Installing an IBM Cluster 1350

Using

Extreme Cluster Administration Toolkit (xCAT)

February 2007

Contents

1. Introduction

The Extreme Cluster Administration Toolkit (xCAT) is a collection of mostly script based tools to build, configure, administer, and maintain Linux clusters. This document describes how to implement a Linux cluster on an IBM Cluster 1350 using xCAT v1.2.0 and other third party software. Software versions referenced in this document may be out of date, check for the latest versions before continuing.

xCAT is for use by IBM and IBM Linux cluster customers. xCAT is copyright © 2000-2007 IBM corporation. All rights reserved. Use and modify all you like, but do not redistribute. No warranty is expressed or implied. IBM assumes no liability or responsibility.

2. Tested configurations

The following hardware architectures and operating system versions have been tested with xCAT v1.2.0:

Operating System	**Supported Versions**	**Platform**
RedHat	RHEL4 ES U4, RHEL4 WS U4, RHEL3 AS U4, RHEL3 WS U4, RHEL3 ES U4	**x86 (i386, i486, i586, i686), x86_64 (**64-bit EM64T & Opteron)
Fedora	Core 1, Core 2, Core 3	**x86, x86_64**
CentOS	3.4	**x86, x86_64**
SuSE Linux	SuSE 8.1, 8.2, 9.0, 9.1, 9.2 SLES8, SLES8 SP1, SLES8 SP2a, SLES8 SP3, SLES9 SP1, SLES9 SP2, SLES9 SP3	**x86, x86_64**
RedHat	RedHat 7.2 RHEL3 AS U4, RHEL3 ES U4, RHEL3 WS U4, RHEL4	**Itanium 1 and 2 (IA64)**
SuSE Linux	SLES 8, SLES8 SP2, SLES8 SP3 SLES9, SLES9 SP1	**Itanium 1 and 2 (IA64)**
RedHat	RHEL3 AS U4 RHEL4 AS, ES, WS	**PPC64 (IBM JS20 only)**
SuSE Linux	SLES8 SP3aa SLES9 SP1	**PPC64 (IBM JS20 only)**

Table (i) xCAT Supported OS distributions and platforms for IBM Cluster 1350

Note:- **PPC64** Node install tested only, however should work as management node.

The configuration examples shown in this document may need to be altered to suit any variances in the cluster and architecture, but the examples should give a good general idea of what needs to be done. Please do not use this document verbatim as an implementation guide. This document should be used as a reference to a custom implementation. Use the man pages, source and other documentation that is available to figure out why certain design or configuration choices are made and what different choices may be made. Because IBM Cluster 1350 clusters are preconfigured from manufacturing, this document covers very little of the hardware configuration that is required to implement a cluster.

Additional documentation including hardware installation and configuration is available as a RedBook at http://publib-b.boulder.ibm.com/Redbooks.nsf/9445fa5b416f6e32852569ae006bb65f/7b1ce6b3913cafb386256bdb007595e8?OpenDocument&Highlight=0,SG24-6623-00.

See http://www.redbooks.ibm.com/redbooks.nsf/Redbooks?SearchView&Query=linux+cluster&SearchMax=4999 for additional information about implementing a cluster.

3. Overview of xCAT

xCAT is set of Perl/UNIX shell scripts and programs necessary for system administrators to implement complex cluster systems using a set of heterogeneous hardware components, as well as to perform the daily cluster administration/maintenance tasks efficiently. xCAT is primarily designed and tested on IBM server systems, however, it has been used with other vendor systems as well with equal success.

The features in xCAT are a result of the requirements met in hundreds of real cluster implementations. xCAT has emerged as a powerful cluster installation tool over the years, resulting in a modular toolkit that represents best practices in cluster management and a flexibility that enables it to change rapidly in response to new customer requirements to work with many cluster topologies and architectures.

xCAT works well with the following cluster types:

1. **High Performance (HPC):** such as computing physics, seismic, CFD, FEA, weather, bioinformatics and other simulations.
2. **Horizontal Scaling (HS):** such as Web farms.
3. **Administrative:** A very convenient platform, although non-traditional, to install and administer a number of Linux machines.
4. **Microsoft Windows and other Operating Systems:** With xCAT's cloning and imaging support, it can be used to rapidly deploy and conveniently manage clusters with compute nodes that run Windows or any other operating system.

xCAT's current features:

1. Supports Any OS on compute nodes via OS-agnostic imaging.
2. Hardware Control
   - Remote Power control (on/off state) through IBM Management Processor Network, BMC, and/or APC Master Switch.
   - Remote software reset (rpower).
   - Remote Network BIOS/firmware update and configuration on IBM hardware.
   - Remote OS console with pluggable support for a number of terminal servers.
   - Remote POST/BIOS console through the IBM Management Processor Network and with terminal servers.
3. Boot Control Ability to remotely change boot type (network or local disk) with syslinux.
4. Automated parallel install using scripted RedHat kickstart, SuSE Linux autoyast, on IA32, x86_64, PPC, and IA64.
5. Automated parallel install using imaging with other Linux distributions, Windows, and other operating systems.
6. Automated network installation with supported PXE NICs, with Etherboot or BootP on supported NICs without PXE.
7. Remote Monitoring
   - Hardware alerts and email notification with IBM's Management Processor Network and SNMP alerts.

Remote vitals such as fan speed, temperature, and more with IBM's Management Processor Network.

Monitoring remote hardware event logs with IBM's Management Processor Network/IPMI Interface.

8. Administration utilities

   Parallel remote shell, ping, rsync, and copy.

   Remote hardware inventory with IBM's Management Processor Network.

9. Software Stack

   PBS and Maui schedulers to build scripts, documentation, automated setup, extra related utilities, and deep integration.

   Myrinet to automate setup and installation.

   MPI to build scripts, documentation, and automated setup for MPICH, MPICH-GM, and LAM.

10. Usability

    Command line utilities for all cluster management functions.

    Single operations can be applied in parallel to multiple nodes with very flexible and customizable group/range functionality.

    Flexible support for various user defined node types.

11. Diskless support using warewulf.

4. Getting the xCAT software distribution

This section explains where and how to get the xCAT software distribution.

1. Download the latest version of xCAT. Three of the five required packages can be downloaded from http://www.alphaworks.ibm.com/tech/xCAT/ to the */opt directory*. *T*he fourth file can be downloaded from http://www-rcf.usc.edu/~garrick/ to the */opt directory*. The fifth file can be downloaded from http://www.xcat.org/patch/ to your desktop. The patch file will be decompressed into */opt/xcat* after *./setupxcat* has been run.
2. Download the latest firmware and hardware configuration software from http://publib.boulder.ibm.com/cluster/.

Note: Additional documentation is accessible in the */opt/xcat/doc* folder once the *xCAT .tar* files are decompressed. For assistance with building, maintaining, and administering the xCAT cluster or an xCAT feature request see the xCAT user mailing list, your IBM sales rep, or other IBM point of contact.

5. Understanding cluster components, connections, and architecture

This document is based on a basic 32 node cluster that uses serial terminal servers for out-of-band console access, an APC Master Switch, IBM's Service Processor Network for remote hardware management, Ethernet, and Myrinet as the basis of most of its examples. All network devices that must be statically set in the IBM Cluster 1350 are pre-configured using the manufacturing defaults listed in the “IBM Manufacturing defaults for all items in Cluster 1350 Clusters” table.

The following three examples describe some of the detail of this example cluster:

Components / Rack Layout

The following hardware is positioned in the rack, starting from the bottom and moving towards the top:

Ethernet Switch
node32
... nodes 27 - 31
node26
node25 MPA
node24
... nodes 19 - 24
node18
node17 MPA
Management Node
apc1 APC Master Switch
ts2 Terminal Servers
ts1
Monitor / Keyboard
node16
... nodes 11 - 17
node10
node09 MPA
node08
... nodes 03 - 07
node02
node01 MPA has MPA card
Myrinet Switch

1. The Myrinet switch: Used for high-speed, low-latency inter-node communication. A cluster may not have Myrinet, if the cluster is not running parallel jobs that do heavy message passing.
2. Nodes 1-16: The first 16 compute nodes. Note that every 8th node has an MPA (Managment Processor Adaptor) installed. The configuration may have RSA adapters, ASMA adapters, or BMCs. These cards enable the SPN (Service Processor Network) to function and remote hardware management to be performed. Newer machines do not require a RSA or MPA because they contain a built in BMC (Baseboard Management Controller) which uses the IPMI protocol for management. The BMC is internal hardware.
3. Monitor/Keyboard: This is for local input/output function.
4. Terminal servers: The terminals enable serial consoles from all of the compute nodes to be accessible from the management node. This feature is very useful during system setup and after setup administration. Serial Over LAN (SOL) can be used to emulate a terminal server setup if the cluster does not have a terminal server.
5. APC master switch: This enables remote power control of devices that are not part of the Service Processor Network, such as terminal servers, Myrinet switch, and ASMA adapters.
6. The management node: The management node is where the rest of the nodes are installed from and the cluster is managed.

