xCAT 2 on AIX

Cloning AIX nodes

(using AIX mksysb images)

03/30/2010, PM 01:40:12

1.0 Overview....1
2.0 Cloning AIX nodes (using mksysb images)....2
2.1 Define the HMC as an xCAT node....2
2.2 Discover the LPARs managed by the HMC....2
2.3 Define xCAT cluster nodes....3
2.4 Add IP addresses and hostnames to /etc/hosts....3
2.5 Define xCAT groups (optional)....4
2.6 Create an operating system image....4
2.7 Define xCAT networks....6
2.8 Create additional NIM network definitions (optional) 7
2.9 Gather MAC information for the install adapters....8
2.10 Set up customization scripts (optional) 9
2.10.1 Add NTP setup script 10
2.10.2 Add secondary adapter configuration script....10
2.10.3 Configure NIM to use nimsh and SSL....10
2.11 Create NIM client & group definitions....11
2.12 Initialize the AIX/NIM nodes....12
2.13 Open a remote console (optional)....12
2.14 Initiate a network boot....13
2.15 Verify the deployment....13
3.0 Cleanup....14
3.1 Removing NIM machine definitions....14
3.2 Removing NIM resources 14

1.0 Overview

This “How-To” illustrates how to use AIX **mksysb** backup images to clone AIX nodes.

A **mksysb** image is the system backup image created by the AIX **mksysb** command. You can use this image to install other machines or to restore the machine that was the source of the mksysb.

The process described below uses xCAT features to automatically run the necessary AIX and NIM commands.

NIM is an AIX tool that enables a cluster administrator to centrally manage the installation and configuration of AIX and optional software on machines within a networked environment. This document assumes you are somewhat familiar with NIM. For more information about NIM, see the *IBM AIX Installation Guide and Reference*. (http://www-03.ibm.com/servers/aix/library/index.html)

Before starting this process it is assumed you have completed the following.

- An AIX system has been installed to use as an xCAT management node.
- The cluster network is configured. (The Ethernet network that will be used to install the cluster nodes.)
- xCAT and prerequisite software has been installed and configured on the management node.
- LPARs have already been created using the HMC interfaces.

2.0 Cloning AIX nodes (using mksysb images)

2.1 Define the HMC as an xCAT node

The xCAT hardware control support requires that the hardware control point for the nodes also be defined as a cluster node.

The following command will create an xCAT node definition for an HMC with a host name of "*hmc01*". The *groups, nodetype, mgt, username*, and *password* attributes must be set.

mkdef –t node –o hmc01 groups="hmc,all" nodetype=hmc mgt=hmc username=hscroot password=abc123

2.2 Discover the LPARs managed by the HMC

This step assumes that the partitions are already created using the standard HMC interfaces.

Use the **rscan** command to gather the LPAR information. This command can be used to display the LPAR information in several formats and can also write the LPAR information directly to the xCAT database. In this example we will use the "-z" option to create a stanza file that contains the information gathered by **rscan** as well as some default values that could be used for the node definitions.

To write the stanza format output of **rscan** to a file called "*mystanzafile"* run the following command.

rscan –z hmc01 > mystanzafile

This file can then be checked and modified as needed. For example you may need to add a different name for the node definition or add additional attributes and values.

Note: The stanza file will contain stanzas for things other than the LPARs. This information must also be defined in the xCAT database. It is not necessary to modify the non-LPAR stanzas in any way.

The updated stanza file might look something like the following.

```
Server-9117-MMA-SN10F6F3D:
    objtype=node
    nodetype=fsp
    id=5
```

```
    model=9118-575
    serial=02013EB
    hcp=hmc01
    pprofile=
    parent=Server-9458-10099201WM_A
    groups=fsp,all
    mgt=hmc

node01:
    objtype=node
    nodetype=lpar,osi
    id=9
    hcp=hmc01
    pprofile=lpar9
    parent=Server-9117-MMA-SN10F6F3D
    groups=lpar,all
    mgt=hmc

node02:
    objtype=node
    nodetype=lpar,osi
    id=7
    hcp=hmc01
    pprofile=lpar6
    parent=Server-9117-MMA-SN10F6F3D
    groups=lpar,all
    mgt=hmc
```

Note: The **rscan** command supports an option to automatically create node definitions in the xCAT database. To do this the LPAR name gathered by **rscan** is used as the node name and the command sets several default values. If you use the "-w" option make sure the LPAR name you defined will be the name you want used as your node name.

