SGI Linux™ Advanced Cluster Environment Administrator's Guide

007–4228–004

CONTRIBUTORS

Written by Lori Johnson

Production by Glen Traefald

Edited by Rick Thompson

Illustrated by Chris Wengelski

Contributions by Julie Boney, Mark Goodwin, Dean Johnson, Lynne Johnson, Aaron Laffin, Susan Lee, Jenny Leung, Richard Logan, Ken McDonell, David Metcalfe, Jonathan Sparks, Jeff Zurshmeide

New Features in This Guide

This version of ACE includes the following:

- SGI SystemImager
- OpenPBS
- Clarifications to PCP information

Record of Revision

Version	Description
001	January 2000 Original publication.
002	May 2000 Supports Linux ACE 1.2.
003	July 2000 Supports Linux ACE 1.3.
004	October 2000 Supports Linux ACE 1.4.

Contents

Figures

Tables

About This Guide

This document discusses Linux Advanced Cluster Environment (ACE) release 1.4 running on SGI 1400 and SGI 1200 systems. It assumes that the Linux operating system (SuSE 6.4, Red Hat 6.2, or TurboLinux 6.0/6.02) has already been installed on the systems in your cluster. This release also supports SGI ProPack for Linux version 1.4; by using SGI ProPack, you can avoid rebuilding the kernel for use with some products.

This document explains how to perform configuration, installation, and administrative tasks specific to Linux ACE.

Related Publications and Web Sites

The following documents and Web sites contain additional information that may be helpful:

- ACE Web Site: `http://oss.sgi.com/projects/ace`
- ACE man pages:
  - `aceinfo`(1), which displays information about the ACE environment
  - `cping`(1), which performs a cluster-wide `ping`
  - `crsh`(1), which executes a command on all nodes in the cluster
  - `confidence`(1), which tests that the cluster has basic PBS and MPICH functionality.
  - `creboot`(8), which reboots all nodes in the cluster
  - `chalt`(8), which halts all nodes in the cluster
  - `uninstallace`(8), which removes ACE software from the cluster
- SGI ProPack:
  - SGI ProPack site: `http://support.sgi.com/linux/index.html`
  - *SGI ProPack 1.4 for Linux Start Here*
- General Linux Clusters:

  - *How to Build a Beowulf: A Guide to the Implementation and Application of PC Clusters.* Thomas L. Sterling, John Salmon, Donald J. Becker, Daniel F. Savarese. ISBN: 026269218X.
  - How to Build a Beowulf: a Tutorial `http://www.cacr.caltech.edu/beowulf/tutorial/tutorial.html`
  - *Beowulf Installation and Administration HOWTO* `http://buweb.parl.clemson.edu/doc_project/BIAA-HOWTO/Beowulf-Installation-and-Administration-HOWTO.html`

- Portable Batch System (PBS):
  - PBS Web site: `http://www.openpbs.org`
  - *PBS Administrator's Guide* on the installed ACE system in `/usr/doc/OpenPBS-`*version*`/pbs_admin_guide.pdf` (and `.ps`)
- Performance Co-Pilot (PCP) (orderable through SGI):
  - PCP open source Web site: `http://oss.sgi.com/projects/pcp`
  - PCP product web page: `http://www.sgi.com/software/co-pilot`
  - Manuals available from the SGI online Technical Publications Library `http://techpubs.sgi.com/`:
    - *Performance Co-Pilot User's and Administrator's Guide*
    - *Performance Co-Pilot Programmer's Guide*
- Ethernet MPICH:
  - `/usr/doc/mpich-`*version*`/install.ps.gz` for the *Installation Guide for mpich, a Portable Implementation of MPI*
  - `/usr/doc/mpich--`*version*`/guide.ps.gz` for the *User's Guide for* `mpich`*, a Portable Implementation of MPI*
  - `/usr/doc/mpich--`*version*`/index.html` for online help
  - MPICH home page: `http://www-unix.mcs.anl.gov/mpi/mpich`
- Myricom GM MPICH:
  - `/usr/doc/mpich-`*version*`/install.ps.gz` for the *Installation Guide for mpich, a Portable Implementation of MPI*

  - /usr/doc/mpich/*version*/guide.ps.gz for the *User's Guide for* mpich*, a Portable Implementation of MPI*
  - /usr/doc/mpich/*version*/www/index.hmlt for online help
  - Myricom home page: http://www.myri.com/
- SGI SystemImager:
  - *SGI SystemImager Software Installationand Administration Guide*available from the SGI online Technical Publications Library http://techpubs.sgi.com/

    The SGI SystemImager tool is based on the VA SystemImager product. For more information on VA SystemImager, see the following site:

    http://www.systemimager.org/
  - Documentation in the /mnt/cdrom/doc/sgi/ace-extras directory on the CD and (after installation) in /usr/share/doc/ace-extras/sgi-SystemImager-*version*/SIGUI_AG directory
- Digi International: http://www.digi.com/
- Comtrol: http://www.comtrol.com

Note: The third-party documents listed here are not orderable from SGI.

Obtaining Publications

To obtain SGI documentation, go to the SGI Technical Publications Library at http://techpubs.sgi.com.

Conventions

The following conventions are used throughout this document:

Convention	Meaning
`command`	This fixed-space font denotes literal items such as commands, files, routines, path names, signals, messages, and programming language structures.
variable	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.

Reader Comments

If you have comments about the technical accuracy, content, or organization of this document, please tell us. Be sure to include the title and document number of the manual with your comments. (Online, the document number is located in the front matter of the manual. In printed manuals, the document number can be found on the back cover.)

You can contact us in any of the following ways:

- Send e-mail to the following address:

  `techpubs@sgi.com`

- Use the Feedback option on the Technical Publications Library World Wide Web page:

  `http://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:

  Technical Publications
  SGI
  1600 Amphitheatre Pkwy., M/S 535
  Mountain View, California 94043–1351

- Send a fax to the attention of "Technical Publications" at +1 650 932 0801.

We value your comments and will respond to them promptly.

Introduction

This document discusses the Linux Advanced Cluster Environment (ACE) software. It explains how to perform configuration, installation, and administrative tasks.

Note: Examples in this document may not show the current RPM numbers.

Terminology

This document uses the following terminology:

- *Node*, which is an SGI 1400 or SGI 1200 computer. A node is a host.
- *Execution node*, which is a node on which a job is run.
- *Head node*, which is the node that performs administrative tasks for the cluster and is where PBS requests are submitted. Users log into the head node. A head node can also be an execution node, but there can be only one head node.
- *Cluster*, which is made up of one head node and multiple execution nodes.
- *Remote administration node*, which is the node from which Lconsole administrative tasks are performed.
- *Serial multiplexer*, which is a hardware and software product that allows a remote machine to perform administrative tasks . The Digi International EtherLite 16 is the default product available from SGI; the Comtrol RocketPort is an alternative product required for use with the Hoover cluster manager.

Linux ACE supports throughput and capability clusters:

- *Throughput* clusters are used to solve a set of problems simultaneously, using different execution nodes for different problems.
- *Capability* clusters are used to solve a large problem across multiple execution nodes.

Assumptions

The instructions in this document assume the following:

- You have experience with system administration tasks.
- You know the IP addresses and hostnames of each node in the cluster. (The hostname is the output of the `hostname`(1) command.)
- The base Linux operating system (OS) — Red Hat version 6.2, SuSE version 6.4, or TurboLinux 6.0/6.0.2 or later – has already been installed on the systems in your cluster. For more information, see the base OS documentation.

Contents of the Linux ACE Release

This section discusses:

- "Products"
- "Space Requirements"

Products

Linux ACE contains the following products:

- An administration tool (mutually exclusive):
  - Lconsole command-line serial console
  - Hoover cluster manager and serial console
- Portable batch system (PBS)
- Performance Co-Pilot (PCP)
- Message Passing Interface (MPI) software packages that supports parallel programming across a network of computer systems through a technique known as message passing (mutually exclusive):
  - MPICH message-passing library for Ethernet
  - Myricom's MPICH message-passing library and Myricom's GM driver
- Drivers for the EtherLite 16 (EL16) serial multiplexer

- Extra software
- Evaluation software

You will use the `installace` utility to install the packages on the head and execution nodes you specify.

Administration Tools

This section discusses Lconsole and Hoover.

Lconsole Command-line Serial Console

The Lconsole utility provides a serial port console that lets you do the following:

- Use serial lines to get the boot prompt in case of a network failure
- Perform a remote reset in case of hardware failure

Note: Lconsole requires that you use a serial multiplexer. You should not install both Lconsole and Hoover.

For information about using Lconsole, see Chapter 11, "Lconsole Utility", page 69.

Hoover Cluster Manager and Serial Console

Hoover is a graphical user interface (GUI) for administration tasks on the nodes or cluster. It lets you monitor node status and perform certain system administration functions such as the following:

- Power cycle or reset the node
- View the System Event Log (SEL)
- Open a serial console to the node

By default, Hoover is organized into two tabbed panels: the top panel is the status and navigation window and the bottom panel contains data panels that result from actions in the top window, such as requesting a SEL for a node.

You should not install both Lconsole and Hoover.

Note: If you intend to use Hoover, you should install the Comtrol RocketPort serial port concentrator, not the Digi EtherLite 16 device.

The Hoover binaries are located in `/usr/bin`.

For more information, see the following:

- `/usr/doc/vacm-hoover-`*version*`/`
- `/usr/doc/vacm-`*version*`/`
- `vacm.sourceforge.net`
- `oss.sgi.com/projects/vacm`

PBS Batch Scheduling Tool

The Portable Batch System (PBS) is developed by Veridian Systems.

The binaries for PBS are installed in the `/usr/local/sbin` and `/usr/local/bin` directories. The configuration files are in the `/usr/spool/pbs` directory.

For more information, see the following:

- *PBS Administrator's Guide* found in `/usr/doc/OpenPBS-`*version*`/pbs_admin_guide.pdf`
- `http://www.openpbs.org/`

Performance Co-Pilot (PCP)

Performance Co-Pilot (PCP) is a framework and services to support system-level performance monitoring and performance management. The PCP infrastructure provides a unifying abstraction for all of the interesting performance data in a system. It also allows client applications to easily retrieve and process any subset of that data.

PCP utilizes a client-server architecture to provide the following:

- Tools for visualizing all hosts in the cluster. For details, see the manual pages for `clustervis`(1) and `pmgcluster`(1).
- Tools for instrumenting MPI applications so that they can export MPI library call counters and other performance statistics into the PCP framework, and a tool for

visualizing this data. For details, see the manual pages for `pcp_mpi`(3), `mpimon`(1), `mpivis`(1), `pmdampi`(1), and `pmfindash`(1), and the examples in the `/usr/share/pcp/demos/pcp_mpi` directory.

- A PCP agent for extracting performance statistics from one or more Myrinet switches and exporting this into the PCP framework, and a tool for visualizing this data. For details, see the manual pages for `myrinetmon`(1), `pmdamyrinet`(1) and the `/var/pcp/pmdas/myrinet/README` file.
- Centralized real-time monitoring of the performance on distributed nodes
- Extensible collection framework where new sources of performance data can be added using a plug-in architecture
- Monitoring tools providing textual, 2-D and 3–D graphical interfaces
- Intelligent rule-based filtering may be used to implement alarms and automated monitoring
- Integrated archive logging and replay for all monitoring tools

The binaries for PCP are installed in `/usr/bin` (public commands) and `/usr/share/pcp/bin` (administrative and private commands). Configuration files reside in `/var/pcp` and log files are in `/var/log/pcp`. The file `/etc/pcp.conf` describes in greater detail where the components of the PCP installation are installed.

For more information, see `http://oss.sgi.com/projects/pcp`.

The version of PCP packaged with Linux ACE includes features that are derived from the IRIX version of PCP. For more information, see:
`http://www.sgi.com/software/co-pilot/`

Detailed instructions for using and configuring the PCP monitoring tools are available in the *Performance Co-Pilot User's and Administrator's Guide.* Users who wish to develop their own performance monitoring tools and/or their own PCP agents, should read the *Performance Co-Pilot Programmer's Guide.* Both of these books are available online from the SGI Technical Publications Library: `http://techpubs.sgi.com/`

These documents describe the IRIX version of the PCP product, which differs only slightly from the Linux version.

Table 1-1 IRIX and Linux Differences

Directory	IRIX	Linux
`rc`/`startup` scripts	`/etc/init.d`	`/etc/rc.d/init.d`
Private PCP binaries	`/usr/pcp/bin`	`/usr/share/pcp/bin`
Shared PCP files (shareable for diskless)	`/usr/pcp`	`/usr/share/pcp`
Directory of manual pages	`/usr/share/catman`	`/usr/man`
PCP logs	`/var/adm/pcplog`	`/var/log/pcp`
PCP documentation	`/var/pcp`	`/usr/doc/pcp-`*Version*
Directory for PCP demos and examples	`/var/pcp/demos`	`/usr/share/pcp/demos`
`magic`, as used by `file`(1)	`/etc/magic`	`/usr/share/magic`

Source code examples for using PCP are installed in the `/usr/share/pcp/demos` directory. In particular, the `/usr/share/pcp/demos/pcp_mpi` directory contains an example MPI application that has been instrumented to export MPI library call counts and statistics into the PCP framework. Once this is set up, you can use the `mpivis`(1) tool to visualize the MPI call rates and performance of MPI applications that have been instrumented in this way. For detailed instructions, see the `mpivis`(1), `pmfindash`(1), and `pcp_mpi`(3) manual pages.

For a summary of the differences between this version of PCP and the previous release, see the following files:

- `/usr/doc/pcp-`*version*`/CHANGELOG`
- `/usr/doc/pcp-pro-`*version*`/CHANGELOG`
- `/usr/doc/pcp-pro-`*version*`/CHANGELOG`

Message Passing

You can only install one of the following:

- MPICH message-passing library for Ethernet
- Myricom's enhanced MPICH message-passing library and Myricom's GM driver

Binaries are located in the `/usr/bin` directory. Header files are located in the `/usr/include` directory.

The Ethernet and Myricom versions use the same binaries, header files, and documentation, but they have different libraries:

- Ethernet MPICH libraries are in `/usr/build/LINUX/ch_p4/lib`
- Myricom MPICH libraries are in `/usr/build/LINUX/ch_gm/lib`

Note: Installing `mpich` or `gm-mpich` will overwrite some of the header files used by `lam`, another MPI product.

MPICH for Ethernet

Ethernet MPICH is produced by Argonne National Laboratory.

For more information, see the following:

- `/usr/doc/mpich-`*version*`/install.ps.gz` for the *Installation Guide for mpich, a Portable Implementation of MPI*
- `/usr/doc/mpich--`*version*`/guide.ps.gz` for the *User's Guide for mpich, a Portable Implementation of MPI*
- `/usr/doc/mpich--`*version*`/index.html` for online help

GM MPICH for Myrinet

The Myricom MPICH release has two products from Myricom:

- MPICH
- `gm` drivers

For more information, see the following:

- `/usr/doc/mpich-`*version*`/install.ps.gz` for the *Installation Guide for mpich, a Portable Implementation of MPI*
- `/usr/doc/mpich/`*version*`/guide.ps.gz` for the *User's Guide for mpich, a Portable Implementation of MPI*
- `/usr/doc/mpich/`*version*`/www/index.hmlt` for online help

- `http://www.myri.com/`

Extra Software: SGI SystemImager

SGI SystemImager enables system administrators to easily replicate an entire Linux system onto a set of Linux machines, even when the machines have unformatted and unpartitioned disks. The GUI makes it easy to keep the hard drives of those systems synchronized after initial replication.

If you are a system administrator who manages mostly identical machines, you will find the SystemImager tool useful in any or all of the following situations:

- You are managing a new cluster with preinstalled or nonpreinstalled machines.
- You need to add preinstalled or non-preinstalled machines to a cluster.
- You need to recover a machine or group of machines from a crashed hard disk.
- You are installing or upgrading software.
- You need to change configurations (for example, to modify system files).
- You need to resynchronize clients after files have been added or deleted.

