domain at the same time.

Note:

It is possible that a node may crash with this panic string when the `max_aliasid` limit has not been reached. If this occurs, you should also reboot the node and file a problem report with pointers to this crash file and any other crashes that occur on the domain at the same time.

4.16 LSF: bjobs -l Times Out

The `bjobs -l` command times out when executed from nodes with no external network interfaces.

4.17 LSF: Delay Between Resource Allocation and Job Start-Up

Sometimes there is a delay between the time LSF allocates an RMS resource on a domain and the start time of the RMS job. Other than this delay, the job will run and complete normally. This delay is caused by CPU-intensive application programs running on the node (usually member 1) on which the LSF daemons run.

If you frequently encounter this problem, contact your HP Customer Support representative for a solution.

4.18 LSF: Imprecise Pending Reason When the Node Required by a Job is Configured Out

If slot reservation is enabled, a job that specifies a configured-out node in the RMS `base` topology option (to specify the starting point of an RMS allocation) remains pending with the following reason:

```
Not enough free cpus to satisfy job requirements
```

4.19 LSF: Reservation and User Limits

When LSF user limits are set, LSF reserves fewer slots than user limits allow.

4.20 LSF: Reservation for Jobs with [hosts=1]

A job with the `[hosts=1]` locality span string does not reserve slots if multiple partitions exist.

4.21 LSF: Slow Update of MXJ=!

When an LSF cluster is started, the value of the maximum job slots indicator (MXJ) on some HP AlphaServer SC hosts may be set to `infinite`. It may take a few minutes before the MXJ value changes to `!`.

4.22 MPI: Invalid DMA May Be Seen With Onset of Paging

For MPI `AllToAll` operations at higher process counts, invalid Direct Memory Accesses (DMAs) may be reported by the `libelan` libraries, leading to the applications being aborted. The problem is related to the onset of paging associated with comms buffers having increased from 2GB to 4GB to handle the process count. The workaround is to tune the application to a level where the memory usage for the node will remain below the threshold that will cause applications to page. Therefore, if problems are experienced on systems with 4GB per node, the application should be tuned to a memory usage below approximately 3400MB per node.

4.23 MPI: Large Number of MPI Communicators/Not Enough Global Memory

A problem can occur when an MPI job creates a large number of MPI communicators. A new ELAN_GROUP is created for each MPI communicator. These groups require large amounts of global memory, which has to be pre-allocated at program startup. If there is not enough global memory to create a new group, a message similar to the following is displayed:

```
ELAN_EXCEPTION @ 1: 5 (Memory exhausted)
elan_createSubGroup(): Failed to allocate global Vaddr for subgroup
```

The message may relate to global `vaddr` space, global main memory or global elan memory.

There are several possible solutions to this problem. In most cases, the best solution is to revisit the application and make sure that it really does need all the communicators it creates. If you cannot reduce the number of communicators being created, any one of the following solutions should enable the job to run:

- Disable the use of subgroups within MPI. This solution works in all cases, but also results in the collectives routines no longer being optimized.

  To disable the use of subgroups within MPI, either set the environment variable MPI_USE_LIBELAN_SUB=0, or use the `--elan-coll-sub 0` command line option to any MPI program.

- Decrease the amount of global memory that groups use, by setting either the LIBELAN_GROUP_RBUFSIZE environment variable or the LIBELAN_GROUP_CBUFSIZE environment variable. These variables default to 64KB and 4KB respectively.

- Increase the amount of global memory that is reserved at program startup by setting the LIBELAN_GALLOC_SIZE environment variable, which defaults to 64MB. As the global address space is allocated from the elan3allocator, you should also increase the value of the LIBELAN_ALLOC_SIZE environment variable, which defaults to 200MB.

4.24 MPI: LIBELAN_WAITTYPE Variable Must Be Set to “POLL”

By default, the LIBELAN_WAITTYPE variable is set to POLL. Do not set this variable to a different value — some applications may hang if this setting is changed.

4.25 MPI: Sending Large Messages Without Buffering

Very often in MPI applications, very large messages are sent between MPI processes. If the application does not arrange any buffering itself (using `MPI_Bsend()` calls), any needed buffering must be carried out by the system software (the Elan software).

The Elan software uses the notion of a “biggest message size” to decide whether the system buffers should be relied on to buffer a specific message, whenever a matching receive is not synchronously posted. A “biggest message size” of 64KB is the default, but this can be increased or decreased at runtime by setting the LIBELAN_TPORT_BIGMSG environment variable.

For message sizes that exceed the “biggest message size” limit, buffering is not permitted and the Elan software requires that the send and receive processes synchronize with one another. For some applications, this synchronization process can cause the application to hang: each process is in either an unmatched `MPI_Send()` or unmatched `MPI_Recv()` call that can never complete. To fix this, set LIBELAN_TPORT_BIGMSG to a value greater than 64KB.

