HP AlphaServer SC

Release Notes

September 2004

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 November 2003 for HP AlphaServer SC Version 2.6.
Operating System and Version:	HP Tru64 UNIX Version 5.1B, Patch Kit 4
Software Version:	Version 2.6 (Update Kit 1)
Maximum Node Count:	1024 nodes
Node Type:	HP AlphaServer ES45 HP AlphaServer ES40 HP AlphaServer DS20L

Legal Notices

The information contained herein is subject to change without notice.

The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

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 2003, 2004 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

UNIX is a registered trademark of The Open Group.

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 DCE/DFS Are Not Supported . . . 1–1
1.2 Documentation Changes . . . 1–2
1.3 Install/Upgrade: Enhanced Restart Capability . . . 1–2
1.4 LSF Changes . . . 1–2
1.5 New Quadrics SHMEM Functions . . . 1–3
1.6 Operating System . . . 1–4
1.7 RMS Changes . . . 1–5
1.8 Upgrade: Upgrade Shortcut Is Supported . . . 1–7

2 Problem Solutions Delivered

2.1 Problem Solutions Delivered in HP AlphaServer SC Version 2.6 (UK1) . . . 2–1
2.1.1 libelan . . . 2–1
2.1.2 LSF . . . 2–2
2.1.3 RMS . . . 2–4
2.1.4 SCFS . . . 2–6
2.2 Problem Solutions Delivered in Patches Since HP AlphaServer SC Version 2.6 . . . 2–6
2.2.1 CFS . . . 2–6
2.2.2 Diagnostics . . . 2–7
2.2.3 Elan . . . 2–7
2.2.4 File Systems . . . 2–7
2.2.5 ICS/Elan . . . 2–8
2.2.6 libelan . . . 2–8
2.2.7 LSM . . . 2–9
2.2.8 MPI . . . 2–9
2.2.9 Operating System . . . 2–9
2.2.10 PFS . . . 2–10
2.2.11 RMS . . . 2–11
2.2.12 Security . . . 2–11
2.2.13 shmem . . . 2–12

3 Restrictions and Problems Identified

3.1 Documentation Errors . . . 3–1
3.2 Reference (man) Page Errors . . . 3–2

4 General Restrictions

4.1 Booting and Shutting Down Nodes — Reminder . . . 4–1
4.2 CAA Daemon Location . . . 4–1
4.3 CAA Placement Policy for srad . . . 4–2
4.4 cfengine Command . . . 4–2
4.5 Disabled Daemons . . . 4–2
4.5.1 AdvFS GUI . . . 4–3
4.5.2 clu_mibs Daemon . . . 4–3
4.5.3 Insight Manager Daemon . . . 4–3
4.5.4 Printing Daemon . . . 4–3
4.5.5 SNMP and Environmental Information Monitoring Daemons . . . 4–3
4.5.6 SysMan Station Daemon . . . 4–4
4.5.7 X Windows Login Daemon . . . 4–4
4.6 Do Not Directly Mount or Unmount SCFS and PFS File Systems . . . 4–4
4.7 Do Not Rebuild the Kernel Before Creating the Domains . . . 4–4
4.8 Do Not Include Blank Lines in the /etc/routes File . . . 4–4
4.9 DRD: Do Not Increase drd_remote_maxphys . . . 4–4
4.10 GPL Code . . . 4–4
4.11 IP Over Elan Not Supported . . . 4–4
4.12 Lazy Swap is Not Supported . . . 4–5
4.13 Local Area Transport Protocol Not Loaded . . . 4–5
4.14 LSF: All LSF Hosts Must be Upgraded to HP AlphaServer SC Version 2.6 (UK1) . . . 4–5
4.15 LSF: Chunk Jobs Are Not Supported . . . 4–5
4.16 LSF: Host Preference Support Restrictions . . . 4–5
4.17 LSF: Kernel-Level Checkpointing Not Supported and Restricted User-Level Checkpointing Support . . . 4–6
4.18 LSF: LSF Administrator Must Be Known User on Nodes Running LSF Daemons . . . 4–6
4.19 LSF: Preemption Preferences . . . 4–6
4.20 LSF: Restricted Usage . . . 4–6
4.21 LSF: "span[ptile=*n*]" Resource String Not Supported . . . 4–7
4.22 Rack Names Must be Populated and Unique . . . 4–7
4.23 Restriction on Using addvol . . . 4–7
4.24 Readline and History Support Removed from rmsquery . . . 4–7
4.25 RMS: Support for prun -r and prun -dr Commands . . . 4–7
4.26 Routing: routed is Not Supported . . . 4–8
4.27 Upgrade: Do Not Leave the System Partially Upgraded . . . 4–8
4.28 Use /tmp — Not /usr/tmp . . . 4–8
4.29 64-Bit Shmem is Not Supported . . . 4–8

5 Known Problems in HP AlphaServer SC Version 2.6 (UK1)

5.1 /etc/binlog.conf File no Longer a CDSL . . . 5–1
5.2 Booting: “sed: Function s/-netmask-/255.255.0.0 cannot be parsed” Error . . . 5–1
5.3 Booting: “The boottime cluster check found a potential problem” Error . . . 5–1

5.4 CMF: Daemon May Need to be Manually Restarted 5–2
5.5 Elan Kernel Comms: Reference Pages Not Updated 5–2
5.6 ICS: Error Message: icssvr_daemon_from_pool: invalid IPL 4 5–2
5.7 Installation/Upgrade: Benign Errors 5–2
5.8 Installation/Upgrade: clu_quorum Check 5–3
5.9 Installation/Upgrade: TCP/IP Ephemeral Range 5–3
5.10 Install: Cannot Install Certain Layered Products 5–3
5.11 Install: ifaccess.config File Set Up Only for Head Node in Domain 5–3
5.12 Install: Possible Errors Seen with Tru64 UNIX Patch 1276.00 5–3
5.13 I/O May Hang on HSG80 in Multibus Mode 5–4
5.14 Kernel: Nodes May Crash with “ics_elan: timeout on free tmds” String 5–4
5.15 Kernel: Nodes May Crash with “malloc_internal: invalid link address may occur” String . . 5–4
5.16 LSF: bjobs Command Does Not Work Correctly When Wild Card Specifier is Used in Job Name 5–5
5.17 LSF: bjobs -l Times Out 5–5
5.18 LSF: Delay Between Resource Allocation and Job Start-Up 5–5
5.19 LSF: Imprecise Pending Reason When the Node Required by a Job is Configured Out 5–5
5.20 LSF: Incorrect Status for Job Submitted with Invalid Command Using prun 5–6
5.21 LSF: Incorrect Pending Reason When the Node Required by a Job is Marked for Draining . 5–6
5.22 LSF: Job Continues Pending Even after Drain Decision is Cancelled 5–6
5.23 LSF: Job May be Lost when RMS Partition Crashes if LSF is in Suspended State 5–6
5.24 LSF: Larger than Expected RMS atime 5–6
5.25 LSF: LSF Not Supported on Dual-Management Server Configuration 5–6
5.26 LSF: Reservation and User Limits 5–6
5.27 LSF: Reservation for Jobs with [hosts=1] 5–7
5.28 LSF: Slow Update of MXJ=! 5–7
5.29 LSF: SQL Delete Query May Fail Because of an Excessive Number of Conditions 5–7
5.30 LSF: Suspended Jobs May be Aborted 5–7
5.31 LSF: Unnecessary Slot Reservation with MCONT/SNODE and Base 5–7
5.32 MPI: Invalid DMA May Be Seen With Onset of Paging 5–7
5.33 MPI: Large Number of MPI Communicators/Not Enough Global Memory 5–8
5.34 MPI: Problem When a Subgroup is the Same Size as COMM_WORLD 5–9
5.35 MPI: Problem with sping.c Test 5–9
5.36 MPI: LIBELAN_WAITTYPE Variable Must Be Set to “POLL” 5–9
5.37 MPI: Sending Large Messages Without Buffering 5–9
5.38 Networks: Configuration of Default Gateway on Peripheral Hardware 5–10
5.39 Networks: Quiet Management Networks can be Misinterpreted by niffd 5–10
5.40 NFS: “Stale File Handle” or “Cannot Overwrite Existing File” Errors 5–11
5.41 Performance Visualizer: pvis Command Causes Stack Trace 5–11
5.42 Performance Visualizer: pvis Command Not Removed 5–12
5.43 PFS: Can Create, and Write to, Files on a Full PFS File System 5–12
5.44 PFS: Changing the Default File Mapping May Not Take Effect Globally 5–12
5.45 PFS: Do Not Delete an Active PFS Mount Point 5–12
5.46 PFS: Do Not Mount a PFS on a PFS Directory 5–12