For more information, see *SGI SystemImager Software Installationand Administration Guide*. To view this document, install the documentation RPM files found in `/mnt/cdrom/ace-extras/systemimager`. They will be installed in `/usr/share/doc/ace-extras/sgi-SystemImager-`*version*`/SIGUI_AG`.

The SGI SystemImager tool is based on the VA SystemImager product. For more information on VA SystemImager, see the following site:

`http://www.systemimager.org/`

Evaluation Software

ACE provides additional software on an evaluation basis. These products are made available for evaluation purposes and as such require procurement of a evaluation license from the respective companies. Agreement has been made with each company to provide at least a 30–day evaluation license.

Note: SGI does not provide support for evaluation software.

For details, see the following file:

```
/mnt/cdrom/ace-demos/README.TXT
```

Space Requirements

To perform the installation, you must have sufficient amounts of free space in `/tmp` and `/usr` on each node in the cluster. For example, to install the `ethernet` or `myrinet` package sets, you should have at least the following amounts of free space on each node in the cluster:

- 11 Mbytes in `/tmp`
- 26 Mbytes in `/usr`

Table 1-2 shows the amount of space required by each package.

Note: You will install one of the following:

- `mpich`
- `gm-mpich`

If you use Myrinet, you should also install the `gm` driver package.

Table 1-2 Approximate Space Requirements Per Package (in Mbytes)

RPM Package	Head Node	Execution Node
`gm`	3.5	3.5
`gm-mpich`	1.7	1.7
`hoover`	2.3	(none)
`lconsole`	0.2	(none)
`mpich`	2.2	2.2
`pbs-client` (mom)	(none)	7.7
`pbs` (server)	13.0	(none)
`pcp`	2.0	2.0
`pcp-ace`	0.2	0.2

RPM Package	Head Node	Execution Node
pcp-pro	11.3	11.3
sgi-ace-tools	0.02	(none)
sgi-acedocs	1.1	(none)
hoover	2.5	(none)
sysimager	23	(none)

Ensure that you have enough disk space before installing Linux ACE.

The following sections provide more information about the products supplied with ACE. You should use the individual product documentation found in /usr/doc. However, you do not need to perform individual product installations because the products are installed as part of the Linux ACE installation procedure. You will install individual RPMs only if you want to customize the installation.

Topology

There are three supported topologies for Linux ACE:

- A head node and multiple execution nodes
- A node that is both the head node and an execution node, plus other execution nodes
- A single node that is both the head node and the execution node (used for testing purposes)

In all of these cases, the nodes must be SGI 1400 or SGI 1200 systems. You can use an Ethernet or Myricom network. If you want to use the Lconsole serial-port console feature, then you must also use a serial multiplexer; if you want to use the Hoover cluster manager, you must use a Rockport multiplexer.

The following figures depict the three topologies.

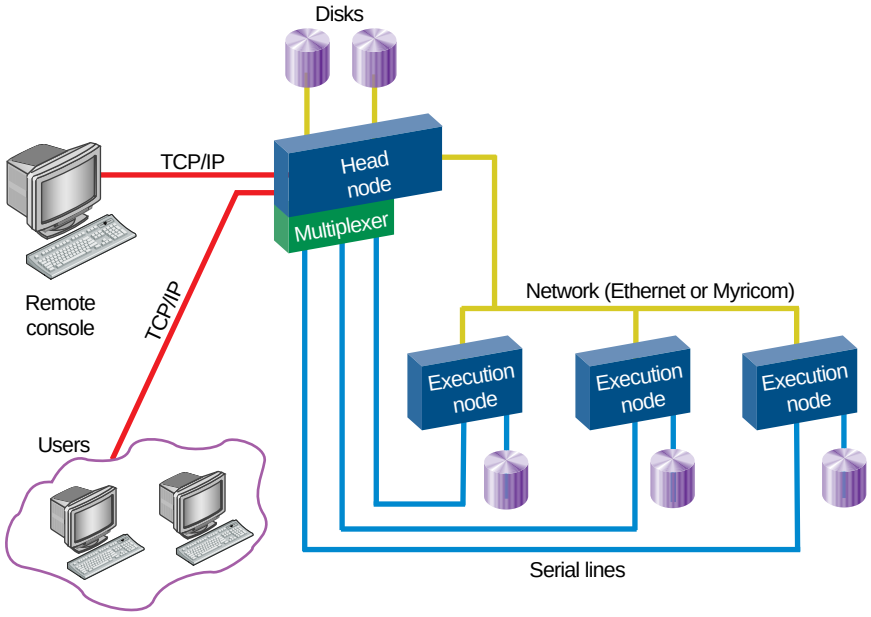

Figure 2-1 Head Node and Multiple Execution Nodes

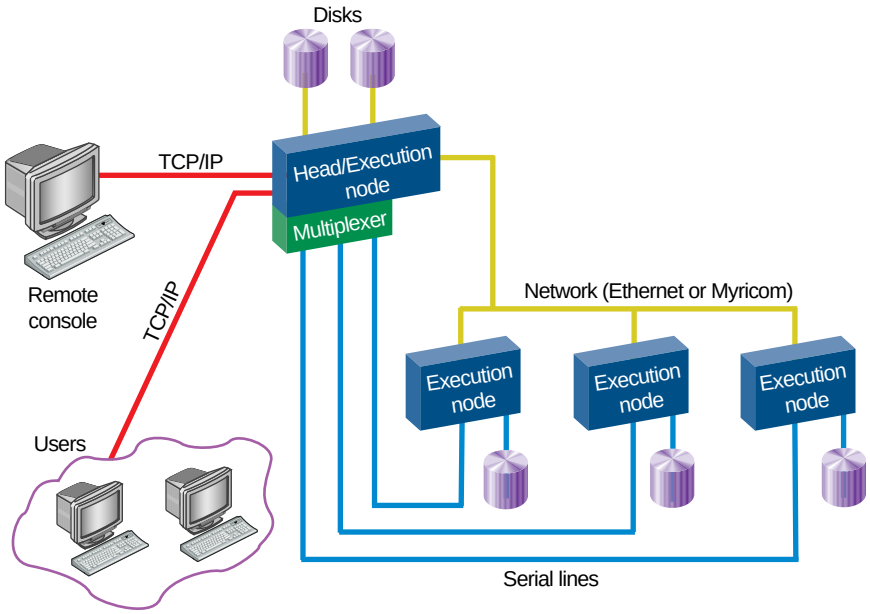

Figure 2-2 Head/Execution Node and Multiple Execution Nodes

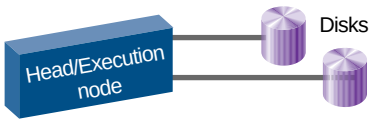

Figure 2-3 Single Node

Preinstallation Requirements

The following are required to install ACE software:

- "TCP/IP Connectivity"
- "Remote Execution for `root`"
- "Installation of Additional Base OS Packages"

TCP/IP Connectivity

You must supply the following information:

- For each node:
  - Hostname
  - Principle IP address
  - Gateway
- For every interface:
  - IP address
  - Netmask
  - Broadcast
  - Device
  - Network address

How you do this depends upon your base OS:

- For Red Hat, use the `control-panel`(8) or `linuxconf`(8) utility. You can also directly edit files in the `/etc/sysconfig/network` directory.
- For SuSE, use the `yast`(8) utility. You can also directly edit the `/etc/rc.config` file; if you do this, execute the `/sbin/SuSEconfig` command upon completion.

The following section describe modifications that must be made regardless of base OS, and an example of the specific changes that must be made for the Red Hat OS.

Modification of Standard Configuration Files

Modify the following standard configuration files on each node in the cluster.

/etc/conf.modules

Enter the following information in the /etc/conf.modules file:

alias *device_name_of_ethernet_card network_driver*

For example:

```
alias eth0 eeprol00
```

/etc/hosts

The /etc/hosts file on each node must contain at least the IP address and name of the local node. If you do not use a network information service (NIS), it must also contain the IP address and name of each node in the cluster. (For more information about NIS, see the ypbind(8) man page.)

For example, if not using NIS:

- On node node01:

```
#/etc/hosts for node01
192.168.0.1     node01 node01.acme.com
192.168.0.2     node02 node2 node02.acme.com
```

- On node node02:

```
#/etc/hosts for node02
192.168.0.2     node02 node02.acme.com
192.168.0.1     node01 node1 node01.acme.com
```

For more information, see the *Linux Networking-HOWTO*:

http://techpubs.engr.sgi.com/library/dynaweb_docs/linux/usr/HOWTO/Net-HOWTO.html

/etc/hosts.equiv or $HOME/.rhosts

To enable users other than root to use MPICH and PBS, you must allow rsh(1) and rcp(1) access to all nodes in the cluster. There are two methods to achieve this:

- Enter the hostnames of each node in the cluster in the `/etc/hosts.equiv` file for each node in the cluster. (The hostname is the output of the `hostname`(1) command.) You can choose to perform this action within the `installace` script.
- Have each user configure the `$HOME/.rhosts` file on all nodes in the cluster to allow access from all nodes in the cluster. Ensure that the file is only readable and writable by that user for security purposes.

Example of Base OS-Specific Files for Red Hat Only

The files listed in this section pertain to the Red Hat base OS. Other systems will have different files that must be modified.

/etc/HOSTNAME

Enter the fully qualified hostname of the node in the `/etc/HOSTNAME` file.

For example:

- On node `node01`:

```
#/etc/HOSTNAME
node01.acme.com
```

- On node `node02`:

```
#/etc/HOSTNAME
node02.acme.com
```

/etc/resolv.conf

Enter the IP address of each name server to be searched in the `/etc/resolv.conf` file.

`search` *domainname.type*
`nameserver` *node_IP_address*

For example:

```
#/etc/resolv.conf
search acme.com
nameserver 192.168.0.31
nameserver 192.168.0.2
nameserver 192.168.0.1
```

For more information, see the resolver(8) man page.

/etc/securetty

To ensure that root will have rsh(1) access, remove or rename the /etc/securetty file if it exists.

/etc/sysconfig/network

Enter the following information in the /etc/sysconfig/network file:

Note: The HOSTNAME value here overrides the value in /etc/HOSTNAME.

```
#/etc/sysconfig/network
HOSTNAME=hostname
DOMAINNAME=domainname.type
GATEWAY=gateway_IP_address_for_cluster
GATEWAYDEV=interface_card_name_(normally_eth0)
NISDOMAIN=NIS_name
```

Note: A *gateway* is a node on which the routed(8) daemon runs with the -s option to supply routing information.

For example:

- On node node01:

  ```
  #/etc/sysconfig/network
  HOSTNAME=node01.acme.com
  DOMAINNAME=acme.com
  GATEWAY=192.168.0.1
  GATEWAYDEV=eth0
  NISDOMAIN="acmepark"
  ```

- On node node02:

  ```
  #/etc/sysconfig/network
  HOSTNAME=node02.acme.com
  DOMAINNAME=acme.com
  GATEWAY=192.168.0.2
  ```

```
GATEWAYDEV=eth0
NISDOMAIN="acmepark"
```

Note: If you add `NISDOMAIN`, it will set the network information service (NIS) domain name the next time the system is rebooted if `ypbind` is turned on. See the `ypbind`(8) and `chkconfig`(8) man pages.

To reset the NIS domain name immediately, enter the following at the command line:

`domainname` ***NIS_name***

For more information, see the `domainname`(1) man page.

/etc/sysconfig/network-scripts/ifcfg-eth0

Enter the following in the `/etc/sysconfig/network-scripts/ifcfg-eth0` file:

`DEVICE=`*device_name_(normally_eth0)*
`IPADDR=`*IP_address_of_this_node*
`NETMASK=`*netmask_to_reserve_for_subdividing*
`NETWORK=`*IP_address_of_the_network*
`BROADCAST=`*IP_to_represent_broadcasts*
`ONBOOT=`*yes_(start_on_boot_process)* | *no*

Note: The *netmask* is used to interpret and define the network portion (subnets included) of the Internet address. This mask normally contains the bits corresponding to the standard network part as well as the portion of the host part that has been assigned to subnetworks.

For example:

- On node `node01`:

```
#/etc/sysconfig/network-scripts/ifcfg-eth0
DEVICE=eth0
IPADDR=192.168.0.1
NETMASK=255.255.255.0
NETWORK=192.168.0.3
BROADCAST=192.168.0.4
ONBOOT=yes
```

- On node `node02`:

```
#/etc/sysconfig/network-scripts/ifcfg-eth0
DEVICE=eth0
IPADDR=192.168.0.2
NETMASK=255.255.255.0
NETWORK=192.168.0.3
BROADCAST=192.168.0.4
ONBOOT=yes
```

For more information, see the `ifconfig`(8) man page.

Remote Execution for `root`

The `installace` program requires `root` to use the `rsh`(1), `rcp`(1), and `rsync`(1) commands from the machine used to install all nodes in the cluster.

The `installace` script requires that the `root` user has `rsh`(1) and `rcp`(1) access to all nodes in the cluster. This is required for installing the RPM Package Manager packages on the remote node and for verifying system configurations.

To enable users other than `root` to use MPICH and PBS, you must allow `rsh`(1) and `rcp`(1) access to all nodes in the cluster:

- For MPICH, the normal process startup mechanism for an ethernet device on networks is `rsh`. You can select alternatives for MPICH, but doing so will require a source rebuild. For more information, see the *Installation Guide for mpich, a Portable Implementation of MPI*.
- PBS uses `rcp` to stage input and output files to and from remote destinations. Like MPICH, the user must have `rsh` permissions on all nodes in the cluster. For more information, see the *PBS Administrators Guide*.

The `pcp`, `pcp-ace`, and `pcp-pro` packages do not require any special access privileges because they use their own TCP/IP transport services and do not use `rsh`(1) or `rlogin`(1).

Installation of Additional Base OS Packages

In order to use the full capabilities of the ACE packages, you must install additional packages from the base OS release. For information about this process, see the base OS release documentation.

Table 3-1 lists the requirements.

Table 3-1 Base OS Packages Required for ACE Packages

ACE Package	Base OS Package Required
gm	bash, sh
gm-mpich	csh, ksh, rsh, sh
lconsole	gtk, sh
mpich	csh, ksh, rsh, sh
pbs (known as pbs-server in installace)	sh, tcl, wish
pbs-mom (known as pbs-client in installace)	sh
pcp	sh, snmp
pcp-ace	sh, mpi, textutils (version 2.0e or greater)
pcp-pro	libstdc++, sh XFree86
sgi-ace-tools	sh
sgi-acedocs	sh

For example, suppose you want to use the ethernet option to installace to install the mpich , pbs-client, pbs-server, pcp, pcp-pro, pcp-ace, and lconsole set of packages. (See "Install ACE using Interactive Mode", page 24.) To fully implement these packages, you should install the following from your base OS release:

```
csh
gtk
ksh
libstdc++
rsh
```

```
sh
tcl
wish
```

Table 3-1 shows the most important dependencies; it does not list every dependency. For a complete list, use the following command:

rpm -qR ***ace_package_name***

For example:

```
# rpm -qR pbs-mom
/bin/sh
ld-linux.so.2
libc.so.6
libdl.so.2
/bin/sh
libc.so.6(GLIBC_2.0)
libc.so.6(GLIBC_2.1)
```

Verification of Connectivity and Remote Access

Before installing ACE software, verify that you have TCP/IP connectivity and remote access for `root`.

Verify TCP/IP Connectivity

To verify TCP/IP connectivity, enter the following from the installation machine:

ping -c1 ***nodename***

Repeat this for each node in the cluster.

For example, the following output shows that nodes `node02` and `node01` and have TCP/IP connectivity:

```
[root@node01 /root]# ping -c1 node01
PING node01 (192.168.0.1) from 192.168.0.1 : 56 data bytes
64 bytes from 192.168.0.1: icmp_seq=0 ttl=255 time=0.1 ms

--- node01 ping statistics ---
```

```
1 packets transmitted, 1 packets received, 0% packet loss
round-trip min/avg/max = 0.1/0.1/0.1 ms
[root@node01 /root]# ping -c1 node02
PING node02 (192.168.0.2) from 192.168.0.1 : 56 data bytes
64 bytes from 192.168.0.2: icmp_seq=0 ttl=255 time=0.3 ms

--- node02 ping statistics ---
1 packets transmitted, 1 packets received, 0% packet loss
round-trip min/avg/max = 0.3/0.3/0.3 ms
[root@node01 /root]#
```

Verify Remote Access for `root`

To verify that the `root` user can use the `rsh`(1) command from the machine used for installing to all nodes in the cluster, execute an `rsh` command such as the following:

rsh ***nodename*** **uname**

Repeat this for each node in the cluster.