Increasing LIBELAN_TPORT_BIGMSG beyond the default value of 64KB can itself have unforeseen consequences. Now that the Elan software is buffering larger messages, it may run out of buffer storage space, as indicated by an error message similar to the following:

```
elan3_alloc: Exhausted ALLOC 452f10 type 71434240 allocator space (0) : wanted
209707008 bytes
```

By default, the amount of buffer storage space is set at 200MB. However, this can be reset at runtime by setting the LIBELAN_ALLOC_SIZE environment variable.

4.26 Networks: Configuration of Default Gateway on Peripheral Hardware

The management network of the HP AlphaServer SC system contains many peripheral devices with TCP/IP address configuration; for example, terminal servers, Fibre Channel switches, Ethernet switches, HP AlphaServer SC Interconnect control processors, and SANworks Management Appliances.

The *HP AlphaServer SC Installation Guide* recommends that the default gateway be configured as `10.128.0.1`, which corresponds to the first node of the first domain. This configuration is well suited to an HP AlphaServer SC system with no management server or an HP AlphaServer SC system with a standalone management server.

However, for an HP AlphaServer SC system with a clustered management server, you must configure the default gateway differently. For such systems, you should set the default gateway on peripheral devices to be the management network address of one of the members of the clustered management server; for example, `10.128.101.1`.

This is necessary to guarantee that communication with such devices will work correctly with the cluster alias of the clustered management server, during the installation and lifetime of the system.

4.27 Networks: Quiet Management Networks can be Misinterpreted by niffd

When installing an HP AlphaServer SC system with no management server or an HP AlphaServer SC system with a clustered management server, the management Ethernet network will have limited traffic initially — all other nodes will be at the SRM prompt. Once you have created a single-node `atlasD0` domain (or a single-node `atlasms` cluster), the `niffd` daemon may falsely detect that the `ee0` interface is `down` and deconfigure the interface. A message will be written to the console log on the affected node. The workaround is to manually boot a second node in the cluster — the `niffd` daemon will no longer think the network is idle.

The misinterpretation will not happen if there is any other traffic on the management network; for example, terminal-server traffic, and so on. However, some smaller systems and test systems may have terminal servers on the external network, so that the terminal servers can be shared with other corporate equipment.

4.28 NFS: “Stale File Handle” or “Cannot Overwrite Existing File” Errors

In certain circumstances, an NFS client may encounter `Stale File Handle` errors for a file system that is NFS (SCFS) served by either a domain or a clustered management server, even though sufficient time has passed so that any cached information should have been invalidated. Fortran applications may report this as a `Cannot overwrite existing file` error.

This problem may occur if the files in question have been removed by a command running in the NFS (SCFS) serving domain, on a node other than the node that is the `cfsmgr` for the served file system.

The following example shows how an application might see this problem:

1. An application spanning multiple domains, including a file-serving (FS) domain, writes out checkpoint data, with one checkpoint file per process.
2. Process 0, which is located on a node in the FS domain but not on the node that is the `cfsmgr` for either the SCFS file system or the root component of the PFS file system, collects the checkpoint files into an archive and removes the checkpoint data files.
3. The application continues and checkpoints again some time later, at which point the process on the client domains might encounter one of the above errors.

If you encounter this problem, contact your HP Customer Support representative for details on how to work around this issue.

4.29 Performance Visualizer: pvis Command Causes Stack Trace

Occasionally, the `pvis` command can cause a stack trace when accessing older versions of the `$HOME/.pvis` initialization file, as shown in the following example:

```
# pvis -setup
# pvis
Error in startup script:
    while executing
"ppm_cluster_set ${type}_updateperiod $period"
    (procedure "periodicChangeType" line 36)
    invoked from within
"periodicChangeType disks $gRefresh_update_period(disks)"
    (procedure "Refresh_update" line 5)
    invoked from within
"Refresh_update"
    (file "/usr/bin/pvis" line 825)
#
```

The solution is to delete the `$HOME/.pvis` file and allow the `pvis` command to recreate it.

4.30 Performance Visualizer: pvis Command Not Removed

The `pvis` command has not yet been removed. If you have used `scpvis` and then try to use `pvis`, it will fail. Delete the `$HOME/.pvis` file and try the `pvis` command again.

4.31 PFS: Can Create, and Write to, Files on a Full PFS File System

When creating a file, PFS randomly selects the base component file system that stores the start of the file. As long as there is sufficient space on this component file system or subsequently accessed components to store that data written to the file, then writes to this file will succeed. However, the `df` command will report the file system as full, and any write that resolves to a full component file system will fail with ENOSPC (no space left on device).

4.32 PFS: Changing the Default File Mapping May Not Take Effect Globally

If you use the PFSIO_SETDFLTMAP ioctl call to change the default file mapping for a PFS file system, this change may not take effect immediately. To ensure that all nodes see the change, either perform the PFSIO_SETDFLTMAP ioctl call on every node that has the PFS file system mounted, or unmount and remount the PFS file system after making this change.