2.3 Define xCAT cluster nodes

The information gathered by the **rscan** command can be used to create xCAT node definitions.

Since we have put all the node information in a stanza file we can now pass the contents of the file to the **mkdef** command to add the definitions to the xCAT database.

cat mystanzafile | mkdef –z

You can use the xCAT **lsdef** command to check the definitions (ex. "*lsdef –l node01*"). After the node has been defined you can use the **chdef** command to make any additional updates to the definitions, if needed.

2.4 Add IP addresses and hostnames to /etc/hosts

Make sure all node hostnames are added to /etc/hosts. Refer to the section titled "Add cluster nodes to the /etc/hosts file" in the following document for details. (http://xcat.svn.sourceforge.net/viewvc/xcat/xcat-core/trunk/xCAT-client/share/doc/xCAT2onAIX.pdf)

2.5 Define xCAT groups (optional)

XCAT supports both *static* and *dynamic* node groups. See the section titled "xCAT node group support" in the "xCAT2 Top Doc" document for details on using xCAT groups. (http://xcat.svn.sourceforge.net/viewvc/xcat/xcat-core/trunk/xCAT-client/share/doc/xCAT2top.pdf)

Note: The *dynamic* node group support is only available in xCAT 2.3 and beyond.

2.6 Create an operating system image

Before we can use a **mksysb** image to install a set of AIX nodes we need to create (or get) a **mksysb** image and define it as a NIM resource. Either a pre-existing image must be provided or some other NIM client machine must be provided as a source for creating the image.

Note: xCAT requires that SSL and SSH software be installed on the cluster nodes. Make sure this software has been installed on the system that is used to create the **mksysb** backup. (The xCAT xCATaixSSL.bnd and xCATaixSSH.bnd installp bundles files may be of use when installing this software.)

In this example we'll assume that you have created NIM installation resources and have already installed and configured an AIX node using NIM. See the xCAT "How-To" "Installing AIX nodes" for details on how to install AIX standalone machines using the "rte" installation method.

Note: It is also possible to use an existing **mksysb** image. This process will be covered in a future update to this document.

Once the target node is ready you can use the xCAT **mknimimage** command to create an xCAT *osimage* definition as well as the required NIM installation resources.

An xCAT *osimage* definition is used to keep track of a unique operating system image. A different *osimage* definition should be created for each unique operating system that will be used in the cluster. (For example, you may have a specific image for your "I/O" nodes and another for your "compute" nodes.) The definition will be used by other xCAT commands to help automate the installation of the cluster nodes.

In order to use NIM to perform a remote network boot of a cluster node the NIM software must be installed, NIM must be configured, and some basic NIM resources must be created.

The **mknimimage** command will handle all the NIM setup as well as the creation of the xCAT *osimage* definition. It will not attempt to reinstall or reconfigure NIM if that process has already been completed. See the **mknimimage man** page for additional details.

Note: If you wish to install and configure NIM manually you can run the AIX **nim_master_setup** command (Ex. "*nim_master_setup -a mk_resource=no -a device=<source directory>*").

By default, the **mknimimage** command will create the NIM resources in subdirectories of /install. Some of the NIM resources are quite large (1-2G) so it may be necessary to increase the files size limit.

For example, to set the file size limit to "unlimited" for the user "root" you could run the following command.

/usr/bin/chuser fsize=-1 root

It may also be necessary to increase the amount of free space available in the file system that will contain the NIM resources.

When you run the command you must provide the name of the target NIM machine to use to create the **mksysb** image. You must also provide a SPOT resource that is the same level as the software running on the node. (This is the SPOT resource that you originally used to install the target node.)

