



**SGI® Altix® UV Systems Linux®
Configuration and Operations Guide**

007-5629-001

COPYRIGHT

© 2010, SGI. All rights reserved; provided portions may be copyright in third parties, as indicated elsewhere herein. No permission is granted to copy, distribute, or create derivative works from the contents of this electronic documentation in any manner, in whole or in part, without the prior written permission of SGI.

LIMITED RIGHTS LEGEND

The software described in this document is "commercial computer software" provided with restricted rights (except as to included open/free source) as specified in the FAR 52.227-19 and/or the DFAR 227.7202, or successive sections. Use beyond license provisions is a violation of worldwide intellectual property laws, treaties and conventions. This document is provided with limited rights as defined in 52.227-14.

TRADEMARKS AND ATTRIBUTIONS

Altix, NUMALink, OpenGL, Performance Co-Pilot, SGI, the SGI logo, SGI ProPack, SHMEM, Silicon Graphics, Supportfolio, and XFS are trademarks or registered trademarks of Silicon Graphics International Corp. or its subsidiaries in the United States and other countries.

Intel and Xeon are trademarks or registered trademarks of Intel Corporation. Linux is a registered trademark of Linus Torvalds in several countries. Novell and SUSE are registered trademarks of Novell, Inc., in the United States and other countries.

All other trademarks mentioned herein are the property of their respective owners.

Record of Revision

Version	Description
001	June 2010 Original publication.

Contents

About This Guide	ix
Related Publications	ix
Additional Reading	ix
Linux System Administration	x
Intel Compiler Documentation	x
Other Intel Documentation	x
Open-Source Documents	x
Obtaining Publications	xi
Conventions	xi
Reader Comments	xii
1. SGI LK License Facility	1
Overview	1
LK License Considerations	1
Obtaining Host Information	2
Obtaining the License Keys from SGI	3
Installing the License Keys	3
Verifying the License Keys	3
Setting Up the LK Floating Licenses 1kd Daemon	4
Using the LK Floating Licenses Facilities	4
2. Configuring Your System	7
CPU Frequency Scaling	7
System Partitioning	8
007-5629-001	v

Overview	8
Advantages of Partitioning	10
Create a Large, Shared-memory Cluster	10
Provides Fault Containment	10
Allows Variable Partition Sizes	10
Provide High Performance Clusters	11
Limitations of Partitioning	11
Supported SSI	12
Installing Partitioning Software and Configuring Partitions	14
Partitioning Software	14
Partitioning a System	15
Making Array Services Operational	20
Floating Point Assist Warnings from Applications	21
Unaligned Access Warnings	22
3. System Operation	25
SGI Altix UV System Control Network	25
Connecting to the UV System Controller Network Using the CMC	26
Power on and Booting an Altix UV System from Complete Power Off	26
Power off an Altix UV System	28
Power NMI to Drop into KDB	28
Upgrading System BIOS	28
Index	31

Procedures

Procedure 2-1	Setting up Networking Between Partitions	15
Procedure 2-2	Partitioning a System Into Four Partitions	16
Procedure 2-3	Making Array Services Operational	21

About This Guide

This guide is a reference document for people who manage the operation of SGI Altix UV systems running SGI ProPack 7 for Linux Service Pack 1 operating system. It explains how to perform general system configuration and operations under the Linux operating system used with SGI Altix UV 100 and UV 1000 series systems. For information about general Linux system administration, see the “Additional Reading” section of this Preface.

This manual applies to SGI Altix UV 100 and 1000 systems. For the SGI Altix UV 10 system, see the *SGI Altix UV 10 System User's Guide*. For SGI Altix 4000 series systems, see the *Linux Configuration and Operations Guide*. For more information on SGI Altix ICE 8200 and SGI Altix ICE 8400 systems, see the *SGI Tempo System Administrator's Guide*, and their respective hardware guides.

SGI Management Center (SMC) software running on the system management node (SMN) provides a robust graphical interface for system configuration, operation, and monitoring. For more information on the SMC, see *SGI Management Center System Administrator's Guide*.

This manual contains the following chapters:

- Chapter 1, "SGI LK License Facility" on page 1
- Chapter 2, "Configuring Your System" on page 7
- Chapter 3, "System Operation" on page 25

Related Publications

For a list of manuals supporting SGI Linux releases, see the *SGI ProPack 7 for Linux Service Pack 1 Start Here*.

Additional Reading

The following sections describe publications that contain additional information that may be helpful in the administration of your system.

Linux System Administration

Linux system administration information is available on your system at the following location:

`/usr/src/linux/Documentation`

Documentation for SUSE Linux Enterprise Server 11 is located on your system in the `/docu/en` directory.

Linux Documentation Project (LDP) also contains useful documentation at: <http://tldp.org/>.

Intel Compiler Documentation

Documentation for the Intel compilers is located on your system in the `/docs` directory of the directory tree where your compilers are installed. If you have installed the Intel compilers, the following documentation is available:

- *Intel C++ Compiler User's Guide* (`c_ug_lnx.pdf`)
- *Intel Fortran Compiler User's Guide* (`for_ug_lnx.pdf`)
- *Intel Fortran Programmer's Reference* (`for_prg.pdf`)
- *Intel Fortran Libraries Reference* (`for_lib.pdf`)

Other Intel Documentation

The Intel Developer Center contains useful information about Intel technologies at <http://developer.intel.com/design/>

Open-Source Documents

The following open-source document may be useful to you.

- *Debugging with DDD User's Guide and Reference Manual* provides information on using the DataDisplayDebugger (DDD). It is available at the following location:

<http://www.gnu.org/manual/ddd/pdf/ddd.pdf>

Obtaining Publications

You can obtain SGI documentation in the following ways:

- See the SGI Technical Publications Library at: <http://docs.sgi.com>. Various formats are available. This library contains the most recent and most comprehensive set of online books, release notes, man pages, and other information.
- Online versions of the *SGI ProPack 7 for Linux Service Pack x Start Here*, the SGI ProPack 7 release notes, which contain the latest information about software and documentation in this release, the list of RPMs distributed with SGI ProPack 7, and a useful migration guide, which contains helpful hints and advice for customers moving from earlier versions of SGI ProPack to SGI ProPack 7, can be found in the `/docs` directory on the SGI ProPack 7 for Linux CD.

The SGI ProPack 7 release notes get installed to the following location on a system running SGI ProPack 7: `/usr/share/doc/sgi-propack-7/README.txt`.

- You can view man pages by typing `man title` on a command line.