See Chapter 8 of the *HP AlphaServer SC System Administration Guide* for details on how to use the `pfsmgr` command to unmount and remount PFS file systems, or for information about the PFS ioctls.

4.33 PFS: Do Not Delete an Active PFS Mount Point

Deleting an active PFS mount point may lead to an inconsistent PFS environment. If a PFS file system is actively mounted on other nodes within a domain, but is not mounted on the current node, it is possible to delete the PFS mount point. After doing so, applications cannot access any PFS files on any node within the domain, as the path to the PFS does not exist.

To recover from this situation, perform the following tasks:

1. Use the `pfsmgr offline` command to unmount the PFS on all members of the domain.
2. Re-create the mount point.
3. Use the `pfsmgr online` command to remount the PFS on all members of the domain.

4.34 PFS: Do Not Mount a PFS on a PFS Directory

It is not possible to mount a PFS file system on a mount point that resides on a PFS file system.

4.35 PFS: File Locking is Not Supported

File locking is not supported on PFS file systems.

4.36 PFS: Link Operations May Fail on a PFS File System

If you expect to compile and link programs on a PFS file system, set up the PFS file system with a default file stripe count of 1. The PFS file system itself will use multiple component file systems, but only one component file system will be used per file.

4.37 PFS: PFS File System Cannot be Deleted if SCFS is Broken

A PFS file system cannot be deleted until all underlying component file systems are mounted by at least one domain. If one of the underlying SCFS file systems cannot be mounted (for example, because the underlying AdvFS domain has been deleted), the `pfsmgr delete` command cannot be used to delete the PFS.

The process described in this section can be used to delete the PFS manually. The example assumes the PFS set is called `pdata`, mounted at `/pdata` with components `/c1/a`, `/c1/b`, `/c2/c`, `/c2/b`, where `/c1` and `/c2` are SCFS file systems. The example assumes that `/c2` is the broken SCFS and that you cannot mount it.

1. Get the PFS set name for the `/pdata` mountpoint, as follows:

```
# rmsquery "select pfs_set from sc_pfs where mountpoint='/pdata'"
pdata
#
```

2. Get the names of the PFS components, as follows:

```
# rmsquery "select component_fs_path from sc_pfs_components where \
pfs_set='pdata'"
/c1/a
/c1/b
/c2/c
/c2/d
#
```

3. Place the underlying SCFS file systems online, as follows:

```
# scfsmgr online /c1
```

4. When the file system is mounted, delete the contents of the component directories listed from step 2. In this example, this is `/c1/a` and `/c1/b`. You cannot delete `/c2/c` or `/c2/d`, as `/c2` is not mountable.

5. Delete records related to the PFS from the database, as follows:

```
# rsmquery "delete from sc_pfs_filesystems where pfs_set='pdata'"
# rsmquery "delete from sc_pfs_components  where pfs_set='pdata'"
# rsmquery "delete from sc_pfs  where pfs_set='pdata'"
# rsmquery "delete from sc_pfs_mount  where pfs_set='pdata'"
```

The PFS file system is now deleted. Later, if you manage to recreate the broken SCFS file system, make sure that all component directories are empty before attempting to recreate the PFS file system.

4.38 PFS: pfsmgr show Displays “new” State

The `pfsmgr show` command may show that the state of the PFS file system is `new` instead of `online` or `offline`. This happens in the following situations:

- Someone else is running the `pfsmgr create` command.

  This is a transitory situation.

- You aborted a previous `pfsmgr create` command while it was in the middle of its operations.

  This situation will persist until you correct the problem by creating a PFS of the same name. Normally, you cannot create another PFS with the same name; however, in this situation, the `pfsmgr create` command will work.

  If you are unable to use the same name, delete the relevant entries from the `sc_pfs`, `sc_pfs_components`, and `sc_pfs_filesystems` tables in the SC database. If you only delete a record from the `sc_pfs` table, the `pfsmgr show` command will work, but a subsequent `pfsmgr create` command may fail.

4.39 PFS: ro (read-only) Access is Not Supported

A PFS file system created with the `ro` (read-only) access mode specified is still writable. The `rw` field in the `sc_pfs` table contains the value `rw`, regardless of the access mode specified in the `pfsmgr create` command. At present, no workaround exists for this problem, as it is not possible to create a PFS file system using read-only SCFS components. Also, although it is possible to change the underlying SCFS components to be read-only after the PFS file system has been created, if you do so, the PFS file system will subsequently fail to mount.