For example, the following output shows that `root` on node `node04` has `rsh` capability to nodes `node02` and `node01`:

```
[root@node04 /root]# rsh node02 uname
Linux
[root@node04 /root]# rsh node01 uname
Linux
```

Installing ACE Software

Note: To perform an upgrade, see Chapter 9, "Upgrading an Existing Cluster", page 55.

You can install Linux ACE using the `installace` utility in either interactive or batch mode. After installing, you can run a simple script to verify that PBS and MPICH are working properly.

The `installace` utility will install the selected Linux ACE software on the specified nodes, and will optionally perform some configuration tasks required for PBS and MPICH. It also creates a file named `/etc/ace/nodes` on the head node that contains the list of nodes in the cluster; the first node listed is the head node.

To get help for `installace`, use the `-h` option. If you want debug information, use the `-v` option.

Note: All nodes in the cluster must be running the Linux operating system in order to use the `installace` utility.

The `installace` utility creates the following files:

- `/etc/ace/nodes`, which contains the hostnames of all nodes in the cluster. The first node listed is the head node.
- `/etc/ace/installace.log`, which contains details of the actions and errors encountered during the install and configuration process.
- `/etc/ace/installace.conf`, which contains the installation configuration information you supply to `installace`. You can use this file to regenerate the cluster. (This file has the same format as the template `install.conf` file.)
- `/etc/ace/manifest`, which contains a list of those RPMs that have been installed on each node.
- `/etc/ace/gmpi.conf`, which contains the `gm` configuration file generated by `installace`. This file is only generated when installing `gm-mpich`.

Install ACE using Interactive Mode

This section tells you how to use `installace` in interactive mode and provides an example.

Interactive Mode Instructions

To install using interactive mode, do the following:

1. Log in as `root` on the host from which you want to perform the installation. Typically, you will do this from the head node.
2. Insert the Linux ACE CD-ROM.
3. Mount the CD-ROM by entering the following:

   # **`/bin/mount /dev/cdrom /mnt/cdrom`**

4. Change to the `/mnt/cdrom` directory:

   # **`cd /mnt/cdrom`**

   Note: You must be in this directory. If you are not, `installace` will not function properly.

5. Start the `installace` utility:

   # **`./installace`**

   If you want to see debug information, use the `-v` option to `installace`.

6. Choose the products you want to install. No matter which package you choose, additional packages for ACE documentation (`sgi-acedocs-print` and `sgi-acedocs`) and tools (`sgi-ace-tools`) will also be installed on the head node .

Installation Entry	Packages to be Installed
ethernet	mpich, pbs-client, pbs-server, pcp, pcp-pro, pcp-ace, lconsole
myrinet	gm, gm-mpich, pbs-client, pbs-server, pcp, pcp-pro, pcp-ace, lconsole
hoover	hoover
lconsole	lconsole
gm	gm
gm-mpich	gm-mpich
mpich	mpich
pbs	pbs-client, pbs-server
pcp	pcp, pcp-pro, pcp-ace
systemimager	systemimager

Note: The following packages are mutually exclusive:

- lconsole and hoover
- mpich and gm-mpich

If you want to choose products individually, you can do so by listing them after the ACE> prompt, separating them with white space.

Examples:

- To install the set of products appropriate to an Ethernet environment (mpich, pbs, pcp, and lconsole), enter the following:

  ```
  ACE> ethernet
  ```

  Note: You can abbreviate the ethernet and myrinet entries to a unique set of characters. For example, eth.

- To install just pcp and lconsole, enter the following:

  ```
  ACE> pcp lconsole
  ```

7. To install the software on multiple nodes, you must provide their hostnames (the output from the hostnames(1) file). Choose the method you prefer:

   - Option 1, enter the hostnames individually at the ACE> prompt. For example:

```
Enter your choice (1, 2, 3, or 4).
ACE> 1

    Enter the hostname of the head node.
    The head node is the host from which users will launch
    execution of parallel jobs.
ACE> node01

    Enter the hostnames for each execution node in the
    cluster.  After you have entered all hostnames,
    enter '.' or <control-D> to end the list.
ACE> node02 node03 node04
ACE> .
```

   Note: If you want to use the head node as an execution node for PBS, you can enter the node name here.

   - Option 2, specify a file containing the list of hostnames. To use this method, you must first create a file that contains a list of the cluster nodes, where the first node listed is the head node. For example, you could create a file named /tmp/ace.config that would contain the following (where node01 is the head node):

```
node01
node02
node03
node04
```

     You can insert comments in the file by using # as the first character on the line.

     You would enter the following:

```
Enter your choice (1, 2, 3, or 4).
ACE> 2

     Enter the full pathname of the file containing your
     list of hostnames.  The first node in the list must
```

```
     be the head node (from which users will launch execution
     of parallel jobs).

Enter the full pathname of the file containing hostnames.
ACE> /tmp/ace.config
```

- Option 3, specify a pattern for hostnames. This option is useful if you have several nodes that have the same hostname except for a number in sequential order. For example, suppose that you have 7 nodes named `node01.acme.com` through `mode07.acme.com`. You would enter the following:

```
Enter your choice (1, 2, 3, or 4).
ACE> 3

    Enter the basename of the hostname of the first node in your
    cluster.  For example, if your hostnames have the form
    "base1.domain.com", enter 'base'.

ACE> node

    Now enter the domain name of each node in your cluster.
    For example, if your hostnames have the form
    "base1.domain.com", enter 'domain.com'.

ACE> acme.com

    Now enter the number at which your node numbering starts.
    For example, if 'base1.domain.com' is the name of your
    head
 node, enter '1'.

ACE> 01

    How many nodes are there in your cluster?  For example,
    if 'base1.domain.com' is the name of your head node and
    you have 8 nodes in your cluster, enter '8'.

ACE> 7
```

- Option 4, install software only on this node. To install only on the node where `installace` is running, enter the following:

  ```
  Enter your choice (1, 2, 3, or 4).
  ACE> 4
  ```

The `installace` utility will verify that all of the nodes are accessible.

8. Specify whether or not the head node should also be configured as an execution node.

Note: This question is only asked if you install PBS and you have not already listed the head node as an execution node.

For example, if the head node will be used only for job submission, enter the following:

```
Should the head node headnode_name also be configured as an execution node?
ACE (y/n)> n
```

9. If you want to use the `installace`(8) utility to allow `rsh`(1).`rsync`(1), and `rcp`(1) access for user accounts to all nodes in the cluster using the `/etc/hosts.equiv` files, enter 1 or 2. For example:

```
     1) Add the specified hostnames to the existing /etc/hosts.equiv files.
        If the file does not exist, create it.
     2) Replace the current /etc/hosts.equiv file with a new one
        containing only the specified cluster hostnames.  The current
        /etc/hosts.equiv will be saved as /etc/hosts.equiv.old.
     3) Do not change the existing /etc/hosts.equiv file.

Enter choice (1, 2, or 3).
ACE> 2
```

For more information, see "`/etc/hosts.equiv` or `$HOME/.rhosts`", page 14.

Interactive Mode Example

The following example shows an entire interactive installation.

Note: The package versions shown in the following output may not match the released system.

```
[root@ace08 /root]# /bin/mount /dev/cdrom /mnt/cdrom
[root@ace08 /root]# cd /mnt/cdrom
[root@ace08 cdrom]# ./installace
[root@ace08 cdrom]# ./installace

    Advanced Cluster Environment v. 1.4 for Linux i686 includes the
    following products:

        pbs       PBS batch-queuing system
        pcp       Performance Co-Pilot
        mpich     MPICH message-passing library - ethernet
        gm-mpich  Myricom's MPICH message-passing library
        gm        Myricom's GM driver
        lconsole  Command-line serial console
        hoover    Hoover cluster manager & serial console

    Enter "ethernet" to select:
        mpich pbs pcp lconsole

    Enter "myrinet" to select:
        gm gm-mpich pbs pcp lconsole

    Or enter the individual products separated by white space.

Please enter all desired products separated by white space.
ACE> ethernet

    Note:  When providing a hostname, use the fully qualified name
    such as "foo.domain.com". If the hostname is resolved on all
    nodes, you can abbreviate it to "foo".

    The installace utility can modify hosts.equiv files on your nodes.
    These files are used by MPICH, GM-MPICH, and PBS.
```

```
    You can choose to provide the information required to edit these files
    or avoid this action by entering the appropriate number:

         1)  Enter the hostnames individually
         2)  Specify a file containing the list of hostnames
         3)  Specify a pattern for hostnames (NAME1.domain.com)
         4)  Install software only on this node

Enter choice (1, 2, 3, or 4).
ACE> 1

    Enter the hostname of the head node.
    The head node is the host from which users will launch
    execution of parallel jobs.
Enter the hostname of the head node.
ACE> ace08

    Enter the hostnames for each execution node in the
    cluster.  After you have entered all hostnames,
    Enter '.' or <control-D> on a new line to end the list.

ACE> ace07
ACE> .

Should the head node ace08 also be configured
as an execution node?
ACE (y/n)> y

    The install process can update the /etc/hosts.equiv file on all
    the hosts in the cluster.  The /etc/hosts.equiv will allow access
    to the hosts.  If /etc/hosts.equiv file is not being used, each user
    must update their .rhosts file to allow rcp/rsh type access to all
    cluster nodes.  Enter one of the following:

      1) Add the specified hostnames to the existing /etc/hosts.equiv files.
         If the file does not exist, create it.
      2) Replace the current /etc/hosts.equiv file with a new one
         containing only the specified cluster hostnames.  The current
         /etc/hosts.equiv will be saved as /etc/hosts.equiv.old.
      3) Do not change the existing /etc/hosts.equiv file.
```

```
Enter choice (1, 2, or 3).
ACE> 1

    Checking rsync/rcp/rsh access ...

    Checking disk space requirements ...

    Installing selected ACE products on node(s)
      ace-tools ........... [yes]
      ace-docs ............ [yes]
      pcp ................. [yes]
      pbs ................. [yes]
      mpich ............... [yes]
      gm .................. [no]
      gm-mpich ............ [no]
      lconsole ............ [yes]
      hoover .............. [no]

     Installing head node software on ace08
    Installing package sgi-ace-tools-1.4-1 on ace08 ...
    Installing package sgi-acedocs-1.4-1 on ace08 ...
    Installing package sgi-acedocs-print-1.4-1 on ace08 ...
    Installing package mpich-1.2.1-1 on ace08 ...
MPICH postinstall script
    Installing package pcp-2.1.9-12 on ace08 ...
    Installing package pcp-pro-2.1.5-2 on ace08 ...
    Installing package pcp-ace-1.3.0-4 on ace08 ...
Creating MPI wrapper
...Done, output in /var/tmp/pcp-ace-1.3.0-4-mpi_wrapper
    Package textutils-2.0e-3 already installed on ace08, skipping.
    Installing package lconsole-1.1-5 on ace08 ...
    Installing package OpenPBS-2-3 on ace08 ...
PBS preinstall script
PBS postinstall script

    Installing execution node software on ace07
    Installing package mpich-1.2.1-1 on ace07 ...
MPICH postinstall script
    Installing package pcp-2.1.9-12 on ace07 ...
    Installing package pcp-pro-2.1.5-2 on ace07 ...
    Installing package pcp-ace-1.3.0-4 on ace07 ...
```

```
Creating MPI wrapper
...Done, output in /var/tmp/pcp-ace-1.3.0-4-mpi_wrapper
    Package textutils-2.0e-3 already installed on ace07, skipping.
    Installing package OpenPBS-mom-2.3 on ace07 ...
PBS mom preinstall script
PBS mom postinstall script

    Configuring selected ACE products on node(s) ...
    Configuring PBS on head node ace08 ...
    Configuring PBS on ace07 ...
    All PBS nodes registered as seen via /usr/local/bin/pbsnodes
    Starting PCP daemons ...
    Starting PCP on ace08 ...
    Starting PCP on ace07 ...

    Post configuration ...
    Adding to the /etc/hosts.equiv file on all nodes ...
    Placed node information in ace08:/etc/ace/nodes
    Placed log information in ace08:/etc/ace/installace.log
    Placed manifest information in ace08:/etc/ace/manifest
    Installed ACE documentation in ace08:/usr/doc/sgi/ace-1.4
    Placed install configuration file in ace08:/etc/ace/installace.conf

    This ACE release also contains third-party evaluation
    software. See /mnt/cdrom/ace-demos/README.TXT for
    more information.

    The installation/configuration process has been completed!
```

Install ACE using Batch Mode

This section tells you how to use `installace` in batch mode and provides an example.

Batch Mode Instructions

To install the Linux ACE software using batch mode, do the following:

1. Log in as `root` on the host from which you want to perform the installation. Typically, you will do this from the head node.
2. Insert the Linux ACE CD-ROM.
3. Mount the CD-ROM by entering the following:

   # **/bin/mount /dev/cdrom /mnt/cdrom**

4. Change to the `/mnt/cdrom` directory:

   # **cd /mnt/cdrom**

   Note: You must be in this directory. If you are not, `installace` will not function properly.

   ---

5. Copy the `install.conf` file template from the CD-ROM (use the `-i` option to prompt before an overwrite):

   # **/bin/cp -i install.conf /tmp/install.conf**

6. Change the permissions on the copy of `install.conf`:

   # **chmod 644 /tmp/install.confj**

7. Edit the `/tmp/install.conf` file so that it contains the required information. Enclose variable settings within double quotation marks. See the comments for more information about each item. For an example of a modified file, see "Batch Mode Example" below.

   Note: Do not install both `lconsole` and `hoover`. These packages are mutually exclusive.

   ---

8. Start the `installace` utility:

   # **./installace /tmp/install.conf**

   If you want to see debug information, use the `-v` option to `installace`.

   You will see output similar to the interactive mode as the `installace` utility proceeds through the installation. If an error occurs, the process will fail.

Batch Mode Example

This section shows you a modified `install.conf` file and example output.