7. Nodes 17-32: The rest of the compute nodes with a Management Processor card every 8th node.
8. Ethernet switch

Networks

All IBM Cluster 1350 network devices that must be statically set are preconfigured with the manufacturing defaults listed in the “IBM Manufacturing defaults for all items in Cluster 1350 Clusters” table.

The following table lists the networks that are used in the rest of this document's examples.

Notes:

1. The listing of attached devices to the right.
2. The external network is the organization's main network. In this example, only the management node has connectivity to the external network.
3. The Ethernet switch hosts both the cluster and management network on separate VLANs.
4. The cluster network connects the management node to the compute nodes. A private class B network that has no connectivity to the external network is recommended during configuration. This is often the easiest way to configure the cluster and a good practice if the configuration might grow to more than 254 nodes.
5. The management network is a separate network used to connect all devices associated with cluster management such as terminal servers, BMC, and ASMA cards, to the management node.
6. Parallel jobs use the message passing network for interprocess communication. Our example uses a separate private class B network over Myrinet. If Myrinet is not being used, this network could be the same as the cluster network. For example, any required message passing could be done over the cluster network.

6. Installing the operating system on the management node

This section covers the steps necessary to install Linux on the management node.

Red Hat Enterprise Linux

The first step in building an xCAT cluster is installing Linux on the management node. See the following URLs for more information on each particular management node type:

xSeries x3650
http://www-304.ibm.com/jct01004c/systems/support/supportsite.wss/docdisplay?lndocid=MIGR-64497&brandind=5000008
xSeries x3550:
http://www-304.ibm.com/jct01004c/systems/support/supportsite.wss/docdisplay?lndocid=MIGR-64455&brandind=5000008
xSeries x3755:
http://www-304.ibm.com/jct01004c/systems/support/supportsite.wss/docdisplay?lndocid=MIGR-65505&brandind=5000008
xSeries x3655:
http://www-304.ibm.com/jct01004c/systems/support/supportsite.wss/docdisplay?lndocid=MIGR-66266&brandind=5000008
xSeries x3455:
http://www-304.ibm.com/jct01004c/systems/support/supportsite.wss/docdisplay?lndocid=MIGR-65672&brandind=5000008

Notes:

1. Your management node may require specific drivers. Please consult the machine specific setup above for instructions.
2. For more detailed setup instructions or troubleshooting assistance on individual IBM servers, refer to the link above.
3. Before you install the OS on the management node it is recommended to disable all PCI adapters in the BIOS initially. The adapters can be enabled again once the OS is installed.

Create and Configure RAID Devices if Necessary

If you are using LSI, HostRaid, or ServeRAID devices in the management node, use the “LSI/HostRaid/ServeRAID flash/configuration” CD to update the LSI/HostRaid/ServeRAID firmware to the latest version and define the RAID volumes. If other nodes exist with hardware RAID, update and configure them now. The latest firmware for the particular RAID type can be downloaded from the IBM server support website http://www.ibm.com/server/support.

NIS Notes

Note: If you plan on interacting with an external NIS server make sure the server supports MD5 passwords and shadow passwords. If it does not support these features, do not turn them on during the install of the management node.

Partition Notes

File System Type should be set up as *ext3*.
Recommended minimum drive partitioning scheme for the management node:

/boot (200 MB)
SWAP (1.5 x physical memory, not to exceed 2GB)
/var (2GB)
/ (the rest of the disk)

Select **No firewall**. This is for xCat installation purposes and can be changed after the configuration of the cluster has completed.

Select **Disable Selinux**.

Select **CUSTOMIZED SOFTWARE PACKAGES TO BE INSTALLED** from the software selection menu.

Scroll down to **Miscellaneous**.
Select the **Everything** option.

Note: If this is the first time installing Red Hat 4, everything is a check box at the end of the selection.

Install all 5, RHWS4 CDs.

You will be prompted during reboot to install and load the Extras CD.

It is recommended to create a normal user other than *root* during the install.

Start the newly installed system by rebooting and then login as root.

Open a terminal

```
>updatedb
```

SuSE Enterprise Linux

If you are using SuSE Linux on your management node, follow the instructions specific to your particular management node type on the following websites to install the SLES operating system:

xSeries x3650:
http://www-304.ibm.com/jct01004c/systems/support/supportsite.wss/docdisplay?lndocid=MIGR-64686&brandind=5000008

xSeries x3755:
http://www-304.ibm.com/jct01004c/systems/support/supportsite.wss/docdisplay?lndocid=MIGR-65511&brandind=5000008

xSeries x3655:
http://www-304.ibm.com/jct01004c/systems/support/supportsite.wss/docdisplay?lndocid=MIGR-66267&brandind=5000008

xSeries x3455:
http://www-304.ibm.com/jct01004c/systems/support/supportsite.wss/docdisplay?lndocid=MIGR-65662&brandind=5000008

Turn off unwanted services (general)

To turn off some of the network services turned on by default during the installation process, use the following commands:

> To view installed services:
> *chkconfig --list | grep ':on'*
>
> To turn off a service:
> *chkconfig --level 0123456 <service> off*

Turn off unwanted services (specific)

Use the following commands to turn off all unnecessary services:

```
chkconfig --level 0123456 autofs off
chkconfig --level 0123456 isdn off
chkconfig --level 0123456 iptables off
chkconfig --level 0123456 ip6tables off
chkconfig --level 0123456 rhnsd off
chkconfig --level 0123456 rawdevices off
chkconfig --level 0123456 kudzu off
chkconfig --level 0123456 FreeWnn off
chkconfig --level 0123456 arptables_jf off
chkconfig --level 0123456 canna off
chkconfig --level 0123456 cups off
chkconfig --level 0123456 hpoj off
chkconfig --level 0123456 alsasound off
```

7. Configuring networking on the management node

This section describes network setup on the management node.

Notes:

1. If you are using the onboard e1000/bcm5700 network interface card, download the latest driver sources from http://publib.boulder.ibm.com/cluster/ and build the kernel module for the driver.
2. A USB storage device for use with the IBM eServer 326 and xSeries 336 servers will be required.

Remove *tg3* driver

```
>service network stop
>lsmod | grep tg3
>rmmod tg3
```

Download *<Driver>.src.rpm*

```
>rpm –ivh <Driver>.src.rpm
>cd /usr/src/redhat/SOURCES/
>tar –zxvf  <Driver>.tgz
>make
>make install
>insmod <Driver.ko>
```

Configure the network devices using the Neat utility.

```
>neat
```

```
      For example, in the Neat utility select:
                  ENT0
                  Statically Set IP
                  IP 172.20.0.1
                  Subnet Mask 255.255.0.0
                  OK

                  ENT1
                  Statically Set IP
                  IP 172.30.0.1
                  Subnet Mask 255.255.0.0
                  OK

                  New (will be BMC and alias for ENT0)
                  Broadcom ETH 0:1
                  IP 172.29.0.1
                  SM 255.255.0.0
                  OK
```

Reboot your machine and enable PCI devices in BIOS, the devices must now configure manually.

```
>Init 6
```

After the machine reboots, log back in as *root* and configure the PCI Network Devices using the Neat utility. See example below.

```
New (will be external connection)
Select devxxxx
Automatically obtain IP address setting
Automatically obtain DNS info from provider
Forward
Apply

ACTIVATE each device

>Service Network Restart
```

Create your */etc/hosts* file

>vi /etc/hosts

Note: This file is provided on the Cluster 1350 configuration disks from manufacturing, which can be found in your ship group. Make sure all devices are entered, such as terminal servers, switches and hardware management devices.

