Installing RedHat Enterprise Linux 3 on Cluster 1350 using xCAT

October 2005
Mark Weber

Installing RedHat Enterprise Linux 3 on Cluster 1350 using xCAT

Send additions and corrections to **ShaddGa@us.ibm.com.**.

Always check for newer software versions and be aware of the stability of these newer versions.

1. Introduction

xCAT is a collection of mostly script based tools to build, configure, administer, and maintain Linux clusters.

This document describes how to implement Linux cluster on IBM xSeries hardware using xCAT and other third party software. It covers the latest version of xCAT - v1.1RC1.2.0 with

x86 (i386, i486, i586, i686) supported distributions for IBM e1350

Red Hat Enterprise Linux ES 3 U3*
Red Hat Enterprise Linux WS 3 U3*

x86_64 (Opteron and EMT64) supported distributions for IBM e1350

Red Hat Enterprise Linux ES 3 U3*
Red Hat Enterprise Linux WS 3 U3*

x86 (i386, i486, i586, i686) distributions supported by xCAT:

Red Hat 7.2
Red Hat 7.3
Red Hat 8.0
Red Hat 9

Red Hat Enterprise Linux AS 2.1
Red Hat Enterprise Linux AS 2.1 U2
Red Hat Enterprise Linux AS 2.1 U3
Red Hat Enterprise Linux ES 2.1*
Red Hat Enterprise Linux WS 2.1*

Red Hat Enterprise Linux AS 3
Red Hat Enterprise Linux ES 3*
Red Hat Enterprise Linux WS 3*
Red Hat Enterprise Linux AS 3 U1
Red Hat Enterprise Linux ES 3 U1*
Red Hat Enterprise Linux WS 3 U1*
Red Hat Enterprise Linux AS 3 U2
Red Hat Enterprise Linux ES 3 U2*

Red Hat Enterprise Linux WS 3 U2*
Red Hat Enterprise Linux AS 3 U3
Red Hat Enterprise Linux ES 3 U3*
Red Hat Enterprise Linux WS 3 U3*
Red Hat Enterprise Linux AS 3 U4*
Red Hat Enterprise Linux ES 3 U4*
Red Hat Enterprise Linux WS 3 U4*

Red Hat Enterprise Linux AS 4*
Red Hat Enterprise Linux ES 4*
Red Hat Enterprise Linux WS 4*

Red Hat Fedora Core 1*
Red Hat Fedora Core 2*
Red Hat Fedora Core 3*

CentOS 3.3 (Treat as RHAS3U3)
CentOS 3.4 (Treat as RHAS3U4) (CD and DVD)

SuSE 8.1*
SuSE 8.2*
SuSE 9.0*
SuSE 9.1*
SuSE 9.2* (DVD Version only, non DVD missing KSH, 32-bit EM64T & Opteron Tested)
SuSE SLES8
SuSE SLES8 SP1
SuSE SLES8 SP2a
SuSE SLES8 SP3
SuSE SLES9
SuSE SLES9 SP1
SystemImager
Partimage

x86_64 (Opteron and EMT64) distributions supported by xCAT:

Red Hat Enterprise Linux AS 3*
Red Hat Enterprise Linux ES 3*
Red Hat Enterprise Linux WS 3*
Red Hat Enterprise Linux AS 3 U1*
Red Hat Enterprise Linux WS 3 U1*
Red Hat Enterprise Linux AS 3 U2*
Red Hat Enterprise Linux ES 3 U2*
Red Hat Enterprise Linux WS 3 U2*
Red Hat Enterprise Linux AS 3 U3* (64-bit EM64T & Opteron Tested)

Red Hat Enterprise Linux ES 3 U3* (64-bit EM64T & Opteron Tested)
Red Hat Enterprise Linux WS 3 U3* (64-bit EM64T & Opteron Tested)
Red Hat Enterprise Linux AS 3 U4* (64-bit EM64T & Opteron Tested)
Red Hat Enterprise Linux ES 3 U4* (64-bit EM64T & Opteron Tested)
Red Hat Enterprise Linux WS 3 U4* (64-bit EM64T & Opteron Tested)
Red Hat Enterprise Linux AS 4*
Red Hat Enterprise Linux ES 4*
Red Hat Enterprise Linux WS 4*
Red Hat Fedora Core 1*
Red Hat Fedora Core 2*
Red Hat Fedora Core 3* (64-bit EM64T & Opteron Tested)

CentOS 3.3 (Treat as RHAS3U3) (64-bit EM64T & Opteron Tested)
CentOS 3.4 (Treat as RHAS3U4) (64-bit EM64T & Opteron Tested) (CD and DVD)

SuSE 9.0*
SuSE 9.1*
SuSE 9.2* (DVD Version only, 64-bit EM64T & Opteron Tested)
SuSE SLES8
SuSE SLES8 SP2
SuSE SLES8 SP3
SuSE SLES9 (64-bit EM64T & Opteron Tested)
SuSE SLES9 SP1 (64-bit EM64T & Opteron Tested)
SystemImager
Partimage

IA64 (Itanium 1 and 2) distributions supported by xCAT

Red Hat 7.2
Red Hat Enterprise Linux AS 2.1 U2*
Red Hat Enterprise Linux AS 3*
Red Hat Enterprise Linux ES 3*
Red Hat Enterprise Linux WS 3*
Red Hat Enterprise Linux AS 3 U1*
Red Hat Enterprise Linux WS 3 U1*
Red Hat Enterprise Linux AS 3 U2*
Red Hat Enterprise Linux ES 3 U2*
Red Hat Enterprise Linux WS 3 U2*
Red Hat Enterprise Linux AS 3 U3*
Red Hat Enterprise Linux ES 3 U3*
Red Hat Enterprise Linux WS 3 U3*
Red Hat Enterprise Linux AS 3 U4*
Red Hat Enterprise Linux ES 3 U4*

Red Hat Enterprise Linux WS 3 U4*

Red Hat Enterprise Linux AS 4*
Red Hat Enterprise Linux ES 4*
Red Hat Enterprise Linux WS 4*

SuSE SLES8
SuSE SLES8 SP2
SuSE SLES8 SP3

SuSE SLES9*
SuSE SLES9 SP1*

PPC64 (IBM JS20 only) distributions supported by xCAT:

Red Hat Enterprise Linux AS 3 U2*
Red Hat Enterprise Linux AS 3 U3*
Red Hat Enterprise Linux AS 3 U4*
Red Hat Enterprise Linux AS 4*
Red Hat Enterprise Linux ES 4*
Red Hat Enterprise Linux WS 4*

SuSE SLES8 SP3aa*
SuSE SLES9*
SuSE SLES9 SP1*

PPC64 Node install tested only, however should work as management node.

You will need to adjust the configuration examples shown in this document to suit your particular cluster and architecture, but the examples should give a good general idea of what needs to be done. Please don't use this document verbatim as an implementation guide. You should rather use it as an inspiration to your own implementation. Use the man pages, source and other documentation that is available to figure out why certain design/configuration choices are made and how you can make different choices. Because IBM e1350 Clusters are preconfigured from manufacturing this document covers only a very little of the hardware connectivity, cabling, etc. 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 . If you're serious about implementing a cluster and learning how things work, you should read the RedBook in addition to this document. http://www.redbooks.ibm.com/redbooks.nsf/Redbooks?SearchView&Query=linux+cluster&SearchMax=4999 .

2. Understanding What xCAT Does - Feature/Functionality Hierarchy

This section explains what you can do with xCAT, why xCAT is designed the way it is, and presents a feature/functionality hierarchy.