Modified `install.conf` File

The following example shows user changes to the `install.conf` file in bold. This example shows the use of a node as both the head node and an execution node.

```
#
# This file contains Advanced Cluster Environment (ACE)
# configuration information that can be used by installace.
#
# This file must be /bin/sh parsable.
#
#
# Advanced Cluster Environment for Linux, includes the
# following products:
#
#       pbs       PBS batch-queuing system
#       pcp       Performance Co-Pilot
#       gm        Myricom's GM driver
#       gm-mpich  Myricom's enhanced MPICH message-passing library
#       mpich     MPICH message-passing library
#       lconsole  Command-line serial console
#       hoover    Hoover cluster manager & serial console

#
#    Enter "ethernet" to select:
#        mpich pbs pcp lconsole
#
#    Enter "myrinet" to select:
#        gm gm-mpich pbs pcp lconsole
#
# List all desired products (separated by commas or white space)
# in the 'Products' variable.

Products="eth"


#
# ACE needs to know about all the nodes in the cluster.
```

```
# There are two ways to specify this information:
# 1) Create a separate file containing all the nodes in the cluster.
#    The first node listed must be the head node.
#    If you want to use this option, list the full pathname in
#    he 'Nodes_filename' variable below.
#
# 2) Define the head and execution nodes in this configuration file.
#    To do this, fill in the 'Head_node' and the 'Execution_nodes'
#    variables below.
#

#
# List the full path to the file containing all cluster nodes.
# The first node listed must be the head node.
# Leave empty if you want to specify nodes below.
#

Nodes_filename=""


#
# List the head node.  The head node is the host from which
# users will submit PBS requests and launch the execution of parallel
# jobs.  Use the fully qualified name such as "foo.domain.com".
#
# If you have left 'Nodes_filename' variable empty and you do not
# fill in a node for 'Head_node', only the software appropriate to
# execution nodes will be installed.  In this case, you must manually
# configure PBS on the head node for the execution nodes.

Head_node="ace08"


#
# NOTE: If you have left 'Nodes_filename' variable empty and
#       you do not list a node for 'Head_node', only the software
#       appropriate to execution nodes will be installed.  In this case,
#       you must manually configure PBS for each new execution node
#       by doing the following:
#
#       1) Add the name of the head node to /usr/spool/pbs/server_name
```

```
#          on the execution nodes.
#       2) Edit /usr/spool/pbs/mom_priv/config on the execution nodes
#          to add the following:
#              $logevent 0x1ff
#              $clienthost
#       3) Edit the /usr/spool/pbs/server_priv/nodes file on the head node
#          and add the name of the execution nodes.

#
# List all execution nodes in the cluster.  They can be comma or
# white space, including newline, separated.
# Use the fully qualified name such as "foo.domain.com".
#
# If you want the head to also be an execution node, you
# must also list it in the 'Execution_nodes' variable.
#

Execution_nodes="ace08 ace07"


#
# The install process can update the /etc/hosts.equiv file on all
# the nodes in the cluster.  The /etc/hosts.equiv file allows access
# to the nodes.  The MPICH, PBS, and GM products use rsh/rcp to access
# other nodes.  If /etc/hosts.equiv is not updated, all users will
# have to update their .rhosts file on all nodes.
# Set the 'Hosts_equiv_file' variable below to one of
# the following values:
#
#      "add") Add the specified nodes to the existing /etc/hosts.equiv files.
#             If /etc/hosts.equiv does not exist, installace will create one.
#  "replace") Replace the current /etc/hosts.equiv file with a new one
#             containing only the specified nodes.  The current
#             /etc/hosts.equiv will be saved as /etc/hosts.equiv.old.
# "nochange") Do not change the existing /etc/hosts.equiv file.
#             If the variable is not set, the default will be "nochange".

Hosts_equiv_file="replace"
```

Batch Output

When using the preceding file, you would get the following output:

```
# /bin/cp -i install.conf /tmp/install.conf
# chmod 644 /tmp/install.conf
# vi /tmp/install.conf
# ./installace /tmp/install.conf
Using install config file '/tmp/install.conf'.

    Checking rsync/rcp/rsh access ...

    Checking disk space requirements ...

    Installing selected ACE products on node(s)
      ace-tools ........... [yes]
      ace-docs ............ [yes]
      pcp ................. [yes]
      pbs ................. [yes]
      mpich ............... [yes]
      gm .................. [no]
      gm-mpich ............ [no]
      lconsole ............ [yes]
      hoover .............. [no]

    Installing head node software on ace08
    Installing package sgi-ace-tools-1.4-1 on ace08 ...
    Installing package sgi-acedocs-1.4-1 on ace08 ...
    Installing package sgi-acedocs-print-1.4-1 on ace08 ...
    Installing package mpich-1.2.1-1 on ace08 ...
MPICH postinstall script
    Installing package pcp-2.1.9-12 on ace08 ...
    Installing package pcp-pro-2.1.5-2 on ace08 ...
    Installing package pcp-ace-1.3.0-4 on ace08 ...
Creating MPI wrapper
...Done, output in /var/tmp/pcp-ace-1.3.0-4-mpi_wrapper
    Package textutils-2.0e-3 already installed on ace08, skipping.
    Installing package lconsole-1.1-5 on ace08 ...
    Installing package OpenPBS-2-3 on ace08 ...
PBS preinstall script
PBS postinstall script
```

```
    Installing execution node software on ace07
    Installing package mpich-1.2.1-1 on ace07 ...
MPICH postinstall script
    Installing package pcp-2.1.9-12 on ace07 ...
    Installing package pcp-pro-2.1.5-2 on ace07 ...
    Installing package pcp-ace-1.3.0-4 on ace07 ...
Creating MPI wrapper
...Done, output in /var/tmp/pcp-ace-1.3.0-4-mpi_wrapper
    Package textutils-2.0e-3 already installed on ace07, skipping.
    Installing package OpenPBS-mom-2-3 on ace07 ...
PBS mom preinstall script
PBS mom postinstall script

    Configuring selected ACE products on node(s) ...
    Configuring PBS on head node ace08 ...
    Configuring PBS on ace07 ...
    All PBS nodes registered as seen via /usr/local/bin/pbsnodes
    Starting PCP daemons ...
    Starting PCP on ace08 ...
    Starting PCP on ace07 ...

    Post configuration ...
    Replacing /etc/hosts.equiv file on all nodes ...
    Placed node information in ace08:/etc/ace/nodes
    Placed log information in ace08:/etc/ace/installace.log
    Placed manifest information in ace08:/etc/ace/manifest
    Installed ACE documentation in ace08:/usr/doc/sgi/ace-1.4
    Placed install configuration file in ace08:/etc/ace/installace.conf

    This ACE release also contains third-party evaluation
    software. See /mnt/cdrom/ace-demos/README.TXT for
    more information.

    The installation/configuration process has been completed!
```

Note: Special software is not installed on `node01` when it is used as both an execution node an a head node. However, configuration will be performed to make it both a head node and an execution node.

PBS Status

After you complete the installation, PBS is set up with one queue. To change this, see the *PBS Administrator's Guide.*

Rebuilding the Kernel when using Myrinet without SGI ProPack

If you want to use the Myrinet drivers but you are not running SGI ProPack for Linux, you must rebuild the kernel in order to make the header files and modules match. You do not need to run the rebuilt kernel.

The following sections describe the following:

- "Standard Rebuild Procedure"
- "Shortcut when the Source Matches the Running Kernel", page 42

Standard Rebuild Procedure

To rebuild the header files and modules required for Myrinet drivers, do the following:

1. Log in as `root`.
2. Get the current kernel version number:

   # **uname -r**

3. Set the current working directory to the desired kernel version:

   # **cd /usr/src/*kernel_version***

4. Copy the kernel configuration file corresponding to the kernel version:

   # **cp configs/kernel-*kernel_version*.config .config**

5. Rebuild the kernel configuration files:

   # **make oldconfig**

6. Rebuild the kernel source and modules:

   # **rm -f include/linux/modules/***
   # **make dep**
   # **make clean**
   # **make modules**

7. Rebuild the Myrinet RPM:

   # **cd mnt/cdrom/SRPMS**
   # **export GM_BUILD_ROOT=/usr/src/*kernel_version***
   # **rpm --rebuild --target *architecture package_version*.src.rpm**

   Watch for the RPM name at the end of the build output. For example:

   `/usr/src/redhat/rpms/i686/gm-smp-1.2-0.src.rpm`

   This RPM target name will be an argument for the next step.

8. Install the Myrinet RPM:

   # **rpm -ihv *package_version*.*architecture*.rpm *package***

   For example:

   # **rpm -ihv /usr/src/redhat/RPMS/i686/gm-smp-1.2-0.i686.rpm**

Shortcut when the Source Matches the Running Kernel

If the kernel source matches the runing kernel, but you have rebuilt the kernel, you can use this shortcut:

1. Log in as `root`.
2. List the library modules:

   # **ls /lib/modules/**

3. Get the current kernel version number:

   # **uname -r**

4. Rebuild the Myrinet RPM:

   # **cd mnt/cdrom/SRPMS**
   # **export GM_BUILD_ROOT=/usr/src/*kernel_version***
   # **rpm --rebuild --target *architecture package_version*.src.rpm**

   Watch for the RPM name at the end of the build output. For example:

   `/usr/src/redhat/RPMS/i686/gm-smp-1.2-0.i686.rpm`

   This RPM target name will be an argument for the next step.

5. Install the Myrinet RPM:

   # **rpm -ihv *package_version*.src.rpm *package***

For example, to build gm for an i686 architecture system:

```
# ls /lib/modules/
2.2.14-5.0  2.2.14-5.0smp
# uname -r
2.2.14-5.0smp
# cd mnt/cdrom/SRPMS
# export GM_BUILD_ROOT=/lib/modules/2.2.14-5.0smp
# rpm --rebuild --target i686 gm-smp-1.2-0.src.rpm
...
# rpm -ihv /usr/src/redhat/RPMS/i686/gm-smp-1.2-0.i686.rpm
```

Installing the Myrinet Driver (gm)

After the you have used the `installace` utility to install the `gm` package (individually or as part of the `myrinet` set of packages), you must manually install the driver into the kernel modules directory and start the driver. Thereafter, the driver will automatically be restarted after each boot. Some steps must be completed on all nodes, others on only the head node.

Note: If you are not, you must rebuild the drivers to match the your kernel. See Chapter 5, "Rebuilding the Kernel when using Myrinet without SGI ProPack", page 41.

Steps on All Nodes

Do the following on each node:

1. Log in as `root`.
2. Install the `gm` package according to the directions in Chapter 4, "Installing ACE Software", page 23.
3. Change to the `/usr/local/gm` directory:

   # **cd /usr/local/gm**

4. Run the `GM_INSTALL` utility. For example:

```
# ./GM_INSTALL
Making device files in /dev.
ifconfig myri0 down - in case it was up
myri0: unknown interface: No such device
Removing any existing gm driver.
rmmod: module gm not loaded
Adding new GM driver.

tail /var/log/messages
Dec 10 10:02:39 node01 kernel: GM:Using MCP: 'L4 4K' m:524288 l:453328
v:401-403 pl:4096 cl:64
Dec 10 10:02:40 node01 kernel: GM:Allocated IRQ18
```

```
Dec 10 10:02:40 node01 kernel: GM:Initialized network driver myri0
Dec 10 10:02:40 node01 kernel: GM:gm: driver loaded, 1 unit initialized

Done
```

Steps on the Head Node Only

Do the following:

1. Log in to the head node as `root`.
2. Change to the `/usr/local/gm/sbin` directory:

   # **cd /usr/local/gm/sbin**

3. Start the Myrinet GM `mapper`(8) process with the `active.args` argument, and redirect the standard output and standard error to a file. For example:

   ```
   # ./mapper active.args > /dev/null  2>&1 &
   ```

4. To test whether the board is up and running, change to the `/usr/local/gm/bin` directory and run the `.gm_board_info` command. The following shows example output:

```
# cd /usr/local/gm/bin
# ./gm_board_info
GM build ID is "1.1.1 root@node01 Fri Dec 10 11:25:36 CST 1999."


Board number 0:
  lanai_clockval    = 0x90479047
  lanai_cpu_version = 0x0403
  lanai_board_id    = 00:60:dd:7f:ec:0d
  lanai_sram_size   = 0x00100000 (1024K bytes)
  fpga_version      = "Mon Aug 10 11:46:37 1998"
  more_version      = ""
  board_type        = 0x0001 (GM_MYRINET_BOARD_TYPE_1MEG_SRAM)
  bus_type          = 0x0002 (GM_MYRINET_BUS_PCI)
  product_code      = 0x0025
  serial_number     = 20375
    (should be labeled: "M2F-PCI32c-20375")
LANai time is 0x0e0775ce5 ticks, or about 28 minutes since reset.
This is node 1 (node01).
Board has 8 ports and has space for 4657 nodes/routes.
```

```
Route table for this node follows:
gmID MAC Address                              Hostname Route
---- ---------------- ------------------------------ --------------------
   1 00:60:dd:7f:ec:0d                            node01 80 (this node) (mapper)
   2 00:60:dd:7f:de:4c                             node02 b8
```

Monitoring Myrinet Packet Traffic

If you have a 16–port Myrinet switch, you can monitor Myrinet packet traffic on the switch using the `myrinetmon`(1) tool, which is part of the `pcp-ace` package in Performance Co-Pilot. (The 4-port Myrinet switch does not have an ethernet port, which means that it cannot be monitored via SNMP and therefore it cannot be monitored with the `myrimon` tool.)

The`/var/pcp/pmdas/myrinet/README` file and the `myrinetmon`(1) and `pmdamyrinet`(1) man pages provide detailed instructions that tell you how to configure the switch to allow monitoring and how to set up the necessary PCP agents.

Figure 6-1 shows an the `myrinetmon` tool monitoring a 16 port switch (of which only two ports are connected).

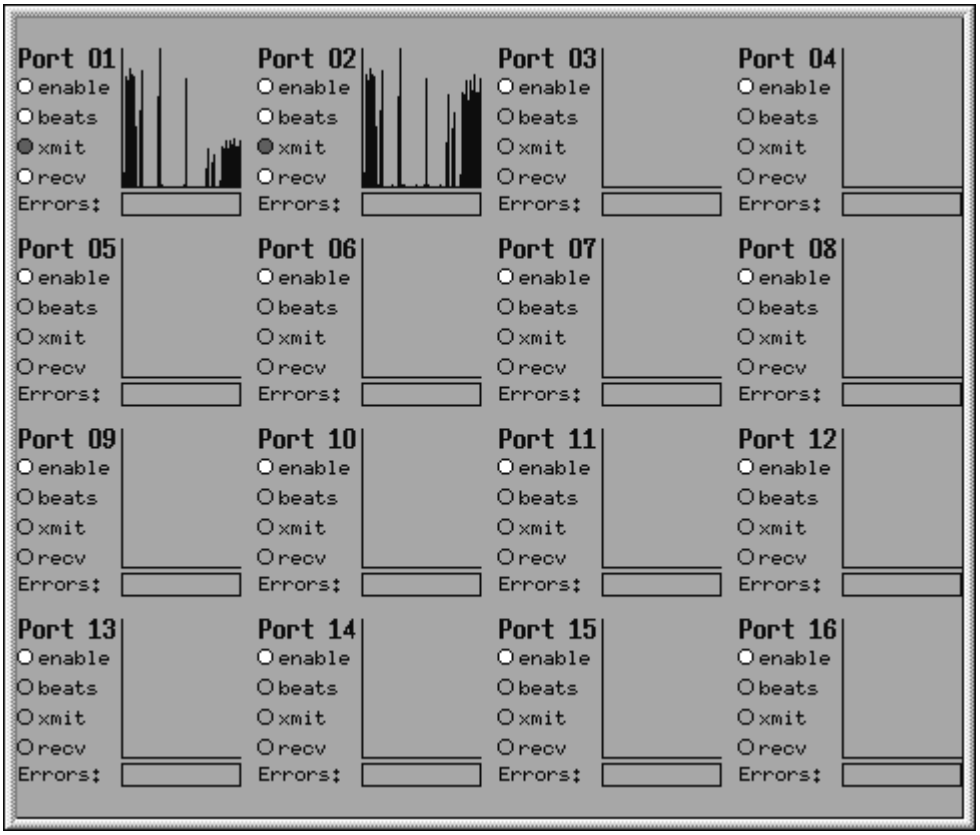

Figure 6-1 Example `myrinetmon` Display

Configure gm-mpich

In order to use gm-mpich, every MPICH user must have a gm configuration file. The default file is $HOME/.gmpi/conf.

The configuration file must contain the following information:

number_of_mpi_process
node port [*board*]
node port [*board*]
...

The fields are as follows:

- *number_of_mpi_processes* is the number of MPI process. For example, for a 2–node cluster, if you want to run an 8 PE MPI jo, then it should be set to 8 and there should be 8 node/port combinations specified, using different ports
- *node* is the name of the node.
- *port* is the port number used on the node. Port 2 is for users, port 3 is for Ethernet. You can specify ports 4 through 7.
- *board* is the board number. This field is assumed to be 0. If you have multiple boards, you must change this number.

Comments (those lines that begin with #) and blank lines are ignored.

The following example, in the case with multiple boards on node node02:

```
# The first non-commented line must contain the number of
# nodes in the cluster:
3
# The following should contain the hostname and port for
# each interface:
red.sgi.com 2
node02.sgi.com 2
node02.sgi.com 2 1
```

When installing the gm-mpich package, the installace utility creates a file named /etc/ace/gmpi.conf on every node. The installace utility assumes that every node is using port 2. Users can copy this file to $HOME/.gmpi/conf on every node;

this is done automatically for the root user. To use a different file location, a user can set the GMPICONF environment variable to the file as follows:

```
# export GMPICONF=/etc/ace/gmpi.conf
```

This file must be accessible on all nodes. When using PBS, this environment variable must be set in the PBS request's environment.

