xCAT 2 BladeCenter HowTo

09/16/2009

Table of Contents

1.0 Introduction

Before preceding to setup your BladeCenter with this document, you should first read xCATtop for information on downloading and installing xCAT on your Management Node.

This document provides step-by-step instructions on setting up an example stateful or stateless cluster for a BladeCenter. Our example will be installed with Fedora 8, x86_64.

2.0 Download Linux Distro ISOs and Create Repository

2.1 Create Fedora repository

1. Download Fedora ISOs or load your OS's DVD's of the appropriate architecture (e.g. x86_64, ppc) and place in a directory:

```
mkdir /root/xcat2
cd /root/xcat2
export BASEURL=ftp://download.fedora.redhat.com/pub/fedora/linux/releases/8
wget $BASEURL/Fedora/x86_64/iso/Fedora-8-x86_64-DVD.iso
```

2. Run copycds to setup the install directory for the node diskfull/diskless boots. The copycds commands will copy the contents of to /install/fedora8/<arch>.

```
cd /root/xcat2
copycds Fedora-8-x86_64-DVD.iso
```

3. Create the *.repo file

```
cd /etc/yum.repos.d
Create fedora.repo with contents:

[fedora]
name=Fedora $releasever - $basearch
baseurl=file:///install/fedora8/x86_64
enabled=1
```

4. Install createrepo (not needed on SLES):

```
yum install createrepo
```

5. Run createrepo

```
cd /install/fedora8/x86_64
createrepo .
```

2.2 Create SLES repository

On SLES, copy the SLES ISO images to the Management Node.

```
mkdir /iso
```

```
copy SLES11-DVD-ppc-GM-DVD1.iso
 to /iso/
mkdir /iso/1
cd /iso
mount -o loop SLES11-DVD-ppc-GM-DVD1.iso 1


zypper ar file:///iso/1 sles11
```

3.0 Set Up Services on the Management Node

3.1 Set Up networks Table

All networks in the cluster must be defined in the networks table. When xCAT was installed, it ran makenetworks, which created an entry in this table for each of the networks the management node is on. Now is the time to add or update any other networks needed to the networks table. Use either the tabedit or the chtab command.

```
#netname,net,mask,mgtifname,gateway,dhcpserver,tftpserver,nameservers,dynamicrange
   ,nodehostname,comments,disable
"mnet","9.114.47.224","255.255.255.224","eth0",,,"9.114.47.250","9.114.47.250,9.11
   4.8.1",,,,
,"192.168.122.0","255.255.255.0","virbr0",,,"192.168.122.1","9.114.47.250,9.114.8.
   1",,,,


For example to add a dynamic range for dhcp to eth0 network (mnet) :

chtab netname=mnet  networks.dynamicrange=9.114.47.233-9.114.47.234

tabdump networks

#netname,net,mask,mgtifname,gateway,dhcpserver,tftpserver,nameservers,dynamicrange
,nodehostname,comments,disable
"mnet","9.114.47.224","255.255.255.224","eth0",,"9.114.47.250","9.114.47.250","9.1
14.47.250","9.114.47.233-9.114.47.234",,,
"virb","192.168.122.0","255.255.255.0","virbr0",,,"192.168.122.1","9.114.8.1,9.114
.8.2",,,,
```

You can have xCAT ignore any table entry by setting the **disable** attribute. For example, if you have a public network defined, and you want to disable the entry for the public network (connected to the outside world):

```
chtab net=9.114.88.160 networks.netname=public networks.disable=1
```

Set the domain name in the site table:

```
chtab key=domain site.value=cluster.net   # domain part of the node hostnames
```

3.2 Set Up DHCP

The dynamic ranges for the networks were set up already in section 3.1 Set Up networks Table . Now you should define the dhcp interfaces in site table if you want to limit which NICs dhcpd will listen on. We use this weird value because our MN uses eth4 to communicate with the service nodes, and the service nodes use eth1 to communicate with the compute nodes.

The interface is

chtab key=dhcpinterfaces site.value='<node or nodegroup>|nic;<node or nodegroup>|nic;...>

For example: if you set dhcpinterfaces as in the example, only eth1 will be setup for the management node. Note only xcatmn , the management node is not defined in the database; all other entries should be defined nodes or nodegroups.

```
chtab key=dhcpinterfaces site.value='xcatmn|eth1'
tabdump -d site  will give more information on the dhcpinterfaces attribute.
```

Add the relevant networks to DHCP:

```
makedhcp -n
```

Restart DHCP:

```
service dhcpd restart
```

3.3 Set Up NTP

To enable the NTP services on the cluster, first configure NTP on the management node and start ntpd.

