Scyld Beowulf Series 30

Quick Installation Guide

Scyld Software

Scyld Beowulf Series 30: Quick Installation Guide
by Scyld Software
Series 30cz-1 Edition
Published February 2006

Table of Contents

Preface

Congratulations on purchasing the most scalable and configurable Linux clustering software on the market, *Scyld Beowulf*™. This guide describes how to install the *Scyld Beowulf* software and have a scalable cluster running in just a few minutes.

While this document contains all of the information needed to get your system running, additional guides and reference manuals are available. The full Installation Guide contains detailed information about installing and configuring your cluster. The *Administrator's Guide* and *User's Guide* describe how to configure, use, maintain and update the cluster. The *Programmer's Guide* and *Reference Guide* describe the commands, architecture and programming interface for the system. All of the documentation may be viewed using a browser from the last CD in the set (this CD is an autorun CD, and should bring up a browser on either Windows or Linux).

Feedback

We welcome any reports on errors or difficulties that you may find. We also would like your suggestions on improving this document. Please direct all comments and problems to: support@scyld.com.

When writing your e-mail, please be as specific as possible, especially with errors in the text. Please include the chapter and section information. Also, mention in which version of the manual you found the error. This version is *Series 30cz-1*, published February 2006.

Chapter 1. Scyld Beowulf System Overview

The *Scyld Beowulf* Series 30cz-1 streamlines the processes of configuring, running, and maintaining a Linux cluster using a group of off-the-shelf computers connected through a private network. The front-end "head node" computer in the cluster distributes computing tasks to the other machines known as the "compute nodes", in a parallel architecture.

System Components and Layout

The head node is configured with a full Linux installation. Each machine in the cluster is installed with a Network Interface Controller (NIC) communicating with an internal cluster network. In order for the head node to communicate with an outside network, it needs two NICs, one for the private internal cluster network, and the other for the outside network. It is suggested that the head node be connected to an outside network so multiple users can access the cluster from remote locations as shown in Figure 1-1.

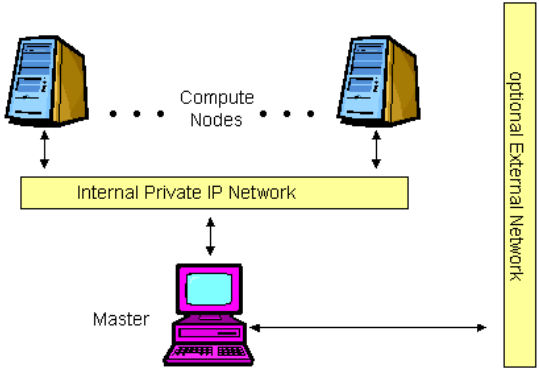

Figure 1-1. Cluster Configuration

In addition to the NIC(s), a booting mechanism is needed. The compute nodes do not boot a full distribution themselves. Instead they boot from the network using PXE boot, or optionally from a CD or floppy disk, or some other supported media that contains a small boot image (see the *Administrator's Guide* for a list of compute node boot options).

For any Beowulf system, hardware selection is based upon the price/performance ratio. Scyld recommends the following components for use with this release of the *Scyld Beowulf* distribution:

Recommended Components

Processors

Intel® Pentium® IV, Intel® Xeon®, AMD Opteron™, single- or multi-core.

Architecture

one, two, or four sockets per motherboard

Physical Memory

1,024 MBytes (1 GByte) or more preferred, minimum 512 MBytes

Operating System

Scyld Beowulf (this release)

Network Interface Controllers (NIC)

Gigabit Ethernet (Fast Ethernet minimum) PCI—X or PCI-Express adapters (with existing Linux driver support) in each node for the internal private IP network. The head node requires an additional NIC for connecting the cluster to the external network. This NIC should be selected based on the network infrastructure (e.g., Fast Ethernet if the external network you are connecting the cluster to is Fast Ethernet). For a list of the latest supported NICs contact Scyld.

