xCAT 2 cookbook for Linux on IBM System P

Date: 04/01/2009

Contents

Table of Contents

1. Introduction

This cookbook introduces how to use the xCAT2 to install Linux on the IBM power series machines.
The power series machines have the following characteristics:

1. May have multiple LPARs (an LPAR will be the target machine to install an operating system image on, i.e. the LPAR will be the compute node);
2. The Ethernet card and SCSI partition can be virtual devices;
3. An HMC or IVM is used for the HCP (hardware control point)

xCAT supports two types of installations for compute nodes: Diskfull installation (Statefull) and Diskless (Stateless). xCAT also supports hierarchical management clusters where one or more service nodes are used to handle the installation and management of compute node. Please refer to xCAT2advanced.pdf for hierarchical usage.
Based on the two types of installation, the following installation scenarios will be described in this document:

1. Install a stateful compute node
2. Install a stateless compute node

To provide the easier understanding of the installation steps, this cookbook provides an example to introduce the xCAT management operations:

```
The management node:
      Arch: an LPAR on a p5/p6 machine
      OS: Red Hat 5.2
      Hostname: pmanagenode
      IP: 192.168.0.1
      HCP: HMC

The management Network:
     Net: 192.168.0.0
     NetMask: 255.255.255.0
     Gateway: 192.168.0.1
     Cluster-face-IF: eth1
     dhcpserver: 192.168.0.1
     tftpserver: 192.168.0.1
     nameservers: 192.168.0.1

The compute nodes:
      Arch: an LPAR on a p5/p6 machine
      OS: Red Hat 5.2
      HCP: HMC
```

```
      Hostname: pnode1 - this node will be installed in
      stateful
      IP: 192.168.0.10
      Cluster-face-IF: eth0

      Hostname: pnode2 - this node will be installed in
      stateless
      IP: 192.168.0.20
      Cluster-face-IF: eth0

The Hardware Control Point:
     Name:  hmc1
     IP: 192.168.0.100

xCAT version:
     xCAT-2.1+
```

2. Install xCAT 2 on the Management node

Before preceding to setup your pLlinux Cluster, you should first read xCATtop for information on downloading and installing xCAT on your Management Node.

3. Setup the management node

3.1. [Power 5] Workaround the atftpd issue

The tftp client in the open firmware of p5 is only compatible with tftp-server instead of atftpd which is required by xCAT2. So we have to remove the atftpd first and then install the tftp-server. This is not required for p6 or later.

3.1.1. Remove atftp

```
service tftpd stop
rpm --nodeps -e atftp
```

3.1.2. Install the tftp server needed by xCAT, and restart it

[RH]:

```
yum install tftp-server.ppc
```

[SLES]:

```
zypper install tftp-server.ppc
```

3.1.3. Restart the tftp server

Notes: make sure the entry "disable=no" in the /etc/xinetd.d/tftp.

```
service xinetd restart
```

3.2. Setup common attributes for xCAT in the database

3.2.1. Add the default account for system

```
chtab key=system passwd.username=root passwd.password=cluster
```

3.2.2. Add the default account for hmc

```
chtab key=hmc passwd.username=hscroot passwd.password=abc123
```

3.3. Define the compute nodes

The definition of a node is stored in several tables of the xCAT database.

You can use **rscan** command to discover the HCP to get the nodes that managed by this HCP. The discovered nodes can be stored into a stanza file. Then edit the stanza file to keep the nodes which you want to create and use the mkdef command to create the nodes definition.

3.3.1. Gather Node information using the rscan command

3.3.1.1. Define HMC as an xCAT node

First, define the hardware control point as a node object.