Conventions

The following conventions are used throughout this document:

Convention	Meaning
<code>command</code>	This fixed-space font denotes literal items such as commands, files, routines, path names, signals, messages, and programming language structures.
<i>variable</i>	Italic typeface denotes variable entries and words or concepts being defined.
user input	This bold, fixed-space font denotes literal items that the user enters in interactive sessions. (Output is shown in nonbold, fixed-space font.)
[]	Brackets enclose optional portions of a command or directive line.

... Ellipses indicate that a preceding element can be repeated.

Reader Comments

If you have comments about the technical accuracy, content, or organization of this publication, contact SGI. Be sure to include the title and document number of the publication with your comments. (Online, the document number is located in the front matter of the publication. In printed publications, the document number is located at the bottom of each page.)

You can contact SGI in any of the following ways:

- Send e-mail to the following address:
techpubs@sgi.com
- Contact your customer service representative and ask that an incident be filed in the SGI incident tracking system.
- Send mail to the following address:
SGI
Technical Publications
46600 Landing Parkway
Fremont, CA 94538

SGI values your comments and will respond to them promptly.

SGI LK License Facility

SGI Foundation Software 2 Service Pack 1 (SP1) contains a licensing mechanism called LK. LK was developed by SGI for SGI products only. LK and the LK floating licenses facilities are described in this chapter.

Overview

The `lkd` program is the SGI LK license manager. It serves counted licenses for use by `localhost` and `remotehost`. Applications linked with LK and using counted licenses use the local `lkd` daemon to proxy requests out to the final daemon. Counted licenses are known by the `ssn` (system serial number) serving that key. The daemon maintains an `[ssn,ip or name,port]` pair so it can process LKN (LK network) requests with the remote daemon. The applications never talk to the remote daemon; they only send requests to the local daemon via a UNIX socket. The request is processed locally by the local daemon if it is its own `ssn` or proxy sent out by the request to the remote daemon. For usage information, see the `lkd(8)` man page.

This chapter covers the following topics:

- "LK License Considerations" on page 1
- "Obtaining Host Information" on page 2
- "Obtaining the License Keys from SGI" on page 3
- "Installing the License Keys" on page 3
- "Verifying the License Keys " on page 3
- "Setting Up the LK Floating Licenses `lkd` Daemon" on page 4
- "Using the LK Floating Licenses Facilities" on page 4

LK License Considerations

The LK licensing daemon (`lkd`) manages counted licenses. Therefore, it manages licenses with a count value greater than or equal to one (1). The LK license references

the server serial number (*ssn*) of the server; this is the *lk_hostid* of the LK license manager. See the *lk(5)* man page for more information. For example:

```
# A counted ( 5 seats ) nodeless license served by ssn 0011432bc7b3
product=acme, version=1.0, \
  count=5@0011432bc7b3, begDate=0, \
  expDate=0, licenseID=0, \
  key=0iK0myC0F9vFgVJqJBsIBBU/ZpwTYIoG
```

From an application license perspective using managed licenses, can run on the following:

- Every host
This is called a nodeless managed license. This is the most commonly issued license.
- One host
This is called a nodelock managed license.

In all cases, the server is locked to a specific *ssn* as indicated by the *count=xxx@ssn* notation.

Obtaining Host Information

When you order SGI Foundation Software 2 SP1, you will receive an entitlement ID. You must submit the system host ID, host name, and entitlement ID when requesting your permanent LK license key.

To obtain the host information for a server-capable administration node, execute the following command (you must have the *lkSGI* RPM from the SGI Foundation Software 2 SP1 installed):

```
/usr/sbin/lk_hostid
```

For example, the following shows that the serial number is *N0000302* and the license ID is *e000012e*:

```
server-admin % /usr/sbin/lk_hostid
N0000302 e000012e socket=8 core=16 processor=16
```

Obtaining the License Keys from SGI

To obtain your license keys, see information provided in your customer letter and the following web page:

<http://www.sgi.com/support/licensing>

Installing the License Keys

To install the license keys, copy them into the `/etc/lk/keys.dat` file. A sample `keys.dat` file is, as follows:

```
server-admin% cat keys.dat
product=XVM_PLEX_LINUX, version=5.000, count=0, \
begDate=1216648987, expDate=0, licenseID=e000012e, \
key=Ea5Yc+lvxqykBXc931CGMHv5cmOCSfns, \
info='XVM PLEX STD 5 M',attr='16 CPU', \
vendor='Silicon Graphics, Inc.',ref_id='070706'
```

Verifying the License Keys

You can use the `lk_verify -A` command to verify LK licenses. To see more output, use the `-v` option. For example:

```
server-admin% lk_verify -A -vvv
lk_check          All          All : total found=1

1 /etc/lk/keys.dat:005          product=XVM_PLEX_LINUX, version=5.000, count=0, be
expDate=0, licenseID=e000012e, key=Ea5Yc+lvxqykBXc931CGMHv5cmOCSfns, \
info='XVM PLEX STD 5 M',attr='16 CPU', vendor='Silicon Graphics, Inc.', \
ref_id='070706'
Verdict:                SUCCESS. Nodelock.
                        Available since 372 days on 21-Jul-2008 09:03:07.
                        No End Date.
Attribute 1 of 4 : info=XVM PLEX STD 5 M
Attribute 2 of 4 : attr=16 CPU
Attribute 3 of 4 : vendor=Silicon Graphics, Inc.
Attribute 4 of 4 : ref_id=070706

lk_check          All          All : total matched=1
```

Setting Up the LK Floating Licenses `lkd` Daemon

In order for the LK floating licenses facilities to operate correctly, both the client and server need to have this common setup:

- The following configuration setting: `chkconfig lkd on`
- Share the same `/etc/lk/keys.dat` (or equivalent) either via distribution mechanism (`rsync`) or NFS, or similar
- Allow TCP/IP port 11106 traffic between the server and client hosts

The client must also define the license server “`ssn host_or_ip`” pair in either one of these locations:

- `/etc/lkd.map` (see `lkd.map(5)`)
- And/or embedded in the license file itself in the form of:

```
#! 080069010203 192.168.1.1
```
- And/or an NIS `lkd.map`. See `/usr/share/doc/lk-2.0/nis/readme.txt` for how to setup this map. For more information, see `lkd.map(5)` for more information.

Using the LK Floating Licenses Facilities

In order to use LK floating licenses facilities, the `lkd` daemon has to run in two places:

- The license server
The host acting as the license server
- The license client(s)

All the hosts where applications utilizing LK floating licenses capabilities

Upon start-up, `lkd` determines the licenses to serve by using the standard lookup LK mechanism, as described in `lk(5)` man page. The daemon will only serve licenses belonging to its local `hostid` as determined by `lk_hostid(1)` program. LK client applications use the same mechanism. However, the daemon does not monitor license file changes. This means `lkd` need to be restarted to account for license file changes.