4.40 PFS/SCFS: Mounts and Unmounts Rely on scrun Command

The SCFS and PFS file-system management system relies on the `scrun` command. If the `scrun` command fails on a member of a domain, the PFS file system on that member will fail to mount or unmount. In addition, the failure of the `scrun` command can affect the underlying component SCFS file systems. Before mounting an SCFS file system, the system checks whether the PFS is already mounted — it uses the `scrun` command to perform this check. If the `scrun` command fails on any node, the check cannot be performed. The SCFS file system is not mounted. As the SCFS file system is not mounted, the PFS file system is also not mounted.

When file systems fail to mount due to `scrun` failures, events are posted. Since failure of the `scrun` command can affect several file systems, you will see events for each individual file system. However, only one event is posted to report the failure of the `scrun` command. For example, the following events show how a failure of the `scrun` command causes the failure of a PFS and its two underlying SCFS component file systems:

```
08/23/02 14:27:57  atlas32      scfs            mount.failed
Must not mount /d_comp1: failed to unmount pfs_fileset

08/23/02 14:27:57  atlas32      scfs            mount.failed
Must not mount /d_comp2: failed to unmount pfs_fileset

08/23/02 14:27:57  atlasD1     scfs            script.error
scrun failed: atlas35

08/23/02 14:28:00  atlasD1     pfs             comp.not_mounted
cannot mount pdata: component /d_comp1 is not mounted
```

4.41 PFS/SCFS: Mount Status Not Updated After Shutdown

The mount status of PFS and SCFS file systems may not be automatically updated when nodes or domains are shut down or crash. Run the `scfsmgr sync` command to force an update of the current status.

4.42 RMS: Detection and Reporting of PSU or Fan Failures on HP AlphaServer ES45 Nodes

On HP AlphaServer ES45 nodes, the environmental monitoring daemon will only report a failure affecting the last remaining fan or power supply units (PSU). As a consequence, RMS will not log any change in the operational status of the cooling fans or power supplies, or generate environment events for single fan or PSU failures.

4.43 RMS: Event Manager Exits Due to Low Memory Availability

On management servers with 256MB memory, the event manager can occasionally exit when failing to `malloc` memory while processing new events.

The problem is exacerbated when there is a high node count, or if RMS is stopped on the management server for some time while events queue up from nodes.

The workaround is to delete unhandled events, by running the following commands:

```
# rcontrol stop server eventmgr
# rmsquery "update events set handled=1"
# rcontrol start server eventmgr
```

Once the unhandled events are gone, the event manager should be able to run without problems unless there is a barrage of new events.

The ultimate solution is to increase the physical memory on the management server — 1MB should be considered to be the minimum value and a good investment.

4.44 RMS: “failed to allocate elan capability” Error

Some capabilities of a process may not be removed when the process dies. This problem manifests itself as jobs failing to start, with the following message:

```
failed to allocate elan capability
```

4.45 RMS: prun and Non-Global File Systems

When you run a program though `prun`, it expands the executable pathname and uses that name everywhere. In addition, it attempts to run the program in the same working directory as that in which the `prun` itself is running (that is, it changes to the same directory). Where the current working directory is a globally mounted file system, this does not create any problems.

However, if the file system is not globally mounted, several problems may occur:

- If the executable is in the non-globally mounted file system, RMS will fail to find it and will be unable to execute it.
- RMS fails to change (`cd`) to the working directory.
- The program attempts to read or write data files that are not present on all nodes.

This issue affects any non-globally mounted file systems. For example:

- `/local` — this is a Context-Dependent Symbolic Link (CDSL) to a locally mounted file system. Within a domain, this is mounted by each node, but between domains two problems can occur:
  - The file system is not present on every domain (this happens when each domain has a different number of nodes).
  - The file (executable or data file) is not present on each domain.
- `/tmp` — this is a CDSL to a locally mounted file system. The same problems can occur as for `/local`.
- NFS file systems — the following problems can occur:
  - The file system is not mounted by each domain. Typically, the program will fail to start or fail shortly after starting.
  - The NFS file system is unresponsive. Typically, the program will appear to hang.

- /usr — although /usr is present on all nodes, each domain has its own copy. If a software product is not installed on all domains, prun will be unable to execute the program on those nodes on which the product is not installed.

4.46 RMS: prun Fails with “failed to get capability” Error

Occasionally, prun operations can fail with the following message:

```
prun: Error: loader 2 failed to start its processes
```

If running in verbose mode, prun will report the following message:

```
loader[atlas2]: failed to get capability
```

At this point, the console log for the relevant node will report the following message:

```
Unable to obtain requested swap space
```

These error messages appear because RMS is not cleanly reporting the ENOMEM error situation. The solution is to reduce the memory requirements of the application or increase the available swap space.

4.47 RMS: prun Fails with “failed to create capability: Invalid argument” Error

This error may occur on a dual-rail system. One cause of this error is that the second HP AlphaServer SC Elan adapter card on one node within the RMS partition has ceased to function correctly — perhaps because of a disconnected cable. The rmsd daemon on the node will only scan the HP AlphaServer SC Elan adapter card status on daemon startup; therefore, the partition believes that there are still two rails on all nodes. The problem can be localized by using the railmask parameter and running jobs on individual nodes in the partition. Alternatively, following a stop and restart of RMS on the nodes in the partition, the rinfo -nl command will correctly report the number of working rails on each node in the partition.