The following command will create an xCAT node definition for an HMC with a host name of "*hmc1*". The *groups, nodetype, mgt, username*, and *password* attributes will be set.

```
mkdef -t node -o hmc1 groups=hmc,all  nodetype=hmc  mgt=hmc
            username=hscroot  password=abc123
```

3.3.1.2. Discover the LPARs managed by HMC

Run 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 "node.stanza*"* run the following command.

```
rscan -z hmc01 >  node.stanza
```

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 stanza file will 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

pnode1:
        objtype=node
        nodetype=lpar,osi
        id=9
        hcp=hmc1
        pprofile=lpar9
        parent=Server-9117-MMA-SN10F6F3D
        groups=all
        mgt=hmc

pnode2:
        objtype=node
        nodetype=lpar,osi
        id=7
        hcp=hmc1
        pprofile=lpar6
        parent=Server-9117-MMA-SN10F6F3D
        groups=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.*

For a node which was defined correctly before, you can use the "lsdef –z [nodename]> node.stanza" command to export the definition into the node.stanza, and use command "cat node.stanza | chdef -z" to update the node.stanza according to your need.

3.3.1.3. Define xCAT node using the stanza file

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

```
cat node.stanza | mkdef –z
```

3.3.2. Update the attributes of the node

3.3.2.1. Set the resource attributes of the node

```
chdef -t node -o pnode1  netboot=yaboot  tftpserver=192.168.0.1
   fsserver=192.168.0.1 monserver=192.168.0.1 xcatmaster=192.168.0.1
   installnic="eth0" primarynic="eth0"
```

Note: Please make sure the attributes "installnic" and "primarynic" are set up by the correct Ethernet Interface of compute node. Otherwise the compute node installation may hang requesting information from an incorrect interface.

3.3.2.2. Set the type attributes of the node

```
chdef -t node -o pnode1 os=<os> arch=ppc64 profile=compute.ppc64
```

Note: The <os> can be rh, centos*, fedora*, sles*. (where * is the version #) For example, the <os> can be rhels5.2 or sles11.*

3.3.2.3. Modify the table postscripts

This only needs to be done when you want to run the my_script after installing the compute node.

```
chdef -t node -o pnode1 postscripts=my_scrip
```

3.4. Setup the Services and Definition

The part should be experienced, when there are compute nodes, hmcs, networks defined/ deleted.

3.4.1. Create the xCAT network object

Create the networks that used for cluster management:

```
mkdef -t network -o net1 net=192.168.0.0 mask=255.255.255.0
   gateway=192.168.0.1  mgtifname=eth1  dhcpserver=192.168.0.1
   tftpserver=192.168.0.1 nameservers=192.168.0.1
```

3.4.2. Setup the DNS

3.4.2.1. Setup /etc/hosts with entries for all you nodes, hmcs, fsps

```
127.0.0.1             localhost
192.168.0.1           pmanagenode
192.168.0.10          pnode1
192.168.0.20          pnode2
192.168.0.100         hmc1
```

3.4.2.2. Setup the nameserver

Add following lines into /etc/resolv.conf

```
search cluster.net
nameserver 192.168.0.1
```

3.4.2.3. Setup the DNS attributes in the Site table

Setup local machine as nameserver:

```
chdef -t site nameservers=192.168.0.1
```

Setup the external nameserver:

```
chdef -t site forwarders=9.114.1.1
```

Setup the local domain name:

```
chdef -t site  domain=cluster.net
```

3.4.2.4. Setup DNS configuration

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

3.4.3. Configure conserver

```
makeconservercf
service conserver restart
```

3.4.4. Check rcons

```
rcons pnode1
```

If it works ok, you will get into the console interface of the pnode1.

3.4.5. Update the mac table with the address of the node(s)

If there's only one Ethernet adapter on the node or you have specified the installnic or primarynic attribute of the node, using following command can get the correct mac address.

```
getmacs pnode1
```

But, if there're more than one Ethernet adapters on the node, and you don't know which one has been configured for the installation process, you have to specify more parameters like this for Lpar to try to figure out an available interface by ping operation:

```
getmacs pnode1 -S 192.168.0.1 -G 192.168.0.1 -C 192.168.0.10
```

The output looks like following:

```
pnode1:
 Type   Location Code   MAC Address     Full Path Name  Ping Result
   Device Type

ent U9133.55A.10E093F-V4-C5-T1 f2:60:f0:00:40:05 /vdevice/l-
   lan@30000005  virtual
```

And the Mac address will be written into the xCAT mac table.