Also, note that the license file may contain floating licenses for several server. For example:

```
# floating license server by ssn=00a0d1e4071d for product 'a'
product=a, version=1, count=5@00a0d1e4071d, begDate=0, expDate=0, \
    licenseID=0, key=/sHK7L45x8aiHy9VRt+YWx/c8KTYxWgd, \
    vendor="Silicon Graphics, Inc.",info="vendor_info"

# floating license server by ssn=00a0d1e4080b for product 'b'
product=b, version=1, count=5@00a0d1e4080b, begDate=0, expDate=0, \
    licenseID=0, key=fs5rUFED3nBdzAGw6UDsmb5yw2wioNOh, \
    vendor="Silicon Graphics, Inc.",info="vendor_info"

# a nodeless license for product 'c'
product=c, version=1, count=5, begDate=0, expDate=0, licenseID=0, \
    key=Pl8UFMVeNBo/SiiG6cbBb4CDchO5R2z8, \
    vendor="Silicon Graphics, Inc.",info="vendor_info"

# a nodelock license for product 'd'
product=d, version=1, count=, begDate=1248886680, \
    expDate=0, licenseID=d1e4071d, \
    key=f5DvhYYFuvHoX2FVoI/ixvcBITbnE9pn
```

The LK licensing daemon (`lkd`) exchanges LK license requests of clients using TCP/IP official `sgi-lk` port 11106. You can use the `-L` option to specify a different port. When using this option, make sure this is reflected in the `ssn` mapping. See the `lkd.map(5)` man page for more information. The `lkd` daemon also satisfies local LK license request using a UNIX TCP socket file in `/var/run/lkd/socket.SGI.server`. Client applications never talk to remote license server via TCP/IP. Instead, all requests are made through the local daemon who will either satisfy the request locally or act as a proxy, and handle the request to the remote license server via TCP/IP. Consequently, it means there are no direct LK exchange between the application and remote LK daemon over TCP/IP.

When an LK request is granted to a client application via a `lk_get(3)` call, the daemon decrements the count for the granted key. The license server then starts monitoring (pings), at periodic interval (`-P` option), the client daemon to make sure the client process ID (PID) is still present and not a "zombie" process. See the `lkdc(8)` man page for more information.

Similarly, when the application release the key using `lk_free(3)`, the daemon increments the original granted count for the application. Should a client application

hang or crash, the license server acts as if the application crashed, hence "taking back" the granted count.

The `lkd` daemon can also report information about granted keys, and so on. This is enabled by the `-c` option. This is normally specified (by default) in `/etc/sysconfig/lkd` file. See the `lkdc(8)` man page for more information.

The `lkd` daemon also maintains a journal of license activities (`-j` option). This file should be cleaned and/or rotated appropriately from time to time. To turn off the journal option, just specify `-j /dev/null`.

Configuring Your System

This chapter provides information on configuring your system and covers the following topics:

- "CPU Frequency Scaling" on page 7
- "System Partitioning " on page 8
- "Making Array Services Operational" on page 20
- "Floating Point Assist Warnings from Applications" on page 21
- "Unaligned Access Warnings" on page 22

CPU Frequency Scaling

CPU frequency scaling is disabled by default on your SGI Altix UV 100 or Altix UV 1000 series system. This is accomplished by adding the `acpi-cpufreq` file to the `/etc/modprobe.d` directory.

An example is, as follows:

```
admin:/etc/modprobe.d # cat acpi-cpufreq

# comment out the following line to enable CPU frequency scaling

install acpi-cpufreq /bin/true
```

To enable CPU frequency scaling, log into your Altix UV system (`ssh root@hostname`) and remove the `acpi-cpufreq` file in the `/etc/modprobe.d` directory. If your system is partitioned, you need to perform this on each partition.

If you decide to enable CPU frequency scaling on your system, SGI highly recommends that you set the default scaling governor to `performance` using the following script:

```
maxcpu=`grep processor /proc/cpuinfo | awk '{print $3}' | tail -1`

for cpu in `seq 0 $maxcpu`
```

```
do  
  
    cpufreq-set -c $cpu -g performance  
  
done
```

Note: In order to enable the Intel processor's Turbo boost feature, CPU frequency scaling has to be enabled.

System Partitioning

This section describes how to partition an SGI ProPack server and contains the following topics:

- "Overview" on page 8
- "Advantages of Partitioning" on page 10
- "Limitations of Partitioning" on page 11
- "Supported SSI" on page 12
- "Installing Partitioning Software and Configuring Partitions" on page 14

Overview

A single SGI ProPack for Linux server can be divided into multiple distinct systems, each with its own console, root filesystem, and IP network address. Each of these software-defined group of processor cores are distinct systems referred to as a *partition*. Each partition can be rebooted, loaded with software, powered down, and upgraded independently. The partitions communicate with each other over an SGI NUMALink connection called *cross-partition communication*. XPNET is the TCP/IP interface for cross-partition communication over NUMALink. Collectively, all of these partitions compose a single, shared-memory cluster.

Direct memory access between partitions, sometimes referred to as *global shared memory*, is made available by the XPC and XPMEM kernel modules. This allows processes in one partition to access physical memory located on another partition.

The benefits of global shared memory are currently available via SGI's Message Passing Toolkit (MPT) software. For more information on MPT, see the *Message Passing Toolkit (MPT) User's Guide*.

Partition discovery software allows all of the partitions to know about each other.

Partition firewalls provide memory protection for each partition. XPMEM software uses firewall code to open up a portion of memory so that it can be accessed by CPU cores in other partitions. Firewalls have kernel, BIOS, and hardware components. The kernel component is XPMEM. XPMEM allocates some memory and makes a BIOS call passing BIOS the memory address. The BIOS calls allow the UV hub hardware to change the memory protections on each of the cache lines in the block of memory. Hardware memory directory bits indicate which nodes have access to that cache line. Nodes in the partition always have access to memory in their own partition. Nodes outside the partition do not have access, unless the memory is opened up by XPMEM.

A *coherence domain* is the extent to which a CPU core is able to coherently load and store cacheable memory. Usually, a coherence domain is the same as an operating system partition or single-system image. With XPMEM, the coherence domain can be expanded to include memory in other partitions. *Expanding the coherence domain* is part of the process of opening up the memory in a remote partition.

A *heartbeat* mechanism allows each partition to determine the state of all partitions in the system. Each partition increments its own heartbeat which is read by other partitions. As long as the local partition keeps incrementing its heartbeat, the other partitions know that it is still operational.