2.1 Understanding What Drives xCAT's Design and Architecture

xCAT's architecture and feature set have two major drivers:
Real world requirements - The features in xCAT are a result of the requirements met in hundreds of real cluster implementations. When users have had needs that xCAT or other cluster management solutions couldn't meet, xCAT has risen to the challenge. Over the last few years, this process has been repeatedly applied, 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 requirements and work with many cluster topologies and architectures.

Unmatched Linux clustering experience - The people involved with xCAT's development have used xCAT to implement many of the world's largest Linux clusters and a huge variety of different cluster types. The challenges faced during this work has resulted in features that enable xCAT to power all types of Linux clusters from the very small to the largest ever built.

2.2 Understanding What Types of Clusters xCAT is Good For

xCAT works well with the following cluster types:

HPC - High Performance Computing Physics, Seismic, CFD, FEA, Weather, and other simulations; Bioinformatics work
HS - Horizontal Scaling Web farms, etc.
Administrative Not a traditional cluster, but a very convenient platform to install and administer a number of Linux machines
Windows or other OSes 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 OS
Other xCAT's modular toolkit approach makes it easy to adjust for building any type of cluster.

2.3 Understanding xCAT's Features

A list of xCAT's current features follows:

1. OS/Distribution support Any OS on compute nodes via OS agnostic imaging support
2. Hardware Control Remote Power control (on/off/state) via IBM Management Processor Network, BMC and/or APC Master Switch
3. Hardware Control Remote software reset (rpower)
4. Hardware Control Remote Network BIOS/firmware update and configuration on IBM hardware

5. Hardware Control Remote OS console via pluggable support for a number terminal servers
6. Hardware Control Remote POST/BIOS console via IBM Management Processor Network and via terminal servers.
7. Boot Control Ability to remotely change boot type (network or local disk) with syslinux.
8. Automated installation Parallel install via scripted RedHat kickstart, SuSE autoyast, on ia32, x86_64, ppc, and ia64
9. Automated installation Parallel install via imaging with other Linux distributions, Widows, or other OSes
10. Automated installation Network installation with supported PXE NICs, via etherboot, or BootP,on supported NICs without PXE
11. Monitoring Hardware alerts and email notification with IBM's Management Processor Network and SNMP alerts
12. Monitoring Remote vitals (fan speed/temp/etc...) with IBM's Management Processor Network
13. Monitoring Remote hardware event logs with IBM's Management Processor Network/IPMI Interface
14. Administration Utilities Parallel remote shell, ping, rsync, and copy
15. Administration Utilities Remote hardware inventory with IBM's Management Processor Network
16. Software Stack PBS and Maui schedulers - Build scripts, documentation, automated setup, extra related utilities, and deep integration
17. Software Stack Myrinet - automated setup and installation
18. Software Stack MPI - Build scripts, documentation, automated setup for MPICH, MPICH-GM, and LAM
19. Usability Command line utilities for all cluster management functions
20. Usability Single operations can be applied in parallel to multiple nodes with a very flexible and customizable group/range functionality
21. Flexibility Support for various user defined node types
22. Diskless support via warewulf

3. Getting the xCAT Software Distribution

This section explains where and how you can get the xCAT software distribution.
3 of the 4 packages required can be located at
http://www.alphaworks.ibm.com/tech/xCAT/
The forth package can be located at
http://www-rcf.usc.edu/~garrick/xcat-dist-oss-1.2.0-RC1.tgz.

4. Getting Other Required Software

Firmware and Hardware Configuration Software

http://publib.boulder.ibm.com/cluster/1350Apr05.htm
NOTE: if you only have to flash one or two systems you can down load the image file to you management node then >dd if=/root/boot.img of=/dev/fd0 bs=10k count=144 To make a bootable

floppy

5. Reading Related Documentation

There's quite a bit of related documentation available. You should read it. It's all accessible in the /opt/xcat/doc folder once you have decompressed your xcat tar files.

6. Getting Help

If you need assistance with building, maintaining, or administering your xCAT cluster, or you have an xCAT feature request, try the xCAT-user mailing list or contact your IBM sales rep or other IBM point of contact.

7. Understanding Cluster Components and the Example Cluster's Architecture

This document uses a basic 32 node cluster that uses serial terminal servers for out-of-band console access, an APC Master Switch and IBM's Service Processor Network for remote hardware management, ethernet, and Myrinet as the basis of most of its examples. The following three examples describe some of the detail of this example cluster:

7.1 Components / Rack Layout

Here you see how the hardware is positioned in the rack. Starting from the bottom and moving towards the top, we have:

The Myrinet switch: Used for high-speed, low-latency inter-node communication. Your cluster may not have Myrinet, if you aren't running parallel jobs that do heavy message passing, or if it doesn't fit in your budget.

Nodes 1-16: The first 16 compute nodes. Note that every 8th node has an MPA (Managment Processor Adaptor) installed. You 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.

Monitor/Keyboard: You know what this is.

Terminal servers: The terminal enable serial consoles from all of the compute nodes to be accessible from the management node. You will find this feature very useful during system setup and after setup administration. SOL (Serial Over Lan) can be used to emulate a terminal server setup if the cluster does not have a terminal server.

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

APC master switch: This enables remote power control of devices that are not part of the Service Processor Network.. terminal servers, Myrinet switch, ASMA adapters, etc.

The management node: The management node is where we install the rest of the nodes from, manage the cluster, etc.

Nodes 17-32: The rest of the compute nodes.. again with Management Processor cards every 8th node.

Ethernet switch: Finally, at the top, we have the ethernet switch.

7.2 Networks

For IBM e1350 all network devices that must be statically set are pre configured as per the manufacturing defaults listed in the "IBM Manufacturing defaults for all items in e1350 Clusters" table.

Here you see the networks that are used in this document's examples. Note the listing of attached devices to the right. Important things to note are:
The external network is the organization's main network. In this example, only the management node has connectivity to the external network.

The ethernet switch hosts both the cluster and management network on separate VLANs.

The cluster network connects the management node to the compute nodes. We use a private class B network that has no connectivity to the external network. This is often the easiest way to do things and a good thing to do if you think your cluster might grow to more than 254 nodes. You may have a requirement to place the compute nodes on a network that is part of your external network.

The management network is a separate network used to connect all devices associated with cluster management... terminal servers, BMC, ASMA cards, etc. to the management node.

Parallel jobs use the message passing network for interprocess communication. Our example uses a separate private class B network over Myrinet. If you are not using Myrinet, this network could be the same as the cluster network. i.e. You could do any required message passing over the cluster network.

7.2.1 IBM Manufacturing defaults for all items in e1350 Clusters.

e1350 Linux Cluster Manufacturing Defaults (5A)(userids/passwords/IP addresses)

Hostnames and IP Addressing Scheme

This table shows the network addressing / hostnames used to identify the various e1350 cluster components.