5.47 PFS: File Locking is Not Supported . . . 5–13
5.48 PFS: Link Operations May Fail on a PFS File System . . . 5–13
5.49 PFS: PFS File System Cannot be Deleted if SCFS is Broken . . . 5–13
5.50 PFS: pfsmgr show Displays “new” State . . . 5–14
5.51 PFS: ro (read-only) Access is Not Supported . . . 5–14
5.52 PFS/SCFS: Mounts and Unmounts Rely on scrun Command . . . 5–14
5.53 PFS/SCFS: Mount Status Not Updated After Shutdown . . . 5–15
5.54 RMS: Detection and Reporting of PSU or Fan Failures on HP AlphaServer ES45 Nodes . . . 5–15
5.55 RMS: Draining Nodes Problem . . . 5–15
5.56 RMS: Event Manager Exits Due to Low Memory Availability . . . 5–15
5.57 RMS: “failed to allocate elan capability” Error . . . 5–16
5.58 RMS: Jobs May be Aborted . . . 5–16
5.59 RMS: Jobs Occasionally Hang on Systems . . . 5–16
5.60 RMS: Jobs Occasionally Report an Error Status . . . 5–16
5.61 RMS: max_per_proc_stack_size Setting in the proc stanza in etc/sysconfigtab Should be Same For Each node in RMS Partition . . . 5–16
5.62 RMS: prun and Non-Global File Systems . . . 5–17
5.63 RMS: prun Fails with “failed to get capability” Error . . . 5–17
5.64 RMS: prun Fails with “failed to create capability: Invalid argument” Error . . . 5–18
5.65 RMS: rmsd Memory Size . . . 5–18
5.66 scevent: <EvmITEM_CLUSTER_NAME> in NFS Events . . . 5–19
5.67 SCFS: File System Status May Be Set To “Unknown”. . . . 5–19
5.68 SCFS: File System May Not Mount If File Server is in Single-User Mode . . . 5–20
5.69 SCFS: Fortran Formatted I/O May Perform Poorly for FAST File Access. . . . 5–20
5.70 SCFS: May Require Reboot to Release Kernel Resources for Reuse . . . 5–20
5.71 SCFS: scfsmgr export or scfsmgr destroy Print Error Message Twice . . . 5–21
5.72 SCFS: SysMan Menu Refresh Button Behaves Incorrectly . . . 5–21
5.73 SCFS: Changing the max_buf and max_buf_total Attributes of the scfs_client kernel Subsystem . . . 5–21
5.74 SC Monitor: Cannot Monitor HSV110 Cells with Multiword Names . . . 5–21
5.75 SC Monitor: HSG Must Have CCL . . . 5–21
5.76 SC Monitor: HSV RAID Systems May Not be Monitored . . . 5–22
5.77 SC Monitor: May Not Mark Objects As Being Stale . . . 5–22
5.78 SC Viewer: “Can't find class p3_desktop.Desktop” Error . . . 5–22
5.79 SC Viewer: Elan Status May Be Incorrect. . . . 5–22
5.80 SC Viewer: “/usr/bin/scviewer: cd: too many arguments” Error . . . 5–23
5.81 Shutting Down: Do Not Use “shutdown -sh now” . . . 5–23
5.82 SRA: Installing on a non-RAID Cluster Disk and Generic Boot Disk . . . 5–23
5.83 SRA: No Sanity Check on the sra shutdown Command . . . 5–24
5.84 SRA: Output Uses the Term “Cluster” . . . 5–24
5.85 SRA: Rerunning the sra setup Command . . . 5–24
5.86 SRA: sra cookie Command Is Only Accurate on rmshost . . . 5–25
5.87 SRA/Sysman: Updating the /etc/hosts File . . . 5–25
5.88 Sysman: Display Error When Configuring NIC On Other Node . . . 5–25

5.89 TotalView: Defining the Path 5–25
5.90 TotalView: Zombie Resources 5–25
5.91 TotalView: TotalView Client May Not Communicate with TotalView Server 5–26
5.92 Upgrade: Benign Error 5–26
5.93 Upgrade: “could not locate Tru64 UNIX in RIS database” Error 5–26
5.94 Upgrade: Member-Specific Crontab Files Not Saved 5–27
5.95 Upgrade: “Unexpected exit while: Building boot disk” Error 5–27
5.96 Upgrade: Upgrade Fails if HP AlphaServer SC-Specific UNIX Accounting is in Place 5–27
5.97 Zombie Resources May Not Be Removed 5–28

About These Release Notes

This document describes new and changed features and problems solved in Version 2.6 (UK1) 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 (UK1):

- Chapter 1: New and Changed Features
- Chapter 2: Problem Solutions Delivered
- Chapter 3: Restrictions and Problems Identified
- Chapter 4: General Restrictions
- Chapter 5: Known Problems in HP AlphaServer SC Version 2.6 (UK1)

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.
“ ”	Quotation marks are used to highlight words or phrases that have a specific meaning or significance in the context in which they are used.
...	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 (UK1) 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 (UK1) since Version 2.6.

Note:

While all of the new and changed features listed here are new in HP AlphaServer SC Version 2.6 (UK1) since Version 2.6, some features were first released as patches for Version 2.6.

All functionality that is listed as being provided in a patch is also provided in Version 2.6 (UK1).

This chapter is organized as follows:

- DCE/DFS Are Not Supported (see Section 1.1 on page 1–1)
- Documentation Changes (see Section 1.2 on page 1–2)
- Install/Upgrade: Enhanced Restart Capability (see Section 1.3 on page 1–2)
- LSF Changes (see Section 1.4 on page 1–2)
- New Quadrics SHMEM Functions (see Section 1.5 on page 1–3)
- Operating System (see Section 1.6 on page 1–4)
- RMS Changes (see Section 1.7 on page 1–5)
- Upgrade: Upgrade Shortcut Is Supported (see Section 1.8 on page 1–7)

1.1 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.

1.2 Documentation Changes

The following documentation has been updated:

- *HP AlphaServer SC Installation Guide*
- *HP AlphaServer SC Release Notes*

Note:

Refer to Section 3.1 on page 3–1 for information about documentation errors that have been identified since HP AlphaServer SC Version 2.6.