Each partition has a *partition page* that stores its heartbeat. A partition page is a page of physical memory that contains information about the local partition and whose address is known by the other partitions in the system.

Reset fences are special hardware mechanisms built into the NUMALink ports of hubs and routers that prevent resets from propagating into a partition. Reset fences are set at the partition boundaries. BIOS sets up the reset fences after it does NUMALink discovery.

The global reference unit (GRU) no-fault code allows a partition to access a remote partition safely. If the remote access fails, the GRU no-fault code cleans up the GRU so it can be reused.

All of the partitions in a partitioned system have the same *system serial number*. The system serial number is stored in the system controller.

It is relatively easy to configure a large SGI Altix system into partitions and reconfigure the machine for specific needs. No cable changes are needed to partition or repartition an SGI Altix machine. Partitioning is accomplished by commands sent to the system controller. For details on system controller commands, see the *SGI UV System Software Controller User's Guide*.

Advantages of Partitioning

This section describes the advantages of partitioning an SGI ProPack server as follows:

- "Create a Large, Shared-memory Cluster" on page 10
- "Provides Fault Containment" on page 10
- "Allows Variable Partition Sizes" on page 10
- "Provide High Performance Clusters" on page 11

Create a Large, Shared-memory Cluster

You can use SGI's NUMALink technology and the XPC and XPMEM kernel modules to create a very low latency, very large, shared-memory cluster for optimized use of Message Passing Interface (MPI) software and logically shared, distributed memory access (SHMEM) routines. The globally addressable, cache coherent, shared memory is exploited by MPI and SHMEM to deliver high performance.

Provides Fault Containment

Another reason for partitioning a system is fault containment. In most cases, a single partition can be brought down (because of a hardware or software failure, or as part of a controlled shutdown) without affecting the rest of the system. Hardware memory protections prevent any unintentional accesses to physical memory on a different partition from reaching and corrupting that physical memory. For current fault containment caveats, see "Limitations of Partitioning" on page 11.

Allows Variable Partition Sizes

Partitions can be of different sizes, and a particular system can be configured in more than one way. For example, a 128-processor system could be configured into four partitions of 32 CPU cores each or configured into two partitions of 64 CPU cores each. (See "Supported Configurations" for a list of supported configurations for system partitioning.)

Your choice of partition size and number of partitions affects both fault containment and scalability. For example, you may want to dedicate all 64 CPU cores of a system to a single large application during the night, but then partition the system in two 32 processor systems for separate and isolated use during the day.

Provide High Performance Clusters

One of the fundamental factors that determines the performance of a high-end computer is the bandwidth and latency of the memory. The SGI NUMAflex technology gives an SGI ProPack partitioned, shared-memory cluster a huge performance advantage over a cluster of commodity Linux machines (white boxes). If a cluster of N white boxes, each with M CPUs is connected via Ethernet or Myrinet or InfiniBand, an SGI ProPack system with N partitions of M CPUs provides superior performance because of the significantly lower latency of the NUMALink interconnect, which is exploited by the XPNET kernel module.

Limitations of Partitioning

Partitioning can increase the reliability of a system because power failures and other hardware errors can be contained within a particular partition. There are still cases where the whole shared memory cluster is affected; for example, during upgrades of hardware which is shared by multiple partitions.

If a partition is sharing its memory with other partitions, the loss of that partition may take down all other partitions that were accessing its memory. This is currently possible when an MPI or SHMEM job is running across partitions using the XPMEM kernel module.

Failures can usually be contained within a partition even when memory is being shared with other partitions. XPC is invoked using normal shutdown commands such as `reboot(8)` and `halt(8)` to ensure that all memory shared between partitions is revoked before the partition resets. This is also done if you remove the XPC kernel modules using the `rmmmod(8)` command. Unexpected failures such as kernel panics or hardware failures almost always force the affected partition into the KDB kernel debugger or the LKCD crash dump utility. These tools also invoke XPC to revoke all memory shared between partitions before the partition resets. XPC cannot be invoked for unexpected failures such as power failures and spontaneous resets (not generated by the operating system), and thus all partitions sharing memory with the partition may also reset.

Supported SSI

See the *SGI UV 1000 System User's Guide* for information on configurations that are supported for system partitioning. The SGI Altix UV 1000 system sizes range from 2 to 128 blades (16 to 2048 cores) in a single system image (SSI). See Figure 2-1 on page 13.

The maximum number of processor cores in an SSI is 2048. The following describe the minimum and maximum metrics within an SSI:

- one partition
- one to four racks
- one to eight individual rack units (IRUs) with maximum of two IRUs per rack
- one to eight Base I/O (only one Base I/O has the capability to boot the system)
- two to 128 compute blades
- two to 128 Altix UV Hubs (one Hub on each compute blade)
- two to 256 processor sockets (one socket on memory expansion blade, two sockets on compute blade)
- 16 to 2048 processor cores (up to 4096 threads with Hyper-Threading enabled)
- eight to 2048 DDR3 memory DIMMs (16 DIMMs maximum per compute blade)
- Up to 16 terabytes (TBs) with up to 4 TB per rack (using 8 GB DIMMs)

Currently, the Linux operating system only supports 2048 cores/threads.

Note: The terms single system image (SSI) and partition can be used interchangeably.