The default-railmask attribute specifies a default value for RMS allocations that do not specify a railmask. If a rail fails, you should modify the default-railmask attribute so that the failed rail is not used. For more information about the default-railmask attribute, see the *HP AlphaServer SC RMS Reference Manual*.

Another potential cause for this error is a situation where one node in the system considers the size of the interconnect to be different to the majority decision of the other nodes. A possible cause for this is where there has been a recent hardware problem with the broadcast tree route through the interconnect. You should first isolate the offending node with prun for reducing sets of nodes, and then restart rmsd on the affected node. If the problem persists,

you should run the `elan_level_test` diagnostic to test all levels within the affected interconnect node-level chassis. If the installation is a federated one, you should also test the uplinks to the top-level chassis.

4.48 RMS: rmsd Memory Size

Faulty elan cards may cause the virtual memory (VSZ/RSS) used by `rmsd` to increase abnormally. If this occurs, you should run elan diagnostics and investigate the hardware issue.

4.49 scevent: <EvmITEM_CLUSTER_NAME> in NFS Events

Sometimes, the domain name is missing in NFS events — instead, the `name` field contains the string `<EvmITEM_CLUSTER_NAME>`.

Frequently, there is a CFS event just after the NFS event, associated with the same transition (for example, an NFS mount is followed by CFS reporting the node that serves the file system). The CFS event usually has the correct domain name.

4.50 SCFS: File System Status May Be Set To “Unknown”

A problem with one file system can affect the status information of all other file systems on the domain.

If any one step of an `scfsmgr sync` command times out, the status of all SCFS and PFS file systems on that domain is set to `unknown`. This is not necessarily an indication that all the file systems are down (unmounted). You can manually check whether file systems are mounted by verifying whether the corresponding mount points are present and contain data. When you have identified the file system or systems that failed to mount, and corrected the problems that may have caused the failure, you can use the `scfsmgr sync_domain` command to mount the file system.

The following example shows this procedure for a system that has three domains, `atlasD0` (an FS domain), `atlasD1`, and `atlasD2`:

1. Check whether the mount points for each file system are present and contain data on each domain, as shown in the following example, where `scfs1` is an SCFS file system:

```
atlasms# scrun -d all "mount |grep /scfs1"
atlasD0: domain1#scfs1 on /scfs1 type advfs (rw)
atlasD1: atlasD0-atlas1:/scfs1 on /scfs1 type nfs (v3, rw, udp, hard, intr)


atlasms# scrun -d all "ls /scfs1"
atlasD0:  .tags
atlasD0:  quota.group
atlasD0:  quota.user
```

```
atlasD1:  .tags
atlasD1:  quota.group
atlasD1:  quota.user
atlasD2: exited with status 1
```

Run these commands for each file system.

If a mount point is not present or is empty on a domain, this indicates that the file system failed to mount failed on that domain; in this example, the output indicates that the `scfs1` file system may have failed to mount on `atlasD2`.

2. Try to force the file system to mount in the domain where the mount failed, as follows:

```
atlasms# scfsmgr try_mount atlasD2 /scfs1
```

If the command executes successfully, proceed to step 3.

If the command fails, examine the output of the command to identify any possible reasons for the failure, then correct any possible problems.

3. Run the following command to correct the status of all file systems on the domain:

```
atlasms# scfsmgr sync_domain atlasD2
```

4. Verify that the file system has mounted successfully, as follows:

```
altasms# scrun -d atlasD2 "ls /scfs1"
atlasD2:  .tags
atlasD2:  quota.group
atlasD2:  quota.user
```

4.51 SCFS: File System May Not Mount If File Server in Single-User Mode

If the preferred SCFS file server is up, other viable file servers will not try to mount the file system. However, this may be a problem if the preferred server is only booted to single-user mode.

4.52 SCFS: Fortran Formatted I/O May Perform Poorly for FAST File Access

If an application uses Fortran formatted I/O for data files that are opened for FAST access on an SCFS file system, the performance may be very poor. This is because Fortran formatted I/O statements can generate I/O operations with very small sizes, which do not work efficiently on an SCFS file system. In many cases, performance can be significantly improved by using buffered I/O, as follows:

- Recompile the application with the `-assume buffered_io` compiler option.
- Specify the `BUFFERED='YES'` flag when opening the file in the Fortran code.

See the Fortran compiler documentation for the implications of using buffered I/O within your application.

4.53 SCFS: May Require Reboot to Release Kernel Resources for Reuse

If you create and destroy a number of SCFS file systems on a file-serving node, a reboot may be required to release some kernel resources before running MPI jobs.