For example, to create an *osimage* named "610sysb" using a node named "node27" as the target source you could run the following command.

mknimimage -m mksysb -n node27 610sysb spot=610spot

(Creating the NIM resources could take a while!)

Note: In some cases it may also be necessary to increase the file size limit on the target node.

By default the command will create NIM *mksysb* and *bosinst_data* resources. When the command completes it will display the *osimage* definition which will contain the names of all the NIM resources that were created. The naming convention for the NIM resources that are created is the *osimage* name followed by the NIM resource type, (ex. " *610sysb_bosinst_data*").

You can also specify alternate or additional resources on the command line using the "attr=value" option, ("<nim resource type>=<resource name>"). For example, if you want to include a NIM *resolv_conf* resource named *my_resolv_conf* you could run the command as follows.

mknimimage -m mksysb -n node27 610sysb spot=610spot resolv_conf= my_resolv_conf

Any additional NIM resources specified on the command line must be previously created using NIM interfaces. (Which means NIM must already have been configured previously.)

Note: Another alternative is to run **mknimimage** without the additional resources and then simply add them to the xCAT *osimage* definition later. You can add or change the *osimage* definition at any time. When you initialize and install the nodes xCAT will use whatever resources are specified in the *osimage* definition.

Once the initial *osimage* definition is created you can change it by using the **chdef** command. The xCAT *osimage* definition can be listed using the **lsdef** command and removed using the **rmnimimage** command. See the man pages for details.

In some cases you may also want to modify the contents of the NIM resources. For example, you may want to change the *bosinst_data* file or add to the *resolv_conf* file etc. For details concerning the NIM resources refer to the NIM documentation.

You can list NIM resource definitions using the AIX **lsnim** command. For example, if the name of your SPOT resource is "*610image*" then you could get the details by running:

lsnim -l 610image

To see the actual contents of a NIM resource use "*nim -o showres <resource name>*". For example, to get a list of the software installed in your SPOT you could run:

nim -o showres 610image

Note: The **mknimimage** command will take care of the NIM installation and configuration automatically, however, you can also do this using the standard AIX support. See the AIX documentation for details on using the **nim_master_setup** command or the SMIT "eznim" interface.

2.7 Define xCAT networks

Create a network definition for each network that contains cluster nodes. You will need a name for the network and values for the following attributes.

net	The network address.
mask	The network mask.
gateway	The network gateway.

In our example we will assume that all the cluster node management interfaces and the xCAT management node interface are on the same network. You can use the xCAT **mkdef** command to define the network.

For example:

mkdef -t network -o net1 net=9.114.113.224 mask=255.255.255.224 gateway=9.114.113.254

Note: The xCAT definition should correspond to the NIM network definition. If multiple cluster subnets are needed then you will need an xCAT and NIM network definition for each one.

2.8 Create additional NIM network definitions (optional)

For the processs described in this document we are assuming that the xCAT management node and the LPARs are all on the same network.

However, depending on your specific situation, you may need to create additional NIM network and route definitions.

NIM network definitions represent the networks used in the NIM environment. When you configure NIM, the primary network associated with the NIM master is automatically defined. You need to define additional networks only if there are nodes that reside on other local area networks or subnets. If the physical network is changed in any way, the NIM network definitions need to be modified.

The following is an example of how to define a new NIM network using the NIM command line interface.

Step 1

Create a NIM network definition. Assume the NIM name for the new network is "clstr_net", the network address is "10.0.0.0", the network mask is "255.0.0.0", and the default gateway is "10.0.0.247".

```
nim -o define -t ent -a net_addr=10.0.0.0 -a snm=255.0.0.0 -a
        routing1='default 10.0.0.247' clstr_net
```

Step 2

Create a new interface entry for the NIM "master" definition. Assume that the next available interface index is "2" and the hostname of the NIM master is "xcataixmn". This must be the hostname of the management node interface that is connected to the "clstr_net" network.

```
nim -o change -a if2='clstr_net xcataixmn 0' -a cable_type2=N/A master
```