Next set the ntpservers attribute in the site table. Whatever time servers are listed in this attribute will be used by all the nodes that boot directly from the management node.

If your nodes have access to the internet you can use the global servers:

```
chtab key=ntpservers site.value=0.north-america.pool.ntp.org,
1.north-america.pool.ntp.org,2.north-america.pool.ntp.org,
   3.north-america.pool.ntp.org
```

If the nodes do not have a connection to the internet (or you just want them to get their time from the management node for another reason), you can use your Management Node as the NTP server.

```
chtab key=ntpservers site.value=xcatmn
```

To set up NTP on the nodes, add the setupntp postinstall script to the postscripts table. See section 4.8, Set Up Postscripts to be Run on the Nodes. Assuming you have a group named compute:

```
chtab node=compute postscripts.postscripts=setupntp
```

3.4 Set Up DNS

Note: The DNS setup here is done using the non-chroot DNS configuration. This requires that you first remove the bind-chroot rpm (if installed) before proceeding:

```
rpm -e bind-chroot-9.5.0-16.a6.fc8
```

Set nameserver, and forwarders in the site table:

```
chtab key=nameservers site.value=9.114.47.250 # IP of mgmt node
chtab key=forwarders site.value=9.114.8.1,9.114.8.2 # site DNS servers
```

Make sure your /etc/hosts file is setup on the Management Node. See .

Run:

```
makedns
```

Set up /etc/resolv.conf:

```
search cluster.net
nameserver 9.114.8.1
```

Start DNS:

```
service named start
chkconfig --level 345 named on
```

3.5 Define AMMs as Nodes

xCAT requires the AMM management module. It does not support MM's.
The nodelist table contains a node definition for each management module and switch in the cluster. For example:

```
chtab node=bca01 nodelist.groups=mm
chtab node=swa01 nodelist.groups=nortel,switch

tabdump nodelist

     .
     .
"bca01",mm,,,
"swa01",”nortel,switch”,,,
```

Also define the hardware control attributes for the management modules:

```
chtab node=mm nodehm.mgt=blade
chtab node=mm mp.mpa=bca01
```

Verify:

```
lsdef mm
```

```
Object name: bca01
    groups=mm
    mgt=blade
    mpa=bca01
    status=alive
```

3.6 Set Up Password Table

Add needed passwords to the passwd table to support installs. Note the "system" password will be the password assigned to the root id during the installation. The "blade" password will be used for communication to the management module (e.g. rspconfig)

```
chtab key=system passwd.username=root passwd.password=cluster
chtab key=blade passwd.username=USERID passwd.password=PASSW0RD
```

3.7 Set Up AMMs

Note: currently the network settings on the MM (both for the MM itself and for the switch module) need to be set up with your own customized script. (Eventually, this will be done by xCAT through lsslp, finding it on the switch, looking in the switch table, and then setting it in the MM. But for now, you must do it yourself.) After setting the network settings of the MM and switch module, then:

```
rspconfig mm snmpcfg=enable sshcfg=enable
rspconfig mm pd1=redwoperf pd2=redwoperf
rpower mm reset
```

Test the ssh set up with:

```
psh -l USERID mm info -T mm[1]
```

TIP for SOL to work best telnet to nortel switch (default pw is "admin") and type:

```
/cfg/port int1/gig/auto off
```

Do this for each port (I.e. int2, int3, etc.)

3.7.1 Update the AMM Firmware, If Necessary

Updating AMM Firmware can be done through the web GUI or can be done in parallel with ssh. To do it in parallel using psh:

Download Firmware from http://www-304.ibm.com/systems/support/supportsite.wss/docdisplay?brandind=5000008&lndocid=MIGR-5073383

```
cd /tftpboot/
unzip ibm_fw_amm_bpet36k_anyos_noarch.zip
# Perform update
psh -l USERID mm "update -i 11.16.0.1 -l CNETCMUS.pkt -v -T mm[1]"
```

```
# Reset the AMM, they will take a few minutes to come back online
psh -l USERID mm "reset -T mm[1]"
```

You can display the current version of firmware with:

```
psh -l USERID mm "info -T mm[1]" | grep "Build ID"
```

3.8 Start Up TFTP

```
service tftpd restart
```

3.9 Other Services

An HTTP server is needed for node installation (diskful), and an FTP server is needed for the nodes to access the postscripts and credentials. Both of these services should be set up automatically when xCAT is installed.

4.0 Define Compute Nodes in the Database

Note: For table attribute definitions run "tabdump -d <table name>". In some of the following table commands, you can use regular expressions are used so that a single row in the table can represent many nodes when dealing with large clusters. See http://xcat.sf.net/man5/xcatdb.5.html for a description of how to use regular expressions in xCAT tables, and see http://www.perl.com/doc/manual/html/pod/perlre.html for an explanation of perl regular expressions.