1.3 Install/Upgrade: Enhanced Restart Capability

In earlier versions of HP AlphaServer SC software, when you restart an upgrade or perform an installation, the HP AlphaServer SC software expected the nodes to be in certain states, that is, system up, single user, or SRM. If the upgrade process determined that the nodes were not in these states, an error message was displayed and the upgrade process would stop. The install process would attempt to fix the state of the lead node, but it would leave the non-lead nodes in incorrect states, which would cause problems subsequently. In HP AlphaServer SC Version 2.6 UK1, both the install and upgrade processes will set all the nodes to the correct state.

1.4 LSF Changes

The following changes have been made to Platform Computing Corporation's LSF® software ("LSF") in HP AlphaServer SC Version 2.6 (UK1):

- By default, the `brun` command ignores topology options. If you specify `LSF_HPC_EXTENSION=`*`BRUN_WITH_TOPOLOGY`* in the `lsf.conf` file, the `brun` command attempts to run the job with the requested topology. If a topology request can be satisfied for a `brun` job (for example, `RMS[RMS_SLOAD;nodes=1]`), `brun` preserves the topology request. LSF allocates the resource according to the request and tries to run the job with the requested topology. If allocation fails because of a topology request that cannot be satisfied, then the job is re-queued and it is returned to a pending state.
- The following LSF HPC extension has been added to LSF: `DISP_RES_USAGE_LIMITS`. The `bjobs` command displays resource usage limits configured in the queue in addition to job-level limits.

To enable Platform LSF HPC extensions, set the LSF_HPC_EXTENSIONS parameter in the lsf.conf file.The syntax is LSF_HPC_EXTENSIONS=*extension_name ...*. See section 4.11.8 of the *HP AlphaServer SC System Administration Guide* for more information on LSF HPC Extensions.

- By default, the LSF scheduler can contact the RLA running on any host for RMS allocation requests. LSB_RLA_HOST_LIST defines a list of hosts that restrict which RLAs the LSF scheduler can contact. If LSB_RLA_HOST_LIST is configured, you must list at least one host per RMS partition for the RMS partition to be considered for job scheduling. Listed hosts must be defined in lsf.cluster.cluster_name. Host names are separated by spaces. The default value for LSB_RLA_HOST_LIST is Undefined. The syntax is LSB_RLA_HOST_LIST="*host_name" ...*.
- The default value for LSB_RLA_UPDATE, which specifies how often the LSF adapter for RMS (RLA) should refresh its RMS information map, has been amended to 60 seconds.
- Platform LSF HP AlphaServer SC Version 2.5 and Version 2.6 hosts cannot co-exist in the same cluster. You must upgrade all HP AlphaServer SC hosts to Version 2.6. There are no incompatibilities between Version 2.6 and Version 2.6 (UK1). However, Platform recommends that the Master should run the latest version, that is, Version 2.6 (UK1), to take advantage of problem fixes in the LSF scheduler.
- The following install.config options are obsolete. You do not need to specify them when running lsfinstall:
  - LSF_ENABLE_EXTSCHEDULER="Y"
  - CONFIGURE_LSB_RLA_PORT="Y"
  - LOGDIR="path"

1.5 New Quadrics SHMEM Functions

The Quadrics SHMEM implementation has been extended to include two new storage allocation functions:

```
void *shmalloc(size_t size)
int shmfree(void *ptr)
```

where *size* is restricted to be less than, or equal to, the total size of the symmetric heap. This size of the symmetric heap is controlled by the environment variable *SHMEM_SYMMETRIC_HEAP_SIZE* (default 512MB).

The functions *shmalloc* and *shmfree* must be called with the same parameters, and in the same sequence, by each process in a parallel program. A barrier is implied.

The function *shmalloc* will not return in any process until all processes have called it with the same size. All processes will return NULL if this requirement is not met. All processes will return NULL if any process fails to allocate the memory, or map it into the Elan.

The function *shmalloc* will allocate regions of memory from the symmetric heap at the same address in each process of a parallel program and will establish Elan translations such that the memory allocated can be used in subsequent communication calls.

The function *shmfree* must be called with the same address in every process. The region of memory from the symmetric heap will only be freed when all processes have called the *shmfree* function. All processes will return an error if this requirement is not met. All processes will return an error if any process fails to free the memory.

1.6 Operating System

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

In HP AlphaServer SC Version 2.6 (UK1), the supported operating system is Tru64 UNIX Version 5.1B and associated Patch Kit 4.

Note that Tru64 UNIX patch kit distribution has been changed as individual patches have been grouped together into service packs. This means that the number of patches visible to the user has decreased from approximately 320 to approximately 64, and therefore patch installations will run significantly faster compared with previous installations.

Note:

A small number of reference pages were not included in the Tru64 UNIX Version 5.1B Patch Kit 4. However, they are now available in an online downloadable tar file. For more information about obtaining the tar file and installing the missing reference pages, refer to the online *Technical Update for Tru64 UNIX Version 5.1B and Higher (August 2004)* available at:

- http://h30097.www3.hp.com/docs/updates/V51B/TITLE.HTM.

The *Tru64 UNIX Version 5.1B Patch Kit 4 Patch Summary and Release Notes* are now available in PDF and HTML formats respectively at the following URLs:

- http://h30097.www3.hp.com/docs/patch/51B/bl25/ReleaseNotes.pdf
- http://h30097.www3.hp.com/docs/patch/51B/bl25/HTML/TITLE.HTM

1.7 RMS Changes

This section describes the new changes made to Resource Management System (RMS) functionality in HP AlphaServer SC Version 2.6 UK1 since Version 2.6.

- Support has been added for listing the fields of each table in the RMS database using rmsquery `-v` tables.
- A new bootlog table feature has been added. A new record is created for each node when the boot time changes and it is updated periodically (every 5 minutes) by the `rmsd`.
- Added support for Elan statistics collection. The number of Elan operations and the total number of bytes transferred are collected in the Elan library. This information is written to a shared memory segment created by RMS for each job. The rmsloader process periodically updates the statistics for the job.

  RMS periodically collates this information along with the `utime`, `stime`, and more details and updates the accounting statistics record for the resource request. For example:

```
sql> select name,exfers,ebytes from acctstats
...
1051  4407874          84674129872
1052  485886           25417434960
1053  485886           25417434960
```

  These statistics are the sum over jobs and processes. This information is available to the user through the `rinfo` command while the job is running. For example:

```
$ rinfo -rl 1054
RESOURCE     PARTITION    CPUS   PRI   USERNAME     PID          PROJECT
1054         parallel     12     50    user1        991231       default

ETIME        ATIME        UTIME        STIME        PAGEFLTS     MEM
00:40        08:00        05:42        00:50        0            190

ELAN OPS      BYTES TRANSFERRED
307886        8556348640

STATUS        HOSTNAMES
allocated     atlas[0,2-3]
```

  The update rates are quite slow, typically 30-90 seconds between samples depending on the setup of a machine.