4.54 SCFS: scfsmgr export or scfsmgr destroy Print Error Message Twice

If either the `scfsmgr export` command or the `scfsmgr destroy` command fails, the resulting error message is printed twice.

4.55 SCFS: SysMan Menu Refresh Button Behaves Incorrectly

When using SysMan Menu to manage SCFS, the Refresh button changes the currently selected item.

4.56 SC Monitor: Cannot Monitor HSV110 Cells with Multiword Names

If the name of a cell within the HSV110 RAID system is composed of more than one word, SC Monitor will interpret it as two different cells and will fail to monitor that cell.

For example, a cell named `SCHV01` will be monitored by SC Monitor. However, if the cell is named `SCH V01`, SC Monitor will fail to monitor that particular cell within the HSV110.

4.57 SC Monitor: HSG Must Have CCL

If the Command Console LUN (CCL) of an HSG80 RAID system is not visible as an `scp` device, the `scmonmgr detect` process does not find the HSG device; therefore, the HSG device will not be scanned.

The solution is to make sure that the CCL is visible as an `scp` device. To do so, you must set the Small Computer System Interface (SCSI) version of the HSG controllers to SCSI-3.

The following example shows how to do this, if you know the disk device name:

```
# /usr/sbin/hwmgr -view devices
HWID:    Device Name         Mfg                 Model        Location
 ------------------------------------------------------------------------------
    :
    :
    :
 150:    /dev/disk/dsk14c    DEC                 HSG80        bus-1-targ-0-lun-2
```

```
      .
      .
      .
# /usr/lbin/hsxterm5 -F /dev/disk/dsk14c 'SET THIS SCSI_VERSION=SCSI-3'
# /usr/lbin/hsxterm5 -F /dev/disk/dsk14c 'SET OTHER SCSI_VERSION=SCSI-3'
```

Restart both controllers and run the `hwmgr -scan scsi` command so that the `scp` device is created. You must then rerun the `scmonmgr detect` command.

4.58 SC Monitor: HSV RAID Systems May Not be Monitored

The HSV element manager on SANworks Management Appliance sometimes fails to respond. This means that the HSV RAID system will not be monitored.

When this happens, the monitor status of the HSV is `stale`, but the monitor status of the associated SANworks Management Appliance is `normal`.

4.59 SC Monitor: May Not Mark Objects As Being Stale

Normally, if a node is shut down, SC Monitor sets the monitored status of all objects being monitored by that node to `stale`. The monitor status is shown by the `scviewer` and `scmonmgr` commands. However, if you have a redundant management server, SC Monitor fails to set the status of its monitored objects to `stale` — instead, the status remains `normal`.

4.60 SC Viewer: “Can't find class p3_desktop.Desktop” Error

This error indicates that the DISPLAY environment variable may not be set.

4.61 SC Viewer: Elan Status May Be Incorrect

The presence of an HP AlphaServer SC Elan adapter card in the `elans` database table is only tested by the `rmsd` daemon at startup. In a new network configuration, the `elans` table is empty — subsequent population by SC Viewer goes unnoticed until the `rmsd` daemon restarts.

You should restart the `rmsd` daemon on all nodes, to ensure that SC Viewer reports the correct status of the HP AlphaServer SC Elan adapter cards.

4.62 SC Viewer: “/usr/bin/scviewer: cd: too many arguments” Error

If you try to run SC Viewer while there is an error in your `.cshrc` file, SC Viewer will not start, and you will see at least two error messages. One or more of these messages will be from the `.cshrc` file. The final message is:

```
/usr/bin/scviewer: cd: too many arguments
```

To correct this problem, delete the lines that are in error from your `.cshrc` file; SC Viewer will then run correctly.

4.63 Shutting Down: Do Not Use “shutdown -sh now”

HP does not recommend using the `shutdown -sh now` command. Use the `sra shutdown` command or the `shutdown -h now` command instead.

4.64 SRA: Installing on a non-RAID Cluster Disk and Generic Boot Disk

The installation process uses the disk location identifier, rather than the UNIX disk name, to identify the RAID cluster disk and the RAID generic boot disk. The identifier information is stored in the SC database. The default identifier values for the cluster disk and generic boot disk are IDENTIFIER=1 and IDENTIFIER=2 respectively.

If your HP AlphaServer SC system does not use RAID for the system storage, remove these identifiers using the `sra edit` command — press Return to enter a blank value when prompted to enter a new value for the identifier, as shown in the following example:

```
atlasms# sra edit
sra> sys
sys> edit image cluster-first

Id   Description                                Value
------------------------------------------------------------------
[0  ] Image Role                                 cluster
[1  ] Image name                                 first
[2  ] UNIX device name                           dsk3
[3  ] SRM device name                            #
[4  ] Disk Location (Identifier)                 IDENTIFIER=1
[5  ] root partition size (%)                    5
[6  ] root partition                             b
[7  ] usr partition size (%)                     50
[8  ] usr partition                              g
[9  ] var partition size (%)                     45
[10 ] var partition                              h
------------------------------------------------------------------

Select attributes to edit, q to quit
eg. 1-5 10 15
```

```
probe = probe for value

edit? 4
Disk Location (Identifier)              [IDENTIFIER=1]
new value?

Disk Location (Identifier)              []
Correct? [y|n] y
sys>
```

You can also remove these identifiers by using the `sra setup` command — see Section 5.2 or Section 6.2 in the *HP AlphaServer SC Installation Guide*.