Verify the Installation

After completing the installation, configuring Myrinet MPICH (if you installed the `gm-mpich` package), and installing the `gm` driver, you can run the `confidence` script to verify that the installation process has completed successfully.

What `confidence` Verifies

The `confidence` script checks the following:

- That the user can use the `rsh`(1) command among all nodes (in any combination) within the cluster
- That the current working directory is accessible on all nodes
- If `gm-mpich` was installed, that the `gm` driver is installed on all nodes
- If `gm-mpich` was installed, that the `gm` configuration file exists on all nodes
- That all nodes are available to PBS
- That the user can submit a PBS request and get output back
- If `mpich` was installed, that MPICH's `tstmachines`(1) command can be run successfully
- That a user can run a simple MPI job that uses all execution nodes in the cluster
- That a user can submit a PBS request that will run a simple MPI job that uses all execution nodes in the cluster

Using `confidence`

You must run `confidence` on the head node.

Do the following:

1. Log in to the head node.

2. Change to a directory that is NFS mounted to all nodes in the cluster and that has an identical path on each node. For example:

   # **cd /nfs_shared/tmp**

   Note: The `confidence` test of MPICH requires an NFS mounted directory. If you do not have such a directory, you can use the `-n` option to the `confidence` script. However, if you use this option, MPICH's `tstmachines`(1) script will not be executed.

3. Execute the `confidence` script.

The following example shows output for a successful test on a two-node cluster when `confidence` is run from an NFS mounted directory:

```
[root@ace08 ~]# cd /nfs_shared/tmp
[root@ace07 tmp]# /usr/local/ace/bin/confidence
Testing that user has rsh access to and from all hosts.
Shipping a script to each host that will rsh to all other hosts.
PASS  All hosts can rsh to every other host.
INFO  hostnames match ACE node names.
INFO  Current Working Dir is NFS mounted on all hosts.
PASS  All 2 nodes configurated and available to PBS.

Testing that PBS can execute a simple script ...
Adjusting PBS configuration to allow root to submit requests.
Submitting test job to default PBS queue. qsub output:
0.ace08
Waiting up to 60 seconds for PBS request to finish ...
PASS  PBS request has completed.
Checking that output matches cluster node list ...
PASS  PBS qsub output contains node ace07 output.
Restoring PBS acl_roots configuration.
Basic PBS test complete.

Assuming /nfs_shared/tmp/ace-conf.13221 is NFS accessible on all nodes.
Testing MPICH communication using tstmachines ...
PASS  MPICH's tstmachines returned successfully.

Testing that you can run an mpitst program.
Compiling mpi test binary mpitst.
```

```
Running mpich/gm_mpich /nfs_shared/tmp/ace-conf.13221/mpitst test program ...
PASS  mpitst test program ran on all 2 nodes.
Hello World! ace07 is 1 of 2
Hello World! ace08 is 0 of 2
Basic mpi test complete.

Testing that an mpi program can be run via PBS ...
Adjusting PBS configuration to allow root to submit requests.
Submitting mpi job to default PBS queue. qsub output:
1.ace08
Waiting up to 60 seconds for PBS request to finish ...
PASS  The submitted mpitst test program ran on all 2 nodes.
Restoring PBS acl_roots configuration.
confidence: No errors detected.  Success!!!
```

To get help for `confidence`, use the `-h` option. If you want debug information, use the `-v` option.

Remove the CD-ROM

At this point, you can remove the CD-ROM.

Upgrading an Existing Cluster

To upgrade an existing cluster, do the following:

1. Wait until all PBS running requests have completed.
2. Save the PBS server and queue structures. For example, enter the following:

   ```
   # qmgr -c "print server" > /tmp/server.dat
   ```

3. Shut down the cluster services:

   a. Stop the following PBS daemons on the head node:

   ```
   # crsh "uname -n; killall -9 pbs_server"
   # crsh "uname -n; killall -9 pbs_sched"
   ```

   b. Stop the the following PBS daemon on the execution node and on the head node if so configured:

   ```
   # crsh "uname -n; killall -9 pbs_mom"
   ```

   c. Stop the PCP daemon on all nodes:

   ```
   # crsh "uname -n; killall -TERM pmcd"
   ```

   d. For Myrinet clusters:

      i. Stop `gmapper`:

      ```
      # killall -9 gmmapper
      ```

      ii. Stop any Myrinet interfaces. For example:

      ```
      # ifconfig myri0 down
      ```

      iii. If there are outstanding references to the driver, unload it. For example:

      ```
      # lsmod gm
      Module                  Size  Used by
      gm                    257784   0  (unused)

      # rmmod gm
      ```

4. Save the following configuration files and directories:
   - Head node:
     - `/etc/ace/nodes`
     - `/etc/ace/manifest`
     - `/etc/ace/gmpi.conf`
     - `/usr/spool/pbs/server_name`
     - `/usr/spool/pbs/server_priv/nodes`
     - `/usr/spool/pbs/mom_priv/config`
     - `/var/Lconsole/.IC`
     - `/usr/lib/vacm/vacm_configuration`
   - Each execution node:
     - `/usr/spool/pbs/server_name`
     - `/usr/spool/pbs/server_priv/nodes`
     - `/usr/util/machines/machines.LINUX` (for `gm-mpich` and for `mpich` ACE 1.0)
     - `/usr/share/machines.LINUX` (for `mpich` ACE 1.1 and later)
5. If you are using Myrinet, do one of the following:
   - (recommended) Upgrade to SGI ProPack 1.4 and SuSE 6.4, Red Hat 6.2, or TurboLinux 6.0/6.0.2. See *SGI ProPack 1.4 for Linux Start Here* and the OS documentation.
   - Build from the `gm` source. See Chapter 12, "Building Products from Source RPMs", page 79.
6. Upgrade the ACE software by using the `installace` utility. See Chapter 4, "Installing ACE Software", page 23. You can use the `/etc/ace/nodes` file that you copied in the previous step (such as `/etc/ace/nodes.old`) as input to the host file question.
7. Reconfigure PBS. For example:

   # **`qmgr < /tmp/server.dat`**

8. Restore the saved configuration files and directories
9. For Myrinet:
   a. Reload Myrinet drivers:

      ```
      # cd /usr/local/gm
      # scripts/gm_install
      ```

   b. Start the mapper process:

      ```
      # scripts/gmmapper start
      ```

      Note: The Myrinet RPM writes the driver startup to the `/etc/rc.d/rc.local` script.

10. For Ethernet:
    a. Move the saved copy of `/usr/util/machines/machines.LINUX` to `/usr/share/machines.LINUX` .
11. Run the `confidence` tool. See Chapter 8, "Verify the Installation", page 51.

Using and Enhancing Performance Co-Pilot (PCP)

Performance Co-Pilot (PCP) is an integrated suite of system-level performance monitoring tools. PCP provides a unifying abstraction for all of the interesting performance data in a system or cluster of systems, and allows client applications to easily retrieve and process any subset of that data using a single API. A client-server architecture allows multiple clients to monitor the same host, and a single client to monitor multiple hosts, such as those in a cluster. This enables centralized monitoring of distributed processing jobs across the cluster.

The PCP tools shipped with ACE include the Performance Monitoring Collection Daemon (`pmcd`), several of its optional agents, and a powerful suite of 2D and 3D visualization tools. The optional post-install configuration steps for PCP and the cluster visualization tools provided with PCP are discussed in the following sections.

This chapter discusses the following:

Configuring PCP for Remote Display

When you install the PCP product, the `pcp`, `pcp-ace`, and `pcp-pro` packages are installed on each node, including the head node. Some sites may have security policies that use the head node as a firewall for controlling access to the execution nodes and therefore not allow direct socket connections from outside the cluster. You can still use PCP tools such as `clustervis`, `pmgcluster`, and `pmchart` in this scenario by remotely logging in to the head node and displaying the monitoring

information back to your remote workstation, as shown in Figure 10-1. See the `/usr/doc/pcp-pro-`*Version*`/README` file for other possibilities.

Note: The PCP monitor tools can be used on either an IRIX or Linux desktop to monitor both IRIX and Linux hosts that are running the PCP collection Daemon (PMCD). The major difference is that the `clustervis` and `pmgcluster` tools that are part of the Linux `pcp-pro` and `pcp-ace` packages are not part of the IRIX PCP product.

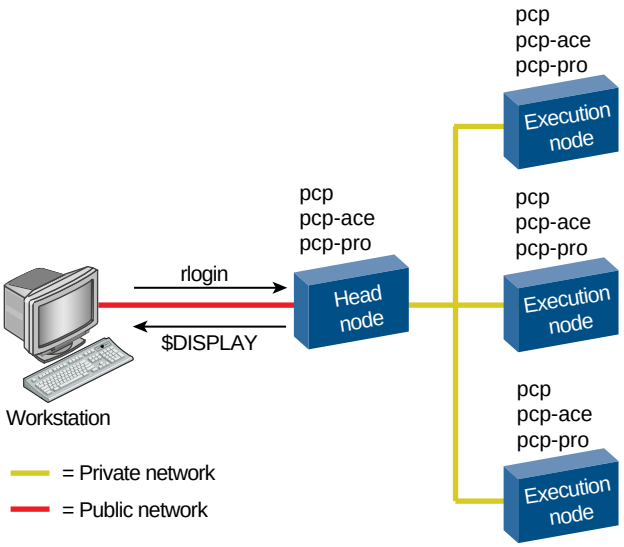

Figure 10-1 Default PCP Topology

The `pcp-ace` and `pcp-pro` packages on the head node require that the `pcp` collector daemon is running on each execution node; once the `pcp` RPM is installed, the `pcp` collector daemon will start automatically on each reboot.

Note: The `installace` utility automatically starts PCP. However, to manually start PCP, enter the following:

```
# /etc/rc.d/init.d/pcp start
```

However, if you have public network access to the execution nodes, you may want to install the `pcp`, `pcp-ace`, and `pcp-pro` packages on a workstation that is not part of the cluster. Doing so will result in better performance, especially if the workstation has hardware-accelerated graphics. Figure 10-2 shows this topology.

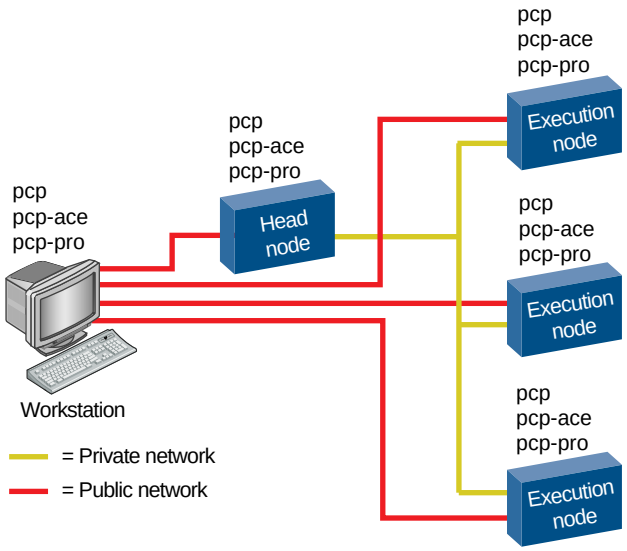

Figure 10-2 Preferred PCP Topology with a Public Network to the Execution Nodes

To install the required packages on a Linux workstation, do the following:

1. Log into the workstation as `root`.
2. Insert the Linux ACE CD-ROM in the workstation's driver.
3. Mount the CD-ROM by entering the following:

   ```
   # /bin/mount /dev/cdrom /mnt/cdrom
   ```

4. Change to the `/mnt/cdrom` directory:

   ```
   # cd /mnt/cdrom
   ```

5. Enter the following to install the packages (line break added here for readability):

```
# rpm -i RPM/pcp*.i386.rpm \
RPM/pcp-pro*.i386.rpm RPM/pcp-ace*.i386.rpm
```

The /etc/ace/nodes file is used by the PCP monitor tools to identify the hostname of all nodes in the cluster. This file would normally already exist on the head node, but if the PCP tools are installed on a workstation other than the head node, you may need to configure the file manually (such as by copying it from the head node). By default, the pmgcluster tool (2D display) and the clustervis tool (3D display) will display performance data for all hosts named in this file. See the clustervis(1)and pmgcluster(1) manual pages for more information.

Configuring your X Server for 16-Bit Display

The PCP monitoring tools that use a 3D display (mpivis, dkvis, osvis, clustervis, webvis and weblogvis) will work best if there is a 16 bpp or better visual available. An 8–bpp visual will still work, but some annoying colormap flashing may occur. You can use the xdpyinfo tool to determine what visuals are available.

Most X servers can be run with the -bbp 16 option; see the Xserver(1)man page for details. The default is usually 8 bpp.

Using PCP Visualization Tools

The clustervis(1), mpivis(1), and pmgcluster(1) tools are the most common starting point for monitoring ACE clusters with PCP. When using clustervis or mpivis, you can click on blocks or base planes in the scene and then select another tool via the **Launch** menu to zoom in on interesting items of cluster/system activity.

The pmgcluster and pmgsys tools are designed to provide an overview of cluster/system activity in a very compact window. To monitor the fine-level performance details on a particular host, use the pmchart(1) tool.

Another important feature of the PCP visualization tools is their ability to "record" and then later "play back" the performance data. This allows retrospective analysis of the performance characteristics of activity across the cluster and is often the only effective way to tune distributed applications. The easiest way to use this facility is to use the **File/Record** menu option in clustervis(1), mpivis(1), or pmchart(1) and

the use the pmafm(1) tool to replay the resulting archives. You can also use the pmlogger(1) tool to create customized archives if necessary.

Note: The pmgadgets(1) tool and its front-end tools(pmgcluster, pmgsys, and others) cannot record or play back archives.

Using PCP to Visualize MPI Jobs

The mpivis(1) tool, and its associated agent pmdampi(1), provide a powerful way to monitor distributed MPI jobs.

Figure 10-3 shows an example mpivis display. This example shows a five-node cluster with nine CPUs in total. Nodes 01, 02 and 03 have MPI ranks that are blocked in barrier wait, while nodes 00 and 03 each have a rank sending and node 04 has both its ranks receiving. Clearly, the receiving ranks on node 04 may represent a bottleneck.