IP Address	Hostname	Component
172.20.0.1	mgt.cluster.com	management node eth0 (cluster vlan)
172.30.0.1	mgt1.cluster.com	management node eth1(management vlan)
172.29.0.1	bmc	mgt node alias eth0:1 (cluster vlan)
172.29.101.1	bmc001	e325/e326/x336/x346 node bmcs
172.20.1.1	storage001	storage nodes
172.30.2.1	triton001	FAStT storage controllers
172.20.101.1	node001	compute nodes
172.40.101.1	hca001	HCA cards ib0 (cluster vlan)
172.20.4.1	user001	usernodes
172.30.10.1	myri001 ib001	Myrinet or InfiniBand TopSpin TS120 Voltaire 9024
172.30.20.1	ts001	Terminal Servers (MRV LX-32, LX-48)
172.30.30.1	rsa001	RSA cards

172.30.50.1	cisco3508-001 smc8624-001 smc8648-001 cisco3750-001	Cisco 3500, 3700 Series or SMC switches
172.30.60.1	apc001	APC
172.30.70.1	rcm001	RCM
172.30.80.1	cisco6503-001 Cisco6509-001 Force 10	Cisco 4000 series or 6500 series switches or Force 10
172.30.101.1	sm001/mm001	bladecenter switch/management modules

Compute Node IP Addressing: Example:.rack 1, node 1 = 172.20.101.1

	Network	Rack Number	Node
Rack1	172.20	101	1 to 84
Rack2	172.20	102	1 to 84

Node numbering increases from bottom of rack upward and from left to right for bladecenters.

e325 BMC IP Addressing: Example:.rack 1, node 1 = 172.29.101.1

	Network	Rack Number	Node
Rack1	172.29	101	1 to 40
Rack2	172.29	102	1 to 40

Bladecenter Switch/Management Module IP Addressing

Network	BC Number	Bay Location Number
172.30	101	1 to 4 (SM bay 1 - SM bay 4) or, 5 (MM ext) or, 6(MM int)
172.30	102	1 to 4 (SM bay 1 - SM bay 4) or, 5 (MM ext) or, 6(MM int)

Examples

172.30.104.3 is the switch module fitted to bay 3 in bladecenter 4
172.30.106.5 is the external port (eth0) for the management module in bladecenter 6
172.30.106.6 is the internal port (eth1) for the management module in bladecenter 6

InfiniBand HCA IP Addressing: Example:.rack 1, hca 1 = 172.40.101.1

	Network	Rack Number	HCA
Rack1	172.40	101	1 to 84
Rack2	172.40	102	1 to 84

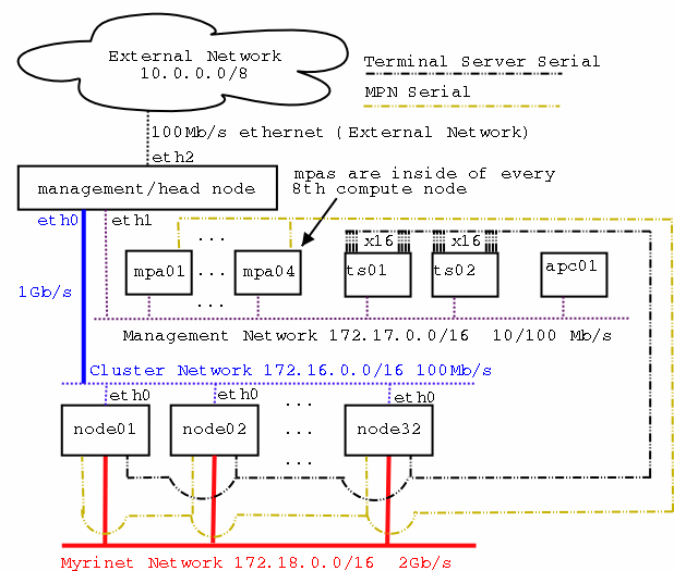
HCA numbering increases from bottom of rack upward and from left to right for bladecenters.

Manufacturing default userids and passwords.

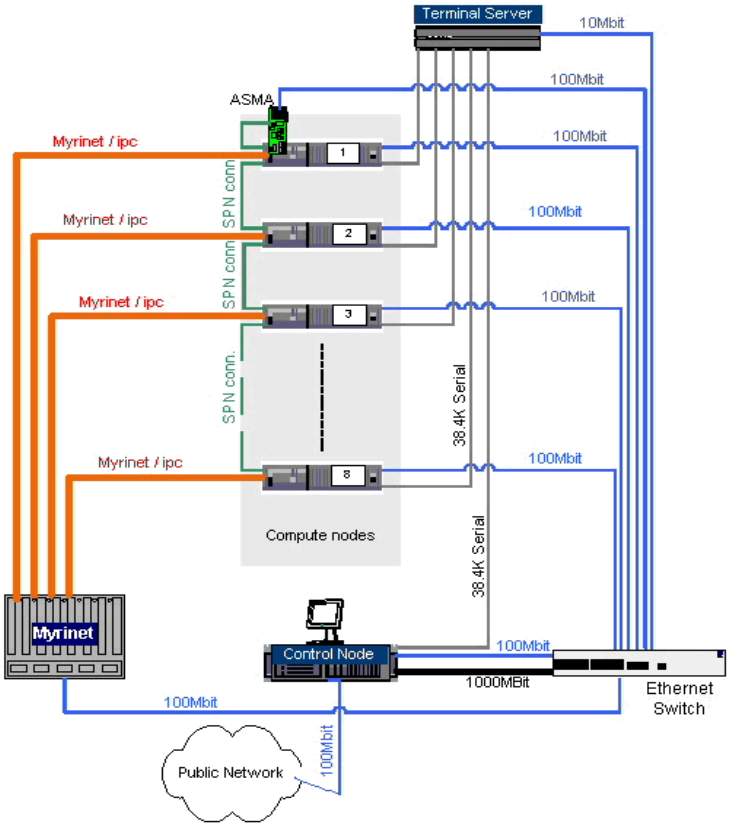
Hardware:	USERID	Password	Comments
APC Switch	apc	apc	defaults
ITouch Terminal	access	system	defaults
InReach Terminal	InReach	access	defaults
Cisco Switch	(see comment)	ibm1350	just press Enter for userid
SMC Switch	admin	admin	defaults
FAStT	(see comment)	infiniti	just press Enter for userid
RSAs	USERID	PASSW0RD	(note the 0 in passw0rd is a zero)
TopSpin	super	super	defaults
Voltaire	(none)	123456	defaults
Force 10	(none)	(none)	defaults

Operating System:	USERID	Password
management server	root	ibm1350
nodes	admin	cluster
	root	cluster

7.3 Connections



7.3.1 Another Connections Diagram



7.4 Other Architecture Notes

Other notes about this architecture (and areas where yours may differ and you may need to make adjustments to this document's examples):
The compute nodes have no access to the external network.
The compute nodes get DNS,DHCP, and NIS services from the management node.
NIS is used to distribute username/passwd information to the compute nodes and the management node is the NIS master.
The management node is the only node with access to the management network.
PBS and Maui are used to schedule/run jobs on the cluster.
Users can only access compute nodes when the scheduler has allocated nodes to them and then only with ssh.
Jobs will use MPICH or LAM for message passing.

8. Configuring the Ethernet Switch

This section describes configuring the ethernet switch. The examples art based on the Cisco 3750 the commands associated with the hardware you have selected in your e1350/ Cluster may vary slightly from this documentation. please consult the documentation that came in your ship group for details on your specific switch.

NOTE: All switches associated with IBM e1350 should be set up from manufacturing with default settings. See 7.2

8.1 Setup VLANs and Configure Ethernet Switches

If you have separate subnets for the management and compute networks, like in our example, you need to setup VLANs on the ethernet switches. Experience has shown that many strange problems are solved with the introduction of VLANS. Use VLANs to separate the ports associated with the management, and cluster subnets. A set of somewhat random notes for configuring VLANs on different switches and setting up "spanning-tree portfast" on ciscos is available here.