The following is an example of the */etc/hosts* for the example cluster:

Note: It is recommended to insert the fully qualified domain name before the short name.

```
#  Localhost
127.0.0.1                localhost.localdomain localhost
########## Management Node ###################
# cluster interface (eth0) GigE
172.20.0.1  mgmt1.mydomain.com       mgmt1
# management interface (eth1)
172.30.0.1  mgmt2.mydomain.com       mgmt2
# external interface (eth2)
10.0.0.1     external.mydomain.com         external
########## Management Equipment ##############
# RSA adapters. You might have ASMA cards instead
172.30.30.1    rsa001.mydomain.com       rsa001
172.30.30.2    rsa002.mydomain.com       rsa002
172.30.30.3    rsa003.mydomain.com       rsa003
172.30.30.4    rsa004.mydomain.com       rsa004
# Terminal Servers
172.17.2.1    ts01.mydomain.com        ts01
172.17.2.2    ts02.mydomain.com        ts02
# APC Master Switch
172.17.3.1    apc1.mydomain.com        apc01
# Myrinet Switch's Ethernet management port
172.17.4.1    myri01.mydomain.com       myri01
# Ethernet Switch
```

```
172.17.5.1    Ethernet01mydomain.com  Ethernet01c
172.16.5.1    Ethernet01.mydomain.com   Ethernet01
########## Compute Nodes #####################
172.20.101.1    node01.mydomain.com       node01
172.30.10.1    node01-myri0.mydomain.com node01-myri0
172.20.101.2    node02.mydomain.com       node02
172.30.10.2    node02-myri0.mydomain.com node02-myri0
172.20.101.3    node03.mydomain.com       node03
172.30.10.3    node03-myri0.mydomain.com node03-myri0
172.20.101.4    node04.mydomain.com       node04
172.30.10.4    node04-myri0.mydomain.com node04-myri0
172.20.101.5    node05.mydomain.com       node05
172.30.10..5    node05-myri0.mydomain.com node05-myri0
172.20.101.6    node06.mydomain.com       node06
172.30.10.6    node06-myri0.mydomain.com node06-myri0
172.20.101.7    node07.mydomain.com       node07
172.30.10.7    node07-myri0.mydomain.com node07-myri0
172.20.101.8    node08.mydomain.com       node08
….
```

Verify the management node’s network setup by:

1. Pinging all network interfaces (refer to the manufacturing defaults for verification.)
2. Pinging other devices on all of the subnets, including the cluster, management, and any external devices.
3. Pinging and route through your gateway.

8. Installing xCAT

Follow the steps below to install xCAT on the management node.

1. Download the latest version of xCAT to the */opt* directory if you have not already done so, as described in section 4.
2. Unpack xCAT in to */opt/*.

```
cd /opt
tar -xzvf xcat-dist-core.tgz
tar -xzvf xcat-dist-doc.tgz
tar -xzvf xcat-dist-ibm.tgz
tar -xzvf xcat-dist-oss.tgz
```

3. Use the following commands to setup xCAT.

   ```
   >export XCATROOT=/opt/xcat
   >cd $XCATROOT/sbin
   >./setupxcat
   ```

Enable time services (xntpd) on management node

```
>mv -f /etc/ntp.conf /etc/ntp.conf.ORIG
```

Create a new */etc/ntp.conf*:

```
>server 127.127.1.0
>fudge 127.127.1.0 stratum 10
>driftfile /etc/ntp/drift
```

Set time, date, and time zone with setup:

```
>chkconfig  - - level 2345 ntpd on
>service ntpd restart
```

Note: It can take a few minutes before *ntpd* is working.

```
>ntpdate -q localhost
```

If NTP is working you should see output similar to the following:

```
server 127.0.0.1, stratum 2, offset -0.000002, delay 0.02570
22 Jan 08:04:24 ntpdate[14540]: adjust time server 127.0.0.1 offset -0.000002 sec
```

If not working you will receive the following output:

```
no server suitable for synchronization found
```

Add the xCAT Man Pages to *$MANPATH* by adding the following line to */etc/man.config*:
>MANPATH /opt/xcat/man

Test the man pages by entering:
> man site.tab

9. Setup xCAT

This section describes some of the xCAT configuration necessary for the 32 node example cluster. If configuring a cluster that differs from this example, you may have to change some settings. xCAT configuration files are located in */opt/xcat/etc*. You must setup these configuration files before proceeding.

Copy the configuration files to their required location.

Note: If this is an IBM Cluster 1350 the configuration files can be found in the ship group. Only copy the samples if the configuration files are not available.

```
> mkdir /install
> cp /opt/xcat/samples/etc/* /opt/xcat/etc
```

Create a custom configuration by editing */opt/xcat/etc/** to work with the cluster. Read the man pages "*man site.tab*", to learn more about the format of these configuration files. More detailed information on some of these files can be found in some of the later sections. The following are examples that work with the example 32 node cluster.

To find documented examples of the *.tab* files, go to the */opt/xcat/samples/etc* directory.

Note: If you have installed Java you may use the xTablePad or xTableWizard table generators in the */opt/bcat/lib* directory to generate your tab configuration files. You may also use this application on a Windows machine but if the files are edited on the Windows machine, the formatting may be wrong.

Required tables:
site.tab
nodehm.tab
nodelist.tab
nodepos.tab
noderes.tab
nodetype.tab
passwd.tab
postscripts.tab
postdeps.tab
snmptrapd.conf
networks.tab
mac.tab (loaded with non-collectable MACs, such as terminal servers, switches, and RSAs.)
mp.tab

mpa.tab

Required tables for clusters with terminal servers or SOL (Server Over LAN):
conserver.tab
conserver.cf

Required tables for clusters using Ethernet switches to collect MAC addresses (use the correct table for your switch):
cisco.tab
summit48i.tab
blackdiamond.tab
switch.tab

Required tables for clusters using IPMI management:
ipmi.tab

Required table for APC Master Switch:
apc.tab

Required table for APC Master Switch Plus:
apcp.tab

Required table for xCAT flash support:
nodemodel.tab

Required table for EMP support:
emp.tab

Required table for Baytech support:
baytech.tab

Required table for xCAT GPFS support:
gpfs.tab

Table for IPMI support. Required for systems having a different IPMI IP address than node address (for example, the IBM e325):
ipmi.tab

site.tab

```
# /opt/xcat/etc/site.tab
# site.tab control most of xCAT's global settings.
# man site.tab for information on what each field means.
# this example uses 'c' as a subdomain private to the cluster and
# 10.0.0.1 as the corp DNS server (forwarder).
rsh                 /usr/bin/ssh
rcp                 /usr/bin/scp
gkhfile             /opt/xcat/etc/gkh
tftpdir             /tftpboot
```

```
tftpxcatroot          xcat
# modify domain to match your domain name
domain                mydomain.com
dnssearch             mydomain.com
```

# nameserver - Comma delimited list of DNS name servers IP addresses, use your #management node IP address. (172.16.n.100)

```
nameservers           192.16.100.1
forwarders            10.0.0.1
```

# nets - Comma delimited list of DNS network and netmask pairs colon delimited or #NA. Required only if this cluster contains a primary DNS server. This list determines what */etc/hosts* entries are used to create the primary DNS server files.

```
nets
        172.16.0.0:255.255.0.0,172.17.0.0:255.255.0.0,172.18.0.0:255.255.0.0
dnsdir                /var/named/chroot
dnschroot             yes
```

#dnsallow - Comma delimited list of DNS network and netmask pairs colon #delimited or NA. Required only if this cluster contains a primary or secondary DNS server. This list determines the access permissions for primary and secondary DNS servers contained within this cluster.

```
dnsallow      172.16.0.0:255.255.0.0,172.17.0.0:255.255.0.0,172.18.0.0:255.255.0.0
```

#domainaliasip - IP address aliased to cluster DNS domain name or NA. Required only if this cluster contains a primary DNS server. Use your management IP address

```
domainaliasip         172.16.100.1
```

#mxhosts - Comma delimited list of FQDN mail exchange hosts for this cluster or #NA. Required only if this cluster contains a primary DNS server.   Each node will be assigned *mxhosts* as the MX records for that host.

```
mxhosts                       mydomain.com,man-mydomain.com
```

#mailhosts - Comma delimited list of mail hosts aliases for this cluster or NA. Required only if this cluster contains a primary DNS server. Each host listed in *#mailhosts* will be aliased as *mailhost* for the purpose of providing the cluster with a single host name for all mail.

```
mailhosts             man-c
```

```
#master - Master host/node name
master                man-c
```