4.65 SRA: No Sanity Check on the sra shutdown Command

The `sra shutdown` command does not check whether you are shutting down nodes in a logical order. For example:

- It allows you to shut down the node on which you are running the `sra` command.
- It allows you to shut down the node that serves `sra` executables and configuration files.

Both of these examples mean that the `sra` command itself would be terminated before it had completed operating on other nodes. See Chapter 2 of the *HP AlphaServer SC System Administration Guide* for more information on how to correctly shut down the system.

4.66 SRA: Output Uses the Term "Cluster"

In the HP AlphaServer SC documentation set, clusters of nodes within HP AlphaServer SC systems are called "domains" (except in the case of clustered management servers). However, the terms "cluster" and "domain" are used interchangeably in output and screen dialog for the `sra` command.

4.67 SRA: Rerunning the sra setup Command

If you rerun the `sra setup` command and choose to reuse the existing database but change the number of boot disks, the partition sizes remain unchanged — for example, if you reduce the number of boot disks to one (so that you no longer configure an alternative boot disk), swap will remain at 15% instead of increasing to 30%.

As a workaround, use the `sra edit` command when the `sra setup` command has completed: use the `sys` menu to `edit image boot-first` or `edit image boot-second` or both, and change the swap, tmp, and local partitions to suit your preference.

4.68 SRA: sra cookie Command Is Only Accurate on rmshost

The `sra cookie` command displays the current state of the `msqld` cookie mechanism. However, this command is only accurate when run on the node running the `msqld` daemon (that is, the `rmshost` system). On other nodes in the system, it will report that the cookies are disabled. Apart from the misleading report, there are no side effects to this problem — cookies are indeed enabled on these nodes once cookies are enabled on the `rmshost` system.

4.69 SRA/Sysman: Updating the /etc/hosts File

When using the `sra edit` command to update the `/etc/hosts` file, the `update hosts` subcommand modifies only the section between `#sra start` and `#sra end`.

Several `sysman` commands also modify the `/etc/hosts` file. However, the `sysman` command moves all comments to the end of the `/etc/hosts` file. If you subsequently update the `/etc/hosts` file using the `sra edit` command, the `update hosts` subcommand cannot find the existing entries and so creates new entries.

HP recommends that you do not use the `sysman` command to update the `/etc/hosts` file — either use the `sra edit` command as described in Chapter 16 of the *HP AlphaServer SC System Administration Guide*, or edit the file manually.

However, there are other `sysman` commands that modify the contents of the `/etc/hosts` file, for which no such "non-`sysman`" workaround is available (for example, configuring the IP address on an additional NIC card on a node). HP recommends that you keep a copy of the `/etc/hosts` file, and use this copy to manually reorganize the contents of the original `/etc/hosts` file after a `sysman` command has altered it.

4.70 Sysman: Display Error When Configuring NIC On Other Node

When you configure a network interface card for a node other than the node on which Sysman is running, Sysman sometimes shows the network devices for the Sysman node.

4.71 TotalView: Defining the Path

You must put the path for TotalView in your `.cshrc` file (if using `csh`). It is not sufficient to define the path in the shell or in the `/etc/csh.login` file when you launch TotalView.

4.72 TotalView: Zombie Resources

When exiting from TotalView, defunct processes can be created that do not exit cleanly, and possibly retain an elan context. This causes resources to go into a `zombie` state, and the CPUs to remain allocated.

4.73 TotalView: TotalView Client May Not Communicate with TotalView Server

The TotalView server runs on the compute node. The TotalView client is the interactive GUI. The TotalView client may not communicate with the TotalView server; this problem only occurs if you are running the TotalView client on the management server and one or more of the domains does not have an Ethernet interface onto the external network. In this scenario, the TotalView server daemon will launch on each node allocated by RMS, but full end-to-end TCP conversation with the management server will not be possible.

This is because the TotalView server communicates with the TotalView client *via* the management network (`10.x.x.x`), but the TotalView client communicates with the TotalView server *via* the external network — the TotalView client communication is sent to the cluster alias, and never reaches its true destination.

The simplest workaround is to start `gated` on the node running the TotalView client process. This will cause the node to listen to the routing information protocol (RIP) and know that it is possible to reach the cluster alias on the `10.x.x.x` network interface, because each member advertises that the cluster alias is available through its `10.x.x.x` interface.