8.2 Connecting to the Switch and Setting IP Address and Password

Cisco:
Connect the management node's COM1 to the switch's console port and...
> cu -l /dev/ttyS0 -s 9600

8.3 Login in and enable: See manufacturing default sec. 7.2 for Username and Password

Assign an IP to the default VLAN:
cisco> conf t
cisco> int vlan1
cisco> ip address 172.17.5.1 255.255.0.0
cisco> exit

Allow telnet access and set telnet password to 'cisco':
cisco> conf t
cisco> line tty 0 15
cisco> login password cisco cisco> exit

Set enable and console passwords:
Set enable password:
cisco> conf t
cisco> enable password cisco
cisco> exit

Set console password:
cisco> conf t
cisco> line vty 0 4
cisco> password cisco
cisco> exit

Extreme Networks:
> config vlan Default ipaddress 172.16.5.1 255.255.0.0

8.4 Changing Port Settings

Cisco:
Setup 'spanning-tree portfast':
Without this option, DHCP may fail because it takes too long for a port to come online after a machine powers up. Do not set spanning-tree portfast on ports that will connect to other switches. Do the following on each port on your switch:
cisco> conf t
cisco> int Fa/1
cisco> spanning-tree portfast
cisco> exit
cisco> int Fa/2
cisco> spanning-tree portfast
cisco> exit
etc.

8.5 Setting up Remote Logging

Here we have the switch send all its logging information to the management node's management interface. (We'll enable remote sysloging on the management node later).
Cisco:
cisco> conf t
cisco> logging 172.17.5.1
cisco> exit

Extreme Networks:
> config syslog 172.17.5.1

8.6 Setting up VLANs

Cisco:

for each interface you want to put in a VLAN Fa0/1 .. Fa0/32, and gig ports { # clearly pseudo-code
  cisco> interface Fa0/1
  cisco> switchport mode access
  cisco> switchport access vlan 2
  cisco> exit
}

Extreme Networks:

> create vlan man
> config vlan man tag 2
> config vlan man ipaddress 172.17.5.1 255.255.0.0
> config Default delete port 1,2,3,4 # the ports you want in the managemnet VLAN
> config man add port 1,2,3,4
> show vlan

8.7 Notes on VLANs with Multiple Switches

Cisco:

 cisco> configure terminal
 cisco> interface Gi0/1
 cisco> switchport mode trunk
 cisco> switchport trunk encapsulation isl # you should prob use the standard encap instead
 cisco> exit

Extreme Networks:

 unconfigure switch
 config Default delete port ports,you,don't,want,in,management,VLAN

 create vlan cluster
 config vlan cluster tag 2
 config cluster add port ports,you,want,in,cluster,VLAN

 show vlan
 save

8.8 Saving your changes

You want to make certain your switch configuration is saved in case the switch is rebooted.

Cisco:

cisco> write mem

Extreme Networks:

extreme> save

9. Installing the OS on the Management Node

NOTE: Before you install to prevent confusion disable all PCI adapters in BIOS.

The first step in building an xCAT cluster is installing Linux on the management node. This is, how to do just that:
x346: http://www-307.ibm.com/pc/support/site.wss/document.do?lndocid=MIGR-57208
x336: http://www-307.ibm.com/pc/support/site.wss/document.do?sitestyle=ibm&lndocid=MIGR-57734
e326: http://www-307.ibm.com/pc/support/site.wss/document.do?lndocid=MIGR-57381

NOTE: Before you install to prevent confusion disable all PCI adapters in BIOS.

NOTE: Your management node may require specific drivers please consult the machine specific setup for instructions.

NOTE: if you find you need detailed setup instructions or troubleshooting assistance on individual IBM servers e326, x336, or x346 please see the above links.

9.1 Create and Configure RAID Devices if Necessary

If you are using LSI/HostRaid/ServeRAID devices in the management node, use the LSI/HostRaid/ServeRAID flash/config CD to update the LSI/HostRaid/ServeRAID firmware to v4.84 and define you RAID volumes. If you have other nodes with hardware RAID, you might as well update and configure them now as well. You can get this CD from http://www.pc.ibm.com/qtechinfo/MIGR-495PES.html.

9.2 NIS Notes

If you plan on interacting with an external NIS server, check if it supports MD5 passwords and shadow passwords. If it doesn't support these modern features, don't turn them on during the install of the management node. I'm not absolutely certain on this point, but it's bitten me hard in the past, so be careful.

9.3 Partition Notes

A good minimum drive partitioning scheme for the management node follows.
/boot (200 MB)
SWAP (1.5 x physical memory not to exceed 2GB)
/ (the rest of the disk)

9.4 Firewall

Select no firewall

Note Default password is ibm1350 as

9.5 Install

Select custom installation. When asked for packages to install choose **EVERYTHING**.

NOTE: If this the first time you have installed RedHat everything is a check box at the end of the selection.

9.6 User

Its a good idea to create a normal user other than root during the install. I usually make an 'ibm' user.

9.7 Bring Up the Newly Installed System

Reboot and login as root.
Open a terminal
>updatedb

9.8 Turn Off Services We Don't Want (General)

You probably want to turn off some of the network services that are turned on by default during installation for security and other reasons...

To view installed services:
> chkconfig --list | grep ':on'

To turn off a service:
> chkconfig --level 0123456 <service> off

9.9 Turn Off Services We Don't Want (Specific)

The following are examples of exactly what services to turn off for a system that works with our example architecture and will have nothing running that isn't necessary:

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

NOTE: This can also be done with the GUI by typing in setup at the prompt and toggling to system services.

9.10 Erase LAM Package

You probably want to remove the RedHat LAM package. It can easily get in the way of the MPI software we install later on, because it's an old version and installs itself in /usr/bin:

>rpm --erase lam

NOTE: Your config may require LAM just make sure you have the latest package built and installed to avoid and problems.

10. Configuring Networking on the Management Node

This section describes network setup on the **management node**.

10.1 e1000/bcm5700 notes

You will want to download the latest **e1000/bcm5700** driver from http://publib.boulder.ibm.com/cluster/1350Apr05.htm website and build it into your kernel. The **e1000/bcm5700** driver supplied with RedHat (tg3) can perform poorly and in some cases will not work in a VLANed environment

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

10.2 Configure Network Adapters

Edit /etc/modules.conf, /etc/sysconfig/network-scripts/*, and /etc/sysconfig/network, to create a network configuration that reflects the cluster's design. The following samples work with the example cluster:

>vi /etc/modules.conf
(change tg3 entries to)
alias eth0 bcm5700
alias eth1 bcm5700
alias eth2 bcm5700

>cd /etc/sysconfig/networking/profiles/default/
(The ifconfig scripts should look like this)

```
# Broadcom Corporation|NetXtreme BCM5721 Gigabit Ethernet PCI Express
DEVICE=eth0
BOOTPROTO=none
HWADDR=<MAC ADDRESS HERE>
ONBOOT=yes
TYPE=Ethernet
NETMASK=255.255.0.0
```

```
IPADDR=172.20.0.1
USERCTL=yes
PEERDNS=yes

# Please read /usr/share/doc/initscripts-*/sysconfig.txt
# for the documentation of these parameters.
# This is the Alias for BMC management network
TYPE=Ethernet
IPADDR=172.29.0.1
DEVICE=eth0:1
HWADDR=<MAC ADDRESS HERE>
BOOTPROTO=none
NETMASK=255.255.0.0
ONBOOT=yes
USERCTL=yes
PEERDNS=yes
```
ONBOOT="yes"

```
# Broadcom Corporation|NetXtreme BCM5721 Gigabit Ethernet PCI Express
DEVICE=eth1
BOOTPROTO=none
HWADDR=<MAC ADDRESS HERE>
ONBOOT=yes
TYPE=Ethernet
NETMASK=255.255.0.0
IPADDR=172.30.0.1
USERCTL=yes
PEERDNS=yes
```

NOTE: Reboot your machine and re enable PCI devices in BIOS, devices must now configure manually.