#homefs - Default global home file system.

```
homefs                man-c:/home
```

#localfs - Default global local file system.

```
localfs                 man-c:/usr/local

pbshome                 /var/spool/pbs
pbsprefix               /usr/local/pbs



#Pbsserver - Name of the node which is running the PBS server.
pbsserver               man-c
scheduler               maui
xcatprefix              /opt/xcat
keyboard                us

#timezone – Current Linux time zone.
timezone                US/Eastern

#offutc - UTC offset.
offutc                  -5
mapperhost              NA

#serialmac - What serial port to use to collect MAC addresses.
serialmac               0
serialbps               9600
snmpc                   public

#snmpd - The IP address to collect SNMP traps.
snmpd                   172.17.100.1
poweralerts             Y

#timeservers - Comma delimited list of IP addresses for nodes to sync their clocks.
timeservers             man-c
logdays                 7
installdir              /install
clustername             Clever-cluster-name

#dhcpver - set this to 3 since we are using DHCP version 3
dhcpver                 2
dhcpconf                /etc/dhcpd.conf

#dynamicr - This is the range of IP addresses assigned for node discovery.  Comment this
out by placing “#” at the beginning of the line
#dynamicr               eth0,ia32,172.30.0.1,255.255.0.0,172.30.1.1,172.30.254.254

#usernodes - A comma delimited list of nodes users are allow to login to.
usernodes               man-c
```

```
#usermaster - The single node that users accounted are added to.
usermaster              man-c

#nisdomain and nismaster.  Set to NA, NIS is beyond the scope of this class.
nisdomain               NA
nismaster               NA

nisslaves               NA
homelinks               NA
chagemin                0
chagemax                60
chagewarn               10
chageinactive           0
mpcliroot               /opt/xcat/lib/mpcli
#End of site.tab
```

nodelist.tab

/opt/xcat/etc/nodelist.tab

```
# nodelist.tab contains a list of nodes and defines groups that can be used in commands.
# Use man nodelist.tab for more information.
node01 all,rack1,compute,myri,mpn1
node02 all,rack1,compute,myri,mpn1
node03 all,rack1,compute,myri,mpn1
node04 all,rack1,compute,myri,mpn1
node05 all,rack1,compute,myri,mpn1
node06 all,rack1,compute,myri,mpn1
node07 all,rack1,compute,myri,mpn1
node08 all,rack1,compute,myri,mpn1
….
node31 all,rack1,compute,myri,mpn4
node32 all,rack1,compute,myri,mpn4
rsa01   nan,mpa
rsa02   nan,mpa
rsa03   nan,mpa
rsa04   nan,mpa
ts01    nan,ts
ts02    nan,ts
myri01  nan
```

mpa.tab

/opt/xcat/etc/mpa.tab

```
#service processor adapter management
#
```

```
#type      = asma,rsa
#name      = internal name (must be unique)
#          internal name should = node name
#          if rsa/asma is primary management
#          processor
#number    = internal number (must be unique and > 10000)
#command   = telnet,mpcli
#reset     = http(ASMA only),mpcli,NA
#dhcp      = Y/N(RSA only)
#gateway   = default gateway or NA (for DHCP assigned)
#
rsa01   rsa,rsa01,10001,mpcli,mpcli,NA,N,NA
rsa02   rsa,rsa02,10002,mpcli,mpcli,NA,N,NA
rsa03   rsa,rsa03,10003,mpcli,mpcli,NA,N,NA
rsa04   rsa,rsa04,10004,mpcli,mpcli,NA,N,NA
```

mp.tab

```
/opt/xcat/etc/mp.tab
# mp.tab defines how the Service processor network is setup.
# node07 is accessed via the name 'node07' on the RSA 'rsa01', etc.
# man asma.tab for more information until the man page to mp.tab is ready
node01 rsa01,node01
node02 rsa01,node02
node03 rsa01,node03
node04 rsa01,node04
node05 rsa01,node05
node06 rsa01,node06
node07 rsa01,node07
node08 rsa01,node08
...
node32 rsa04,node32
```

apc.tab

/opt/xcat/etc/apc.tab

```
# apc.tab defines the relationship between nodes and APC
# MasterSwitches and their assigned outlets. In our example,
# the power for asma1 is plugged into the 1st outlet the
# APC MasterSwitch, etc.
rsa01   apc1,1
rsa02   apc1,2
rsa03   apc1,3
rsa04   apc1,4
ts01    apc1,5
ts02    apc1,6
myri01  apc1,7
```

conserver.cf

/opt/xcat/etc/conserver.cf

```
# conserver.cf defines how serial consoles are accessed.  Our example
# uses the ELS terminal servers and node01 is connected to port 1
# on ts01, node02 is connected to port 2 on ts01, node17 is connected to
# port 1 on ts02, etc.
# man conserver.cf for more information
#
# The character '&' in logfile names are substituted with the console
# name.  Any logfile name that does not begin with a '/' has LOGDIR
# prepended to it.  So, most consoles will just have a '&' as the logfile
# name which causes /var/consoles/ to be used.
#
LOGDIR=/var/log/consoles
#
# list of consoles we serve
#    name : tty[@host] : baud[parity] : logfile : mark-interval[m|h|d]
#    name : !host : port : logfile : mark-interval[m|h|d]
#    name : |command : : logfile : mark-interval[m|h|d]
#
node01:!ts01:3001:&:
node02:!ts01:3002:&:
node03:!ts01:3003:&:
node04:!ts01:3004:&:
node05:!ts01:3005:&:
node06:!ts01:3006:&:
node07:!ts01:3007:&:
node08:!ts01:3008:&:
...
```

```
node32:!ts02:3016:&:
%%
#
# list of clients we allow
# {trusted|allowed|rejected} : machines
#
trusted: 127.0.0.1
```

### *conserver.tab*

/opt/xcat/etc/conserver.tab

# *conserver.tab* defines the relationship between nodes and conserver servers. Our example uses only one conserver on the localhost. Use *man conserver.tab* for more information.

```
node01 localhost,node01
node02 localhost,node02
node03 localhost,node03
node04 localhost,node04
node05 localhost,node05
node06 localhost,node06
node07 localhost,node07
node08 localhost,node08
...
```

nodehm.tab

/opt/xcat/etc/nodehm.tab

```
#
#node hardware management
#
#power     = mp,baytech,emp,apc,apcp,NA
#reset     = mp,apc,apcp,NA
#cad       = mp,NA
#vitals    = mp,NA
#inv       = mp,NA
#cons      = conserver,tty,rtel,NA
#bioscons  = rcons,mp,NA
#eventlogs = mp,NA
#getmacs   = rcons,cisco3500
#netboot   = pxe,eb,ks62,elilo,file:,NA
#eth0      = eepro100,pcnet32,e100,bcm5700
#gcons     = vnc,NA
#serialbios   = Y,N,NA
#
```

```
#node
          power,reset,cad,vitals,inv,cons,bioscons,eventlogs,getmacs,netboot,eth0,gcons,ser
ialbios
#
node01  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node02  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node03  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node04  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node05  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node06  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node07  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node08  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
...
node32  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
rsa01   apc,apc,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
rsa02   apc,apc,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
rsa03   apc,apc,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
rsa04   apc,apc,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
ts01    apc,apc,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
ts02    apc,apc,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
myri01  apc,apc,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
```

noderes.tab

```
/opt/xcat/etc/noderes.tab
#
#TFTP         = Where is my TFTP server?
#              Used by makedhcp to setup /etc/dhcpd.conf
#              Used by mkks to setup update flag location
#NFS_INSTALL  = Where do I get my files?
#INSTALL_DIR  = From what directory?
#SERIAL       = Serial console port (0, 1, or NA)
#USENIS       = Use NIS to authenticate (Y or N)
#INSTALL_ROLL = Am I also an installation server? (Y or N)
#ACCT         = Turn on BSD accounting
#GM           = Load GM module (Y or N)
#PBS          = Enable PBS (Y or N)
#ACCESS       = access.conf support
#GPFS         = Install GPFS
#INSTALL NIC  = eth0, eth1, ... or NA
#
#node/group
          TFTP,NFS_INSTALL,INSTALL_DIR,SERIAL,USENIS,INSTALL_ROLL,ACCT,G
M,PBS,ACCESS,GPFS,INSTALL_NIC
#
compute man-c,man-c,/install,0,N,N,N,Y,Y,Y,N,eth0
```

```
nan     man-c,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA
```

nodetype.tab

nodetype.tab maps nodes to types of installs. The example below only uses one type. For more information, *man nodetype*.