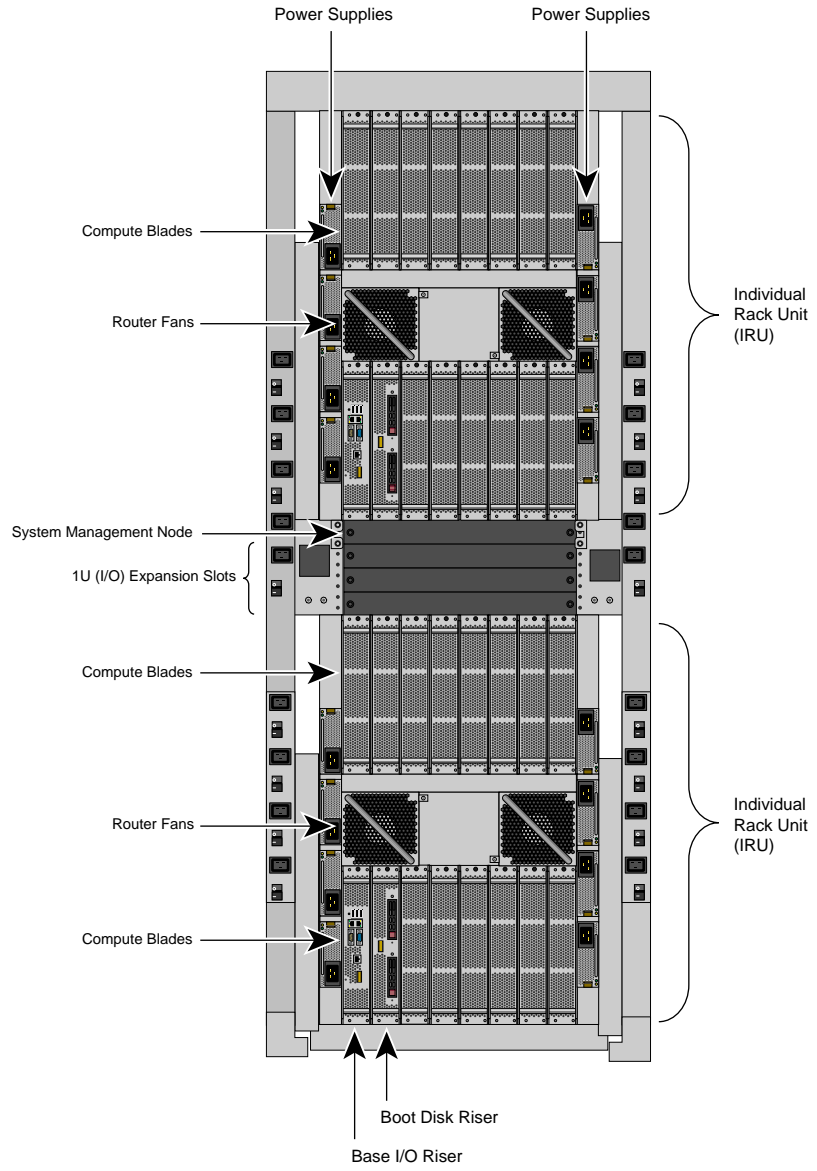


Figure 2-1 SGI Altix UV 1000 System Rack

For additional information about configurations that are supported for system partitioning, see your sales representative.

Installing Partitioning Software and Configuring Partitions

To enable or disable partitioning software, see "Partitioning Software" on page 14, to use the system partitioning capabilities, see "Partitioning a System" on page 15 and "Partitioning a System" on page 15.

This section covers the following topics:

- "Partitioning Software" on page 14
- "Partitioning a System" on page 15

Partitioning Software

SGI ProPack for Linux servers have XP, XPC, XPNET, and XPMEM kernel modules installed by default to provide partitioning support. XPC and XPNET are configured off by default in the `/etc/sysconfig/sgi-xpc` and `/etc/sysconfig/sgi-xpnet` files, respectively. XPMEM is configured on by default in the `/etc/sysconfig/sgi-xpmem` file. To enable or disable any of these features, edit the appropriate `/etc/sysconfig/` file and execute the `/etc/init.d/sgi-xp` script.

On SGI ProPack systems running SLES11, if you intend to use the cross-partition functionality of XPMEM, you will need to add `xpc` to the line in the `/etc/sysconfig/kernel` file that begins with `MODULES_LOADED_ON_BOOT`. Once that is added, you may either reboot the system or issue an `modprobe xpc` command to get the cross-partition functionality to start working. For more information on using `modprobe`, see the `modprobe(8)` man page.

The XP kernel module is a simple module which coordinates activities between XPC, XPMEM, and XPNET. All of the other cross-partition kernel modules require XP to function.

The XPC kernel module provides fault-tolerant, cross-partition communication channels over NUMALink for use by the XPNET and XPMEM kernel modules.

The XPNET kernel module implements an Internet protocol (IP) interface on top of XPC to provide high-speed network access via NUMALink. XPNET can be used by applications to communicate between partitions via NUMALink, to mount file systems across partitions, and so on. The XPNET driver is configured using the `ifconfig`

commands. For more information, see the `ifconfig(1M)` man page. The procedure for configuring the XPNET kernel module as a network driver is essentially the same as the procedure used to configure the Ethernet driver. You can configure the XPNET driver at boot time like the Ethernet interfaces by using the configuration files in `/etc/sysconfig/network-scripts`. To configure the XPNET driver as a network driver see the following procedure.

Procedure 2-1 Setting up Networking Between Partitions

The procedure for configuring the XPNET driver as a network driver is essentially the same as the procedure used to configure the Ethernet driver (`eth0`), as follows:

1. Log in as root.
2. For SGI ProPack systems, configure the `xp0` IP address using `yast2`. For information on using `yast2`, see *SUSE Linux Enterprise Server 11 Administration Guide*. The driver's full name inside `yast2` is SGI Cross Partition Network adapter.
3. Add the network address for the `xp0` interface by editing the `/etc/hosts` file.
4. Reboot your system or restart networking.

The XPMEM kernel module provides direct access to memory located on other partitions. It uses XPC internally to communicate with XPMEM kernel modules on other partitions to accomplish this. XPMEM is currently used by SGI's Message Passing Toolkit (MPT) software (MPI and SHMEM).

Partitioning a System

This section describes how to partition your system. The following example shows how to use chassis manager controller (CMC) software to partition a two rack system containing four IRUs into four distinct systems, use the `uvcon` command to open a console and boot each partition and repartition it back to a single system.

Note: Each partition must have one base I/O blade and one disk blade for booting. `001i01b00` refers to rack 1, IRU 0, and blade00. `r001i01b01` refers to rack 1, IRU 0, and blade01.

Base I/O and the boot disk are displayed by the `config -v` command, similar to the following:

```
r001i01b00 IP93-BASEIO
r001i01b01 IP93-DISK
```

Procedure 2-2 Partitioning a System Into Four Partitions

To partition your system, perform the following steps :

1. Use the `hwcfg` command to create four system partitions, as follows:

```
CMC:rlilc>hwcfg partition=1 "r1i1b*"
CMC:rlilc>hwcfg partition=2 "r1i2b*"
CMC:rlilc>hwcfg partition=3 "r2i1b*"
CMC:rlilc>hwcfg partition=4 "r2i2b*"
```