Network Switch

All compute nodes, and the head node private network NIC, must be connected to a non-blocking Gigabit Ethernet switch for the internal private network. At a minimum, the network switch must match the speed of the network cards. Note that the switch is a critical component for the correct operation and performance of the cluster. In particular, the switch must be able to handle all of the network traffic over the private interconnect, including cluster management traffic,

migrating processes, transferring libraries, and storage traffic, and it must also properly handle DHCP and PXE.

Note: it is sometimes confusing to identify which NIC is connected to the private network. Take care to connect the head node to the private switch through the NIC with the same or higher speed than the NICs in the compute nodes.

Drives

For the head node, we recommend using either SATA or SCSI disks in a RAID 1 (mirrored) configuration. The operating system on the head node requires approximately 3 GB of disk space. We recommend configuring the compute nodes without local disks (diskless).

If the compute nodes do not support PXE boot, a floppy (32-bit architectures) or bootable CD-ROM drive (32- and 64-bit architectures) is required. If local disks are required on the compute nodes, we recommend using them for storing data that can easily be re-created, such as scratch storage or local copies of globally-available data.

If you plan to create boot CDs for your compute nodes, your head node requires a CD-RW or writable DVD drive.

Optional Hardware Components

Gigabit Ethernet with a non-blocking switch serves most users. However, some applications benefit from a lower-latency interconnect. Infiniband is an industry standard interconnect providing low latency messaging support, as well as IP and storage support. Although higher cost than Gigabit Ethernet, Infiniband can be configured as a single universal fabric serving all of the cluster's interconnect needs. More information about Infiniband may be found at the Infiniband Trade Association web site (http://www.infinibandta.org). Scyld supports Infiniband as a supplemental messaging interconnect in addition to Ethernet for cluster control communications.

Assembling the Cluster

The full Scyld Beowulf Scalable Computing Distribution is only installed on the head node. A graphic utility (`BeoSetup`) is available and included on the head node to aid in the cluster configuration process.

Most recent hardware supports network boot (PXE boot), and Scyld recommends the use of PXE boot for booting compute nodes. For nodes that do not support network boot, each compute node requires a floppy disk or CD-ROM drive, with suitable boot media inserted before being powered up.

Software Components

A brief description of the major portions of the *Scyld Beowulf* distribution is given below. For more information, see the *Administrator's Guide* and the *User's Guide*.

- *BeoSetup*: A graphic utility for configuring and administering the Scyld Beowulf cluster.
- *BeoStatus*: A graphic utility for monitoring the status of the Scyld Beowulf cluster.
- *Scyld BeoMaster* The *Scyld BeoMaster* software is an integral part of the *Scyld Beowulf* distribution. It allows processes to be started on compute nodes in the cluster and tracked in the process table on the head node. *BeoMaster* also provides process migration mechanisms to help in the creation of remote processes and removes the need for most binaries on the remote nodes.
- *MPICH/LAM*: Message Passing Interface, modified to work with Scyld Beowulf cluster software.
- *Beowulf utilities*: several utilities to control and display nodes and processes in the Scyld Beowulf cluster.

Chapter 2. Quick Start Installation

The Scyld Beowulf distribution is provided as a set of three CD-ROM discs which include the basic Linux operating system distribution as well as Scyld Beowulf cluster software. The first CD in the series (disc 1) is bootable, and is used to initiate the install on the head node. The last disc in the series also contains product documentation which may be read directly from the disk on any running PC or workstation.

This chapter outlines two simple cases: installing on a head node with network-booted compute nodes, and installing on a head node where the compute nodes must be booted from local media. Refer to the full Installation Guide, Chapter 3, Graphical Installation for other scenarios.

Other machines join the cluster as compute nodes, and require no explicit installation. They boot either by obtaining a boot image over the network using PXE, or with boot media (floppy or CD-ROM) that converts them to a Scyld-developed network boot system. If your hardware does not support PXE boot, a bootable floppy or CD-ROM may be created using the BeoSetup utility on the head node.

Install the Head Node

If you need more information on any of the following steps, the installation procedure is fully documented in the full Installation Guide, Chapter 3, Graphical Installation.