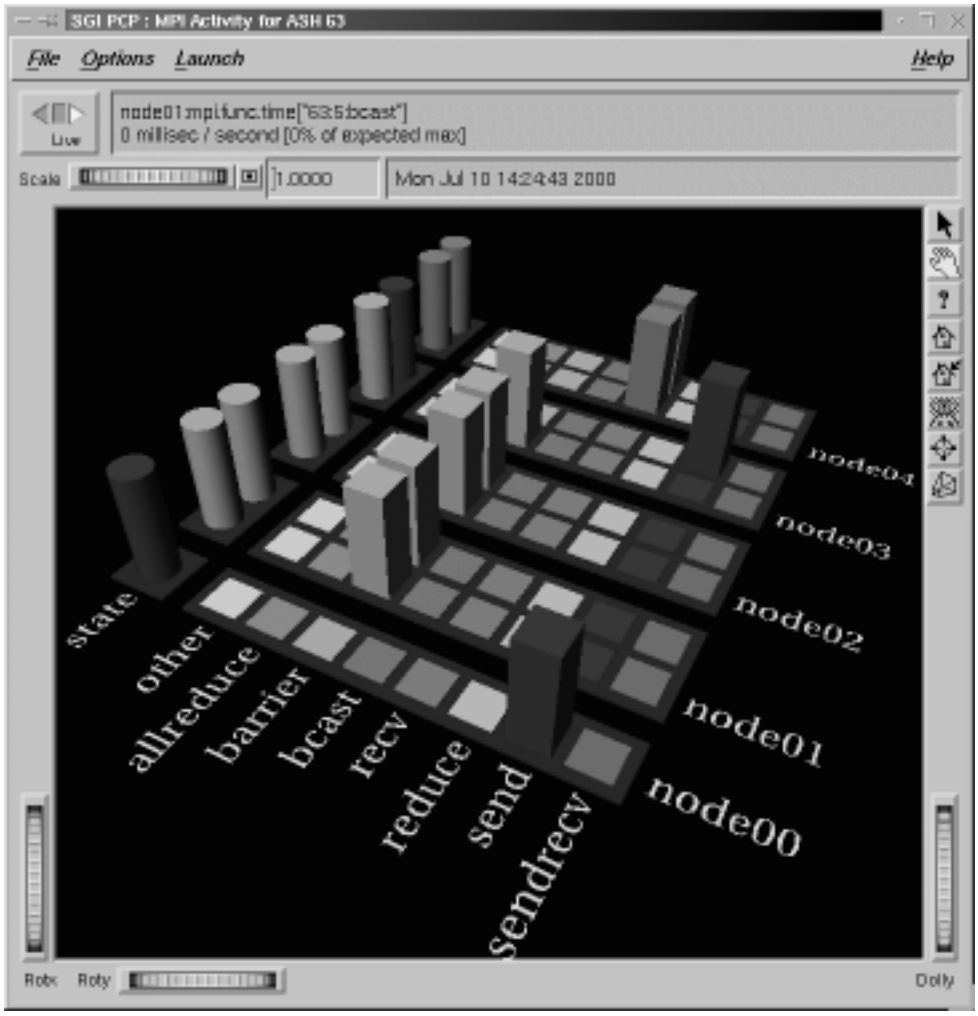

Figure 10-3 Example `mpivis` display

To enable this facility, you must configure the PCP MPI agent on **every** execution node in the cluster. You must also relink your MPI applications with `libpcp_mpi`, which is an instrumented wrapper library for the MPI library. The `libpcp_mpi` library has been engineered to impose very little overhead and can thus be used without significant performance degradation. For detailed installation and

configuration procedures, see the /var/pcp/pmdas/mpi/README file. The pmfindash(1) utility is also extremely useful when using mpivis(1), especially on a shared cluster where multiple users may be running jobs simultaneously.

Using PCP to Visualize Myrinet Switch Traffic

PCP includes an optional agent and visualization tool to monitor Myrinet switch traffic (on 16-port switches only). This is also described in "Monitoring Myrinet Packet Traffic", page 47. For detailed instruction, see the pmdamyrinet(1) and myrinetmon(1) man pages and in the /var/pcp/pmdas/myrinet/README file.

Note: Due to performance reasons, it is not desirable to monitor Myrinet traffic on individual nodes. The gm driver (kernel module) does have a facility for printing packet send/recv counters, but the default configuration has this instrumentation turned off because it lies in the critical path in the code, thus adding unwanted latency to Myrinet packet transmission (which is highly optimized). Interested users may examine the source code for the gm_counters tool (which is found in binary/bin/gm_counters in the gm installation directory) and then construct a PCP agent for exporting this information.

Using PCP to Visualize Cisco Switch Traffic

PCP includes an optional agent that can extract traffic statistics from Cisco switches and routers. To enable this facility, follow the instructions in the /var/pcp/pmdas/cisco/README file and in the pmdacisco(1) and pmgcisco(1) man pages.

When configuring pmdacisco(1), you will be asked several questions, including the host name of the Cisco hardware and which ports you want to monitor. Once you have the Cisco agent configured, you can use the pmgcisco tool on any node (usually the head node) to monitor the Cisco hardware. If you are running pmgcisco on a host other than the host running the Cisco agent, you must use the -h *host* option with pmgcisco. The host given with the -h option should be the host name of the host running the Cisco agent (not the hostname of the Cisco hardware itself).

Monitoring Specific Applications and Processes

PCP provides a thread-safe API for allowing your C/C++, Fortran or Java application to export performance data into the PCP framework. The `pmdatrace`(1) optional agent is used to collect this information and the `pmchart`(1) tool is generally used to monitor applications instrumented in this way. See the `/var/pcp/pmdas/trace/README` file for further instructions. There is also the `pmtrace`(1)command for permitting scripted applications.

The `pmchart`(1) tool can also be used to monitor specific processes running on the cluster. Users should should start their application on one or more nodes across the cluster, and then use `pmchart` to create charts showing one or more of the `proc` metrics. The **File/NewPlot** menu choice opens the **Metric Selection Dialog**, from which you can navigate to the desired metrics. Once a metric has been selected (in the right hand list box), you can use the **info** button to display descriptive text for that metric. Some useful `proc` metrics include the following:

- `proc.psinfo.stime` — system (kernel) CPU time for the process
- `proc.psinfo.cstime` — system (kernel) CPU time for the process & all children
- `proc.psinfo.utime` — user mode CPU time for the process
- `proc.psinfo.cutime` — user mode CPU time for the process & all children
- `proc.memory.*` — memory usage (virtual size, resident size, etc)
- `proc.psinfo.maj_flt` — major page faults

Customizing PCP Visualization Tools

All of the PCP 3D tools are in fact front-end scripts that generate input for the `pmview`(1) application. As such, it is usually fairly simple to construct a custom visualization tool. See the `pmview`(1) man page for a description of the scene-configuration language. Likewise, tools such as `pmgcluster`, `pmgsys`, `pmgcisco` and others are front-end scripts for `pmgadgets`(1). In both cases, the `-V` command line option may be used with a front-end script to print the configuration file before it is fed to the back-end. The result is often a good starting point for a custom front-end tool.

The `pmchart`(1) tool allows "views" to be saved and customized. When monitoring multiple hosts, such as in a cluster, it is often useful to interactively create a desired set of charts, then use the **File/Save** menu option to save that configuration. You can

reload a view with the `-c` command line option or via the **File/OpenView** menu choice. See the `pmchart`(1) man page for further information and a description of the view configuration syntax.

Further Information About PCP

The `PCPIntro`(1) man page provides an overview of PCP and common command line options.

The `pmcd`, libraries, and API features of PCP are now available as Open Source software. See `http://oss.sgi.com/projects/pcp`

The *Performance Co-Pilot User's and Administrator's Guide*and the *Performance Co-Pilot Programmer's Guide* are available online from SGI's Tech Pubs Library. Search for the strings `PCP_UAG` and `PCP_PG`, respectfully.

Questions about PCP should be directed to the PCP Open Source mailing list at `pcp@oss.sgi.com`. To join the mailing list, please use the link on the PCP home page (`http://oss.sgi.com/projects/pcp`).

Lconsole Utility

The Lconsole utility provides a serial port console that lets you do the following:

- Use serial lines to get the boot prompt in case of a network failure.
- Perform a remote reset in case of hardware failure. console can only power cycle a single node at once.

Note: Lconsole requires that you use a serial multiplexer.

Do not attempt to use Lconsole and the Hoover cluster manager together.

Lconsole consists of the following commands, which are installed on the remote administration node (that is, the node to which the serial multiplexer is connected, which may or may not be the head node):

- `lctelnetd`(8), which invokes the Lconsole menu when you log in to the configured port on the remote administration node
- `lcpasswd`(8), which lets you add and delete passwords
- `lclogin`(8), which lets you invoke the Lconsole menu directly when logged in to the remote administration node

The `/var/Lconsole/logs` directory contains Lconsole logs.

This chapter tells you how to do the following:

Configure the Serial Console

You can configure the cluster to do the following:

- Send boot messages to the remote administration node

  Note: Boot messages can go to either the monitor (the default) or the serial console; however, they cannot be seen in both places.

- Permit login to the other nodes on their serial ports

Send Boot Messages to the Remote Administration Node

By default, boot messages are sent to the monitor. To send them to the remote administration node instead, do the following on each nonadministration node:

1. Edit the `/etc/lilo.conf` file so that it redirects all boot messages to the serial console:

   `serial=`*tty#*`,`*baudrate_parity(n|e|o)_bits(7|8)*

   The variables are as follows:

   - *tty#* is the number of the serial port; 0 corresponds to COM1 alias `/dev/ttyS0`.
   - *baudrate* is the baud rate of the serial port in bits per second (b/s).
   - *parity(n|e|o)* is the parity used on the serial line: `n` for no parity, `e` for even parity, `o` for odd parity.
   - *bits(7|8)* is the number of bits in a character. If the *parity* value is `n`, the default is `8`; if the *parity* value is `e` or `o`, the default is `7`.

   For example, to use port `/dev/ttyS0` with a baud rate of 9600 b/s, no parity, and 8 bits per character, enter the following:

   ```
   serial=0,9600n8
   ```

2. Under each `image` area in the `/etc/lilo.conf` file, add the following:

   `append="console=`*serial_device_in_/dev*`"`

For example:

```
append="console=ttyS0"
```

The following example shows an entire `/etc/lilo.conf` file, with added portions highlighted:

```
boot=/dev/hda
map=/boot/map
install=/boot/boot.b
prompt
timeout=50
serial=0,9600n8
image=/boot/vmlinuz-2.2.5-15
        label=linux
        root=/dev/hda1
        initrd=/boot/initrd-2.2.5-15.img
        read-only
        append="console=ttyS0"
image=/boot/bzImage
        label=new
        root=/dev/hda1
        initrd=/boot/initrd-2.2.5-15.img
        read-only
        append="console=ttyS0"
```

For more information, see the `lilo.conf`(5) man page.

Permit Login on the Serial Port

To permit login on the serial port of a nonadministration node, add the following to the `/etc/inittab` file on that node:

device#:run_levels:action:getty_program device speed

These items must match what is in the `/etc/lilo.conf` file.

For example, for the Red Hat base OS:

```
S0:2345:respawn:/sbin/uugetty ttyS0 DT9600
```

This means:

- The device is `/dev/ttyS0`

- It can run in levels 2, 3, 4, and 5
- The `uugetty` program will be executed, and if it dies it will be restarted (`respawn`)
- The device speed is DT9600

For more information, see the `inittab`(5) and `getty`(8) man pages.

Configure the Remote Administration Node

On the remote administration node, do the following:

1. Edit the `/etc/services` file so that it contains the port number to be used for Lconsole. The recommended port is `5000`, but you can choose another free port if you want. For example:

   ```
   # Lconsole
   lctelnet        5000/tcp
   ```

2. Edit the `/etc/inetd.conf` file (which describes the services that will be available through the INETD TCP/IP super server) to add the following line:

   ```
   # Lconsole
   #
   lctelnet        stream  tcp     nowait  root
   /usr/Lconsole/bin/lctelnetd lctelnetd
   ```

   The fields are as follows:

   - Service name: `lctelnet`
   - Socket type: `stream`
   - Protocol: `tcp`
   - Wait status: `nowait`
   - User: `root`
   - Server program: `/usr/Lconsole/bin/lctelnetd`
   - Server program arguments: `lctelnetd`

   For more information, see the `inetd.conf`(8) man page.

3. Create a file named `/var/Lconsole/.LC/`*clustername*`/`*nodename* for each nonadministration node. In the file, add the following:

   - Console port
   - Reset port
   - Logging options

   For example, the file `/var/Lconsole/.LC/acmecluster/node04` could contain the following:

```
#CONSOLE    EMP         USER  BAUD  LOG  DFLTPW
/dev/ttyR0  /dev/ttyR9  root  9600  yes  yes
```

   This means:

   - The console port is `/dev/ttyR0`
   - The emergency management port (EMP) is `/dev/ttyR9`
   - The user who can perform a reset is `root`
   - The baud rate is 9600
   - Logging to a file is turned on
   - The use of a default password (`none`) is turned on (if this field is `no`, then the user must supply a password)

Note: The `/dev/tty` values above are device dependent and may change on site.

Connect to Lconsole

To connect to the Linux console on the remote administration node, enter the following from any workstation on your network:

```
$ telnet hostname port
```

Note: Rather than a hostname, you could enter the node's IP address.

For example, to log into a remote administration node named node04, enter the following:

```
$ telnet node04.acme.com 5000
Trying 192.168.0.4...
Connected to node04.sgi.com.
Escape character is '^]'.

Red Hat Linux release 6.0 (Hedwig)
Kernel 2.2.5-15 on an i686

LConsole Login: root
Password:
def...


Lconsole Release 1.1
Copyright 1995-1998 Silicon Graphics, Inc.All Rights Reserved.



   Main Menu:
   Clusters
   -----------
   1. acmecluster
```

You must then specify your Lconsole user name and password.

Use Lconsole Features

This section tells you how to use and exit from Lconsole.

Operations

The Lconsole utility allows you to perform the following functions on any node in the cluster:

- Connect to the node's console. This is useful if the network is down.
- Reset the hardware

- Power up the node
- Power down the node
- Steal an occupied console line from another user

Note: To use the reset, power up, and power down operations, you must first set the DISPLAY variable to your remote workstation. (Xlib routines are used to handle asynchronous IO routines.) For example:

```
$ export DISPLAY=myworkstation:0
```

For example, to connect to the console for node04, enter the following:

```
           System 'node04':
           Available Operations
           --------------------
           1. Connect to Serial Console
           2. Hardware Reset
           3. Power Up
           4. Power Down
           5. Steal Occupied Console Line

        Enter choice (1 - 5) or '..' for Systems Menu: 1

------------ Connected port=/dev/ttyR1 -----------
  9600 BAUD 8 NONE 1 SWFC=ON  HWFC=OFF
CAR=OFF DTR=ON  RTS=ON  CTS=OFF DSR=OFF
Type ~X to Exit.
Connection Ready.
```

Exit

To exit from Lconsole, you must be at the main menu:

- To leave the node console and return to the Available Operations menu, press ~X
- To leave the Available Operations menu and return to the Available Systems menu, enter .. (dot dot)

- To leave the Available Systems menu and return to the Main Menu, enter `..` (dot dot)
- To exit from Lconsole from the Main Menu, enter `q`.

For example:

```
           System 'node02':
           Available Operations
           --------------------
           1. Connect to Serial Console
           2. Hardware Reset
           3. Power Up
           4. Power Down
           5. Steal Occupied Console Line

        Enter choice (1 - 5) or '..' for Systems Menu: 1

----------- Connected port=/dev/ttyR3 -----------
  9600 BAUD 8 NONE 1 SWFC=ON  HWFC=OFF
CAR=OFF DTR=ON  RTS=ON  CTS=OFF DSR=ON
Type ~X to Exit.
Connection Ready.
Lconsole DINC closing...
~X
------------
        Connect to Console done.


           System 'node02':
           Available Operations
           --------------------
           1. Connect to Serial Console
           2. Hardware Reset
           3. Power Up
           4. Power Down
           5. Steal Occupied Console Line

        Enter choice (1 - 5) or '..' for Systems Menu: ..


        Cluster 'acmecluster':
```

```
        Available Systems
        -----------------
        1. node01
        2. node02
        3. node03
        4. node04

        Enter choice (1 - 4) or '..' for Main Menu: ..


   Main Menu:
   Clusters
   ---------------
   1. acmecluster

   Enter choice 1 or 'q' to Quit: q
Connection closed by foreign host.
```

Add Lconsole Users and Change Passwords

To add an Lconsole user, enter the following as `root`:

```
# /usr/Lconsole/bin/lcpasswd [username] password
Re-enter password: password
```

For example:

```
#  /usr/Lconsole/bin/lcpasswd lhj
Creating console User password for user lhj
New console password: (password_not_shown)
Retype new console password: (password_not_shown)
lcpasswd: Success
```

If you do not supply a *username*, the current login is the default. Passwords are stored in `/usr/Lconsole/adm`.

The same process is used to change the password for an existing user.

Delete Lconsole Users

To delete an Lconsole user, enter the following as `root`:

```
# lcpasswd -d username
```

Building Products from Source RPMs

To build a product from a source RPM Package Manager (SRPM), do the following:

1. Install the SRPM:

   a. Insert the CD-ROM on the head node.

   b. Enter the following:

      # **rpm -ihv /mnt/cdrom/SRPMS/*package_version*.src.rpm *package***

      For example, to install `lconsole`:

      # **rpm -ihv /mnt/cdrom/SRPMS/lconsole-1.1-1.src.rpm lconsole**

      The contents of the SRPM will be installed on the system in the `/usr/src/redhat/SOURCES` directory. The contents of the SRPM may be compressed `tar` files (*tarballs*) and/or patch files.