Step 3

Create routing information so that NIM knows how to get from one network to the other. Assume the next available routing index is "2", and the IP address of the NIM master on the "master_net" network is "8.124.37.24". Assume the IP address on the NIM master on the "clstr_net" network is " 10.0.0.241". This command will set the route from "master_net" to "clstr_net" to be " 10.0.0.241" and it will set the route from "clstr_net" to "master_net" to be "8.124.37.24".

```
nim -o change -a routing2='master_net 10.0.0.247 8.124.37.24'
clstr_net
```

Step 4

Verify the definitions by running the following commands.

```
lsnim -l master

lsnim -l master_net

lsnim -l clstr_net
```

See the NIM documentation for details on creating additional network and route definitions. (*IBM AIX Installation Guide and Reference*. http://www-03.ibm.com/servers/aix/library/index.html)

2.9 Gather MAC information for the install adapters.

Use the xCAT **getmacs** command to gather adapter information from the nodes. This command will return the MAC information for each Ethernet adapter available on the target node. The command can be used to either display the results or write the information directly to the database. If there are multiple adapters the first one will be written to the database.

The command can also be used to do a ping test on the adapter interfaces to determine which ones could be used to perform the network boot. In this case the first adapter that can be successfully used to ping the server will be written to the database.

Before running **getmacs** you must first run the **makeconservercf** command. You need to run **makeconservercf** any time you add new nodes to the cluster.

makeconservercf

Shut down all the nodes that you will be querying for MAC addresses.

rpower aixnodes off

To retrieve the MAC address for all the nodes in the group "*aixnodes*" and write the first adapter MAC to the xCAT database you could issue the following command.

getmacs aixnodes

To display all adapter information but not write anything to the database.

getmacs -d aixnodes

To retrieve the MAC address and do a ping test to determine which adapter MAC to use for the node you could issue the following command. (The ping operation may take a while to complete.)

getmacs -d aixnodes -S 10.14.0.2 -G 10.14.0.2 -C 10.14.0.4

The output would be similar to the following.

 # Type Location Code MAC Address Full Path Name Ping Result Device Type

ent U9125.F2A.024C362-V6-C2-T1 fef9dfb7c602 /vdevice/l-lan@30000002 successful virtual

ent U9125.F2A.024C362-V6-C3-T1 fef9dfb7c603 /vdevice/l-lan@30000003 unsuccessful virtual

From this result you can see that "*fef9dfb7c602*" should be used for this nodes MAC address.

For more information on using the **getmacs** command see the man page.

To add the MAC value to the node definitions you can use the **chdef** command. For example:

chdef -t node node01 mac=fef9dfb7c603

2.10 Set up customization scripts (optional)

xCAT supports the running of customization scripts on the nodes when they are installed.

This support includes:

- The running of a set of default customization scripts that are required by xCAT.

  You can see what scripts xCAT will run by default by looking at the "xcatdefaults" entry in the xCAT "postscripts" database table. (I.e. Run "tabdump postscripts".). You can change the default setting by using the xCAT **chtab** or **tabedit** command. The scripts are contained in the /install/postscripts directory on the xCAT management node.

- The optional running of customization scripts provided by xCAT.

  There is a set of xCAT customization scripts provided in the /install/postscripts directory that can be used to perform optional tasks such as additional adapter configuration.

- The optional running of user-provided customization scripts.

  To have your script run on the nodes:

  1. Put a copy of your script in /install/postscripts on the xCAT management node. (Make sure it is executable.)
  2. Set the "postscripts" attribute of the node definition to include the comma separated list of the scripts that you want to be executed on the nodes. The order of the scripts in the list determines the order in which they will be run. For example, if you want to have your two scripts called "foo" and "bar" run on node "node01" you could use the **chdef** command as follows.

     chdef -t node -o node01 -p postscripts=foo,bar

     (The "-p" means to add these to whatever is already set.)

Note: The customization scripts are run during the boot process (out of /etc/inittab).

2.10.1 Add NTP setup script

To have xCAT automatically set up ntp on the cluster nodes you must add the **setupntp** script to the list of postscripts that are run on the nodes.

To do this you can either modify the "postscripts" attribute for each node individually or you can just modify the definition of a group that all the nodes belong to.