1. Boot the front-end (head node) machine from the Scyld Beowulf distribution CD-ROM labeled Disc 1. The graphical installation process starts after 20 seconds.
2. Follow the on-screen instructions to execute the installation. The first few screens set basic elements, including the default language, keyboard and mouse. For most screens you may accept the defaults.

3. If you need to set up hard disk partitions or bootloader defaults differently from the default, please see the detailed instructions in the full Installation Guide, Chapter 3, Graphical Installation.
4. You must configure the networks, as described here and in the full Installation Guide, Chapter 3, Graphical Installation.

> **Tip:** To proceed with configuring the network, you must know which interface is connected to the public network and which is connected to the private network.

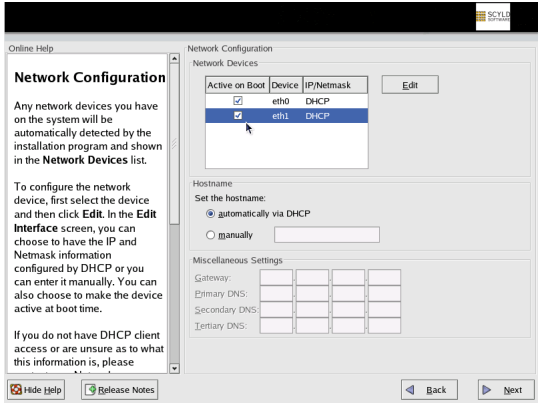

Figure 2-1. Network Interface configuration

For eth0 (or the interface connected to the public network), *DHCP* is selected by default. If your external network is set up to use static IP addresses, select this interface and click **Edit**—your network administrator should provide you with the IP address). Set the *IP Address* and *Netmask*, then click **OK**. If you set a static IP address for the public interface, you must also click *manually* for *Set the hostname* and provide a hostname, gateway and primary DNS IP addresses.

Caution

Note: For eth1 (or the device connected to the private cluster network), *do not* select DHCP. You must select and edit this interface and manually set the IP address (see Figure 2-2). Also check the *Activate on Boot* box to make the specific network device initialized at boot-time.

Note that the head node also functions as a PXE and DHCP server for the cluster. On the Firewall page following, ensure that the interface is set as a trusted interface (check the box under "Allow all traffic to pass for this interface").

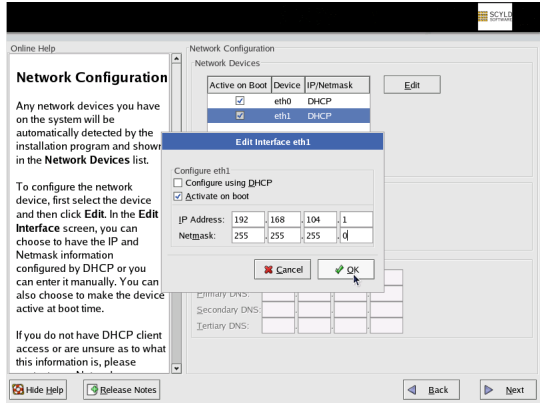

Figure 2-2. Manually set IP Address

For eth1 (or interface connected to the internal private cluster network), you must not choose a dynamic IP address, and we recommend choosing a non-reroutable address (such as 192.168.x.x or 10.x.x.x). Once you have specified the *IP Address*, you must also set your *Netmask* based on the address. Click **OK** to return to the screen described in Figure 2-1.

Configure the network settings for all of the devices listed. Click **Next** to continue.