10.3 /etc/hosts

Edit /etc/sysconfig/network and add/edit:
HOSTNAME=mgt1

>Init 6
After the Mavhine reboots log back in as root

Create your /etc/hosts file.
>vi /etc/hosts
NOTE: this file is provided on the e1350 configuration disks from manufacturing that can be found in your ship group.
Make sure all devices are entered... terminal servers, switches, hardware management devices, etc.
The following is an sample of the /etc/hosts for the example cluster:

Note: It is a good idea 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
172.30.10.8 node08-myri0.mydomain.com node08-myri0
172.20.101.9 node09.mydomain.com node09
172.30.10.9 node09-myri0.mydomain.com node09-myri0
172.20.101.10 node10.mydomain.com node10
172.30.10.10 node10-myri0.mydomain.com node10-myri0
172.20.101.11 node11.mydomain.com node11
172.30.10.11 node11-myri0.mydomain.com node11-myri0

172.20.101.12 node12.mydomain.com node12
172.30.10.12 node12-myri0.mydomain.com node12-myri0
172.20.101.13 node13.mydomain.com node13
172.30.10.13 node13-myri0.mydomain.com node13-myri0
172.20.101.14 node14.mydomain.com node14
172.30.10.14 node14-myri0.mydomain.com node14-myri0
172.20.101.15 node15.mydomain.com node15
172.30.10.15 node15-myri0.mydomain.com node15-myri0
172.20.101.16 node16.mydomain.com node16
172.30.10.16 node16-myri0.mydomain.com node16-myri0
172.20.101.17 node17.mydomain.com node17
172.30.10.17 node17-myri0.mydomain.com node17-myri0
172.20.101.18 node18.mydomain.com node18
172.30.10.18 node18-myri0.mydomain.com node18-myri0
172.20.101.19 node19.mydomain.com node19
172.30.10.19 node19-myri0.mydomain.com node19-myri0
172.20.101.20 node20.mydomain.com node20
172.30.10.20 node20-myri0.mydomain.com node20-myri0
172.20.101.21 node21.mydomain.com node21
172.30.10.21 node21-myri0.mydomain.com node21-myri0
172.20.101.22 node22.mydomain.com node22
172.30.10..22 node22-myri0.mydomain.com node22-myri0
172.20.101.23 node23.mydomain.com node23
172.30.10..23 node23-myri0.mydomain.com node23-myri0
172.20.101.24 node24.mydomain.com node24
172.30.10.24 node24-myri0.mydomain.com node24-myri0
172.20.101.25 node25.mydomain.com node25
172.30.10.25 node25-myri0.mydomain.com node25-myri0
172.20.101.26 node26.mydomain.com node26
172.30.10.26 node26-myri0.mydomain.com node26-myri0
172.20.101.27 node27.mydomain.com node27
172.30.10.27 node27-myri0.mydomain.com node27-myri0
172.20.101.28 node28.mydomain.com node28
172.30.10.28 node28-myri0.mydomain.com node27-myri0

10.4 Verify Management Node Network Setup

You can ping all of the network interfaces (See manufacturing defaults)
You can ping other devices on all of the subnets (cluster, management, external, etc.)
You can ping and route through your gateway

NOTE: For IBM e1350 you should stick with the recommended versions. Updates or workarounds can be found at http://publib.boulder.ibm.com/cluster

11. Installing xCAT

Installing xCAT on the management node is very straight forward.

11.1 Download the Latest Version of xCAT to /opt/xcat

3 of the 4 packages required can be located at www.xcat.org http://www.alphaworks.ibm.com/tech/xCAT/ the forth can be located at http://www-rcf.usc.edu/~garrick/ . The latest verion of xCAT is 1.2.0.

11.2 Unpack xCAT Into /opt/

Copy xcat tgz files to /opt dir
> cd /opt
> tar -xzvf xcat-dist-core-RCx.x.x.tgz
> tar -xzvf xcat-dist-doc-RCx.x.x.tgz
> tar -xzvf xcat-dist-ibm-RCx.x.x.tgz
> tar -xzvf xcat-dist-oss-RCx.x.x.tgz

11.3 Install xCAT

Set up Java if needed

NOTE: A few words about Java:
For some ASMA, RSA, RSA2, and Bladecenter functions xCAT uses IBM's *mpcli* and *mpcli2* utilities (included in the xcat-dist-ibm tarball). Both utilities require Java. The Java included with both tools only work with older RH x86 distributions. SuSE includes a functioning Java for all four xCAT supported architectures (x86, x86_64, IA64, and PPC64) and has been tested. However, RH does not provide a functioning Java. If you wish to install or use a different Java, just install and create a link to `$XCATROOT/java/$ARCH`, where `$ARCH = x86`, `x86_64`, `ia64`, or `ppc64`. E.g.:

Install IBM Java in /usr/ibm/java
>cd /opt/xcat/java
>ln -s /usr/ibm/java x86
>ls -l
>total 1
lrwxrwxrwx 1 root root 13 Jul 21 18:19 x86 -> /usr/ibm/java
b
Some good Java for x86, x86_64, and PPC64:
https://www6.software.ibm.com/dl/lxdk/lxdk-p

Java for IA64? If you find a good one let us know. SuSE includes it. The x86 versions will run on IA64 natively--slow--but works OK for systems management

11.1 Setup xCAT

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

>date MMDDhhmmYY
(i.e) >date 0717050505
Will set the time to 5:05 AM July 17 2005

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 345 ntpd on
>service ntpd start

Test **(NOTE: it can take a few minutes before ntpd is working)**, type:

>ntpdate -q localhost

If working you should receive the following output:

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 (try again later or fix):

no server suitable for synchronization found

NOTE: setupxcat must actually be run after xCAT .tab files are setup later on.

11.2Add xCAT Man Pages to $MANPATH and test out xCAT man pages

Add the following line to /etc/man.config:
>MANPATH /opt/xcat/man

Test out the man pages:
> man site.tab

12. Configuring xCAT

This section describes some of the xCAT configuration necessary for the 32 node example cluster. If your cluster differs from this example, you'll have to make changes. xCAT configuration files are located in /opt/xcat/etc. You must setup these configuration files before proceeding.

12.1 Copy the Config Files to Their Required Location

NOTE: If this is an IBM e1350 cluster you will find your config files in the ship group. Only copy the samples if you don't already have your config files.

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

12.2 Create Your Own Custom Configuration

Edit /opt/xcat/etc/* to suit your cluster. Please read the man pages 'man site.tab', etc., to learn more about the format of these configuration files. There is a bit more detail on some of these files in some of the later sections. The following are examples that will work with our example 32 node cluster...

To find documented examples of the tab files go to /opt/xcat/samples/etc if writing your own.

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 your windows machine but if you edit the files on your windows machine the formatting may be wrong.

You will also need the post install scripts these come standard in /opt/xcat/samples/etc//post. Just copy them to your /opt/xcat/etc dir.