Note: *nodetype.tab* can not contain comments.

/opt/xcat/etc/nodetype.tab

```
node01 compute73
node02 compute73
node03 compute73
node04 compute73
node05 compute73
node06 compute73
node07 compute73
node08 compute73
...
```

Continue until the last node is entered.

passwd.tab

The file *passwd.tab* defines some passwords that will be used in the cluster, *man passwd.tab* for more information

/opt/xcat/etc/passwd.tab

```
cisco         cisco
rootpw        netfinity
asmauser      USERID
asmapass      PASSW0RD
```

ipmi.tab

```
node001     bmc001,"",""
node002     bmc002,"",""
node003     bmc003,"",""
node004     bmc004,"",""
```

Continue until the last node is entered.
Check tabs by running *rpower* and *rbeacon* commands.

10. Configuring the terminal servers

For Cyclades AlterPath ACS see http://www.cyclades.com for configuration instructions.

This section describes setting up ELS and ESP terminal servers and conserver. If the cluster has either ELSes or ESPs, skip the instructions for the terminal server type. Terminal servers enable out-of-band administration and access to the compute nodes, for example watching a compute node's console remotely before the compute node can be assigned an IP address or after the network configuration is lost.

Notes:

1. The hardware associated with the Cluster 1350 serial network will be preconfigured from manufacturing.
2. Please see the last section in this document for Cluster 1350 orders not containing terminal servers.

Learn about conserver

Conserver's website: http://conserver.com/.

Shutdown conserver

Before setting up the terminal servers, make sure that the conserver service is stopped:

>service conserver stop

Setup terminal servers

This section describes how to configure the Equinox ELS terminal server.

conserver.cf setup

Modify */opt/xcat/etc/conserver.cf*

Each node has an entry similar to:

nodeXXX:!tsx:yyyy:&:

Where:
x = Terminal Server Unit number and
yyyy = Terminal Server port + 3000 e.g. node1:!ts1:3001:
& = access node1 via telnet to ts1 on port 3001. 'node1' should be connected to ts1's first serial port.
conserver.cf setup If neede.

Modify */opt/xcat/etc/conserver.cf*

Each node gets a line like:

```
node001:!ts001:7001:&:                (Cyclades)
node002:!ts001:7002:&:
```

Start Conserver

```
> service conserver start
```

Test if Conserver and terminal servers are working.

```
wcons –t <node range>
```

11. Configuring xCAT

This section covers configuring xCAT on the cluster.

A restart of xCAT is required after the *.tab* files are installed

Use the following commands to setup xCAT:

```
>export XCATROOT=/opt/xcat
>cd $XCATROOT/sbin
>./setupxcat
```

If not done during the OS install, edit */etc/selinux/config*.
SELINUX=disabled

Build a DNS server:

```
>makedns master
```

Check DNS with:

```
>host mgt
```

The DNS should return the IP for mgt, *172.20.0.1*.

Enter non-collectable MACs, such as terminal servers, switches, and RSA adapters in *$XCATROOT/etc/mac.tab*.

Notes:

1. Some network devices, such as the APC Master Switch, do not have the MAC address affixed to the unit. Some devices may have the MAC printed on a piece

of paper in the manual. Before installing a device in to the rack, verify the MAC address will be visible when racked. Some network devices have a serial port that may be used to obtain the MAC.

2. Manual non-collectable MAC entries in *mac.tab* do not require a *-eth0* appended, it is optional.

12. DHCP setup and configuration

This section covers installing and configuring DHCP on the cluster.

Collect the MAC addresses of the cluster equipment and place each MAC address that requires DHCP for an IP address into */opt/xcat/etc/<MANAGEMENT_NET>.tab*. See the man page for *macnet.tab*.

Note: If using APC master switches, include their MAC addresses into this file.

Make the Initial *dhcpd.conf* configuration file.

> makedhcp -new

Edit *dhcpd.conf* and check for anything out of the ordinary.

> vi /etc/dhcpd.conf

Use the example below to verify the contents of */etc/dhcp.conf*

```
#xCAT 1.2.0

authoritative;
ddns-update-style none;

option option-128 code 128 = string;
option option-150 code 150 = string;
option option-160 code 160 = string;
option option-192 code 192 = string;
option option-193 code 193 = string;
option option-194 code 194 = string;
option option-195 code 195 = string;

shared-network eth0 {

      filename                "/tftpboot/pxelinux.0";
      subnet 172.20.0.0 netmask 255.255.0.0 {
            max-lease-time                43200;
            default-lease-time            43200;
            option routers                172.20.0.1;
            option subnet-mask            255.255.0.0;
            option nis-domain             "cluster.com";
            option domain-name            "cluster.com";
            option domain-name-servers    172.20.0.1;
            option time-offset            -7;
            range                   172.20.200.1 172.20.255.254;

      } #172.20.0.0/255.255.0.0 subnet_end#

      subnet 172.29.0.0 netmask 255.255.0.0 {
            max-lease-time                43200;
            default-lease-time            43200;
            option routers                172.29.0.1;
            option subnet-mask            255.255.0.0;
            option nis-domain             "cluster.com";
            option domain-name            "cluster.com";
            option domain-name-servers    172.29.0.1;
            option time-offset            -7;

      } #172.29.0.0/255.255.0.0 subnet_end#

} #eth0 network_end#

shared-network eth1 {

      subnet 172.30.0.0 netmask 255.255.0.0 {
            max-lease-time                43200;
            default-lease-time            43200;
            option routers                172.30.0.1;
            option subnet-mask            255.255.0.0;
            option nis-domain       "cluster.com";
            option domain-name            "cluster.com";
            option domain-name-servers    172.30.0.1;
            option time-offset            -7;

      } #172.30.0.0/255.255.0.0 subnet_end#

} #eth1 network_end#

#shared-network all {

#} #all network_end#
```

Notes:

1. After using *getmacs* and then *makedhcp –allmacs*, an entry for each MAC address for each node will be listed in the *dhcp.conf*.
2. Usually DHCP should not run on the network interface that is connected to the rest of the network. In this case, remove the network section from *dhcpd.conf* that corresponds to the external network and then explicitly list the interfaces DHCPD should listen for in */etc/dhcpd.conf*.

Edit */etc/sysconfig/dhcpd*, with something similar to:

DHCPDARGS="eth0 eth1"

Notes:

1. The *dhcpver* field in *$XCATROOT/etc/site.tab* must be set to match the version of dhcpd installed. Generally 2 for older Red Hat and 3 for SUSE Linux and newer Red Hat before you run *makedhcp*. If incorrect, correct and rerun *makedhcp --new --allmac*.
2. *$XCATROOT/etc/networks.tab* must define each network that dhcpd is to support. Let *makedhcp* build it for you the first time, edit and rerun *makedhcp --new --allmac*.

Configure all Ethernet switches, but block DHCP from in and out bound ports that are used to connect the cluster to a production environment. Please read the "xCAT 1.1.0 Redbook", the "cisco2950-HOWTO", and the "force10-HOWTO" found in */opt/xcat/doc* for more information.

Configure all terminal servers. Please read the "terminalserver-HOWTO".

Restart *conserver* if you are using terminal servers or SOL. If you are using an IBM BladeCenter without SOL do not use conserver.

```
>service conserver restart
```

Setup stage boot image.

For x86 and x86_64 type:

```
>cd /opt/xcat/stage
>./mkstage
```

For ia64 type:

```
>cd /opt/xcat/stage
>./mkstage-ia64
```

Collect the MAC addresses of the compute nodes and create entries in *dhcpd.conf* for them.

Prepare to monitor stage2 progress

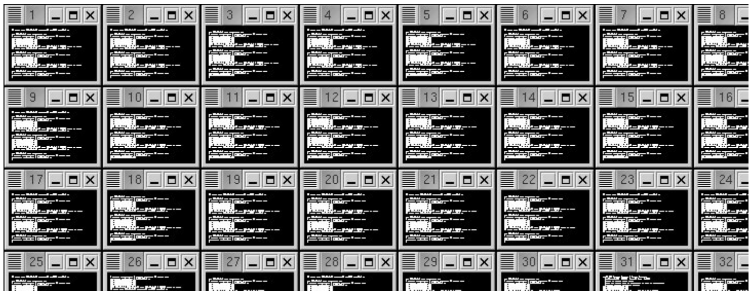
```
> wcons -t 8 compute (or a subset like rack01)
> tail -f /var/log/messages
```

Be aware of system messages, it is a very good way to stay informed about the cluster.