4.1 Set Up the nodelist Table

The nodelist table contains a node definition for each node in the cluster. Nodes can be added to the nodelist table using nodeadd and a node range and automatically be assigned to the all,ls21 and blade groups. For example:

```
nodeadd blade01-blade04 groups=all,ls21,bc01,blade,compute
```

4.2 Set Up the nodehm table

Specify that the BladeCenter management module should be used for hardware management.

```
chtab node=compute nodehm.cons=blade nodehm.mgt=blade  nodehm.serialspeed=19200
   nodehm.serialflow=hard nodehm.serialport=1
```

Check the definition of your blades:

```
lsdef compute
```

```
Object name: blade01
    cons=blade
    conserver=xcatmn
    groups=all,ls21,blade,bc01,compute
    mgt=blade
    serialflow=hard
    serialport=1
    serialspeed=19200
    status=alive
         .
         .
         .
```

Note: if you are using JS blades, do not set serialspeed or serialport.

4.3 Set Up the mp and mpa Table

Specify the slot (id) and mm that each blade has in the mp table.

```
chtab node=blade01  mp.id=1    mp.mpa=bca01
```

Define the username and password for the management module in the mpa table only if you have different passwords for your management modules, otherwise the password will default from the passwd table.

```
chtab mpa=bca01 mpa.username=USERID mpa.password=newpasswd
```

4.4 Set Up Conserver

Now that the nodehm and mp tables are set up, hardware management should work.

```
makeconservercf
service conserver stop
service conserver start
```

Test a few nodes with rpower and rcons.

If you have problems with conserver:

1. Setting the blade bios versions correctly
2. Setting xCAT tables correctly (check your nodehm table).

For #1, check your docs (for example for a hs21 blade):

http://download.boulder.ibm.com/ibmdl/pub/systems/support/system_x_cluster/hs21-cmos-settings-v1.1.htm

4.5 Set Up the noderes Table

The noderes table defines where each node should boot from (xcatmaster), where commands should be sent that are meant for this node, and the type of network booting supported (among other things).

In this case, the management node hostname (as known by the compute node) should be used for xcatmaster of the node.

```
chtab node=compute noderes.netboot=pxe noderes.xcatmaster=xcatmn
   nodehm.serialport=1 noderes.installnic=eth0 noderes.primarynic=eth0
   noderes.nfsserver=xcatmn
```

4.6 Set Up nodetype Table

Define the OS version and the specific set of packages (profile) that should be used for each node. The profile refers to a pkglist and exlist in /opt/xcat/share/xcat/netboot/<os> or /opt/xcat/share/xcat/install/<os>.

```
chtab node=compute nodetype.os=fedora8 nodetype.arch=x86_64
   nodetype.profile=compute nodetype.nodetype=osi
```

4.7 Verify the Tables

To verify that the tables are set correctly, run lsdef on a blade:

```
lsdef blade01

Object name: blade01
    arch=x86_64
    cons=blade
    conserver=xcatmn
    groups=all,ls21,blade,bc01,compute
    id=1
    installnic=eth0
    mgt=blade
    mpa=bca01
    netboot=pxe
    nfsserver=xcatmn
    nodetype=osi
    os=fedora8
    primarynic=eth0
    profile=compute
    serialflow=hard
    serialport=1
    serialspeed=19200
    status=alive
    tftpserver=xcatmn
    xcatmaster=xcatmn
           .
           .
           .
```

4.8 Set Up Postscripts to be Run on the Nodes

xCAT automatically adds the syslog and remoteshell postscripts to the xcatdefaults row of the table. If you want additional postscripts run on the nodes that are shipped with xCAT, for example the ntp setup script:

```
chtab node=compute postscripts.postscripts=setupntp
```

If you want to add your own postscript, then place the postscript (myscript) in the /install/postscripts directory and add to the postscripts table.

```
chtab node=compute postscripts.postscripts=setupntp,myscript
```

4.9 Get MAC Addresses for the Blades

For blades, MACs can either be collected through the boot discovery process or by using the getmacs command:

```
getmacs compute
```

("compute" is the group of all the blades.) To verify mac addresses in table:

```
tabdump mac
```

4.10 Add Compute Nodes to DHCP

Ensure dhcpd is running:

```
service dhcpd status
If not:
service dhcpd start
```

Configure DHCP:

```
makedhcp -a
```

4.11 Setup Blade for net boot

rbootseq <nodename> net,hd

5.0 Diskfull install the Blades

If you want to run the LS21 blades diskfull, statefull, then at this point, simply run:

```
nodeset <nodename> install
rpower <nodename> boot
rcons <nodename>
tail -f /var/log/messages
```

6.0 Build and Boot the Stateless Images on the Blades

If you desire to build stateless images and then boot nodes, instead of installing the blades, then follow these instructions:
Note: you can do both. You can have your blades installed with one image, but stateless boot another image. This is convenient for testing new images.