  The Elan statistics data is also available when a job completes `via` the `-s` option to `prun`. For example:

```
atlas0: prun -s -n12 tping -n1000 0 1m

...
Elapsed time       155.19 secs    Allocated time    1861.98 secs
User time         1368.61 secs    System time         47.43 secs
```

```
Memory                190          Page faults             0
Elan xfers         485886          Elan data        24821713 KB
Status             exited          Cpus used               12
```

- Support has been added for users that will allow them to specify a numeric account ID using either a command-line argument (the `-A` option to `prun` or `allocate`) or the environment variable `RMS_ACCOUNT` when making a request to allocate CPUs. The account ID is written to the accounting statistics record, which enables resource usage to be collated by account ID. If no account ID is specified, the default is `-1`.

  The list of accounts that each user belongs to is specified in the accounts table. Records can be added to this table using rmsquery. For example:

  ```
  # rmsquery "insert into accounts (account,user)values (2, 'user1')"
  ```

  Having added (or removed) records from the accounts table, it is necessary to instruct the partition manager(s) to read it again by running the following command:

  ```
  # rcontrol reload partition = name
  ```

  The following queries can be used to show the accounts that a user belongs to and the users that belong to a given account:

  ```
  $ rmsquery "select account from accounts where user='duncan'"
  1
  4
  3
  2

  $ rmsquery "select user from accounts where account=1"
  user1
  tony
  user2
  daniel
  ```

  Account ID support is enabled with the `account-id-support` attribute. This attribute takes the following values: `disabled`, `warning`, or `strict`. If the attribute does not exist, or if it is set to any other value, then account ID support is disabled. If the `account-id-support` attribute is set to `warning`, then a warning message is generated if the user is not a member of the specified account. The account ID defaults to `-1` under these circumstances. If the `account-id-support` attribute is set to `strict`, then an error message is generated, and the request for CPUs fails if the user is not a member of the specified account.

  If a user who is a member of an account tries to start a job without specifying an account, then the job cannot be run. A warning will be generated that no account ID has been set. If the `account-id-support` attribute is set to `strict`, then unless a user is a member of an account, they cannot run a job. The account ID support mechanism is independent of job scheduling.

1.8 Upgrade: Upgrade Shortcut Is Supported

The upgrade shortcut mechanism that was added in HP AlphaServer SC Version 2.5 (UK1) is supported for upgrading to HP AlphaServer SC Version 2.6 (UK1). The upgrade shortcut functionality has been added in HP AlphaServer SC Version 2.6 (UK1). This allows an easier transition from Patch Kit 3 to Patch Kit 4. This shortcut process should only be used if the upgrade does not require the base operating system version to change. For more information on the upgrade shortcut, see Chapter 4 of the *HP AlphaServer SC Installation Guide*.

2
Problem Solutions Delivered

This chapter lists the problems for which solutions have been delivered since HP AlphaServer SC Version 2.6, and is organized as follows:

- Problem Solutions Delivered in HP AlphaServer SC Version 2.6 (UK1) (see Section 2.1 on page 2–1)
- Problem Solutions Delivered in Patches Since HP AlphaServer SC Version 2.6 (see Section 2.2 on page 2–6)

2.1 Problem Solutions Delivered in HP AlphaServer SC Version 2.6 (UK1)

This section lists the problems solutions that are being delivered for the first time in HP AlphaServer SC Version 2.6 (UK1).

2.1.1 libelan

The following `libelan` problems were fixed:

- Added an `elan_gptrcmp()` function, which can be used to test if buffers are global across the job.
- Added statistics gathering code to the library in a way that can be collated by RMS.
- Problem that caused `tping -f misalign` to hang multi-rail machines because messages smaller than `bigmsg` were being sent with the SYNC-bit set.
- Replaced all `printf %zi` formats with the `%lx` format as Tru64 UNIX does not support the `%zi` format.
- Added an enhancement to the global allocator to handle the out of memory case more effectively to ensure that all processes exit with ENOMEM.

2.1.2 LSF

The following LSF problems were fixed in HP AlphaServer SC Version 2.6 (UK1). Please contact Platform Support (support@platform.com) for more information.

- For Job Control action configuration, the `batchid` in the RMS rcontrol command must include the LSF cluster name. The `rcontrol` command now has the form:
  `rcontrol [suspend | kill | resume] batchid=cluster_name@$LSB_JOBID`

  See section 4.7 in the *HP AlphaServer SC System Administration Guide* for details on using Job Control actions.
- Problem where the `sbatchd` may core dump inside RMS API calls. This is preventable on the LSF side.
- Problem where the job-level `extsched` option was being ignored when `NEW_JOB_SCHED_DELAY=0` was configured at the queue level.
- Problem when both reservation and fairshare are configured for a queue. A reservation would oscillate between two different users.
- Problem within the RLA when CPUs cannot be allocated to a resource. The resource needs to be removed if CPUs cannot be allocated to it.
- Problem with preemption, where in certain circumstances it did not occur as expected.
- Added a few minor improvements to the `CAA lsf.scr` script.
- Changed the default update interval from 600 to 60 seconds. This reduces the recovery time when the `mbschd` requires a new map due to some RMS failure (for example, if allocation fails).
- Problem with job dispatch where LSF could not find the RMS resource for a job. This problem was typically observed in large clusters.
- Problem when the RLA would core dump after a `badmin hrestart` command is issued.
- Problem that the `sbatchd` may core dump if a job had an invalid RMS resource.
- Problem with RLA failover selection in a multi-RMS partition environment.
- Problem where a decision to dispatch a job may be cancelled because of job dependencies, and an RMS allocation has already been created for the job. In the case of a cancelled dispatch decision, the RMS allocation needs to be deleted.
- Problem with orphan allocations. It was possible that after a `mbschd` restart that a job had an allocation created for it, even though the job was not dispatched. Such allocations would be lost after a restart and they would not be removed.

- Problem when users reach their `UJOB_LIMIT`, they receive an incorrect pending reason for any pending jobs.
- Problem that jobs could not be dispatched when the parameter `LSF_STRICT_CHECKING=Y` is contained in the `lsf.conf` file.
- Added enhancements that reduce the likelihood of job fragmentation when a cluster is initially starting up.
- Problem where host preference may not be honored with `RMS_SLOAD` defined as the `extsched` option.
- Problem with jobs pending indefinitely where the predicted start time of jobs pending with reservations had expired.
- Problem where the LSF cluster contains multiple RMS machines and LSF would stop scheduling jobs to one of the RMS machines.
- Problem with jobs pending indefinitely when host partitions are configured within LSF.
- Problem with a memory leak in the RLA daemon.
- Problem with `mbschd` core dumps when `mbschd` can no longer allocate memory.
- Problem with jobs with base specified in the `extsched` option pending indefinitely when host partitions are configured within LSF.
- Problem with the `LSB_RLA_HOST_LIST` feature when the `mbsched` is running outside the RMS environment.
- Problem where `mbatchd` may core dump when the SLA feature is enabled.
- Improved the performance of interactive jobs. The LSF administrator can define the parameter `LSF_CHILD_RES_SLEEPTIME` to control the sleep interval of the child reservation associated with a job.
- Restored compatibility with the `bsub -a` option. When the `mesub` component was introduced, the traditional behavior of the `-a` option was changed.
- Core dump problem with `mbatchd` in a multi-cluster environment after a queue configuration change.
- Problem when host partitions are configured and the `badmin mbdreconfig` command is issued where parallel jobs will pend indefinitely displaying the message: `System is not ready for scheduling after reconfiguration`.
- Problem where job dependencies are ignored when a job is switched to another queue.
- Problem with running the `brun` command on large parallel jobs, which caused the command to fail with a XDR encode/decode error.