5. Proceed through the dialog boxes to configure a firewall, additional language support, and your timezone.
6. You must supply a root password. Refer to the full Installation Guide, Chapter 3, Graphical Installation. An alphanumeric password of at least 8 characters, with special characters is recommended.
7. Review the packages to be installed.
8. At this point, the system is ready to install the software you configured in earlier steps. The installer prompts for additional disks as necessary.
9. After the software is installed to the hard drive, you are prompted to verify the video hardware determined when the installer probed the system. If the default selections are incorrect, see the full Installation Guide, Chapter 3, Graphical Installation. The system prompts you to reboot when necessary.
10. On the subsequent boot, a *Welcome* screen appears with the Scyld Software logo. The next screen enables you to set up your Scyld Beowulf cluster by choosing which ethernet interface to use, choosing the number of nodes in your cluster, and establishing the initial IP address for the cluster. Choose an IP numbering system that will encompass your entire cluster, with room for expansion. For example, you might use the range 192.168.104.10 through 192.168.104.50 for a 30-node cluster, with room to add 10 more nodes later.

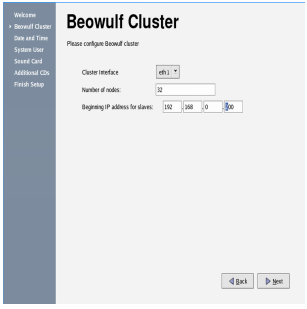

Figure 2-3. Manually set IP Address

11. Continue through the installation screens to set the date and time, set up a system user, and verify the sound card, if one is installed. Scyld Beowulf is now installed and set up on your system. Next, boot and configure the compute nodes to get your cluster up and running.

    Note: if prompted to restart Beowulf services, click *Yes*.

Boot and Configure the Compute Nodes

1. If you are not logged in as root already, log into the head node using the root username and password set up earlier. Start the cluster configuration tool, **beosetup**, by clicking on the Beosetup "Building Blocks" icon at the bottom of the screen (hover the cursor over the tray icon that looks like triangle of yellow blocks, then click). Note, if **beosetup** fails to start, check the syslog for possible errors. You can manually start it by typing **/usr/bin/beosetup** in a terminal window.
2. If your compute nodes can't PXE boot, or if for some reason you don't want them to use PXE, you may create compute node disks. *You only need to perform this step if you want to boot compute nodes from boot media.* You may create either boot floppy disks, or boot CD-ROM disks. You need a CD-RW drive installed on the head node to create boot CDs, and CD-ROM drives on each compute node in order to use them.
   a. Click **Node Floppy** or **Node CD** in the *beosetup* window.
   b. Click **OK** in the *Create BeoBoot Floppy* or **Create BeoBoot CD Image** window to write the node boot image. This image consists of a basic boot image which first boots the node and then downloads the full compute node boot image.
   c. Click **Close** to close the window.
   d. Write the boot image to blank media, and then place the boot media in each compute node. Repeat for each compute node.
3. Boot the compute nodes by powering them on in the order you want them to be numbered, typically one by one from the top of a rack down (or from the

bottom up). You can reorder nodes later if necessary (see the Administrator's Guide).

As compute nodes join the cluster, they are listed in `beosetup` by Ethernet Station (MAC) Addresses and given node numbers in the order they initially contact the head node. After installation is complete, this order may be manually changed, but powering up nodes in the desired order is easier and is recommended.

4. The nodes transition through the boot phases and after a few seconds be shown in the **up** state in `beosetup`. The cluster is now fully operational with diskless compute nodes.

Status of the compute nodes is listed in the `beosetup` window. All compute nodes should show status of *`up`* when ready for use. Note: an *`error`* state may be encountered due to lack of a partition table.

Chapter 3. Cluster Verification Procedure

After you've finished configuring the master and compute nodes of your Scyld Beowulf cluster, the next step is to verify that the cluster is working properly. The following verification procedure is meant to identify common software and hardware configuration problems by running basic administrative and operational commands. When contacting your reseller for support with a new problem, typically the first question asked is if this verification procedure has been run, and what the results were.

bpstat

Entering the command **bpstat** at a shell prompt on the master node displays a table of status information for each node in your cluster. You do not need to be a privileged user to use this command. An example of using this command is shown below.

```
[root@cluster root]# bpstat
Node(s)     Status       Mode        User       Group
5-9         down         ---------- root       root
4           up           ---x--x--x any        any
0-3         up           ---x--x--x root       root
```