6.1 Build the Stateless Image

1. On the management node, check the compute node package list to see if it has all the rpms required.

   ```
   cd /opt/xcat/share/xcat/netboot/fedora/
   vi compute.pkglist compute.exlist    # for ppc64, edit compute.ppc64.pkglist
   ```

   For example to add vi to be installed on the node, add the name of the vi rpm to compute.pkglist. Make sure nothing is excluded in compute.exlist that you need. For example, if you require perl on your nodes, remove ./usr/lib/perl5 from compute.exlist . Ensure that the pkglist contains bind-utils so that name resolution will work during boot.

2. Generate the image:

   ```
   cd /opt/xcat/share/xcat/netboot/fedora/
   ./genimage -i eth0 -n tg3,bnx2 -o fedora8 -p compute
   ```

3. On the management node, edit fstab in the image:

   ```
   export ARCH=x86_64      # set ARCH to the type of image you are building

   cd /install/netboot/fedora8/$ARCH/compute/rootimg/etc
   cp fstab fstab.ORIG
   ```

   Edit fstab. **Change:**

   ```
   devpts  /dev/pts devpts   gid=5,mode=620 0 0
   tmpfs   /dev/shm tmpfs    defaults       0 0
   proc    /proc    proc     defaults       0 0
   ```

```
sysfs   /sys     sysfs    defaults       0 0
```

to (replace $ARCH with the actual value):

```
proc             /proc                 proc         rw 0 0
sysfs            /sys                  sysfs        rw 0 0
devpts           /dev/pts              devpts       rw,gid=5,mode=620 0 0
#tmpfs           /dev/shm              tmpfs        rw 0 0
compute_$ARCH    /                     tmpfs        rw 0 1
none             /tmp                  tmpfs        defaults,size=10m 0 2
none             /var/tmp              tmpfs        defaults,size=10m 0 2
```

Note: adding /tmp and /var/tmp to /etc/fstab is optional, most installations can simply use /. It was documented her to show that you can restrict the size of filesystems, if you need to. The indicated values are just and example, and you may need much bigger filessystems, if running applications like OpenMPI.

4. Pack the image:
```
packimage -o fedora8 -p compute -a $ARCH
```

6.2 Test Boot the Stateless Image

You can continue to customize the image and then you can boot a node with the image:

```
nodeset <nodename> netboot
rpower <nodename> boot
```

You can monitor the install by running:

```
rcons <nodename>
```

7.0 References

- xCAT web site: http://xcat.sf.net/
- xCAT man pages: http://xcat.sf.net/man1/xcat.1.html
- xCAT DB table descriptions: http://xcat.sf.net/man5/xcatdb.5.html
- Installing xCAT on iDataPlex: http://xcat.svn.sourceforge.net/svnroot/xcat/xcat-core/trunk/xCAT-client/share/doc/xCAT-iDpx.pdf
- xCAT2 Linux Cookbook : http://xcat.svn.sourceforge.net/svnroot/xcat/xcat-core/trunk/xCAT-client/share/doc/xCAT2.pdf
- For installing Torque and Moab : http://xcat.svn.sourceforge.net/svnroot/xcat/xcat-core/trunk/xCAT-client/share/doc/xCAT2.pdf

- Using LDAP for user authentication in your cluster: http://xcat.svn.sourceforge.net/svnroot/xcat/xcat-core/trunk/xCAT-client/share/doc/xCAT2.ldap.pdf
- Monitoring Your Cluster with xCAT: http://xcat.svn.sourceforge.net/svnroot/xcat/xcat-core/trunk/xCAT-client/share/doc/xCAT2-Monitoring.pdf
- xCAT on AIX Cookbook: http://xcat.svn.sourceforge.net/svnroot/xcat/xcat-core/trunk/xCAT-client/share/doc/xCAT2onAIX.pdf
- xCAT wiki: http://xcat.wiki.sourceforge.net/
- xCAT mailing list: http://xcat.org/mailman/listinfo/xcat-user
- xCAT bugs: https://sourceforge.net/tracker/?group_id=208749&atid=1006945
- xCAT feature requests: https://sourceforge.net/tracker/?group_id=208749&atid=1006948