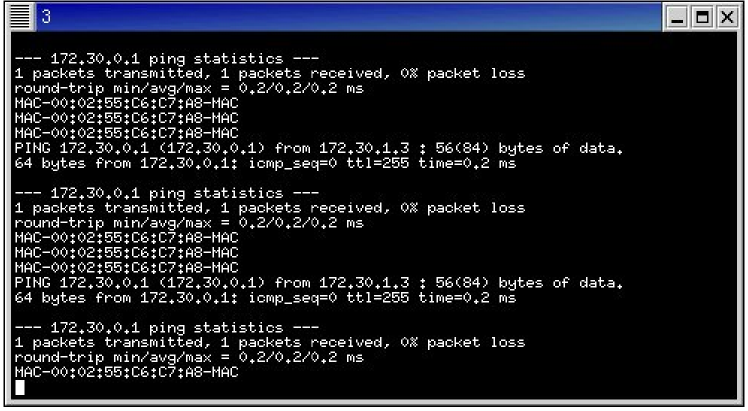
Manually reboot the compute nodes.

During the boot process, the machines should PXE boot syslinux, obtain a dynamic IP address, and then load a Linux kernel and a special RAM disk that contains a script to print the machine's MAC address to the console.

Observe the output of the wcons windows. If the terminal servers are working correctly, the machines boot their kernels and display an image similar to the one below:

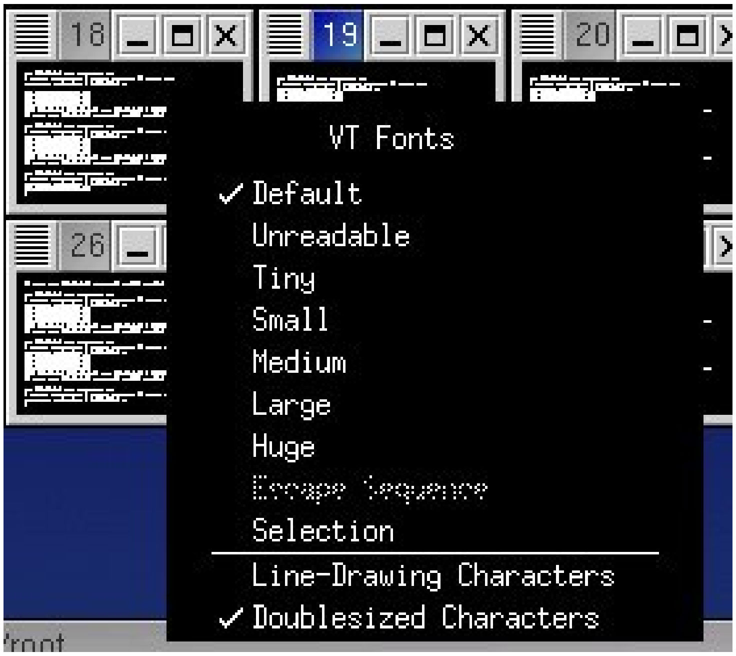


A closeup:

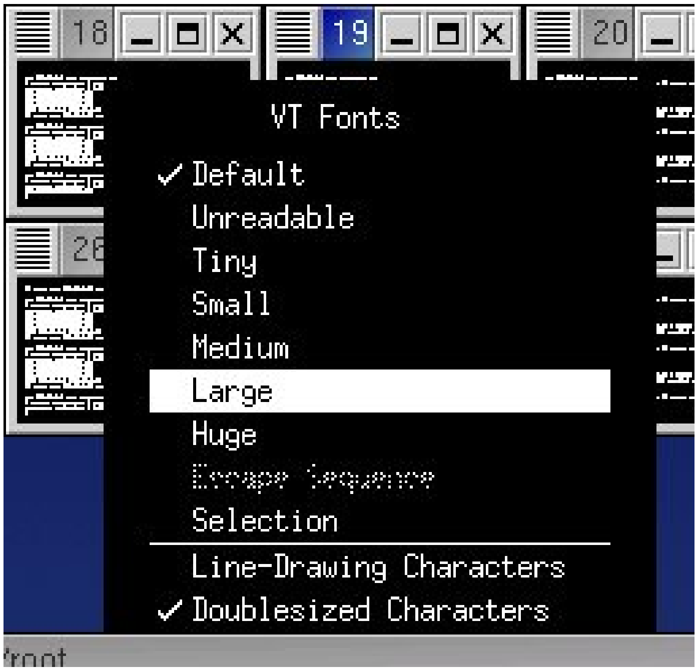


```
--- 172.30.0.1 ping statistics ---
1 packets transmitted, 1 packets received, 0% packet loss
round-trip min/avg/max = 0.2/0.2/0.2 ms
MAC-00:02:55:C6:C7:A8-MAC
MAC-00:02:55:C6:C7:A8-MAC
MAC-00:02:55:C6:C7:A8-MAC
PING 172.30.0.1 (172.30.0.1) from 172.30.1.3 : 56(84) bytes of data.
64 bytes from 172.30.0.1: icmp_seq=0 ttl=255 time=0.2 ms

--- 172.30.0.1 ping statistics ---
1 packets transmitted, 1 packets received, 0% packet loss
round-trip min/avg/max = 0.2/0.2/0.2 ms
MAC-00:02:55:C6:C7:A8-MAC
MAC-00:02:55:C6:C7:A8-MAC
MAC-00:02:55:C6:C7:A8-MAC
PING 172.30.0.1 (172.30.0.1) from 172.30.1.3 : 56(84) bytes of data.
64 bytes from 172.30.0.1: icmp_seq=0 ttl=255 time=0.2 ms

--- 172.30.0.1 ping statistics ---
1 packets transmitted, 1 packets received, 0% packet loss
round-trip min/avg/max = 0.2/0.2/0.2 ms
MAC-00:02:55:C6:C7:A8-MAC
```

The wcons windows are xterms. When viewing a large number of consoles on the screen at the same time, the xterms come up with the “unreadable” font size. Xterms have a feature that allows a user to change the size of the font very easily. This allows the user to enlarge a specific view when a screen of unreadable consoles is displayed. To do this, move the mouse over the text portion of the xterm in question, press and hold **Ctrl+Right click** the mouse. The following menu will be displayed:



Move the mouse down to select a larger font and then release the mouse button as shown:



Using this xterm feature, you can switch to a large font for detailed viewing and back to the smaller font to view all the consoles at once.

Press **Ctrl+E** to access additional terminal functionality.

Use the following command to collect the MACs once all the compute nodes are displaying their MAC addresses out of their serial consoles.

```
> getmacs compute
```

Note: This *.tab* file should be in the configuration files that came with the IBM 1350 cluster.

Use the following command to close the wcons windows.

```
> wkill
```

Manually reboot each node and use the following command to collect MAC addresses:

```
>getmacs <noderange>
          or
>getmacs compute

node1-eth0 00:07:E9:93:F8:DD
node1-eth1 00:00:5A:9A:DB:7C
node2-eth0 00:07:E9:93:F8:DD
node2-eth1 00:00:5A:9A:DB:7C
```

When the message "*Auto merge mac.lst with /opt/xcat/etc/mac.tab(y/n)?*" appears, type **Y**.

Each node will be suffixed with the interface of the collected MAC. Please do not alter.

Notes

1. Do not alter the *mac.tab* entries for collected MACs. It is critical that the stored node names remain untouched. If necessary, you may change the MAC.
2. Multiple *getmacs* commands will corrupt *mac.tab*. Only run one instance at a time.
3. Some operating systems report eth0 and eth1 differently than xCAT *getmac* reports. The settings may need to be reversed manually in *mac.tab*, however this may have a negative impact on other non-switched entries. Verify your settings making corrections.

```
perl -pi -e 's/(nodeprefix.*)-eth0/$1-ethfoo/' mac.tab
perl -pi -e 's/(nodeprefix.*)-eth1/$1-eth0/' mac.tab
perl -pi -e 's/(nodeprefix.*)-ethfoo/$1-eth1/' mac.tab
```

Note: Currently only the serial-based (rcons) method of connecting MACs will collect multiple MAC/node. A future version of xCAT will address this limitation.