4.74 Upgrade: Benign Error

When running the `sc_upgrade_db` command on an HP AlphaServer SC system that does not have an external network IP address assigned to Node 0 in the SC database, the following error will occur:

```
Generating sc_cmf table
Error: Database Upgrade failure: syntax error in expression " & 255"
```

This error is benign, and may be ignored.

4.75 Upgrade: “could not locate Tru64 UNIX in RIS database” Error

After the operating system has been upgraded and the Tru64 UNIX patch kit has been installed, the following error may occur while the HP AlphaServer SC Version 2.6 software is being installed:

```
25/Sep/2003 22:44:37 Attempting to install subsets: QSWNETDIAGS340
25/Sep/2003 22:44:37 QSWNETDIAGS340 now installing
25/Sep/2003 22:44:56 Error - could not locate Tru64 UNIX in RIS database
25/Sep/2003 22:45:03
25/Sep/2003 22:45:03 Installation Complete
25/Sep/2003 22:45:04 SC_UPGRADE: completing sub-state install_load_sc :
Thu Sep
```

This error is benign when it occurs when upgrading a domain; however, it is not benign if it occurs when upgrading a management server. For information on how to correct this error if it occurs when upgrading a management server, please see Chapter 11 of the *HP AlphaServer SC Installation Guide*.

4.76 Upgrade: Member-Specific Crontab Files Not Saved

The upgrade process involves deleting all members except member 1. Member-specific `crontab` files for the deleted members are not restored when the members are re-added.

4.77 Upgrade: "Unexpected exit while: Building boot disk" Error

If a domain is defined as a CS domain in the `sc_domain` table, but the external interface IP address for the lead member of that CS domain is not held in the SC database, the upgrade of the CS domain will fail during the `Post_Installed` state. Specifically, the upgrade will fail to add the non-lead members back into the CS domain. This happens because of a faulty network configuration caused by the missing entries in the SC database.

If this problem occurs, perform the following steps:

1. Update the SC database with the external interface IP address information for those domain members that have external interfaces.
2. Audit the lead member of the CS domain and correct it according to the commands described in the "Validating the Preferred Server Cluster Aliases Configuration" section of Chapter 19 and the "Validating the Default Route Configuration" section of Chapter 22 of the *HP AlphaServer SC System Administration Guide*.
3. Reboot the lead member of the CS domain.
4. Run the `sra upgrade` command again.

The upgrade process will continue to add the non-lead members and the boot disk phase will succeed.

4.78 Upgrade: Upgrade Fails if HP AlphaServer SC-Specific UNIX Accounting is in Place

If the HP AlphaServer SC-specific UNIX accounting setup documented in Chapter 21 of the *HP AlphaServer SC System Administration Guide* is in place, the upgrade command `installupdate` will fail. The accounting configuration should be undone before performing an upgrade.

To check whether the HP AlphaServer SC-specific UNIX accounting setup is in place, run the `/usr/sbin/cdslinvchk` command. The following output indicates that the workaround is in place:

```
Expected CDSL: ./usr/var/adm/acct -> ../cluster/members/{memb}/adm/acct
An administrator or application has modified this CDSL Target to:
/var/cluster/members/{memb}/adm/acct
```

For instructions on how to undo the HP AlphaServer SC-specific UNIX accounting setup, see Chapter 21 of the *HP AlphaServer SC System Administration Guide*.

4.79 Zombie Resources May Not Be Removed

Zombie resources are resource requests that have been killed but have not yet had their CPUs deallocated. When a parallel job completes, RMS executes its cleanup process before allocating the CPUs to another resource request. The cleanup process consists of instructing the `rmsd` daemon on each node that was in use to kill any remaining processes, remove shared memory segments, (optionally) delete corefiles, or execute an epilogue script. This cleanup normally completes quickly and the resource may be displayed briefly by `rinfo` as being in a `zombie` state.

However from time to time, the zombie resource may remain. One cause of zombie resources is that there is a process that cannot be removed on one or more of the nodes associated with the resource request. The CPUs associated with a zombie resource remain allocated while the process remains. You can remove zombie resources by restarting RMS or configuring out the node. However, the process may remain on the affected nodes.

To find the process that cannot be removed, run the following command:

```
# rinfo -rl <resource id of the zombie resource>
```

This displays all processes associated with the resource. RMS will deallocate these CPUs when attempts to end or clean up the processes have failed for 15 minutes (a timeout set in the `cleanup-timeout` attribute). An event outlining this will be logged in the `events` table. When the CPUs have been deallocated, they are available for new resource requests; however, the process may still remain on the affected node.

If you have an uninterruptible process that has not been removed, HP recommends that you examine the kernel logs on the affected nodes. If the kernel logs do not provide an explanation for the hanging processes, HP recommends that you run the `dumpsys` command on the affected nodes, and contact your HP Customer Support representative for assistance.