2. Use the `config -v` command to show the four partitions, as follows:

```
CMC:rlilc> config -v

CMCs:                4
    r001i01c UV1000 SMN
    r001i02c UV1000
    r002i01c UV1000
    r002i02c UV1000

BMCs:                64
    r001i01b00 IP93-BASEIO P001
    r001i01b01 IP93-DISK P001
    r001i01b02 IP93-INTPCIE P001
    r001i01b03 IP93 P001
    r001i01b04 IP93 P001
    r001i01b05 IP93 P001
    r001i01b06 IP93 P001
    r001i01b07 IP93 P001
    r001i01b08 IP93 P001
    r001i01b09 IP93-INTPCIE P001
    r001i01b10 IP93-INTPCIE P001
    r001i01b11 IP93-INTPCIE P001
    r001i01b12 IP93-INTPCIE P001
    r001i01b13 IP93 P001
    r001i01b14 IP93 P001
```

```
r001i01b15 IP93 P001
r001i02b00 IP93-BASEIO P002
r001i02b01 IP93-DISK P002
r001i02b02 IP93-INTPCIE P002
r001i02b03 IP93 P002
r001i02b04 IP93 P002
r001i02b05 IP93 P002
r001i02b06 IP93 P002
r001i02b07 IP93 P002
r001i02b08 IP93 P002
r001i02b09 IP93 P002
r001i02b10 IP93 P002
r001i02b11 IP93 P002
r001i02b12 IP93 P002
r001i02b13 IP93 P002
r001i02b14 IP93 P002
r001i02b15 IP93 P002
r002i01b00 IP93-BASEIO P003
r002i01b01 IP93-DISK P003
r002i01b02 IP93 P003
r002i01b03 IP93 P003
r002i01b04 IP93 P003
r002i01b05 IP93 P003
r002i01b06 IP93 P003
r002i01b07 IP93 P003
r002i01b08 IP93 P003
r002i01b09 IP93 P003
r002i01b10 IP93 P003
r002i01b11 IP93 P003
r002i01b12 IP93 P003
r002i01b13 IP93 P003
r002i01b14 IP93 P003
r002i01b15 IP93 P003
r002i02b00 IP93-BASEIO P004
r002i02b01 IP93-DISK P004
r002i02b02 IP93 P004
r002i02b03 IP93 P004
r002i02b04 IP93 P004
r002i02b05 IP93 P004
r002i02b06 IP93 P004
r002i02b07 IP93 P004
```

```
r002i02b08 IP93 P004
r002i02b09 IP93 P004
r002i02b10 IP93 P004
r002i02b11 IP93 P004
r002i02b12 IP93 P004
r002i02b13 IP93 P004
r002i02b14 IP93 P004
r002i02b15 IP93 P004
```

```
Partitions:      4
partition001 BMCs: 16
partition002 BMCs: 16
partition003 BMCs: 16
partition004 BMCs: 16
```

3. Use can also use the `hwcfg` command to display the four partitions, as follows:

```
CMC:rlilc> hwcfg
NL5_RATE=5.0
PARTITION=1 ..... 16/64 BMC(s)
PARTITION=2 ..... 16/64 BMC(s)
PARTITION=3 ..... 16/64 BMC(s)
PARTITION=4 ..... 16/64 BMC(s)
```

4. To reset the system and boot the four partitions, use the following commands:

```
CMC:rlilc> power on
CMC:rlilc> power reset "p*"
```

Note: In the `power reset "p*"` command, above, quotes are required to prevent shell expansion.

5. Use the `uvcon` command to open consoles to each partition and boot the partitions. Open a console to partition one, as follows:

```
CMC:rlilc> uvcon p1
uvcon: attempting connection to localhost...
uvcon: connection to SMN/CMC (localhost) established.
uvcon: requesting baseio console access at partition 1 (r001i01b00)...
uvcon: tty mode enabled, use 'CTRL-]' 'q' to exit
uvcon: console access established (OWNER)
uvcon: CMC <--> BASEIO connection active
```

```

*****
***** START OF CACHED CONSOLE OUTPUT *****
*****
***** [20100513.215944] BMC r001i01b15: Cold Reset via NL broadcast reset
***** [20100513.215944] BMC r001i01b07: Cold Reset via NL broadcast reset
***** [20100513.215945] BMC r001i01b13: Cold Reset via NL broadcast reset
***** [20100513.215945] BMC r001i01b05: Cold Reset via NL broadcast reset
***** [20100513.215945] BMC r001i01b06: Cold Reset via NL broadcast reset
***** [20100513.215946] BMC r001i01b10: Cold Reset via NL broadcast reset
***** [20100513.215946] BMC r001i01b09: Cold Reset via NL broadcast reset
***** [20100513.215945] BMC r001i01b11: Cold Reset via NL broadcast reset
***** [20100513.215945] BMC r001i01b12: Cold Reset via NL broadcast reset
***** [20100513.215945] BMC r001i01b04: Cold Reset via NL broadcast reset
***** [20100513.215945] BMC r001i01b08: Cold Reset via NL broadcast reset
***** [20100513.215946] BMC r001i01b02: Cold Reset via NL broadcast reset
***** [20100513.215945] BMC r001i01b00: Cold Reset via NL broadcast reset
***** [20100513.215945] BMC r001i01b14: Cold Reset via NL broadcast reset
***** [20100513.215947] BMC r001i01b09: Cold Reset via ICH
***** [20100513.215946] BMC r001i01b12: Cold Reset via ICH
***** [20100513.215947] BMC r001i01b10: Cold Reset via ICH
***** [20100513.215947] BMC r001i01b11: Cold Reset via ICH
***** [20100513.215947] BMC r001i01b02: Cold Reset via ICH
***** [20100513.215947] BMC r001i01b00: Cold Reset via ICH
***** [20100513.215953] BMC r001i01b03: Cold Reset via NL broadcast reset
***** [20100513.220011] BMC r001i01b01: Cold Reset via NL broadcast reset
***** [20100513.220012] BMC r001i01b08: Cold Reset via NL broadcast reset
***** [20100513.220012] BMC r001i01b07: Cold Reset via NL broadcast reset
***** [20100513.220011] BMC r001i01b15: Cold Reset via NL broadcast reset
***** [20100513.220012] BMC r001i01b06: Cold Reset via NL broadcast reset
***** [20100513.220012] BMC r001i01b05: Cold Reset via NL broadcast reset
***** [20100513.220012] BMC r001i01b14: Cold Reset via NL broadcast reset
***** [20100513.220012] BMC r001i01b13: Cold Reset via NL broadcast reset
***** [20100513.220011] BMC r001i01b04: Cold Reset via NL broadcast reset
***** [20100513.220012] BMC r001i01b03: Cold Reset via NL broadcast reset
***** [20100513.220013] BMC r001i01b09: Cold Reset via NL broadcast reset
***** [20100513.220013] BMC r001i01b10: Cold Reset via NL broadcast reset
***** [20100513.220013] BMC r001i01b11: Cold Reset via NL broadcast reset
***** [20100513.220012] BMC r001i01b12: Cold Reset via NL broadcast reset
***** [20100513.220012] BMC r001i01b02: Cold Reset via NL broadcast reset
***** [20100513.220012] BMC r001i01b00: Cold Reset via NL broadcast reset