Exception: IBM BladeCenter mpcli2 and bcmm getmacs methods can collect both MAC addresses.

Note: For IBM BladeCenter please use bcmm method in *nodehm.tab*.

MAC addresses may be collected without a terminal server.
Configure *cisco3500.tab* with an example of the following:
node01 Ethernet01,1
node02 Ethernet01,2
node03 Ethernet01,3
node04 Ethernet01,4

Make *nodehm.tab* have entries like:
nodexx mp,mp,mp,mp,mp,conserver,mp,mp,rcons,cisco3500,bcm5700,vnc

Make sure the switch has a hostname and DNS resolves.
Verify that the nodes plugged into the switch ports match those in *cisco3500.tab.*

Example: *node1 port1 node2 port2*

Make sure you can ping the switch, telnet to it and login. Make sure the password you set on the switch is the same in *passwd.tab*. Put the nodes in
stage2. Power them on and *getmacs* as usual. The *getmacs* command issues the show mac-address-table on the switch and grabs the MACs from it.

For other switches, *switch.tab*, *getmacs.switch.snmp*, and *getmacs.switch* are required. Place *getmacs.switch.snmp* and *getmacs.switch* into the *opt/xcat/lib* directory and make sure they are executable by using *ls –l* to verify.

Place *switch.tab* into the *opt/xcat/etc* directory:

For example SMC alters the *switch.tab* as follows: (see examples in *switch.tab* for SMC and other switches. This will be the future way of setting up switches.)

```
nodexxx        smc8648-001,18,NA
   |                     |      |
   |                     |     smc port number
   |                   smc name-switch number (as named in your other tab files & hosts)
  Node
```

Edit the *nodehm.tab*.

Here is an example of one that is set up for using RCONS as a method for *getmacs* (not necessarily the way yours will look but just an example of how the *nodehm.tab* file may appear):

node1   mp,mp,mp,mp,mp,conserver,mp,mp,rcons,pxe,eepro100,vnc,Y,NA,NA,def

Here is an example of using new *getmacs*:
Edit the appropriate entry to point to the switch scripts (this will be what tells *getmacs* to use *getmacs.switch* script).

node1   mp,mp,mp,mp,mp,conserver,mp,mp,switch,pxe,eepro100,vnc,Y,NA,NA,def

Build */etc/dhcpd.conf* with MAC entries:

makedhcp –allmac

For all IBM xSeries nodes with IBM management processors and the IBM eServer 325 and 326, excluding IBM BladeCenter, read "managementprocessor-HOWTO" found in */opt/xcat/doc* for more information. For IBM BladeCenter, use *mpname noderange*.

nodeset noderange stage3

Reboot each node manually after all MACs collected and DHCP server restarted.

Read the "managementprocessor-HOWTO" and "IBM BladeCenter-NOTES" for information on testing and troubleshooting all nodes management processors if applicable.

Test systems management:

rpower noderange stat

rbeacon noderange on

Note: Not all servers have a blinking light.

Copy the Red Hat Install CD(s) by inserting the CDs and then run *copycds* and follow the prompts.

Example: *copycds <namecd1>.iso, <namecd2>.iso, <namecd3>.iso*

Notes:

1. When the CDs are entered and a prompt for "auto run" appears, select **No**.
2. You may also use *copycds* to copy the contents of one or more *.iso* files.

Copy the "post" files for Red Hat.
Copy install files from the xCAT distribution to the post directory that is used during unattended installs:

```
>cp /opt/xcat/samples/etc/post* /opt/xcat/etc
>cp /opt/xcat/install/rhas4/x86_64/base/compute.tmpl ..
```

Enter the following commands to enable remote logging:

```
> cp /opt/xcat/samples/syslog.conf /etc
> touch /var/log/pipemessages
> service syslog restart
```

Setup snmptrapd

snmptrapd received messages from the SPN.

```
> chkconfig snmptrapd on
> service snmptrapd start
```

The following command creates a SSH keypair for root with an empty passphrase which sets up root's SSH configuration, copying *keypair* and *config* to */install/post/.ssh* so that all installed nodes will have the same root *keypair/config*. This allows you to install and log into nodes.

```
>gensshkeys root
```

Setup NFS and NFS Exports by making */etc/exports* look similar to the following:

```
/install node*(ro,sync,no_root_squash)
/tftpboot node*(ro,sync,no_root_squash)
/usr/local node*(rosync,no_root_squash)
/opt/xcat node*(ro,sync,no_root_squash)
/home node*(rw,sync,no_root_squash)
```

Turn on NFS by using the following commands:

```
> chkconfig nfs on
> service nfs start
> exportfs -ar #         (to source)
> exportfs #             (to verify)
>echo "/install *(ro,async,no_root_squash)" >>/etc/exports
>service nfs restart
```

Notes:

1. If you do not have a Myrinet read the “myrinet-how to” document in */opt/xcat/doc*. For more detailed information, read the “nodeinstall-HOWTO” and “systemimager-HOWTO” for details on node install and diskless installs.

2. If installing a node from disk, use *rinstall* or *winstall*. Only install 32 nodes at a time or use staging. For more information, read the “man” pages on *rinstall* and *winstall*.

13. Installing compute nodes

This section covers the installation of the compute nodes.

Modify the “kickstart” template file if needed and verify the correct version of RedHat is listed.

```
>cd /opt/xcat/install/rhws4/<architecture>/base/
>cp /opt/xcat/install/rhws4/<architecture>/base/compute.tmpl ..
```

The following command makes the nodes PXE boot the RedHat “kickstart” image by altering the files in */tftpboot/pxelinux.cfg/*.

```
> nodeset compute install
```

Prepare to Monitor the Installation Progress

```
> wcons -t 8 compute
```

Note: A subset like rack01 may be substituted.

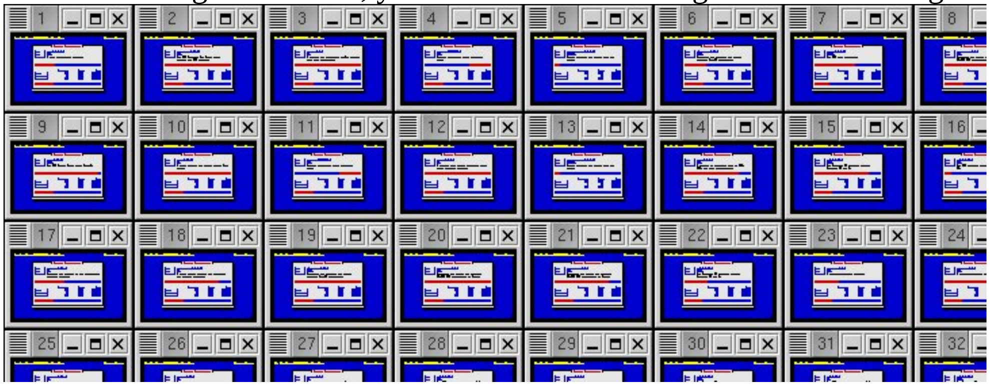
```
> tail -f /var/log/messages
```

Note: Be aware of any warning messages that may appear.

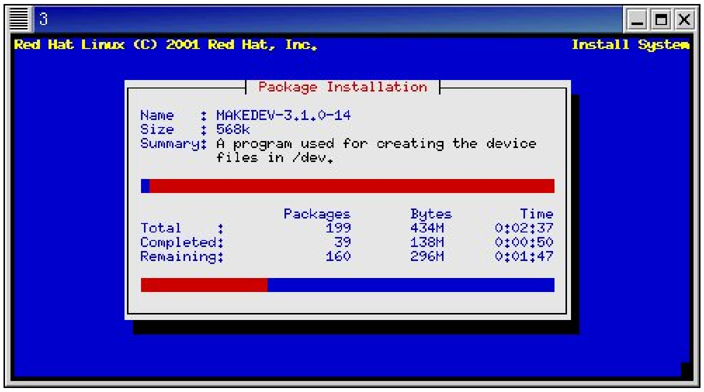
Reboot the Compute Nodes.

```
>rpower compute boot
```

When installing with *wcons*, you should see something like the following:



Close up:



```
3
Red Hat Linux (C) 2001 Red Hat, Inc.                              Install System

                    ┤ Package Installation ├

          Name    : MAKEDEV-3.1.0-14
          Size    : 568k
          Summary: A program used for creating the device
                   files in /dev.

                           Packages      Bytes       Time
          Total    :            199       434M    0:02:37
          Completed:             39       138M    0:00:50
          Remaining:            160       296M    0:01:47
```

Note: To install without terminal servers, Serial Over LAN must be configured.