From the above table generated by **bpstat**, verify that you see `up` listed in the `Status` column for each node you've configured and have powered up. Status will be shown for each possible node in the cluster. The possible node count is based on the number of nodes specified by the `iprange` (see the Preference Settings in **beosetup**). Nodes that have not yet been configured are marked as `down`. If any node in the table contains `boot` in the `Status` column, this state is temporary while the node is booting. Wait 10-15 seconds and try again. If any node in the table contains `error` in the `Status` column, that node is operating but has experienced an initialization problem. As a first step, right click on the node entry in the **BeoSetup** display and select *View*

log to check for error messages. Typical problems are failing network connections, unpartitioned hard disks or unavailable network file systems.

beostatus

Clicking on the **Beostatus** icon on the desktop system tray, or entering the command **beostatus** at a terminal windows on the master node displays a graphical user interface (GUI) program. You do not need to be a privileged user to use this command. The **beostatus** window is shown in Figure 3-2.

Figure 3-1. Beostatus Icon

Node	Up	Available	CPU 0	CPU 1	Memory	Swap	Disk	Network
-1	✓	✓	7/100% (7%)	14/100% (14%)	125/128MB (98%)	25/259MB (10%)	79/960MB (8%)	34 kBps
0	✓	✓	0/100% (0%)	0/100% (0%)	23/65MB (36%)	None	9/36MB (25%)	33 kBps
1	✓	✓	0/100% (0%)	0/100% (0%)	35/65MB (54%)	None	9/36MB (25%)	33 kBps
2	✓	✓	0/100% (0%)	0/100% (0%)	27/65MB (42%)	None	9/36MB (25%)	33 kBps
3	✓	✓	0/100% (0%)	0/100% (0%)	26/65MB (40%)	None	9/36MB (25%)	33 kBps
4	✗	✗	0/100% (0%)	N/A	0 %	None	0 %	0 kBps
5	✓	✓	0/100% (0%)	N/A	34/496MB (7%)	None	9/36MB (25%)	34 kBps
6	✓	✓	1/100% (1%)	N/A	34/496MB (7%)	None	9/36MB (25%)	67 kBps
7	✓	✓	1/100% (1%)	N/A	34/496MB (7%)	None	9/36MB (25%)	65 kBps
8	✓	✓	1/100% (1%)	N/A	34/496MB (7%)	None	9/36MB (25%)	33 kBps
9	✓	✓	0/100% (0%)	N/A	34/496MB (7%)	None	9/36MB (25%)	34 kBps
10	✓	✓	1/100% (1%)	N/A	34/496MB (7%)	None	9/36MB (25%)	72 kBps
11	✓	✓	0/100% (0%)	N/A	34/496MB (7%)	None	9/36MB (25%)	34 kBps
12	✓	✓	0/100% (0%)	N/A	34/496MB (7%)	None	9/36MB (25%)	22 kBps
13	✓	✓	0/100% (0%)	N/A	34/496MB (7%)	None	9/36MB (25%)	33 kBps

Figure 3-2. BeoStatus

The default mode of the Beostatus GUI is known as the "Classic" display. This mode displays specific state and resource usage information on a per-node table format.

Each row in the **beostatus** window corresponds to a different node in the cluster. The following list details the columns in the **beostatus** window:

Node

This is the node's assigned number in the cluster. Compute nodes are numbered starting with zero. Node -1, if shown, is the master node. The total number of node entries shown is set by the `iprange` or `nodes` keywords in the file

`/etc/beowulf/config`, not the number of detected nodes. Inactive node entries display the last reported data in a faded or "grayed" row.

Up

This column gives a graphical representation of the node's status. A green checkmark is shown if the node is up and available. Otherwise, a red 'X' is shown.

State

This column prints the last known state of the node. The information in this column should agree with that reported by both **bpstat** and `BeoSetup`.