For example, if all your nodes belong to the group "compute" then you could add **setupntp** to the group definition by running the following command.

chdef -p -t group -o compute postscripts=setupntp

2.10.2 Add secondary adapter configuration script

It is possible to have additional adapter interfaces automatically configured when the nodes are booted. XCAT provides sample configuration scripts for both Ethernet and IB adapters. These scripts can be used as-is or they can be modified to suit you particular environment. The Ethernet sample is /install/postscript/configeth. When you have the configuration script that you want you can add it to the "postscripts" attribute as mentioned above. Make sure your script is in the /install/postscripts directory and that it is executable.

If you wish to configure IB interfaces please refer to: "xCAT 2 InfiniBand Support" **http://xcat.svn.sourceforge.net/viewvc/xcat/xcat-core/trunk/xCAT-client/share/doc/xCAT2IBsupport.pdf**

Note: Do not forget that the new adapter interface hostnames must be resolvable on the node. To do this you can use the NIM resolve.conf resource to automatically create a resolv.conf file on the nodes when they are installed.

2.10.3 Configure NIM to use nimsh and SSL.

The NIM service handler (**nimsh**), is provided as an optional feature of NIM to be used in cluster environments where the standard **rsh** protocols are not secure enough.

Although **nimsh** eliminates the need for **rsh**, in the default configuration it does not provide trusted authentication based on key encryption. To use cryptographic authentication with NIMSH, you can configure NIMSH to use OpenSSL in the NIM environment. When you install OpenSSL on a NIM client, SSL socket connections are established during NIMSH service authentication. Enabling OpenSSL provides SSL key generation and includes all cipher suites supported in SSL version 3.

In order to facilitate the setup of **nimsh**, xCAT provides a sample customization called "confignimsh" that can be used to configure **nimsh** on the cluster nodes.

This script will also configure **nimsh** to use SSL and will remove the /.rhosts file from the node. If you do not wish to have the .rhosts file removed from the node you must remove those lines from the **confignimsh** script before using it.

This script should only be run on AIX standalone (diskfull) cluster compute nodes. It should NOT be run on the xCAT management node, service nodes or diskless nodes.

The basic processes is:

- Make sure the AIX openssl fileset gets installed on the management node and all the other cluster nodes. (Which should be done in any case.)
- On the xCAT management node run the following command.
  nimconfig -c

  You must also run this command on any service nodes that are being used.
- Add "confignimsh" to the list of scripts you want run on the nodes

  For example, if all your nodes belong to the group "compute" then you could add **confignimsh** to the group definition by running the following command.
  chdef -p -t group -o compute postscripts=confignimsh

After the nodes boot up you can verify that **nimsh** was set up correctly by running a NIM command such as: "nim -o lslpp <nodename>".

To be sure that nimsh is actually using SSL you can run the command: "nimquery -a host=<nodename>".

Example:

> nimquery -a host=xcatn11
host:xcatn11.cluster.com:addr:10.2.0.104:mask:255.255.0.0:gtwy:
10.2.0.200:_pif:en0:_ssl:yes:_psh:no:_res:no:asyn:no:mac:
163D0DDAE202:_sslver:OpenSSL 0.9.8k 25 Mar 2009:

The "_ssl:yes" indicates that **nimsh** is using SSL.

Note: You could also set up **nimsh** at any time using the xCAT **updatenode** command to run the **confignimsh** script on the nodes.

2.11 Create NIM client & group definitions

You can use the xCAT **xcat2nim** command to automatically create NIM machine and group definitions based on the information contained in the xCAT node and group definitions. By doing this you synchronize the NIM and xCAT names so that you can use the same target names when running either an xCAT or NIM command.

To create NIM machine definitions you could run the following command.

xcat2nim -t node –o aixnodes

To create NIM group definitions you could run the following command.

xcat2nim –t group –o aixnodes

To check the NIM definitions you could use the NIM **lsnim** command or the xCAT **xcat2nim** command. For example, the following command will display the NIM definitions of the nodes: *node01*, *node02*, and *node03* (from data stored in the NIM database).

xcat2nim -t node -l –o node01-node03

2.12 Initialize the AIX/NIM nodes

You can use the xCAT **nimnodeset** command to initialize the AIX standalone nodes. This command uses information from the xCAT *osimage* definition and default values to run the appropriate NIM commands.