```

```
***** [20100513.220014] BMC r001i01b09: Cold Reset via ICH
***** [20100513.220014] BMC r001i01b10: Cold Reset via ICH
***** [20100513.220014] BMC r001i01b11: Cold Reset via ICH
***** [20100513.220013] BMC r001i01b12: Cold Reset via ICH
***** [20100513.220013] BMC r001i01b02: Cold Reset via ICH
***** [20100513.220016] BMC r001i01b00: Cold Reset via ICH
***** [20100513.220035] BMC r001i01b14: Cold Reset via NL broadcast reset
***** [20100513.220035] BMC r001i01b06: Cold Reset via NL broadcast reset
***** [20100513.220034] BMC r001i01b15: Cold Reset via NL broadcast reset
***** [20100513.220035] BMC r001i01b05: Cold Reset via NL broadcast reset
***** [20100513.220034] BMC r001i01b01: Cold Reset via NL broadcast reset
***** [20100513.220035] BMC r001i01b07: Cold Reset via NL broadcast reset
. . . . .
Hit [Space] for Boot Menu.
ELILO boot:
```

Note: Use the `uvcon` command to open consoles on the other three partitions and boot them. The system will then have four single system images.

- 6. Use the `hwcfg -c partition` command to clear the four partitions, as follows:

```
CMC:rlilc> hwcfg -c partition
PARTITION=0
PARTITION=0
```

Note: This will take several minutes on large systems.

- 7. To reset the system and boot it as a single system image (one partition), use the following command:

```
CMC:rlilc> power reset "p"
```

Making Array Services Operational

This section describes how to get Array Services operational on your system. For detailed information on Array Services, see chapter 3, “Array Services”, in the *Linux Resource Administration Guide*.

Standard Array Services is installed by default on an SGI ProPack 7 SP1 system. To install Secure Array Services, use the YaST Software Management and use the **Filter->search** function to search for secure array services by name (`sarraysvcs`).

Procedure 2-3 Making Array Services Operational

To make Array Services operational on your system, perform the following steps:

Note: Most of the steps to install array services is now performed automatically when the array services RPM is installed. To complete installation, perform the steps that follow.

1. Make sure that the setting in the `/usr/lib/array/arrayd.auth` file is appropriate for your site.
-



Caution: Changing the `AUTHENTICATION` parameter from `NOREMOTE` to `NONE` may have a negative security impact on your site.

2. Make sure that the list of machines in your cluster is included in one or more array definitions in `/usr/lib/array/arrayd.conf` file.
3. To determine if Array Services is correctly installed, run the following command:

```
array who
```

You should see yourself listed.

Floating Point Assist Warnings from Applications

Some applications can generate an excessive number of kernel `KERN_WARN "floating point assist"` warning messages. This section describes how you can make these messages disappear for specific applications or for specific users.

An application generates a "floating point assist" trap when a floating point operation involves a corner case of operand value(s) that the Itanium processor cannot handle in hardware and requires kernel emulation software to complete.

Sometimes the application can be recompiled with specific options that will avoid such problematic operand value(s). Using the `-ffast-math` option with `gcc` or the `-ftz` option with the Intel compiler might be effective. Other instances can be avoided by recoding the application to avoid denormalized or non-finite operands.

When recoding or recompiling is ineffective or infeasible, the user running the application or the system administrator can avoid having these "floating point assist" syslog messages appear by using the `prctl(1)` command, as follows:

```
% prctl --fpemu=silent command
```

The *command* and every child process of the *command* invoked with `prctl` will produce no "floating point assist" messages. The *command* can be a single application that is producing unwanted syslog messages or it may be any shell script. For example, the "floating point assist" messages for the user as a whole, that is, for all applications that the user may execute, can be silenced by changing the `/etc/passwd` entry of the user to invoke a custom script at login, instead of executing (for instance) `/bin/bash`. That custom script then launches the user's high level shell using the following:

```
prctl --fpemu=silent /bin/bash
```

Even if the syslog messages are silenced, the "assist" traps will continue to occur and the kernel software will still handle the problem. If these traps occur at a high enough frequency, the application performance may suffer and notification of these occurrences are not logged.

The syslog messages can be made to reappear by executing the `prctl` command again, as follows:

```
% prctl --fpemu=default command
```

Unaligned Access Warnings

The section describes unaligned access warnings, as follows:

- The kernel generated unaligned access warnings in syslog and on the console, when applications do misaligned loads and stores. This is normally a sign of badly written code and an indication that the application should be fixed.
- Use the `prctl(1)` command to disable these messages on a per application basis.
- SLES offers a new way allowing system administrators to disable these messages on a system wide basis. This is generally discouraged, but useful for the case where a system is used to run third-party applications which cannot be fixed by the system owner.

In order to disable these messages on a system wide level, do the following as root:

```
echo 1 > /proc/sys/kernel/ignore-unaligned-usertrap
```


System Operation

This chapter describes some basic system operations using the chassis manager controller (CMC) commands and covers the following topics:

- "SGI Altix UV System Control Network" on page 25
- "Connecting to the UV System Controller Network Using the CMC" on page 26
- "Power on and Booting an Altix UV System from Complete Power Off" on page 26
- "Power off an Altix UV System" on page 28
- "Power NMI to Drop into KDB" on page 28
- "Power NMI to Drop into KDB" on page 28
- "Upgrading System BIOS" on page 28

SGI Altix UV System Control Network

SGI Management Center (SMC) software running on the system management node (SMN) provides a robust graphical interface for system configuration, operation, and monitoring. For more information, see *SGI Management Center System Administrator's Guide*.

The SGI Altix UV 1000 and Altix UV 100 system control network is a private, closed network. It is not to be reconfigured in any way different from the standard UV installation, nor is it to be directly connected to any other network. The Altix UV system control network does not accommodate additional network traffic, routing, address naming other than its own schema, and DHCP controls other than its own configuration. The system control network also is not security hardened, nor is it tolerant of heavy network traffic, and is vulnerable to Denial of Service attacks. The system management node acts as a gateway between the UV system control network and any other networks. For more information on the SGI Altix UV system control network, see the *SGI Altix UV 100 System User's Guide* or the *SGI Altix UV 1000 System User's Guide*.

Connecting to the UV System Controller Network Using the CMC

The Altix UV software controller is designed to manage and monitor the individual blades in SGI Altix UV systems. Depending on your system configuration, you can monitor and operate the system from the system management node (SMN) or on smaller systems, such as, the Altix UV 100 from the CMC itself. UV 1000 systems up to 16 racks (four building blocks, also called one super block) can also be controlled and monitored from a CMC in the system. For more information, see the “Connecting to the UV System Controller Network” in the *SGI Altix UV CMC Controller Software User’s Guide*.