>cp /opt/xcat/samples/etc/post* /opt/xcat/etc/

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, e.g. terminal servers, switches, RSAs, etc...)
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`

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 that have a different IPMI IP address than node address (e.g. e325):
`ipmi.tab`

13. Tab examples

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 #determins 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

#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 determins 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 - Your Linux Timezone. Use: US/Eastern
timezone US/Central

#offutc - Your UTC offset. Use: -5
offutc -6

mapperhost NA

#serialmac - What serial port to use to collect MAC addresses. Because we are #using Blade Center
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 address assigned for node discovery. Comment this out with at # in front 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
mpciroot /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. 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
node09 all,rack1,compute,myri,mpn2
node10 all,rack1,compute,myri,mpn2
node11 all,rack1,compute,myri,mpn2
node12 all,rack1,compute,myri,mpn2
node13 all,rack1,compute,myri,mpn2
node14 all,rack1,compute,myri,mpn2
node15 all,rack1,compute,myri,mpn2
node16 all,rack1,compute,myri,mpn2
node17 all,rack1,compute,myri,mpn3
node18 all,rack1,compute,myri,mpn3
node19 all,rack1,compute,myri,mpn3
node20 all,rack1,compute,myri,mpn3
node21 all,rack1,compute,myri,mpn3
node22 all,rack1,compute,myri,mpn3
node23 all,rack1,compute,myri,mpn3
node24 all,rack1,compute,myri,mpn3
node25 all,rack1,compute,myri,mpn4
node26 all,rack1,compute,myri,mpn4

```
node27 all,rack1,compute,myri,mpn4
node28 all,rack1,compute,myri,mpn4
node29 all,rack1,compute,myri,mpn4
node30 all,rack1,compute,myri,mpn4
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
```

```
node08rsa01,node08
node09rsa02,node09
node10rsa02,node10
node11rsa02,node11
node12rsa02,node12
node13rsa02,node13
node14rsa02,node14
node15rsa02,node15
node16rsa02,node16
node17rsa03,node17
node18rsa03,node18
node19rsa03,node19
node20rsa03,node20
node21rsa03,node21
node22rsa03,node22
node23rsa03,node23
node24rsa03,node24
node25rsa04,node25
node26rsa04,node26
node27rsa04,node27
node28rsa04,node28
node29rsa04,node29
node30rsa04,node30
node31rsa04,node31
node32rsa04,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 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 doesn't 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:&:
node09:!ts01:3009:&:
node10:!ts01:3010:&:
node11:!ts01:3011:&:
node12:!ts01:3012:&:
node13:!ts01:3013:&:
node14:!ts01:3014:&:
node15:!ts01:3015:&:
node16:!ts01:3016:&:
node17:!ts02:3001:&:
node18:!ts02:3002:&:
node19:!ts02:3003:&:
node20:!ts02:3004:&:
node21:!ts02:3005:&:
node22:!ts02:3006:&:
node23:!ts02:3007:&:
node24:!ts02:3008:&:
node25:!ts02:3009:&:
node26:!ts02:3010:&:
node27:!ts02:3011:&:
node28:!ts02:3012:&:
node29:!ts02:3013:&:
node30:!ts02:3014:&:
node31:!ts02:3015:&:
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. man conserver.tab for more information.
node01localhost,node01
node02localhost,node02
node03localhost,node03
node04localhost,node04
node05localhost,node05
node06localhost,node06
node07localhost,node07
node08localhost,node08
node09localhost,node09
node10localhost,node10
node11localhost,node11
node12localhost,node12
node13localhost,node13
node14localhost,node14
node15localhost,node15
node16localhost,node16
node17localhost,node17
node18localhost,node18
node19localhost,node19
node20localhost,node20
node21localhost,node21
node22localhost,node22
node23localhost,node23
node24localhost,node24
node25localhost,node25
node26localhost,node26
node27localhost,node27
node28localhost,node28
node29localhost,node29
node30localhost,node30
node31localhost,node31
node32localhost,node32

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
node09  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node10  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node11  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node12  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node13  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node14  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node15  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node16  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node17  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node18  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node19  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node20  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node21  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node22  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node23  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
```

```
node24  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node25  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node26  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node27  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node28  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node29  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node30  mp,mp,mp,mp,mp,conserver,rcons,mp,rcons,pxe,eepro100,vnc,N
node31  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 authencate (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,AC
CT,GM,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

```
/opt/xcat/etc/nodetype.tab
# nodetype.tab maps nodes to types of installs.
```

Our example uses only one type, but you might have a few
different types.. a subset of nodes with GigE, storage nodes,
etc. man nodetype.tab for more information.
########## !!!!!!!!!!!! this file can not contain comments !!!!
########## !!!!!!!!!!!! this file can not contain comments !!!!
########## !!!!!!!!!!!! this file can not contain comments !!!!
node01compute73
node02compute73
node03compute73
node04compute73
node05compute73
node06compute73
node07compute73
node08compute73
node09compute73
node10compute73
node11compute73
node12compute73
node13compute73
node14compute73
node15compute73
node16compute73
node17compute73
node18compute73
node19compute73
node20compute73
node21compute73
node22compute73
node23compute73
node24compute73
node25compute73
ect…

passwd.tab

/opt/xcat/etc/passwd.tab
passwd.tab defines some passwords that will be used in the cluster
man passwd.tab for more information.
cisco	cisco
rootpw	netfinity
asmauser	USERID
asmapass	PASSW0RD

ipmi.tab

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

```
node004     etc …

Check tabs by running some <rpower commands and rbeacon command>
```

14. Configuring the Terminal Servers

NOTE:The hardware associated with e1350 serial network will be preconfigured per IBM e1350 defaults delivered from manufacturing.
NOTE: Please see the last section in this document for e1350 orders not containing terminal servers
This section describes setting up ELS and ESP terminal servers and conserver. Your cluster will probably have either ELSes or ESPs so you can skip the instructions for the terminal server type that is not a part of your cluster. Terminal servers enable out-of-band administration and access to the compute nodes... e.g. watching a compute node's console remotely before the compute node can be assigned an IP address or after the network config gets messed up, etc.

14.1 Learn About Conserver

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

14.2 Shutdown Conserver

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

>service conserver stop

14.3 Setup Terminal Servers

This section describes how to configure the Equinox ELS terminal server. If you're using the ESP terminal servers instead of the ELSes, you'll want to skip this section and skip ahead to 16.4 and follow the ESP instructions.

14.4 conserver.cf Setup

Modify /opt/xcat/etc/conserver.cf
This has already been covered in the configuring xCAT section, but this explains it...

Each node gets a line like:

nodeXXX:!tsx:yyyy:&:

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

14.5 Set ELS's IP Address

For each ELS unit in your cluster...

Reset the ELS to factory defaults. You usually have to push the reset botton. If the button is green, just push it. If the botton is white, you need to hold it down until the link light stops blinking. All the new units have green buttons.

Connect the DB-9 adaptor Equinox part #210062 to the management nodes's first serial port (COM1) and connect a serial cable from the ELS to the DB-9 adapter. You can test that the serial connection is good with:

> cu -l /dev/ttyS0 -s 9600

Hit Return to connect and you should see:

Username>

Unplug the serial cable to have cu hangup and then reconnect it for the next step:
> setupelsip <ELS_HOSTNAME>

Test for success:
> ping <ELS_HOSTNAME>

14.6 Final ELS Setup

After assigning the ELS' IP address over the serial link, use

> setupels <ELS_HOSTNAME>

to finish the setup for each ELS in your cluster. This sets up the terminal server's serial settings. After the serial settings are set, you can not use setupelsip again, because the serial ports have been set for reverse use. A reset of the unit will have to be performed again, if you need to change the IP address.

16.4 Setup ESP Terminal Servers
This section describes how to configure the Equinox ESP terminal server. If you're using ELS terminal servers, as most of the examples in this document do, you should skip this section and use the ELS section instead.

14.7 conserver.cf Setup

Modify /opt/xcat/etc/conserver.cf

Each node gets a line like:

nodeXXX:/dev/ttyQxxyy:9600p:&:

where xx = ESP Unit number and yy = ESP port (in hex) e.g. ttyQ01e0

14.8 Build ESP Driver

Install the RPM (must be 3.03 or later!)

14.9 Startup Configuration

Type /opt/xcat/sbin/updaterclocal (you can run this multiple times without creating problems). You need to run this because the ESP RPM puts evil code in the rc.local file, that forces the ESP to load very last and any other service that needs the ESP to start (e.g. conserver) will fail.

> cp /opt/xcat/rc.d/espx /etc/rc.d/init.d/
> chkconfig espx on

14.10 ESP Driver Configuration

Note the mac address of each ESP and manually create the /etc/eqnx/esp.conf file. All that esp util does is create this file, you can do it yourself and save a lot of time. No need to setup DHCP for the ESPs this way.

> service espx stop
> rmmod espx
> service espx start

14.11Start Conserver

> service conserver start

14.12 Understanding How To Tell if Conserver and Terminal Servers are Working

wcons –t <node range>

Setup xCAT

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

Build a DNS server (this is not an option):

>./makedns master

Check DNS with:
>host mgt1
The dns should return the IP for mgmt1 172.20.0.1

Enter non-collectable MACs in $XCATROOT/etc/mac.tab. (E.g. terminal servers, switches, RSAs, etc...)

NOTE: Some network devices (e.g. APC Master Switch) do not have the MAC address affixed to the unit. Some (e.g. APC Master Switch) have the MAC printed on a piece of

receipt paper and stuffed in the manual. Hopefully you didn't install all the APCs and chuck the manuals in a pile somewhere. The morale of this story is that before you rack anything please verify the that MAC address is visible and will be visible when racked. Very cool network devices (e.g. APC Master Switch and RSA) have a serial port, you can use this to get the MAC.

NOTE: Manual non-collectable MAC entries in mac.tab do not require a -eth0 appended--it's optional.

15. Initial DHCP Setup

15.1 Collect the MAC Addresses of Cluster Equipment

Place the MAC addresses of cluster equipment that needs to DHCP for an IP address into /opt/xcat/etc/<MANAGEMENT_NET>.tab. See the man page for macnet.tab.

If you have APC master switches, put their MAC addresses into this file.

15.2 Make the Initial dhcpd.conf Config File

> makedhcp --new

15.3 Edit dhcpd.conf

Check for anything out of the ordinary
> vi /etc/dhcpd.conf

Verify entries they should look something like this in /etc/dhcp.conf