2. Create a build directory and change directories to it:

   # **mkdir /tmp/build**
   # **cd /tmp/build**

3. Move the tarball and/or patch files to the build directory. For example:

   # **mv /usr/src/redhat/SOURCES/lconsole-1.1.tar.gz /tmp/build**

4. Uncompress the tarball and extract the files:

   # **tar zxvf *tarball***

   For example:

   # **tar zxvf lconsole-1.1.tar.gz**

5. Change to the package's directory. For example:

   # **cd lconsole-1.1**

6. If there are patches that you want to apply, enter the following:

   # **patch -p0 <*patchfile***

   For more information, see the `patch`(1) man page.

7. Build the package:

   # **make**

Adding an Execution Node to the Cluster

If you later want to add an execution node to the cluster, you can manually install packages on the node by using the `rpm`(8) command or you can use the `installace` utility as follows to install software on just that node:

- In interactive mode, bypass installation on the head node by entering a dot (`.`) for the head node name. When prompted for the execution nodes, enter only the name of the new node. See "Interactive Mode Example", page 83.
- In batch mode, bypass installation on the head node by leaving the `Head_node=` field blank. Include only the name of the new node in the `Execution_nodes=` field. See "Batch Mode Example", page 87

For more information about `installace`, see "Interactive Mode Instructions", page 24, and "Batch Mode Instructions ", page 32.

Configuration Tasks

After the software is installed, perform the following configuration tasks on the new execution node and head node, according to the products you are using.

New Execution Node

Do the following on the new execution node:

- Ethernet MPICH:
  - Add all cluster node names to the `/usr/share/machines.LINUX` file
- Myrinet GM-MPICH:
  - Add all cluster node names to the `/usr/util/machines/machines.LINUX` file

- PBS:
  - Edit the `/usr/spool/pbs/mom_priv/config` file and add the following:

    ```
    $logevent 0x1ff
    $clienthost headnode
    ```

  - Edit the `/usr/spool/pbs/server_name` file so that it contains the name of the head node.
  - Start the `pbs_mom` process by entering the following:

    ```
    # /usr/local/sbin/pbs_mom
    ```

- PCP:
  - Start the `pmcd` daemon:

    ```
    # /etc/rc.d/init.d/pcp start
    ```

- Lconsole:
  - See "Send Boot Messages to the Remote Administration Node", page 70
  - See "Permit Login on the Serial Port", page 71

Head Node

Do the following on the head node:

- All products:
  - Edit the `/etc/ace/nodes` file to include the name of the new execution node.
- Myrinet GM-MPICH:
  - Edit the `gm` configuration files (`$HOME/.gmpi/conf`and `/etc/ace/gmpi.conf` by default) according to the directions in Chapter 7, "Configure `gm-mpich`", page 49.
  - Add the node name to the `/usr/util/machines/machines.LINUX` file.
  - Tell users to copy the an existing `gm` configuration file (such as `$HOME/.gmpi/conf`).

- Ethernet MPICH:
  - Add the node name to the /usr/share/machines.LINUX file.
- PBS:
  - Edit the /usr/spool/pbs/server_priv/nodes file and add the name of the new execution node.
  - Start the pbs_mom process by entering the following:

    # **/usr/local/sbin/pbs_mom**
  - If you wish to perform a PBS check, enter the following :

    # **/usr/local/bin/pbsnodes -a**
- Lconsole:
  - See "Configure the Remote Administration Node", page 72

Interactive Mode Example

The following example shows user entries in bold:

```
root@ace07 /root]# /bin/mount /dev/cdrom /mnt/cdrom
[root@ace07 /root]# cd /mnt/cdrom
[root@ace07 cdrom]# ./installace

ACE nodes file (/etc/ace/nodes) already exists.
If you continue with the install, you will be redefining your cluster.
CONTROL-C if you want to stop this install.


    Advanced Cluster Environment v. 1.4 for Linux i686 includes the
    following products:

        pbs       PBS batch-queuing system
        pcp       Performance Co-Pilot
        mpich     MPICH message-passing library - ethernet
        gm-mpich  Myricom's MPICH message-passing library
        gm        Myricom's GM driver
        lconsole  Command-line serial console
```

```
        hoover    Hoover cluster manager & serial console

    Enter "ethernet" to select:
        mpich pbs pcp lconsole

    Enter "myrinet" to select:
        gm gm-mpich pbs pcp lconsole

    Or enter the individual products separated by white space.

Please enter all desired products separated by white space.
ACE> eth

    Note:  When providing a hostname, use the fully qualified name
    such as "foo.domain.com". If the hostname is resolved on all
    nodes, you can abbreviate it to "foo".

    The installace utility can modify hosts.equiv files on your nodes.
    These files are used by MPICH, GM-MPICH, and PBS.

    You can choose to provide the information required to edit these files
    or avoid this action by entering the appropriate number:

         1)  Enter the hostnames individually
         2)  Specify a file containing the list of hostnames
         3)  Specify a pattern for hostnames (NAME1.domain.com)
         4)  Install software only on this node

Enter choice (1, 2, 3, or 4).
ACE> 1
    Enter the hostname of the head node.
    The head node is the host from which users will launch
    execution of parallel jobs.

Enter the hostname of the head node.
ACE> .

    Enter the hostnames for each execution node in the
    cluster.  After you have entered all hostnames,
    Enter '.' or <control-D> on a new line to end the list.
```

```
ACE> ace06
ACE> .

    The install process can update the /etc/hosts.equiv file on all
    the hosts in the cluster.  The /etc/hosts.equiv will allow access
    to the hosts.  If /etc/hosts.equiv file is not being used, each user
    must update their .rhosts file to allow rcp/rsh type access to all
    cluster nodes.  Enter one of the following:

      1) Add the specified hostnames to the existing /etc/hosts.equiv files.
         If the file does not exist, create it.
      2) Replace the current /etc/hosts.equiv file with a new one
         containing only the specified cluster hostnames.  The current
         /etc/hosts.equiv will be saved as /etc/hosts.equiv.old.
      3) Do not change the existing /etc/hosts.equiv file.

Enter choice (1, 2, or 3).
ACE> 1

    Checking rsync/rcp/rsh access ...

    Checking disk space requirements ...

    Installing selected ACE products on node(s)
      ace-tools ........... [yes]
      ace-docs ............ [yes]
      pcp ................. [yes]
      pbs ................. [yes]
      mpich ............... [yes]
      gm .................. [no]
      gm-mpich ............ [no]
      lconsole ............ [yes]
      hoover............... [no]


    Installing execution node software on ace06
    Installing package mpich-1.2.1-1 on ace06 ...
MPICH postinstall script
    Installing package pcp-2.1.9-12 on ace06 ...
    Installing package pcp-pro-2.1.5-2 on ace06 ...
    Installing package pcp-ace-1.3.0-4 on ace06 ...
```

```
Creating MPI wrapper
...Done, output in /var/tmp/pcp-ace-1.3.0-4-mpi_wrapper
    Package textutils-2.0e-3 already installed on ace06, skipping.
    Installing package pbs-mom-2.2-11RH6 on ace06 ...
PBS mom preinstall script
PBS mom postinstall script

    Configuring selected ACE products on node(s) ...
    Configuring PBS on ace06 ...
PBS has not been configured on the head node for
this new execution node.
You must do the following to properly configure PBS:
1) Add the name of the head node to ace06:/usr/spool/pbs/server_name
2) Edit ace06:/usr/spool/pbs/mom_priv/config to add the following:
$logevent 0x1ff
$clienthost 3) Edit the head node /usr/spool/pbs/server_priv/nodes file and
add the execution node ace06.

    Starting PCP daemons ...
    Starting PCP on ace06 ...

    Post configuration ...
    Adding to the /etc/hosts.equiv file on all nodes ...

    This ACE release also contains third-party evaluation
    software. See /mnt/cdrom/ace-demos/README.TXT for
    more information.

    The installation/configuration process has been completed!
```

Batch Mode Example

This section shows an example of a modified `install.conf` file and output.

Modified `install.conf` File

The following example shows user changes (from the template `/mnt/cdrom/install.conf` file) in bold:

```
#
# This file contains Advanced Cluster Environment (ACE)
# configuration information that can be used by installace.
#
# This file must be /bin/sh parsable.
#
#
# Advanced Cluster Environment for Linux, includes the
# following products:
#
#       pbs       PBS batch-queuing system
#       pcp       Performance Co-Pilot
#       gm        Myricom's GM driver
#       gm-mpich  Myricom's enhanced MPICH message-passing library
#       mpich     MPICH message-passing library
#       lconsole  Command-line serial console
#       hoover    Hoover cluster manager & serial console
#
#    Enter "ethernet" to select:
#        mpich pbs pcp lcdinc
#
#    Enter "myrinet" to select:
#        gm gm-mpich pbs pcp lcdinc
#
# List all desired products (separated by commas or white space)
# in the 'Products' variable.

Products="eth"


#
# ACE needs to know about all the nodes in the cluster.
```

```
# There are two ways to specify this information:
# 1) Create a separate file containing all the nodes in the cluster.
#    The first node listed must be the head node.
#    If you want to use this option, list the full pathname in
#    he 'Nodes_filename' variable below.
#
# 2) Define the head and execution nodes in this configuration file.
#    To do this, fill in the 'Head_node' and the 'Execution_nodes'
#    variables below.
#

#
# List the full path to the file containing all cluster nodes.
# The first node listed must be the head node.
# Leave empty if you want to specify nodes below.
#

Nodes_filename=""


# List the head node.  The head node is the host from which
# users will submit PBS requests and launch the execution of parallel
# jobs.  Use the fully qualified name such as "foo.domain.com".
#
# If you have left 'Nodes_filename' variable empty and you do not
# fill in a node for 'Head_node', only the software appropriate to
# execution nodes will be installed.  In this case, you must manually
# configure PBS on the head node for the execution nodes.

Head_node=""


#
# NOTE: If you have left 'Nodes_filename' variable empty and
#       you do not list a node for 'Head_node', only the software
#       appropriate to execution nodes will be installed.  In this case,
#       you must manually configure PBS for each new execution node
#       by doing the following:
#
#       1) Add the name of the head node to /usr/spool/pbs/server_name
#          on the execution nodes.
```

```
#       2) Edit /usr/spool/pbs/mom_priv/config on the execution nodes
#          to add the following:
#              $logevent 0x1ff
#              $clienthost
#       3) Edit the /usr/spool/pbs/server_priv/nodes file on the head node
#          and add the name of the execution nodes.

#
# List all execution nodes in the cluster.  They can be comma or
# white space, including newline, separated.
# Use the fully qualified name such as "foo.domain.com".
#
# If you want the head to also be an execution node, you
# must also list it in the 'Execution_nodes' variable.
#

Execution_nodes="ace06"


#
# The install process can update the /etc/hosts.equiv file on all
# the nodes in the cluster.  The /etc/hosts.equiv file allows access
# to the nodes.  The MPICH, PBS, and GM products use rsh/rcp to access
# other nodes.  If /etc/hosts.equiv is not updated, all users will
# have to update their .rhosts file on all nodes.
# Set the 'Hosts_equiv_file' variable below to one of
# the following values:
#
#      "add") Add the specified nodes to the existing /etc/hosts.equiv files.
#             If /etc/hosts.equiv does not exist, installace will create one.
#  "replace") Replace the current /etc/hosts.equiv file with a new one
#             containing only the specified nodes.  The current
#             /etc/hosts.equiv will be saved as /etc/hosts.equiv.old.
# "nochange") Do not change the existing /etc/hosts.equiv file.
#             If the variable is not set, the default will be "nochange".

Hosts_equiv_file="add"
```

Batch Output

The following shows the output when using the above file:

```
[root@ace08 cdrom]# ./installace /tmp/install.conf


ACE nodes file (/etc/ace/nodes) already exists.
If you continue with the install, you will be redefining your cluster.
CONTROL-C if you want to stop this install.

Using install config file '/tmp/install.conf'.

    Checking rsync/rcp/rsh access ...

    Checking disk space requirements ...

    Installing selected ACE products on node(s)
      ace-tools ........... [yes]
      ace-docs ............ [yes]
      pcp ................. [yes]
      pbs ................. [yes]
      mpich ............... [yes]
      gm .................. [no]
      gm-mpich ............ [no]
      lconsole ............ [yes]
      hoover .............. [no]



    Installing execution node software on ace06
    Installing package mpich-1.2.1-1 on ace06 ...
MPICH postinstall script
    Installing package pcp-2.1.9-12 on ace06 ...
    Installing package pcp-pro-2.1.5-2 on ace06 ...
    Installing package pcp-ace-1.3.0-4 on ace06 ...
Creating MPI wrapper
...Done, output in /var/tmp/pcp-ace-1.3.0-4-mpi_wrapper
    Package textutils-2.0e-3 already installed on ace06, skipping.
    Installing package OpenPBS-mom-2-3-mom on ace06 ...
PBS mom preinstall script
PBS mom postinstall script
```

```
    Configuring selected ACE products on node(s) ...
    Configuring PBS on ace06 ...
PBS has not been configured on the head node for
this new execution node.
You must do the following to properly configure PBS:
1) Add the name of the head node to ace06:/usr/spool/pbs/server_name
2) Edit ace06:/usr/spool/pbs/mom_priv/config to add the following:
$logevent 0x1ff
$clienthost 3) Edit the head node /usr/spool/pbs/server_priv/nodes file and
add the execution node ace06.

    Starting PCP daemons ...
    Starting PCP on ace06 ...

    Post configuration ...
    Adding to the /etc/hosts.equiv file on all nodes ...

    This ACE release also contains third-party evaluation
    software. See /mnt/cdrom/ace-demos/README.TXT for
    more information.

    The installation/configuration process has been completed!
```

Remove an Execution Node

To remove an execution node from the cluster, remove the name of the execution node from the following files on the head node:

- `/usr/spool/pbs/server_priv/nodes`
- `/usr/util/machines/machines.LINUX` for `gm-mpich` (Myrinet)
- `/usr/share/machines.LINUX` for `mpich` (Ethernet)
- `/etc/ace/nodes`
- `/etc/ace/manifest`

For Myrinet `gm-mpich`, you must also remove the name of the execution node from the `gm` configuration file (`$HOME/.gmpi/conf` by default) for each user on each node

---

Note: You must also decrement the node count in the `gm` configuration file. This information is the first noncomment line in the file.

Syncronizing Clocks in the Cluster

To use the xntpd(8) extended network time protocol, install xntpd(8) on each node in the cluster and configure the daemon to synchronize with a master server.