CPU 'x'

The next set of columns show the CPU loads for the node. At a minimum, there will be one column displaying the CPU load for the first processor in each node. Since it is possible to mix uni-processor machines with multi-processor machines in a Scyld Beowulf, the number of CPU load columns is equal to the maximum number of processors for a given node in your cluster. For those nodes that contain less than the maximum number of processors, their columns display N/A.

Memory

This column displays the current memory usage of the node.

Swap

This column displays the current swap space (virtual memory) usage of the node.

Disk

This column displays the current hard disk usage of the node. If the nodes are using a RAMdisk, they will show a maximum of 36MB.

Network

This column displays the current network bandwidth usage of the node. The total amount of bandwidth available is the sum of all network interfaces for the individual node.

Verify that the information shown in the **beostatus** window is correct. The configured nodes that are powered up (those with a green checkmark in the `Up` column) should show expected values in the subsequent usage columns. Assuming there are no active jobs on your cluster, the CPU and Network usage columns should be fairly close to zero. The memory usage columns (Memory, Swap and Disk) should be showing reasonable values.

bpsh

The **bpsh** command is the Beowulf shell command. It is analogous in functionality to both the **rsh** and **ssh** commands. It is used to execute commands on the nodes in your cluster from the master. For example, this command will execute on node number 3:

```
[root@cluster root]$ bpsh 3 ls -al /tmp
```

linpack

HPL is a portable version of the High Performance Linpack benchmark. Run it with all available nodes using the following shell script (wait up to a minute to see its

complete output).

```
[root@cluster root]$ linpack
```

> **Caution**
>
> The linpack script runs a non-optimized version of the HPL benchmark, and is intended for verification purposes only. Do not use the results for performance characterization.

mpi-mandel

The **mpi-mandel** program is a visualizer for the Mandelbrot set. The following command is an example of how to run this program using 4 processors:

```
[root@cluster root]$ NP=4 mpi-mandel  --demo \
                     /usr/share/doc/mpi-mandel-1.0.20a/mandel.fav
```

Chapter 4. Troubleshooting a Scyld Beowulf Installation

Failing PXE Network Boot

If upon initially powering on the compute nodes (or subsequently), the nodes fail to join the cluster (or fail to appear in either the main or *Unknown* panels of `BeoSetup`, the problem may be that they are not finding the head node's DHCP server, or another DHCP server is answering and supplying an IP address.

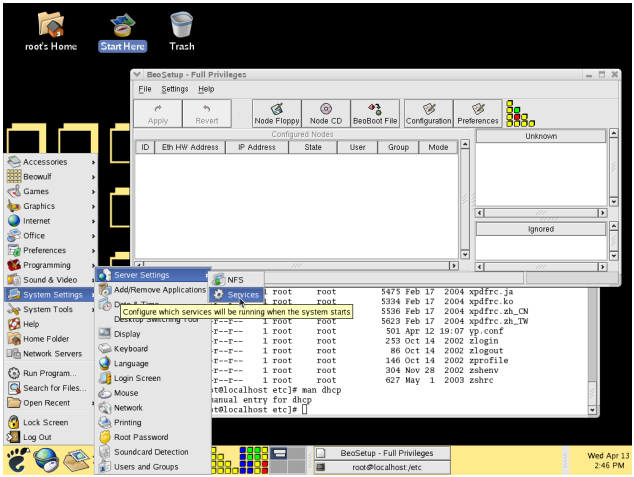

Figure 4-1. Starting Services Applet

Verify that the Beowulf services are started on the head node by opening the `Server Settings` applet (see Figure 4-1), and observing whether the beowulf service is started. If not, click **StartCluster** item under the **File** menu (see Figure 4-2). Then power-cycle a compute node to see if it now joins the cluster.