```
#xCAT 1.2.0-RC1

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#
```

NOTE: once you getmacs and then makedhcp –allmacs you will find an entry for each mac address for each node in the dhcp.conf.

15.4 Important DHCP Note

You probably don't want DHCP running on the network interface that is connected to the rest of the network. Except for in special circumstances, you'll want to remove the network section from dhcpd.conf that corresponds to the external network and then explicitly list the interfaces you want dhcpd to listen on in /etc/dhcpd (leaving out the external interface).

Edit /etc/sysconfig/dhcpd, with something like:

DHCPDARGS="eth0 eth1"

NOTE: 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 and newer Red Hat before you run makedhcp. If incorrect, correct and rerun makedhcp --new --allmac.

NOTE: $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, please block DHCP in and out bound on ports that are used to uplink the cluster to the real world. 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 (only if using terminal servers or SOL, Bladecenter without SOL do not use conserver):

>service conserver restart

15.5 Setup stage boot image:

For x86 and x86_64 type:
>cd /opt/xcat/stage
>./mkstage

For ia64 type:

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

15.6 Collecting MAC Addresses (stage2)

In this section, we 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

You should always be watching messages. It's a very good way to get information about what's happening with your cluster. Watching it is a great habit to get into.

Reboot Compute Nodes

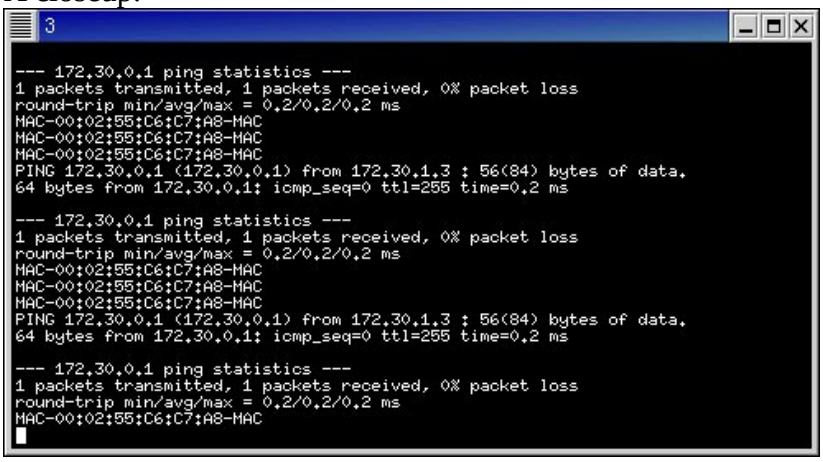
You'll have to do this manually.

When the machine's boot, they should PXE boot syslinux, get a dynamic IP address, and then load a linux kernel and a special RAMdisk that contains a script that prints the machine's MAC address to the console.

Observe Output in wcons Windows

If your terminal servers are working correctly, you should see the machines boot their kernels and then something like this:

A closeup:



Notes on wcons, xterms and Changing Font Size

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 you to change the size of the font very easily. This is very useful when you have a screen of 'unreadable' consoles and you want to zoom in on one to view the output in greater detail.

To do this, move the mouse over the text portion of the xterm in question, hold down the control key, and press the right mouse button. You'll see a menu like the following:

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 'unreadable' font to view all the consoles at once.

Notes on wcons, conserver and 'Ctrl-E .'

This is a placeholder to remind me to document the 'Ctrl-E .' escape sequence that conserver uses to provide a lot of terminal functionality.
> getmacs compute

Collect the MACs

Once you see that all the compute nodes are spitting their MAC addresses out of their serial consoles...

NOTE: This tab file should be in the configuration files that came with your cluster if it is an IBM e1350.

Kill the wcons windows

> wkill

Manually reboot each node. 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

>Auto merge mac.lst with /opt/xcat/etc/mac.tab(y/n)? y

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

NOTE: Do not alter the mac.tab entries for collected MACs. It is critical that the stored node names remain untouched. If necessary changing the MAC is OK.

NOTE: Multiple getmacs commands will corrupt mac.tab. Only run one instance at a time.

NOTE: Some OSes report eth0 and eth1 different than xCAT getmacs collect. You may need to reverse manually in mac.tab. E.g. (this may hose other good non-switched entries, verify before you run commands):

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: Bladecenter mpcli2 and bcmm getmacs methods can collect both MAC addresses.

NOTE: For Bladecenter please use bcmm method in nodehm.tab.

Notes on Collecting MAC addresses 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 pluged into the switch ports match what you put into cisco3500.tab IE 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. What the getmacs command does is issue the show mac-address-table on the switch and grab the macs from it.

15.7 All other switches

You need 3 files, switch.tab, getmacs.switch.snmp, and getmacs.switch
Place **getmacs.switch.snmp** and **getmacs.switch** into the **opt/xcat/lib** directory (MAKE SURE THEY ARE EXCUTABLE..**ls -l** to verify)
Place **switch.tab** into the **opt/xcat/etc** directory:

For example SMC alter 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 method for getmacs (not necessarily the way yours will look but just an example of what piece of the nodehm.tab file you need to change for this to work)

```
            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 e325/e326 (read managementprocessor-HOWTO for more info found in /opt/xcat/doc):
EXCEPTION: Bladecenter (just use mpname noderange).

nodeset noderange stage3

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

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

Test systems management:

rpower noderange stat

rbeacon noderange on (if blinking lights entertain you -- NOTE: not all servers have a blinking light.)