Do the following on each node in the cluster:

1. Install the xntpd(8) package from the base OS.

   For example, using the Red Hat CD, do the following:

   ```
   # /bin/mount /dev/cdrom /mnt/cdrom
   # cd /mnt/cdrom/RedHat/RPMS# /bin/rpm -i xntp*
   ```

2. Edit the /etc/ntp.conf file and add a driftfile and a list of the hostnames to use as xntp servers. (The driftfile contains the difference between the clocks and is required by xntpd.) For example, if external master servers are used, the /etc/ntp.conf file could contain the following:

   ```
   driftfile /etc/ntp.drift
   server mum.sgi.com
   server lilly.sgi.com
   server carnation.sgi.com
   ```

3. Ensure that the nodes have the correct timezone.

   For example, to use the tzselect(8) command to set the timezone to CST, do the following on each node in the cluster:

```
# tzselect
Please identify a location so that time zone rules can be set correctly.
Please select a continent or ocean.
 1) Africa
 2) Americas
 3) Antarctica
 4) Arctic Ocean
 5) Asia
 6) Atlantic Ocean
 7) Australia
 8) Europe
 9) Indian Ocean
10) Pacific Ocean
```

```
11) none - I want to specify the time zone using the Posix TZ format.
2
Please select a country.
 1) Anguilla              18) Ecuador               35) Paraguay
 2) Antigua & Barbuda     19) El Salvador           36) Peru
 3) Argentina             20) French Guiana         37) Puerto Rico
 4) Aruba                 21) Greenland             38) St Kitts & Nevis
 5) Bahamas               22) Grenada               39) St Lucia
 6) Barbados              23) Guadeloupe            40) St Pierre & Miquelon
 7) Belize                24) Guatemala             41) St Vincent
 8) Bolivia               25) Guyana                42) Suriname
 9) Brazil                26) Haiti                 43) Trinidad & Tobago
10) Canada                27) Honduras              44) Turks & Caicos Is
11) Cayman Islands        28) Jamaica               45) United States
12) Chile                 29) Martinique            46) Uruguay
13) Colombia              30) Mexico                47) Venezuela
14) Costa Rica            31) Montserrat            48) Virgin Islands (UK)
15) Cuba                  32) Netherlands Antilles  49) Virgin Islands (US)
16) Dominica              33) Nicaragua
17) Dominican Republic    34) Panama
45
Please select one of the following time zone regions.
 1) Eastern Time
 2) Eastern Time - Michigan - most locations
 3) Eastern Time - Louisville, Kentucky
 4) Eastern Standard Time - Indiana - most locations
 5) Eastern Standard Time - Indiana - Crawford County
 6) Eastern Standard Time - Indiana - Starke County
 7) Eastern Standard Time - Indiana - Switzerland County
 8) Central Time
 9) Central Time - Michigan - Wisconsin border
10) Mountain Time
11) Mountain Time - south Idaho & east Oregon
12) Mountain Time - Navajo
13) Mountain Standard Time - Arizona
14) Pacific Time
15) Alaska Time
16) Alaska Time - Alaska panhandle
17) Alaska Time - Alaska panhandle neck
18) Alaska Time - west Alaska
19) Aleutian Islands
```

```
20) Hawaii
8

The following information has been given:

        United States
        Central Time

Therefore TZ='America/Chicago' will be used.
Local time is now:      Mon Dec  6 22:35:28 CST 1999.
Universal Time is now:  Tue Dec  7 04:35:28 UTC 1999.
Is the above information OK?
1) Yes
2) No
1
America/Chicago
```

Or, you could reset the `/etc/localtime` symbolic link as follows:

```
# rm /etc/localtime
# ln -s ../usr/share/zoneinfo/us/central /etc/localtime
```

4. Ensure that the node system clocks and hardware clocks are not too far off from the correct time.

   Note: The `xntpd` daemon cannot synchronize the clocks if the difference between the master and client is too great.

   a. Set the correct time on the node's system clock by using the `date`(1) command. For example, to set the time to 10:36 AM on Tuesday December 7, 1999, enter the following:

      ```
      # date 1207103699
      Tue Dec  7 10:36:00 CST 1999
      ```

   b. Synchronize the hardware clock with the system clock by using the `setclock`(8) command. For example:

      ```
      # setclock
      ```

c. Verify that the timestamps for the system clock and hardware clock are close to the same by entering the `date`(1) and `clock`(1) commands. For example:

```
# date
Tue Dec  7 10:36:05 CST 1999
# clock
Tue Dec  7 10:36:09 1999  -0.969269 seconds
```

5. Verify that the `driftfile` exists by using the `touch`(1) command. For example:

```
# touch /etc/ntp.drift
```

6. Start the `xntpd`(8) daemon:

   - To start `xntpd` now, enter the following:

     ```
     # /etc/rc.d/init.d/xntpd start
     ```

   - To start `xntpd` at boot time, enter the following:

     ```
     # /sbin/chkconfig xntpd on
     ```

To verify the status of the xntpd(8) daemon, look at the /var/log/messages file. For example:

```
# tail -f /var/log/messages | grep xntpd
Dec  7 10:36:25 green xntpd[9036]: xntpd 3-5.93e Wed Apr 14 20:23:29
EDT 1999 (1)
Dec  7 10:36:25 green xntpd[9036]: tickadj = 5, tick = 10000,
tvu_maxslew = 495, est. hz = 100
Dec  7 10:36:26 green xntpd[9036]: precision = 13 usec
Dec  7 10:36:26 green xntpd[9036]: read drift of 0 from /etc/ntp.drift
Dec  7 10:41:22 green xntpd[9036]: synchronized to 10.162.8.103,
stratum=1
Dec  7 10:41:32 green xntpd[9036]: time reset (step) 9.206565 s
Dec  7 10:41:32 green xntpd[9036]: synchronisation lost
Dec  7 10:46:23 green xntpd[9036]: synchronized to 10.162.1.126,
stratum=2
Dec  7 10:46:52 green xntpd[9036]: synchronized to 10.162.8.103,
stratum=1
```

To synchronize to a clock other than the head node, see the xntp documentation in the /usr/doc directory.

Troubleshooting

This chapter highlights possible solutions to common problems.

No `rsh` Access

If the `installace` script does not have `rsh`(1) access, it may be because a `/etc/securetty` file exists. Move the file to `/etc/securetty.save` and rerun `installace`.

`.rhosts` Errors in `/var/log/messages`

If you see errors related to `.rhosts` in `/var/log/messages`, it may be because the file permissions are not set properly. Change the file permissions so that `group` and `other` do not have write permission.

For example:

```
# chmod 744 .rhosts
```

Problems with PAM

If you run into problems with pluggable authentication modules (PAM), you may need to do one or more of the following:

- Change the `rhost.auth.so` entry from `required` to `sufficient`
- Reorder the `auth` lines in the `/etc/pam.d/rlogin` as follows:
  - Original order:

    ```
    auth required /lib/security/pam_securetty.so
    auth sufficient /lib/security/pam_rhosts_auth.so
    ```

  - Corrected order:

    ```
    auth sufficient /lib/security/pam_rhosts_auth.so
    auth required /lib/security/pam_securetty.so
    ```

For more information about PAM, see the *Red Hat Linux 6.0 The Official Red Hat Linux Installation Guide*.

MPICH Problems

This section describes solutions to common MPICH problems.

Permission Denied

If you see a `Permission denied` message from `mpirun`(1), it probably means that the user does not have permission to use `rsh`(1) to start the process. To test for access, use the `confidence`(1) script. For example:

```
# /usr/local/ace/bin/confidence
```

If you have installed `mpich`, you can also use the `tstmachines` command:

```
$ /usr/sbin/tstmachines
```

The script will display an error and corrective actions if it fails. For more information, see the `tstmachines`(1) man page and the *User's Guide for mpich, a Portable Implementation of MPI*.

No Such File or Directory

If you see a `No such file or directory` error when running `mpirun`(1), make sure that the `mpi` binary is accessible on all nodes in the cluster, at the same location. This can be done by using an NFS mount an an identical path, or by copying the binary to each node in the cluster, using an identical path on each node.

Process Not Running on Desired Nodes

The `/usr/share/machines.LINUX` file contains the nodes configured for use by MPICH. The entries should be listed one node per line. If the node name has the `:#` suffix, it shows the number of MPI processes supported on that node at one time. The number is usually the same as the number of CPUs. The `mpirun` command distributes the MPI process in a orderly manner.

For example:

```
node1:2
node2:2
node3:2
```

The `mpirun` command will put two MPI processes on `node1` before putting any processes on `node2`. If you want to spread the load out across the nodes, you could edit the file to look like the following:

```
node1:1
node2:1
node3:1
node1:1
node2:1
node3:1
```

Both formats support 6 processes.

For a temporary fix, users can create their own a file using this format and run it in place of the default `/usr/share/machines.LINUX` by specifying it using the following option:

`mpirun -machinefile` *mymachinefile*

PBS Problems

This section lists a few common problems when using PBS. For more information, see the *PBS Administrator's Guide.*

Making a Head Node into an Execution Node

To make an head node an execution node, do the following:

1. Edit the `/usr/spool/pbs/server_priv/nodes` file and remove the `:ts` suffix from the node name.

   For example, where `ace04` is the node name:

   - Old: `ace04:ts np=2`
   - New: `ace04 np=2`

2. Restart PBS by entering the following:

   # **/etc/rc.d/init.d/pbs restart**

3. Verify your change by entering the following:

   # **/usr/local/bin/pbsnodes -a**

   If the output for head node contains the following, the head node is now an execution node:

   ```
   ntype = cluster
   ```

Making a Head/Execution Node into a Head Node

To make a node that is both a head node and an execution node into just a head node, do the following:

1. Edit the `/usr/spool/pbs/server_priv/nodes` file and add the `:ts` suffix from the node name.

   For example, where `ace04` is the node name:

   - Old: `ace04 np=2`
   - New: `ace04:ts np=2`

2. Restart PBS by entering the following:

   # **/etc/rc.d/init.d/pbs restart**

3. Verify your change by entering the following:

   # **/usr/local/bin/pbsnodes -a**

   If the output for head node contains the following, the head node is strictly a head node and is no longer also an execution node:

   ```
   ntype = time-shared
   ```

Jobs Fail to Start

If PBS jobs fail to start, it is likely that the execution nodes cannot communicate with the head node. Do the following:

- Check that `pbs_mom` is running on the execution nodes.
- Check that the PBS server node is listed in the `mom_priv/config` file.

- Check that the queues are enabled on the server.

`root` Cannot Submit Jobs

If the `root` user is prevented from submitting jobs, use another user ID. The default installed ACE PBS will not be configured so that `root` can submit requests.

To configure PBS to allow `root` to submit requests, execute the following on the head node:

```
# /usr/local/bin/qmgr -c 'set server acl_roots="root@*" '
```

Getting PBS State Information

To get PBS state information on all nodes, execute the following on the head node:

```
# /usr/local/bin/pbsnodes -a
```

If a node is missing, add it to the `/usr/spool/pbs/server_priv/nodes` file and restart PBS by entering the following on the head node:

```
# /etc/rc.d/init.d/pbs restart
```

If a node is down, ensure that the execution node's `/usr/spool/pbs/server_name` file contains the name of the head node. You must restart `pbs_mom` by entering the following on the execution node:

```
# /etc/rc.d/init.d/pbs_mom restart
```

It may take a few moments for the head node to synchronize up with the execution node.

PCP Errors

PCP applications will return the following error if the PCP daemons have died or have not previously been started:

application`: hinv.ncpu: Unknown metric name`
application`: Failed to get the number of CPUs from host "test"`

To resolve this problem, restart the PCP daemons as follows:

```
# /etc/rc.d/init.d/pcp stop
# /etc/rc.d/init.d/pcp start
```

The PCP daemons must be running on every execution node in the cluster (optionally on the head node).

For information about other problems, see the following files:

- /usr/doc/pcp-*Version*/README
- /usr/doc/pcp-ace-*Version*/README
- /usr/doc/pcp-pro-*Version*/README

NFS Version Defaults to 2

The nfs(5) manpage incorrectly says that the NFS version defaults to version 2; it actually defaults to version 3. This is a problem when integrating NFS into a standard Linux environment. To get around this problem, you must mount the other Linux filesystems as NFS version 2 by using nfsvers=2 in the /etc/fstab file.

For example:

```
voila:/users  /users nfs     exec,dev,suid,rw,nfsvers=2 1 1
```

For more information, see the fstab(5) man page.

Connecting the Digi EtherLite Serial Multiplexer

This appendix tells you how to connect the Digi EtherLite 16 serial multiplexer:

- "Hardware and Software Requirements"
- "Connectivity"
- "DHCP Configuration"
- "EtherLite 16 Driver Installation"

Hardware and Software Requirements

You will need the following:

- EtherLite 16 serial multiplexer
- Two RJ45-DB9 (female) cables (Part No: 9290165) per node (for example, a 16-node cluster requires 32 cables)
- RJ45 Ethernet standard or crossover cable
- EtherLite 16 driver software, which is found on the CD under the RPMS directory called `40002090_2P.rpm`.

  Note: This RPM will build a version of the Digi software to match your kernel source located in `/usr/src/linux`.

- `dhcp` RPM, installed from the Linux base OS distribution

  For example, you could install DHCP as follows if you are using Red Hat base OS:

```
# mount /dev/cdrom /mnt/cdrom
mount: block device /dev/cdrom is write-protected, mounting read-only
# ls /mnt/cdrom/RedHat/RPMS/dhcp-*
/mnt/cdrom/RedHat/RPMS/dhcp-2.0b1pl6-6.i386.rpm
# rpm -ihv /mnt/cdrom/RedHat/RPMS/dhcp-2.0b1pl6-6.i386.rpm
package dhcp-2.0b1pl6-6 is already installed
```

Connectivity

Connect a node's serial console and the EMP port to the EtherLite 16 using two RJ45-DB9 cables.

If the EtherLite 16 is connected to a Ethernet HUB/Switch use a standard RJ45 Ethernet cable. If the unit is directly connected to an Ethernet card, use a crossover cable.

Power up the unit. The power LED will flicker while the system tries to obtain its IP address from a host system by means of the Dynamic Host Configuration Protocol Server (DHCP). After the IP address as been assigned, the link light will be lit.

DHCP Configuration

DHCP allows hosts on a IP network to request and be assigned IP addresses. To enable the host system to serve IP addresses, ensure that the `dhcp` package is loaded and is functioning.

Once installed, configure the `dhcpd`(8) daemon so that the EtherLite 16 can obtain an IP address as follows:

1. Place the EtherLite 16 MAC address, which is printed on the cover of the EtherLite 16, in the `/etc/dhcpd.conf` file; this permits the host system to always assign the same address.

```
# Sample /etc/dhcpd.conf
# (add your comments here)
default-lease-time 7200;
max-lease-time 7200;
option subnet-mask 255.255.255.0;
option broadcast-address 123.112.212.255;
option routers 123.112.212.254;
option domain-name-servers 123.112.186.51, 123.112.186.50;
option domain-name "acme.com";

host el16 {
     hardware ethernet 00:a0:e7:01:15:c4;
     fixed-address 123.112.212.100;
}
```

```
subnet 123.112.212.0 netmask 255.255.255.0 {
      range 123.112.212.101 123.112.212.110;
}
```

2. Start DHCP in one of the following ways:

   - By using `init.d`:

     # **chkconfig dhcpd on**

   - Manually, specifying to display debug information and to run in the foreground:

     # **/usr/sbin/dhcpd -d -f eth1**

     ```
     [root@ace1 /root]# dhcpd -d -f eth1
     Internet Software Consortium DHCPD $Name:  $
     Copyright 1995, 1996, 1997, 1998 The Internet Software Consortium.
     All rights reserved.
     Multiple interfaces match the same subnet: eth0 eth1
     Multiple interfaces match the same shared network: eth0 eth1
     Multiple interfaces match the same subnet: eth0 eth2
     Multiple interfaces match the same shared network: eth0 eth2
     Listening on Socket/eth1/128.162.212.0
     Sending on   Socket/eth1/128.162.212.0
     DHCPDISCOVER from 00:a0:e7:01:15:c4 via eth1
     DHCPOFFER on 128.162.212.100 to 00:a0:e7:01:15:c4 via eth1
     DHCPREQUEST for 128.162.212.100 from 00:a0:e7:01:15:c4 via eth1
     DHCPACK on 128.162.212.100 to 00:a0:e7:01:15:c4 via eth1
     ```

   For more information , see the `dhcpd`(8) man page.

3. Add the EL16 host address and name to the system's `/etc/hosts` file.

4. Power up the EL16. DHCP will assign an IP address to the unit.

5. Check the `/var/log/messages` file for status messages.

6. It may be necessary to add a host-level route to the EtherLite 16 if it is on another interface. For example:

   ```
   [root@ace1 /root]# route add -host 128.162.212.100 dev eth1
   [root@ace1 /root]# rlogin el16
   ```

```
EL-16 EtherLite module

? ver
Product:  EL-16
FW Ver:   V7.2

Ethernet:  00:A0:E7:01:15:C4
IP:        128.162.212.100
GW IP:     128.162.212.254
SN Mask:   255.255.255.0
Lease:     0x1BF6
Boot Host: 128.162.212.42
Bootfile:  el16.prm
TFTP of Bootfile timed-out
? exit
? rlogin: connection closed.
#
```

EtherLite 16 Driver Installation

Follow the instruction in the following file:

`/usr/src/dg/els/drv/linux/release.notes`

Index

R

S

T