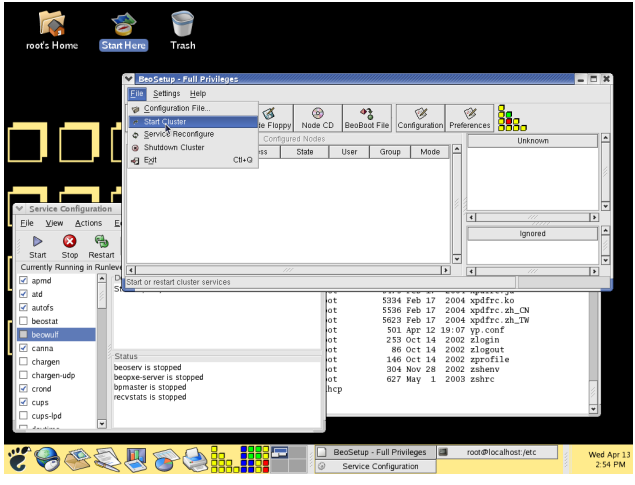

Figure 4-2. Starting Beowulf Services

If you are unable to start the cluster services (or Service Reconfigure), verify that the Master network interface is properly set using the **Configuration** button, Network Properties tab (see Figure 4-3), then start or reconfigure cluster services again. Verify that the Beowulf services have started (see Figure 4-4). Checking the boxes next

to the beostat and beowulf services (in the `Service Configuration` applet) will insure these services start at boot time. Be sure to click **Save** before exiting the applet.

Try booting your compute nodes again.

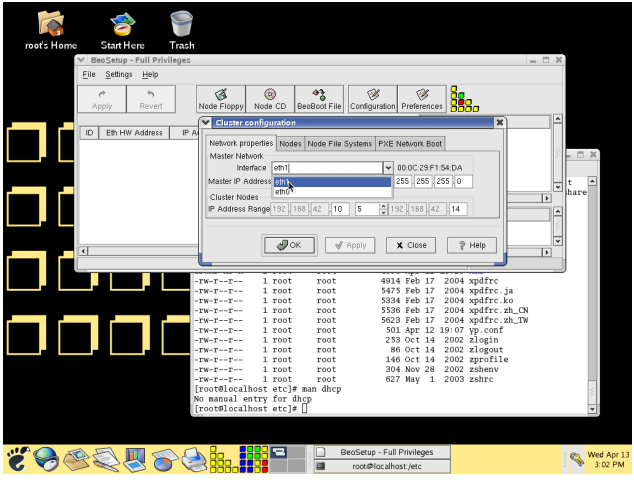

Figure 4-3. Checking Master network

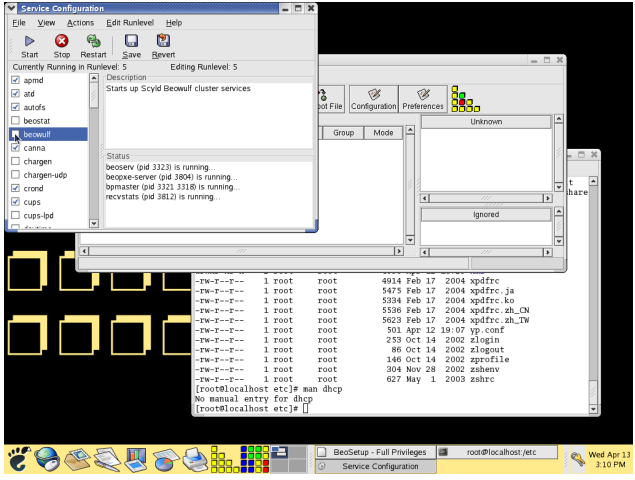

Figure 4-4. Beowulf Services Running

If the compute nodes fail to boot immediately after power-up, but successfully boot later, a common problem is the configuration of a managed switch.

Some Ethernet switches delay forwarding packets approximately one minute after link is established, attempting to verify that no network loop has been created ("span-

ning tree"). This delay is longer than the PXE boot timeout on some servers.

The solution is to disable the spanning tree check on the switch. The parameter is typically named "fast link enable". Note that BeoBoot Stage One was designed to attempt network boot for longer than the spanning tree timeout, thus the observed symptom is a delay booting the computing nodes rather than a failure.

Mixed Uniprocessor and SMP Cluster Nodes

One of the benefits of the Scyld Beowulf system architecture is that it eliminates the problem of unintentionally running different versions of a program over the cluster. The system eliminates version skew among compute nodes.