- Problem where restarting `mbatchd` may result in a job name dependency being ignored.
- Problem where the `bhosts` command displays the wrong values for `NJOBS` and `RSV` when LSF is used in a multi-cluster environment.
- Problem that after a multi-cluster job is stopped using the `bstop` command, it could not be resumed subsequently. This issue would occur after a multi-cluster job was forwarded to the remote cluster, and it was pending for a resource.
- Problem when a job's CWD is too long and the commands `bacct` and `bhist` generate the error: `Bad event format`.
- Problem where `Bhosts` and `Bjobs` do not update slot reservations occasionally.
- Problem where LSF will keep starting the job action terminate script.
- Insufficient reservation may occur when the `Host Partition` parameter and the `General User Limit` parameter are defined.

2.1.3 RMS

The following RMS problems were fixed:

- Problem with local core file directories being owned by the process allocating the resource (for example, LSF) rather than the user on whose behalf the resource is being allocated.
- Problem with RMS sending `SIGXCPU` on expiry of the time limit instead of the signal defined by the warning-signal attribute.
- The partition manager does not update the `configured_nodes` and `cpus` fields of the partitions table.
- Added checks on `batchid` and `project` field widths to the RMS API.
- Problem with the accounting statistics that caused the user CPU time (`utime`) to be incorrect if the job was killed.
- Problem in rmsquery where the `tables` command did not work when the query was run from the command-line as `rmsquery` *`tables`*.
- Added a new feature that allows rmsquery to optionally add table footers in addition to table headers. This is useful for viewing long tables where the table headers have scrolled off the top of the screen. This feature is enabled by adding another `v` option (for example: `rmsquery -vv` *`select ...`*) or, by entering `verbose` twice when run interactively.

- Problem with the partition manager not correctly updating the count for the number of nodes in a partition, when a node is rebooted with more or fewer CPUs than it had on the previous occasion it was booted.
- Problem in the HP AlphaServer SC kit, which resulted in the directory `/usr/opt/rms/usr/opt/rms/msql` being created.
- Problem with `rmsloader` not restoring the default signal behavior for `USR1`, `USR2`, `QUIT`, and `WINCH` as it executes the application processes.
- Problem with the RMS partition manager leaking memory as nodes are configured in and out.
- Problem with the partition manager crashing if accounting statistics records for running jobs were missing.
- Problem with the RMS `mmanager` which sets the partition status to `down` when time changed under it.
- Problem with `rmstbladm -u` where it does not add a sequence to an existing table.
- Made changes regarding how `swmgr` posts events. `swmgr` only posts one inconsistent firmware revision event per module. It posts one event for each error on the broadcast tree.
- Changed `rmsbackup` (and `rmsarchive`) so that they do not email root reports when the `mail-to-users` RMS attribute is empty. Root reports are now appended to the `rmsbackup.log` file.
- Problem with `rmspost` where the arguments were not being read properly, if the colon-separated form was used.
- Added enhancement so that `rmstbladm` will delete only temporary SQL files it creates, as a result of decompressing a backup or an archive.
- The `rmsbuild` usage string has been corrected as follows: the `-p ports` option has been removed and an `-R rails` option has been added.
- Problem with `rmsloader` crashing on 16-way systems running non-uniform numbers of processes per node.
- Problems with the default memory limits on systems with large amounts of memory per node and on systems with no swap space.

2.1.4 SCFS

The following SCFS problem was fixed:

- The default value of the SCFS server tuning parameter `sync_dirty_size` has been changed back to 64MB, which was its value in HP AlphaServer SC Version 2.5 (UK1) and all previous HP AlphaServer SC versions. The `sync_dirty_size` parameter value was changed to 256MB in HP AlphaServer SC Version 2.6 to increase performance of large FAST mode writes. However, subsequent testing has shown that the original default value of 64MB provides better overall FAST mode write throughput for most live environment workloads.

2.2 Problem Solutions Delivered in Patches Since HP AlphaServer SC Version 2.6

This section describes the problem solutions that were delivered in patches since HP AlphaServer SC Version 2.6; 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 (UK1), unless they have been superseded by new or changed functionality. Where available, the associated problem codes are shown.

2.2.1 CFS

The following CFS problems were fixed:

- A deadlock problem on the client, in particular, the CFS client of the NFS client. This problem relates to a situation where there are many I/Os that are queued to the CFS asynchronous I/O daemon threads, and user threads are waiting for the I/O to complete. (100574)
- SCFS/CFS file corruption and permission failures. This problem relates to a cluster client node which had two users accessing a file from an NFS imported filesystem mounted on another node in the cluster. One user had write permission and was writing to the file. The second user had read permission only and was reading the file. The cluster node that mounted the filesystem (not the node running the two user processes) reported an NFS error indicating that a write data transfer had been lost. (105525)
- A `vm_map_lock` hang occurs in `scext_wire_pages` when using SCFS file systems. When this problem occurs, there may also be threads hanging with `vm_fault` errors. (105440)
- Panic: `pg_nwriters going negative`. This relates to a panic in `ubc_page_release`, which was caused by a bug in the CFS code. (105931)

2.2.2 Diagnostics

The following diagnostics problem was fixed:

- Problem in the `elan_level_test` diagnostic tool, which may cause it to follow incorrect routes through the Elan hierarchy, and report non-existent errors.

2.2.3 Elan

The following Elan problems were fixed:

- Certain applications have experienced `Elan memory exhausted` errors when the environment variable `MPI_USE_LIBELAN_SUB` is set to a non-zero value. (105882)
- The LSMS application has been observed to hang when making certain MPI calls. (105884)
- Problem with insufficient `oolmd` memory available for GigE networks. (105459)
- Problem in `ics_elan` that may cause a memory leak in bucket 31 when that bucket is very busy. This would eventually result in a node hang problem. This problem is most frequently observed on busy UBC mode SCFS file servers using GigE as their primary interface. However, in theory, any network interface may trigger this behavior if sufficient traffic is passing over it. (106622)
- Panic: `Elan3: assertion failed: PTBL_LEVEL (ptbl->ptbl_flags)`. This issue relates to a panic encountered in the Elan3 driver when running an application.

2.2.4 File Systems

The following file systems problems were fixed:

- The `mount_advfs` command can crash/KMF the kernel using incorrect mount-cmd or `/etc/fstab` entries. For example, the crash can occur if the `#FileSet` parameter/specification is missing for an AdvFS and CFS file system, in either the `fstab` file or the `mount_advfs` command-line. (106120)
- Compute nodes receive the error code `ENOTCONN` (socket not connected) when attempting to read or write data to a file on an SCFS file system after a server change. Metadata operations are not affected by this problem. (106425)
- Problem in `cfs.mod`, which may lead to an Elan panic: `iovcnt limit exceeded`. (105698)

2.2.5 ICS/Elan

The following ICS/Elan problems were fixed:

- Calls to the `ep_complete_rpc` routine were being made while a large volume of multicast traffic was being transmitted (*via* the `ep_multicast_large` routine). Normally, the `ep_complete_rpc` routine called the Elan driver handler, but occasionally, it failed to call the Elan driver handler.
- Panic: `assertion failed: ((E3_uint32 *) ((unsigned long) queue->QueueEntries + nfptr + queue->QueueEntrySize - queue->QueueBase))[-1]`. This issue relates to an Elan kernel communications assertion failure. (105433)
- Elan kernel communications: `assertion failed: hdr->NumMaps > 0 cm.c`. This issue relates to changes introduced into Elan code for HP AlphaServer SC Version 2.6 to reduce the noise caused by the heartbeat code. This can result in an Elan kernel communications assertion failure in certain circumstances. (105646)