For example, to set up all the nodes in the group "*aixnodes*" to install using the *osimage* named "*610sysb*" you could issue the following command.

nimnodeset -i 610sysb aixnodes

To verify that you have allocated all the NIM resources that you need you can run the "**lsnim –l**" command. For example, to check node "node01" you could run the following command.

lsnim -l node01

The command will also set the "*profile*" attribute in the xCAT node definitions to "*610sysb*". Once this attribute is set you can run the **nimnodeset** command without the "-i" option.

2.13 Open a remote console (optional)

You can open a remote console to monitor the boot progress using the xCAT **rcons** command. This command requires that you have **conserver** installed and configured.

If you wish to monitor a network installation you must run rcons before initiating a network boot.

To configure conserver run:

makeconservercf

To start a console:

rcons node01

Note: You must always run makeconservercf after you define new cluster nodes.

2.14 Initiate a network boot

Initiate a remote network boot request using the xCAT **rnetboot** command. For example, to initiate a network boot of all nodes in the group "*aixnodes*" you could issue the following command.

rnetboot aixnodes

Note: If you receive timeout errors from the **rnetboot** command, you may need to increase the default 60-second timeout to a larger value by setting ppctimeout in the site table:

chdef -t site -o clustersite ppctimeout=180

2.15 Verify the deployment

- You can use the AIX **lsnim** command to see the state of the NIM installation for a particular node, by running the following command on the NIM master:

  lsnim -l <clientname>

- Retry and troubleshooting tips:
  - For p6 lpars, it may be helpful to bring up the HMC web interface in a browser and watch the lpar status and reference codes as the node boots.
  - Verify network connections
  - If the **rnetboot** returns "unsuccessful" for a node, verify that bootp and tftp is configured and running properly.
    - View /etc/bootptab to make sure an entry exists for the node.
    - Verify that the information in /tftpboot/<node>.info is correct.
    - Stop and restart inetd:
      stopsrc -s inetd
      startsrc -s inetd
    - Stop and restart tftp:
      stopsrc -s tftp
      startsrc -s tftp
    -
  - Verify NFS is running properly and mounts can be performed with this NFS server:
    - View /etc/exports for correct mount information.
    - Run the showmount and exportfs commands.
    - Stop and restart the NFS and related daemons:
      stopsrc -g nfs
      startsrc -g nfs
    - Attempt to mount a filesystem from another system on the network.

3.0 Cleanup

The NIM definitions and resources that are created by xCAT commands are not automatically removed. It is therefore up to the system administrator to do some clean up of unused NIM definitions and resources from time to time. (The NIM lpp_source and SPOT resources are quite large.) There are xCAT commands that can be used to assist in this process.

3.1 Removing NIM machine definitions

Use the xCAT **xcat2nim** command to remove all NIM machine definitions that were created for the specified xCAT nodes. This command will not remove the xCAT node definitions.

For example, to remove the NIM machine definition corresponding to the xCAT node named "node01" you could run the command as follows.

xcat2nim -t node -r node01

The **xcat2nim** command is intended to make it easier to clean up NIM machine definitions that were created by xCAT. You can also use the AIX **nim** command directly. See the AIX/NIM documentation for details.

3.2 Removing NIM resources

Use the xCAT **rmnimimage** command to remove all the NIM resources associated with a given xCAT *osimage* definition. The command will only remove a NIM resource if it is not allocated to a node. You should always clean up the NIM node definitions before attempting to remove the NIM resources. The command will also remove the xCAT *osimage* definition that is specified on the command line.

For example, to remove the "610sysb" *osimage* definition along with all the associated NIM resources run the following command.

rmnimimage 610sysb

If necessary, you can also remove the NIM definitions directly by using NIM commands. See the AIX/NIM documentation for details.