14. Serial Over LAN (SOL) setup

Download and install “SMbridge RPM” from http://www.ibm.com/support/docview.wss?uid=psg1MIGR-57729.

This RPM needs to be installed on a management node (or the “crash cart” that has xCAT on it) since it is client software for the BMCs.

For IBM xSeries 336, 346, and eServer 326

Flash the Management Processor (BMC) to the latest version.
Flash BIOS to the latest version.
Remove the power cord for 10 seconds.
Restore the power cord.
Reboot and press **F1** to enter the BIOS configuration.

Configure the BIOS settings for optimum performance and then edit the Devices and I/O Ports as shown below:

Devices and I/O Ports

Serial port A: **Port 3F8, IRQ 4**
Serial port B: **Disabled**
Remote Console **Redirection**

- Remote Console Active: **Enabled**
- Remote Console COM Port: **COM 1**
- Remote Console Baud Rate: **19200**

To use Remote Console Text Emulation: VT100/VT220
Configure the text emulation settings as listed below:

Remote Console Keyboard Emulation: **VT100/VT220**
Remote Console After Boot: **Enabled**
Remote Console Flow Control: **Hardware**

Configure the “Startup” settings as listed below:

From the main menu, select **Startup Sequence** and configure the settings as follows:

First Startup Device: **CD ROM**
Second Startup Device: **Diskette Drive 0**
Third Startup Device: **Network**
Forth Startup Device: **Hard Disk**
Wale On LAN: **Disabled**
Planer Ethernet PXE/DHCP: **Planer Ethernet 1**

Boot Fail Count: **Disabled**

Go to **Advanced Setup** then **CPU Options** and configure the settings as follows:

Hyper-Threading Technology: Disabled

For IBM xSeries 326

Configure the optimum BIOS settings for the IBM xSeries 326 and include the following settings.

From the main menu, select **Console Redirection** and configure the settings as follows:

Console Redirection: **COM A**
Baud rate: **19.2 K**
FIFO Level: **14**
Console Type: **vt100**
Flow Control: **CTS/RTS**
Console Connection: **Direct**
Continue CR After Post: **On**

From the main menu, select **BMC** and configure the settings as follows:

IPMI Spec Version: **1.5**
BMC Firmware Version: **1.11**
Com port on BMC: **CLI**
Change Com port Setting: **No**
Clean System Eventlog: **Disabled**
System Firmware Progress: **Enable**
BIOS Post Watchdog: **Enable**

Additional settings for xSeries 336, 346, and eServer 326

From the main menu, select **Advanced Setup** then **Baseboard Management Controller (BMC) Settings** and configure the settings as follows:

System BMC Serial Port Sharing: **Enabled**
BMC Serial Port Access Mode: **Dedicated**

Save the settings.

Note: If switching to SOL, you must remove the power cord for 5 sec.

Tabs

conserver.cf

Conserver.cf will have to be altered to point to the SOL script for the specific node.

Example:

```
node001:|sol.eServer 326 node001::&:
node002:|sol.xSeries 336 node002::&:
node003:|sol.xSeries 346 node003::&:
```

ipmi.tab

There are a few different ways of approaching this

```
node123         bmc123,"",""

node123         bmc123,
node123         bmc123,USERID,PASSW0RD
```

Note: There is a zero in "PASSW0RD".

If you use quotations (") then you will have to enter " " for the userid and password when you start your *wcons* session.
If you leave the field blank then the userid and password should default to the definitions in the *passwd.tab* file.
We have also used the third example and placed the default userid and password (*USERID,PASSW0RD*). Whatever you put in there will over ride the defaults and that is what you will have to enter on your wcons window for the node you intend to view.

nodehm.tab

Set up the *nodehm.tab* file to point to the *ipmi* tool (uses BMC).

Example:

```
node001
ipmi,ipmi,ipmi,ipmi,ipmi,conserver,NA,ipmi,switch,pxe,bcm5700,vnc,Y,ipmi,NA,19200
```

Note: The *ipmi* parameter in several of the fields. In this example we have also setup the baud rate for 19200. It has to match what is set in the **BIOS Setup** under **Remote Console settings.**

site.tab

RHEL 3.0 and below may cause a problem when using wcons to view the node console. The problem is that the console title will not show the node name and therefore confusion as to which node you are viewing may occur. To correct this you must turn off *bufferedcons* in the *site.tab* file, then the node name will display correctly in the title bar.

Example:

```
Bufferedcons no
```

WCONS

When you run *wcons <nodename>* , the screen will display:

```
connected....

login :

password :
```

Entry for login and password has to be the same as what is configured in the *ipmi.tab* file.

Example:

login : "" and *password : ""* is used as the user ID and password in the previous example of the *ipmi.tab*.

Note: You must have “smbridge RPM” installed.

Verify that the compute nodes installed correctly

```
>pping all
```

Update the SSH global known hosts file

```
> makesshgkh compute
```

15. Clean up

This section covers the final installation steps and test information..

Copy the xCAT initialization files. This will enable some services to start at boot time and change the behavior of some existing services.

```
> cd /opt/xcat/rc.d
> cp atftpd portmap snmptrapd syslog /etc/rc.d/init.d/
```

There are other initialization files in */opt/xcat/rc.d* that may also be used, depending on the installation.

Move unneeded *.tab* files from */opt/xcat/etc/* to a temporary directory.

Test the cluster. Read the man pages for *rvitals*, *rinv*, and *rpower* and then try some of these commands on the cluster.

```
>psh compute date | sort
```

The output here will be a good way to see if SSH/gkh is setup correctly on all of the compute nodes, which is a requirement for most cluster tasks. If a node does not appear here correctly, you must go back and troubleshoot the individual node. Make certain the install process completed correctly by using *makesshgkh* and then test again with *psh*. The *psh* test should pass before continuing.

Additional test commands:

```
> rvitals compute ambtemp
>mpncheck compute
>pping all
>rbeacon ccompute on
```

16. Contributing to xCAT

Join the “xCAT-dev” mailing list and post your suggestions, bug-fixes and code by visiting http://xcat.org/mailman/listinfo/xcat-user.

17. Credits

This document was most recently modified:
01/20/2007
Original author Matt Bohnsack

Send additions and corrections to either Mark Weber (jmweb@us.ibm.com) or **Srihari Angaluri (**sangalu@us.ibm.com), so this document can continue to be improved.

Thanks go out to the following people, who helped this document become what it is today:
Egan Ford for writing xCAT, Jarrod B Johnson, Mike Galicki, Andrew Wray, Chris DeYoung, Mark Atkinson, Greg Kettmann, Jay Urbanski, The people from POSDATA, Kevin Rudd, Tom Alandt, and Tonko L De Rooy for there continuing support and dedication to the development of xCAT,

18. Supporting documentation

Additional documentation may be located in */opt/xcat/doc*.

License
xCAT Support

xCAT Redbooks

xCAT Man Pages

OSS Licenses (Incomplete, WIP)

HOWTOs:

- xCAT Mini HOWTO (1.2.0) (Start Here)
- xCAT HOWTO (1.1.0) (Reference Only)

Hardware HOWTOs:

- Blade Center NOTES (1.1.7.2 and 1.2.0)
- Management Processor HOWTO (1.2.0)
- Stage1 HOWTO (1.2.0)

Switch/Terminal Server HOWTOs:

- Cisco 2950 HOWTO (1.2.0)
- Force 10 HOWTO (1.2.0)
- Myrinet-HOWTO (1.2.0)
- Terminal Server HOWTO (1.2.0)

Management Node HOWTOs:

- SUSE Linux Management Node HOWTO (1.2.0)

Node Install HOWTOs:

- Node Installation HOWTO (1.2.0)
- Imaging HOWTO (1.2.0)
- SystemImager HOWTO (1.2.0)
- Remote Flash HOWTO (1.2.0)
- Windows HOWTO (1.1.0)
- Diskless HOWTO (1.2.0)
- Warewulf HOWTO (1.2.0)

Software HOWTOs:

HPC Benchmark HOWTO (1.2.0)
GPFS HOWTO (1.1.0)

For more information, visit http://www.xcat.org.