2.2.6 libelan

The following `libelan` problems were fixed:

- Problem with outgoing cache performance-related issues in `libelan`. (107975)
- Lack of performance symmetry with `touchbuf` code in `shmalloc()`. This performance issue affects shmem communications only; MPI communications are unaffected. Processes on the first node of an allocation touched all of the region affected by `shmalloc`, thereby loading both the Elan page translations and the intra-node shmem page translations for the entire region affected by `shmalloc`. However, processes on the other nodes failed to touch the region affected by `shmalloc`, thereby deferring the loading of the intra-node shmem page translations until the region affected by `shmalloc` was subsequently used by those processes. (108023)
- A memory corruption problem was encountered when calling `shmalloc` during intra-node shmem. (107489)
- An MPI benchmark caused an SEGV on dual-rail systems. This issue relates to a problem with unmatched messages if the `recv` buffer is larger than the `send` buffer. (105805)
- An MPI benchmark hung at `n=49` unless `MPI_USE_LIBELAN_SUB=0`. This issue relates to a problem with subgroups in the `libelan` library, where nodes in the middle of an allocation used the wrong `libelan` virtual process (that is, the `libelan` process

corresponding to the MPI rank) to perform the hardware broadcast. This problem only manifested itself when the first process of an allocation was not included within the MPI subgroup. (107549)

- MPI issues with fast fourier transform program hang on single or dual-rail systems. (105805)
- Broadcast from read-only memory problem when shared memory is disabled. (105987)

2.2.7 LSM

The following LSM problem was fixed:

- Slow boot when using LSM in HP AlphaServer SC Version 2.6. (101334)

2.2.8 MPI

The following MPI problems were fixed:

- Problem in intra-node MPI that may lead to the following assertion failure: `ng==NULL, file adi2init.c` for MPI jobs that use more than one CPU on the same node. The same jobs will run successfully with one CPU per node (which uses inter-node MPI only). Problem with subgroups that may cause an MPI program using them to core dump during exit. (107348)
- Problem when using MPI subgroups with optimized collectives that may cause an MPI program to hang or core dump. (107658)

2.2.9 Operating System

The following operating system problems were fixed:

- The Tru64 UNIX V5 environment was enhanced to facilitate multipathing and a larger name space for buses, targets, and logical unit numbers than can be represented by a 14-bit minor number. The CAM driver has been updated to use this 14-bit minor device number. This caused a backwards compatibility issue with an application that reads the CAM minor device number with an `fstat()` system call. (100397)
- Sysconfigtab settings for `titan_sys_ps_hotswap` and `titan_sys_fan_hotswap` are not surviving reboot. (102356)
- Panic: `kmf in dyn_hash_remove`. This kernel memory fault panic results from `quotact1` requests on advfs filesets. (105839)
- Panic: `lock_terminate: lock held`. This kernel memory fault panic results from `quotact1` requests on advfs filesets. (105839)

- A performance issue identified in `vm.mod` as a result of running NP benchmarking code on HP AlphaServer SC Version 2.6. (105883)
- Third-party file systems that do not support the `FSOP_GETPAGE` callback are supposed to fall into the NFS server `fallback` read code, but they do not because of an uninitialized pointer. (106835)
- Problem in `vm.mod` kernel memory that caused stack corruption and subsequent KMF panics. Two different panics were generated as follows:
  `trap: invalid memory ifetch access from kernel mode` (105840)
  `panic (cpu 1): bad nofault` (106511)
- Issues with defunct child processes occasionally being left behind when exiting the totalview debugger. These defunct child processes were adopted by `init` and they should have cleared from the process table accordingly. However, in certain circumstances, this did not happen and these defunct child processes would remain indefinitely. (105794)
- Problem with `setId` crashing when an attempt is made to install a subset that has prerequisites, which are not on the system. This problem occurs irrespective of whether `setId` is run from the command-line or *via* menu options.

2.2.10 PFS

The following PFS problems were fixed:

- Panic at `pfs_subr.c:2622` when the importing node goes down. This occurs when a change in the behavior of CFS in Tru64 UNIX V5.1B can cause PFS to panic the other nodes of a cluster, when the importing node goes down unexpectedly for any reason. This particular panic can be recognized by the following message in the console log just before the node goes down:
  `cmn_err: CE_PANIC: pfs: [os/pfs_subr.c:2622] ASSERT(error == 0)`
- Machines containing dash (-) characters in their names were unable to mount PFS file systems after they were upgraded to HP AlphaServer SC Version 2.6. This fix provides `pfsmgr` support for machine names containing dash characters. While the use of dash characters in machine names has never been permitted in any version of HP AlphaServer SC, older versions of the installer did not prevent a user from entering dash characters in machine names. As a result, there exist a few machines that contain dash characters in their names— the machine name was preserved through various upgrades. (106134)

- Newly-created Version 2.6 PFS file systems fail to mount on the file serving domain that hosts their disks, if that domain has an old-style FS domain name, for example, `atlas[A-Z]`. Domains with new-style names, for example, `atlasD4`, and PFS file systems that existed prior to the Version 2.6 upgrade are not affected by this problem. The problem only affects the creation of new Version 2.6 PFS file systems. (106134)

2.2.11 RMS

The following RMS problems were fixed:

- `pmanager` checks the `elites.revision` field in the RMS database before enabling Elan hardware broadcast. The check is redundant since HP AlphaServer SC Version 2.5, because the machine will not function without the correct elite revisions. (105291)
- RMS core-file directories created with incorrect permissions when a resource is created by a root-owned process using `rmsapi` functions. (105292)
- The `batchid` field of the resource RMS database table is occasionally not large enough to store the LSF `batchid` identification strings. This problem occurs because in LSF SC Version 2.6 SSB, the batchid identification string is constructed as `<LSF_CLUSTER_NAME>@<LSF_BATCHID>` and this string can become larger than 16 characters, particularly if the LSF job arrays feature is used. When this happens, LSF stops functioning properly because it fails to submit these jobs.
- Non-privileged (or ordinary) users can perform rmsquery updates and they can update the system database. (106012)
- Problem where a root login that runs an `su - 'user'` command is able to use the `rms rcontrol kill` command to kill another user's jobs. (106670)
- Problem in `pmanager` which may cause it to fail to close the connection to a third-party job scheduler after deallocating a resource that was originally created *via* the RMS API. A `pmanager` error message was displayed where one argument to `printf` was omitted, so the error text was always null, for example:
  `pmamager [parallel]: Warning: Failed to deallocate resource 1104880768: (null).` (107370)

2.2.12 Security

The following security problem was fixed:

- Security issue for `scmountd`. Updated and simplified the report format used by the `scfsmgr status` command, as it contained unnecessary information that made the status report difficult to interpret.

2.2.13 shmem

The following shmem problem was fixed:

- The `scfsmgr show command` does not display listed file systems output in sorted order.

3
Restrictions and Problems Identified

This chapter lists any new problems and restrictions identified since HP AlphaServer SC Version 2.6.