15.8 Copy the RedHat Install CD(s)

>copycds
insert cds and fallow prompts

NOTE: When the cds are entered and you are prompted for auto run select "NO"

You may also use the copy cds to copy the contents of an .iso if you need to do this just type copycds <namecd1>.iso, <namecd2>.iso, <namecd3>.iso, etc

19.2 Copy the 'post' Files for RedHat

Copy some install files from the xCAT distribution to the post directory that is used during unattended installs:

> cd /opt/xcat/post
> find . | cpio -dump /install/post

19.3 Setup syslog

Here we enable remote logging...

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

On RH7.x based installs, you might want to edit /etc/sysconfig/syslog, changing SYSLOGD_OPTIONS and add the -r switch instead of copying the modified rc.d/syslogd. See the note here (but ignore the watchlogd stuff).

19.5 Setup snmptrapd

snmptrapd received messages from the SPN.

> chkconfig snmptrapd on
> service snmptrapd start

15.9 Generate root's SSH Keypair

The following command create's a SSH keypair for root with an empty passphrase, sets up root's ssh configuration, and copies 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

19.7 Setup NFS and NFS Exports

Make /etc/exports look something like the following:

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

Turn on NFS:
> 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

NOTE: if you do not have a Myrinet read the myrinet-how to in /opt/xcat/doc. For more detailed information read the nodeinstall-HOWTO and systemimager-HOWTO for details on node install and diskless installs..

NOTE: Got disk? Install nodes. Use rinstall or winstall. Only install 32 at a time or use staging. Read man pages on rinstall and winstall, e.g.:

16. Installing Compute Nodes

Edit/Generate Kickstart Scripts

Modify kickstart template file if needed. Substitute your version of RedHat for xx...
>cp /opt/xcat/install/rhws3/<architecture>/base/compute.tmpl ..

Nodeset

The following command makes the nodes PXE boot the RedHat kickstart image. (it alters the files in /tftpboot/pxelinux.cfg/)
> nodeset compute install

Prepare to Monitor the Installation Progress

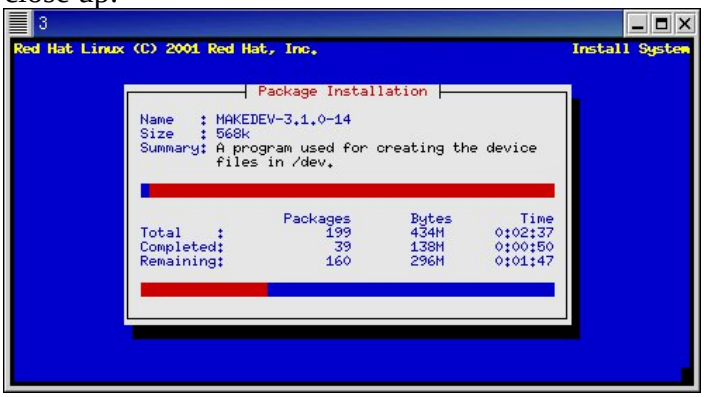
> wcons -t 8 compute (or a subset like rack01)
> tail -f /var/log/messages (you should always be watching messages)

Reboot the Compute Nodes

You might want to do only a subset of 'compute'
>rpower compute boot

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

close up:



Installs with No Terminal Servers

SOL must be set up

17. E1350 Serial Over Lan (SOL) Setup Version 1.0

SMbridge Notes:
Download and install smbridge RPM: This rpm needs to be installed on management node (for mfg that means the crash cart that has xCAT on it) since it is client software for the BMCs.
http://www-1.ibm.com/support/docview.wss?uid=psg1MIGR-57729&rs=0&cs=utf-8&context=HW20Q&dc=D400&q1=sm+bridge&loc=en_US&lang=en&cc=US">here</A>

IBM xSeries x336/x346/x236

Flash Management Processor (BMC) to best recipe.
Flash BIOS to Best Recipe Remove the power cord. Count to 10, restore power cord.
Reboot and press [F1] to enter BIOS.

Bios Setup

Load Best Recipe settings first, then move on to the following:
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

Remote Console Text Emulation: VT100/VT220

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

Startup Options

- Startup Sequence
  - 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

Advanced Setup
- CPU Options
  - Hyper-Threading Technology: Disabled

IBM xSeries x326

BIOS settings from e326

Load Best Recipe settings first then move on to the following

Console Redirection

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

BMC

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

Addtional Settings for x326,x336,x346 (do the above first)

Advanced Setup

- Baseboard Management Controller (BMC) Settings
  - System BMC Serial Port Sharing: Enabled
  - BMC Serial Port Access Mode: Dedicated

Save Settings

If you switched to using 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.e326 node001::&:
node002:|sol.x336 node002::&:
node003:|sol.x346 node003::&:
ipmi.tab

There are a few different ways of approaching this
node123 bmc123,"","""

node123 bmc123,

node123 bmc123,USERID,PASSW0RD (that is a zero in PASSW0RD)
If you use “””” 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 ipmi tool (uses bmc). Below is an 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 BIOS 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 (a relatively new feature) in the site.tab file, the node name will then be displayed correctly in the title bar.
bufferedcons no

WCONS

When you run **wcons <nodename>** a screen will appear with :

connected....

login :

password :

Entry for login and password has to be the same as ipmi.tab file. For example: login : "" and password : "" if you were to use example one of the ipmi.tab above

NOTE: You must have smbridge RPM installed (smbridge notes at top).

Verify that the Compute Nodes Installed Correctly

pping all

Update the SSH Global Known Hosts File

> makesshgkh compute (or, again, a subset of 'compute')

18. Clean Up

Copy xCAT init 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 init files in /opt/xcat/rc.d that you may wish to use, depending on your installation.

Clean Up the Unneeded .tab Files

In /opt/xcat/etc/, move unneeded .tab files somewhere out of the way e.g. rtel.tab, tty.tab, etc.

Testing the cluster

Read the man pages for rvitals, rinv, and rpower, etc. and then try out some of these commands on your cluster.

Test SSH and psh

>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 (a requirement for most cluster tasks). If a node doesn't appear here correctly, you must go back and troubleshoot the individual node, make certain the install happens correctly, rerun makesshgkh, and finally test again with psh. You really must get psh working correctly before continuing.

Commands to test
> rvitals compute ambtemp
>mpncheck compute
>pping all
>rbeacon ccompute on

Contributing to xCAT

Join the xCAT-dev mailing list and post your suggestions, bug-fixes, code, etc.

http://xcat.org/mailman/listinfo/xcat-user

Credits

This document was most recently modified:
06/01/2005
Original author Matt Bohnsack

Send additions and corrections to the editor **ShaddGa@us.ibm.com**, so this document can continue to be improved.

Thanks go out to the following people. They 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,

19. Supporting Documentation 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) <= For 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 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 go to xcat.org.

Filename: C1350_xCAT_R110805.doc
Directory: C:\Cluster 1350\1350 5B
Template: C:\Documents and Settings\weiler\Application Data\Microsoft\Templates\Normal.dot
Title: Extreme Cluster Administration Toolkit
Subject:
Author:
Keywords:
Comments:
Creation Date: 11/8/2005 5:29 PM
Change Number: 9
Last Saved On: 11/16/2005 6:49 PM
Last Saved By: IBM
Total Editing Time: 29 Minutes
Last Printed On: 11/16/2005 6:49 PM
As of Last Complete Printing
Number of Pages: 57
Number of Words: 14,133 (approx.)
Number of Characters: 80,562 (approx.)