Power on and Booting an Altix UV System from Complete Power Off

To boot an SGI Altix UV 1000 or UV 100 system from complete power off, perform the following steps:

1. Make sure the power breakers are on.
2. Establish a serial connection to the CONSOLE on the CMC or establish a network connection to the CMC. For a network connection, your PC must be connected to the CMC (via the network connection) and have its `/etc/hosts` file setup to include the CMCs. See "Connecting to the UV System Controller Network Using the CMC" on page 26.
- 3.

```
ssh root@hostname-cmc
SGI Chassis Manager Controller, Firmware Rev. 0.0.22

CMC:r1i1c>
```

Typically, the default password set out of the factory is `root`. The CMC prompt appears. `CMC:r1i1c` refers to rack 1, IRU 1, CMC.

If the host name is not set up in the PC/workstation’s hosts file, you can simply use the IP address of the CMC, as follows:

```
ssh root@<IP-ADDRESS>
```

4. Power up your Altix UV system using the `power on` command, as follows:

```
CMC:r1i1c> power on
```

Note: You can open a second window on the CMC, `ssh root@hostname-cmc` and use the `uvcon` command to open a console and watch the system power on.

5. Open a second console to the CMC using the `uvcon` command to see the system power on, as follows:

```
ssh root@hostname-cmc
SGI Chassis Manager Controller, Firmware Rev. 0.0.22

CMC:rlilc> uvcon
uvcon: attempting connection to localhost...
uvcon: connection to SMN/CMC (localhost) established.
uvcon: requesting baseio console access at r001i01b00...
uvcon: tty mode enabled, use 'CTRL-]' 'q' to exit
uvcon: console access established
uvcon: CMC <--> BASEIO connection active
*****
*****  START OF CACHED CONSOLE OUTPUT  *****
*****
***** [20100512.143541] BMC r001i01b10: Cold Reset via NL broadcast reset
***** [20100512.143541] BMC r001i01b07: Cold Reset via NL broadcast reset
***** [20100512.143540] BMC r001i01b08: Cold Reset via NL broadcast reset
***** [20100512.143540] BMC r001i01b12: Cold Reset via NL broadcast reset
***** [20100512.143541] BMC r001i01b14: Cold Reset via NL broadcast reset
***** [20100512.143541] BMC r001i01b04: Cold Reset via NL
```

Note: Use `CTRL-] q` to exit the console.

6. Depending upon the size of your system, it can take 5 to 10 minutes for the Altix UV system to power on. When the `shell>` prompt appears, enter `fs0`, as follows:

```
shell> fs0
```

7. At the `fs0` prompt, enter `boot`, as follows:

```
fs0> boot
```

ELILO Linux Boot loader is called and various SGI configuration scripts are run and the SUSE Linux Enterprise Server 11 SP1 installation program appears.

Power off an Altix UV System

To power down the Altix UV system, use the `power off` command, as follows:

```
CMC:r1i1c> power off
==== r001i01c (PRI) ====
```

You can use the `power status` command, to check the power status of your system, as follows:

```
CMC:r1i1c> power status
==== r001i01c (PRI) ====
on: 0, off: 32, unknown: 0, disabled: 0
```

Power NMI to Drop into KDB

To send a nonmaskable interrupt (NMI) signal from the power command to the CMC to drop into the kernel debugger (KDB), use the `power nmi` command, as follows:

```
CMC:r1i1c> power nmi

Entering kdb (current=0xffff8aa3fe11c040, pid 0) on processor 7 due to NonMaskable
  r15 = 0x0000000000000000      r14 = 0x0000000000000000
  r13 = 0x0000000000000000      r12 = 0x0000000000000000
  bp = 0xffffffff81927380      bx = 0xffff8ac1ff11dfd8
  r11 = 0xffffffff8101a2c0      r10 = 0xffff88000bee7fd8
  r9 = 0x00000000ffffffff       r8 = 0x0000000000000000
  ax = 0x0000000000000000       cx = 0x0000000000000000
  dx = 0x0000000000000000       si = 0xffff8ac1ff11dfd8
  di = 0xffffffff81a2b308      orig_ax = 0xffffffffffffffff
  ip = 0xffffffff8100ad42       cs = 0x0000000000000010
  flags = 0x0000000000000246    sp = 0xffff88000bee7ff0
  ss = 0x0000000000000018      &regs = 0xffff88000bee7f58
[7]kdb>
```

Upgrading System BIOS

To upgrade the compute blade PROM, perform the following steps:

1. From the CMC prompt, to show the current PROM level, perform the following command:

```
CMC:r1i1c> showbios  
Flashed on Sat May 1 14:14:45 UTC 2010 was bios.latest.fd (20100429_1603)
```

2. Get the newest PROM image from SupportFolio Online at <http://support.sgi.com/>.
3. Copy the latest BIOS to a directory on the CMC in `/work/bmc/common/`. An example directory is, as follows:

```
CMC:r1i1c> ls  
bios.latest.fd flashbios
```

4. Use the `flashbios` command to flash the compute blade BIOS, as follows:

```
CMC:r1i1c> flashbios  
Using default bios: bios.latest.fd  
Checking processor status on all nodes....  
Done. System is read for BIOS flash update  
Flashing bios bios.lastest.fd (20100429_1603) This will take several minutes.  
...
```

For more information on using CMC software, see the *SGI Altix UV CMC Controller Software User's Guide*.

Index

A

- Array Services
 - making Array Services operational, 20

C

- connecting to the system control network, 26
- CPU frequency scaling, 7

I

- Intel processor Turbo boost feature, 8

L

- license mechanism, 1
 - floating license keys, 4
 - installing license keys, 3
 - LK, 1
 - lkd program, 1
 - verifying license keys, 3

S

- SGI LK license facility, 1

- Single system image
 - maximum partitions, 12
- system control network, 25
 - connecting, 26
- system partitioning, 8
 - advantages, 10
 - allows variable partition sizes, 10
 - creating a large, shared-memory cluster, 10
 - provide high performance clusters, 11
 - provides fault containment, 10
 - configuring the XPNET driver as a network driver, 15
 - how to partition a system, 15
 - installing partitioning software, 14
 - XP kernel module, 14
 - XPC kernel module, 14
 - XPMEM kernel module, 14
 - XPNET kernel module, 14
 - limitations, 11
 - partition, 8
 - partitioning a system, how to, 15
 - setting up networking between partitions, 15
 - supported configurations, 12