3.1 Documentation Errors

The following documentation errors have been identified since HP AlphaServer SC Version 2.6:

- In the *HP AlphaServer SC System Administration Guide*, Chapter 19: Managing the Cluster Alias Subsystem, Table 19-2: Preferred Server Cluster Aliases Configuration, the `/etc/exports.aliases` configuration setting/file table cell description under the **CS Node with External Interface** column should read as follows: No aliases defined.
- In the *HP AlphaServer SC System Administration Guide*, Chapter 21: Managing Domain Members, Section 21.5: Converting a CS Domain to an FS Domain, step 5 should read as follows:

  Add `atlasD1` to the list of FS domains in the SCFS_SRV_DOMS variable in the `/etc/rc.config.common` file on the management server and on all domains, as shown in the following example, where `atlasD0` is an existing FS domain, and `atlasD1` is the new FS domain:

  ```
  atlasms# rcmgr -c set SCFS_SRV_DOMS atlasD0,atlasD1
  atlasms# scrun -d all "rcmgr -c set SCFS_SRV_DOMS atlasD0,atlasD1"
  ```
- In the *HP AlphaServer SC Interconnect Installation and Diagnostics Manual*, Chapter 7: HP AlphaServer SC Interconnect Diagnostic Tests, Section 7.3.3: Defining the SC Viewer Interconnect Configuration, you can ignore step 6c and its associated instructions regarding how to configure the `Module Controller Base IP` address field in the Network Parameters dialog box. This is because in the released SC Viewer product, the `Module Controller Base IP` address field is greyed out and it cannot be modified, because it is not necessary to change the Base IP address.

- In the *HP AlphaServer SC System Administration Guide*, Chapter 4: Managing the Load Sharing Facility (LSF), Section 4.9.1.1: Allocation Type on Page 4-13, the final paragraph of the RMS_SNODE value description should read as follows:

  Use RMS_SNODE on larger clusters where only the number of available job slots matters for job placement decisions. When RMS_SNODE is used with job preemption, LSF may suspend more jobs than necessary. You should use RMS_SLOAD if you use preemptive scheduling.

3.2 Reference (man) Page Errors

The following reference (man) page errors have been identified since HP AlphaServer SC Version 2.6:

- The `vrestore -m` command does not function as documented. In the `vrestore` man page, the `-m` flag is described as follows:

  ```
  -m Does not preserve the owner, group, or modes of each file from the device.
  ```

  When the `vrestore` command is run with the `-m` flag option, the owner, group, or modes should not be restored from the archives. In the following example, tests are run as `root` user:

  ```
  atlas0 # ls -al /vrestore_m_fset1/dir1/file.1
  -rwxrwxrwx   1 tmp_user tmp_grp     1024 Apr 30 09:40
  /vrestore_m_fset1/dir1/file.1

  atlas0 # vdump -f /tmp/vrestore_m_file /vrestore_m_fset1

  atlas0 # vrestore -m -f /tmp/vrestore_m_file -x dir1/file.1
  vrestore: Date of the vdump save-set: Fri Apr 30 10:39:52 2004
  vrestore: Save-set source directory : /vrestore_m_fset1
  vrestore: Target directory : /

  eve0 # ls -al dir1/file.1
  -rwxrwxrwx   1 root     tmp_grp     1024 Apr 30 10:43 dir1/file.1
  eve0 #
  ```

  Note that the user has changed to `root`, but that group (`tmp_grp`) has been incorrectly preserved.

4

General Restrictions

This chapter describes the general restrictions that apply to HP AlphaServer SC Version 2.6 (UK1).

4.1 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.

4.2 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
   ```

4.3 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
```

4.4 cfengine Command

The `cfengine` command is present but not supported in HP AlphaServer SC Version 2.6 (UK1).

4.5 Disabled Daemons

A number of daemons have been disabled in HP AlphaServer SC Version 2.6 (UK1) 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.

4.5.1 AdvFS GUI

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

4.5.2 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.

4.5.3 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.

4.5.4 Printing Daemon

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

4.5.5 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. This is a built-in feature of the RMC functionality. However, if a UNIX controlled shutdown is required by a site, then the `envmond` daemon can be re-enabled on each domain by running the command:

```
scrun -d all 'mv /sbin/init.d/envmon.disabled /sbin/
init.d/envmon'
```

Afterwards, the `envmond` daemon can be started using the command:

```
scrun -n all '/sbin/init.d/envmon start'
```

On startup, the messages indicating that the extensible SNMP master agent is not running can be ignored.

4.5.6 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.

4.5.7 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.

4.6 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`.

4.7 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.

4.8 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.

4.9 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).

4.10 GPL Code

The RMS functionality in HP AlphaServer SC Version 2.6 (UK1) 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.

4.11 IP Over Elan Not Supported

HP AlphaServer SC Version 2.6 (UK1) software does not support IP over Elan for inter-process communication.

4.12 Lazy Swap is Not Supported

Lazy swap is not supported in HP AlphaServer SC Version 2.6 (UK1) — use eager swap only.

4.13 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.

4.14 LSF: All LSF Hosts Must be Upgraded to HP AlphaServer SC Version 2.6 (UK1)

HP AlphaServer SC Version 2.6 (UK1) 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 (UK1).

4.15 LSF: Chunk Jobs Are Not Supported

If the chunk job size is defined in a queue configuration, jobs will still be dispatched as single jobs on an SC machine. This is because LSF chunk jobs are not supported in HP AlphaServer SC Version 2.6 (UK1).

4.16 LSF: Host Preference Support Restrictions

Host preference (for example, `bsub -m hostA`) is only supported for `RMS_SLOAD`, `RMS_SNODE`, and `RMS_MCONT` allocations. However, it is only honored for the `RMS_SLOAD` allocation. With the other allocation types, the hosts will be re-ordered based on RMS ordering. All hosts in the preference list must be within the same host group or RMS partition.

It is recommended that the `RMS_SLOAD` allocation be used on smaller clusters, where the job placement decision should be influenced by host load, or where you want to keep a specific host preference.

4.17 LSF: Kernel-Level Checkpointing Not Supported and Restricted User-Level Checkpointing Support

Application-level checkpointing is supported in HP AlphaServer SC Version 2.6 (UK1). However, kernel-level checkpointing is not available on HP AlphaServer SC systems. User-level checkpointing is supported for sequential jobs only. See section 4.10.10 of the *HP AlphaServer SC System Administration Guide* for details on the limitations on user-level checkpointing.

4.18 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.

4.19 LSF: Preemption Preferences

The topological preemption preference is the only preemption preference supported in HP AlphaServer SC Version 2.6 (UK1).

4.20 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.

4.21 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 (UK1).

4.22 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.

4.23 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.

4.24 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: www.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.

4.25 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 (UK1) when RMS is only installed on the management server.

4.26 Routing: routed is Not Supported

In HP AlphaServer SC Version 2.6 (UK1), 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).

4.27 Upgrade: Do Not Leave the System Partially Upgraded

When upgrading to HP AlphaServer SC Version 2.6 (UK1), 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 (UK1) 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*.

4.28 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.

4.29 64-Bit Shmem is Not Supported

HP AlphaServer SC Version 2.6 (UK1) 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.

5

Known Problems in HP AlphaServer SC Version 2.6 (UK1)

This chapter lists the problems that have been identified in HP AlphaServer SC Version 2.6 (UK1).