One requirement is that cluster nodes must run the same kernel version, typically with the same features and optimization. Uniprocessor machines can run the SMP kernel and SMP machines can run the uniprocessor kernel (although it will use only one processor). The best choice for a mixed cluster is to run the SMP kernel.

The kernel selection is handled at master installation time, based on the type of hardware detected and the response to questions. An SMP kernel is installed if the master is detected as an SMP, or if the *SMP* CPU selection box is checked during installation. There is a similar decision made based on the processor generation, for instance a kernel compiled to use Intel® Pentium® IV features will not work on a Transmeta(tm) Crusoe(tm) processor. If you installed a specialized kernel on a master that now needs to support slave nodes with a different set of features (e.g. uniprocessor master with SMP slave nodes), you must execute the following steps:

1. Mount the Scyld Beowulf CD-ROM on the head node.
2. Change to the directory `mount-point/Scyld/RPMS`, where mount-point is typically /mnt/cdrom.

3. Copy the kernel rpm, `kernel-smp-`*`kernel-version`* to the head node.
4. Install this kernel on the head node, by executing:

   ```
   bash$ rpm -i kernel-smp-version
   ```

5. Reboot the head node and select the name of the SMP kernel from the boot loader prompt or GUI.
6. Make a new Phase 2 image by executing:

   ```
   bash$ /usr/bin/beoboot -2 -n
   ```

Note that rebooting the head node automatically reboots the compute nodes, causing them to automatically use the updated kernel.

Mixed 32- and 64-bit cluster nodes

Mixing 32- and 64-bit nodes is not possible. The head node is migrating processes to the compute nodes. All nodes in the cluster must have the same CPU architecture. If you want to mix Opteron nodes and IA32 (Pentium or Xeon), you must boot the Opteron in 32-bit mode.

Device Driver Updates

Scyld Beowulf releases are tested on many different machine configurations, but it is not possible to provide device drivers for hardware that was unknown at the time of release.

Most unsupported hardware, or device-specific problem are resolved by updating to a newer device driver, but some devices may not yet be supported under Linux. Check with your hardware vendor.

The Scyld Beowulf architecture makes most driver updates simple. Drivers are installed and updated on the head node exactly as with a single machine installation. The new drivers are immediately available to compute nodes, although already-loaded drivers are not replaced.

There are two irregular device driver types that require special actions: disk drivers on the head node, and network drivers on the compute nodes. In both cases the drivers must be available to load additional drivers and programs, and are thus packaged in initial ramdisk images.

Device Driver Notes

Scyld Beowulf uses XFree86 version 4.3.0-1_Scyld for video card support. Any driver compatible with XFree86 will work with the system; check xfree86.org[1] for driver updates and video related trouble shooting information.

The LM sensor subsystem is an optional package that allows monitoring temperature, fan speed and other physical parameters. Before configuring this driver package, check that your chipset is supported. Installing on unsupported chipsets has been known to hang machines during the boot phase. If compute nodes hang during boot, the last line in the node boot log, `/var/log/beowulf/node.x` usually indicates the problem.

Finding Further Information

If you encounter a problem installing your Scyld Beowulf cluster, and you find this guide cannot help you, check the following sources for pertinent information:

- See *Installation Guide, Graphical Install of Front-End Node* on the head node or on the Scyld disc in the installation kit for detailed installation instructions.
- The *Administrator's Guide* is available on the head node or the Scyld disc in the installation kit for a description of more advanced administration and setup options.
- The *Reference Guide* on the head node or on the Scyld disc in the installation kit for a complete technical reference to the Scyld Beowulf software.
- Run the `BeoSetup` application for access to detailed error info regarding the status of booting the compute nodes.

Please visit the Scyld MasterLink™ website at http://www.scyld.com/support.html for the most up to date product documentation and other helpful information about your Scyld Beowulf software.

Notes

1. http://www.xfree86.org
2. http://www.scyld.com/support.html