5.1 /etc/binlog.conf File no Longer a CDSL

On a non-SC system, the `/etc/binlog.conf` file is created as a Context-Dependent Symbolic Link (CDSL) and it is listed as a CDSL in the `/usr/.smdb./OSFBASE540.inv` file. However, for the consolidated logging feature introduced in HP AlphaServer SC Version 2.6, the `/etc/binlog.conf` file is changed to a cluster-wide file rather than a CDSL. This will cause the `/usr/sbin/cdslinvchk` command to report the `/etc/binlog.conf` file as an error. This error can safely be ignored, and the `/etc/binlog.conf` file should be left as a non-CDSL.

5.2 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 (UK1), you can ignore these errors.

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

A number of daemons have been disabled in HP AlphaServer SC Version 2.6 (UK1) as a general performance-enhancement measure. The disabled daemons are listed in Section 4.5 on page 4–2.

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.

5.4 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.

5.5 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 (UK1).

5.6 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.

5.7 Installation/Upgrade: Benign Errors

During installation and upgrade, the Patch Kit 4 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.

5.8 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.

5.9 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.

5.10 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*.

5.11 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*.

5.12 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/
```

5.13 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.

5.14 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.

5.15 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.

5.16 LSF: bjobs Command Does Not Work Correctly When Wild Card Specifier is Used in Job Name

The `bjobs` command does not work as expected when the wild card specifier is used in the job name.

5.17 LSF: bjobs -l Times Out

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

5.18 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.

5.19 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
```

5.20 LSF: Incorrect Status for Job Submitted with Invalid Command Using prun

If an LSF job is submitted using `prun` and an invalid command, for example, `bsub prun bad_cmd`, the job will return a completed status of DONE, which is incorrect.

5.21 LSF: Incorrect Pending Reason When the Node Required by a Job is Marked for Draining

A job requesting a node marked for draining remains pending with the following incorrect reason:

```
Cannot satisfy RMS base node requirement
```

5.22 LSF: Job Continues Pending Even after Drain Decision is Cancelled

A job continues pending even after a drain decision is cancelled.

5.23 LSF: Job May be Lost when RMS Partition Crashes if LSF is in Suspended State

If LSF is in a suspended state when the RMS partition goes down, the LSF job may be lost in some cases.

5.24 LSF: Larger than Expected RMS atime

LSF jobs scheduled from slot reservation queues may result in larger then expected RMS `atime`.

5.25 LSF: LSF Not Supported on Dual-Management Server Configuration

Platform do not support LSF in a dual-management server configuration.

5.26 LSF: Reservation and User Limits

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

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

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

5.28 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 `!`.

5.29 LSF: SQL Delete Query May Fail Because of an Excessive Number of Conditions

In rare circumstances, an SQL delete query performed by the RLA may fail because it has too many conditions specified.

5.30 LSF: Suspended Jobs May be Aborted

If the RMS partition is restarted while suspended LSF jobs exist, these suspended jobs may be aborted.

5.31 LSF: Unnecessary Slot Reservation with MCONT/SNODE and Base

It is a known limitation, where certain jobs will have topology requirements that cannot be satisfied. However, such jobs will reserve slots if a reservation is configured. For example, consider a host that consists of four nodes only: `adam[0-3]`. The following job will run:

```
$ bsub -n 8 -ext 'RMS[RMS_SNODE;base=adam2]' a.out
```

If, however, you typed `adam3` instead of `adam2`, the job would never run:

```
$ bsub -n 8 -ext 'RMS[RMS_SNODE;base=adam3]' a.out
```

However, this job will reserve slots on the `adam3` node.

5.32 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.

5.33 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.

5.34 MPI: Problem When a Subgroup is the Same Size as COMM_WORLD

If the `libelan` subgroup optimizations are disabled and if an MPI subgroup is the same size as COMM_WORLD, then the MPI code will use the COMM_WORLD group for such subgroup communications. However, in doing so, the MPI code does not take into account the fact that the ranks may have been reversed, and hence `MPI_Bcast` will potentially go wrong.

5.35 MPI: Problem with sping.c Test

A problem has been encountered with the `sping.c` test which resides in `/usr/opt/rms/bin` when it is run on three processors. At the time of the HP AlphaServer SC Version 2.6 UK1 release, a fix is not yet available. HP advises that this test is not supported in HP AlphaServer SC Version 2.6 UK1 and it will be fixed in future releases.

5.36 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.

5.37 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.

5.38 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.

5.39 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.

5.40 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.

5.41 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.

5.42 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.

5.43 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).

5.44 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.

5.45 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.

5.46 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.

5.47 PFS: File Locking is Not Supported

File locking is not supported on PFS file systems.

5.48 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.

5.49 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.

5.50 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.

5.51 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.

5.52 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
```

5.53 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.

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

On HP AlphaServer ES45 nodes, the environmental monitoring daemon may 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.

5.55 RMS: Draining Nodes Problem

If a request is made to drain a node which is running jobs, the node will be marked as draining and the drain will complete when all running jobs have been completed. However, if the node is still draining and a new job is requested requiring resources, which spans the node (using the `-N prun` flag), the new job will queue and it will not be serviced.

5.56 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.

5.57 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
```

5.58 RMS: Jobs May be Aborted

It has been observed that when a large number (greater than 170) of multiple users’ small jobs (2 CPUs or less) are submitted and the partition is restarted, some jobs may be aborted.

5.59 RMS: Jobs Occasionally Hang on Systems

If one of the management servers crashes in a system containing clustered management servers, jobs which were running on the machine may occasionally hang.

5.60 RMS: Jobs Occasionally Report an Error Status

It has been observed that if a large number of multiple users’ jobs, are submitted from within a script, some of the jobs will occasionally report an error status. This can be resolved by introducing a small delay before submitting each user’s jobs.

5.61 RMS: max_per_proc_stack_size Setting in the proc stanza in etc/sysconfigtab Should be Same For Each node in RMS Partition

The `max_per_proc_stack_size` setting in the proc stanza in `/etc/sysconfigtab` should be the same for each node in an RMS partition. If this value is different, then it is likely that any one process on one node will obtain a different return from `malloc` than a corresponding process on another node. For parallel programs that rely on a common return, such as the `shmem shmalloc` function call, then a mismatch in this setting can cause the parallel program to abort unexpectedly.

5.62 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.

5.63 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.

5.64 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.

5.65 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.

5.66 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.

5.67 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
```

5.68 SCFS: File System May Not Mount If File Server is 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.

5.69 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.

5.70 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.

5.71 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.

5.72 SCFS: SysMan Menu Refresh Button Behaves Incorrectly

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

5.73 SCFS: Changing the max_buf and max_buf_total Attributes of the scfs_client kernel Subsystem

If the `max_buf` attribute value is changed to a value which is greater than or equal to the value of the `max_buf_total` attribute, SCFS client nodes may crash with the following panic string:

```
SCFS assertion failed scfscli_max_total_buffering > scfscli_replay_max , file sys/
scfs_failover.c
```

Refer to the "Tuning SCFS Client Operations" section of Chapter 7 of the *HP AlphaServer SC System Administration Guide* for information on changing the `max_buf` and `max_buf_total` configurable attributes of the `scfs_client` kernel subsystem.

5.74 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 it 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.

5.75 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.

5.76 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`.

5.77 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`.

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

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

5.79 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.

5.80 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.

5.81 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.

5.82 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*.

5.83 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.

5.84 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.

5.85 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.

5.86 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.

5.87 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.

5.88 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.

5.89 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.

5.90 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.

5.91 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.

5.92 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.

5.93 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 (UK1) 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*.

5.94 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.

5.95 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.

5.96 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*.

5.97 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.