3.4.6. Check rpower is working on the node

```
rpower pnode1 stat
```

3.4.7. Setup dhcp service

3.4.7.1. Setup the dhcp listen interfaces in site table

```
chdef -t site dhcpinterfaces='pmanagenode|eth1'
```

3.4.7.2. [SLES] Check the installation of dhcp-server

On the SLES management node, the dhcp-server rpm may not have been automatically installed. Use following command to check whether it has been installed:

```
rpm -qa | grep -i dhcp-server
```

If it is not installed, installed it manually:

```
zypper install dhcp-server
```

3.4.7.3. Configure the DHCP

Add the relevant networks into the DHCP configuration:

```
makedhcp -n
```

Add the defined node into the DHCP configuration:

```
makedhcp -a
```

Restart he dhcp service:

```
service dhcpd restart
```

Note: Please make sure there is only one dhcpd server can server these compute nodes.

4. Install a Compute Node

4.1. Prepare the installation source

You can use the iso file of the installed OS to extract the installation files. For example, you have a iso file /iso/RHEL5.2-Server-20080430.0-ppc-DVD.iso

```
copycds /iso/RHEL5.2-Server-20080430.0-ppc-DVD.iso
```

Note: If you encounter the issue that the iso cannot be mounted by the copycds command. Make sure the SElinux is disabled.

4.2. Statefull Node installation

4.2.1. Customize the installation profile

xCAT uses KickStart or AutoYaST installation profile and related installation scripts to complete the installation and configuration of the compute node.
You can find the template and sample profiles in following directories:

```
/opt/xcat/share/xcat/install/<os>/
```

Commonly for installing the ppc64 compute node, you can use the compute.ppc64 profile.

If you want to customize the profile for compute node like <profile>.ppc64, you can copy the compute.ppc64 to the following directory, and make your modification base on it.

```
/install/custom/install/<os>/
```

Note: The profile name in the <profile> can be set to certain compute node by following command:

```
chdef -t node -o pnode1 profile=<profile>
```

4.2.1.1. Install the specific packages

If you want to install the specific package like specific.rpm onto the compute node, copy the specific.rpm into the following directory:

```
/install/post/otherpkgs/<os>/<arch>
```

4.2.1.2. [SLES 11]: Add the stunnel rpm from xCAT deps tarball

XCAT requires the stunnel rpm to be installed on the nodes to agent the transport between nodes and management node. The stunnel rpm is not shipped with SLES11, so you need to download the package and put it into the installation directory.

You can get the stunnle package for SLES11 from here:
```
http://xcat.sourceforge.net/yum/xcat-dep/sles11/ppc64
```

```
mkdir -p /install/post/otherpkgs/sles11/ppc64
copy the stunnel package to the /install/post/otherpkgs/sles11/ppc64
```

4.2.2. Set the node status to ready for installation

```
nodeset pnode1 install
```

4.2.3. Use network boot to start the installation

```
rnetboot pnode1
```

4.2.4. Check the installation results

1. Check that ssh service on the node is working and you can login without password
2. If ssh is working but cannot login without password, force exchange the ssh key to the compute node using xdsh:

```
xdsh pnode1 -K
```

After exchanging ssh key, following command should work.

```
xdsh pnode1 date
```

4.3. Stateless node installation

4.3.1. Generate the stateless image for compute node

Typically, you can build your stateless compute node image on the Management Node if it will have the same OS and architecture with the node. If you need another OS image or architecture than the OS installed on the Management Node, you will need a machine that meets the OS and architecture you want for the image and create the image on that node.

4.3.1.1. Make the compute node packaging list

If you want to exclude certain package, add it into the following exlist file:

```
/install/custom/install/<os>/<profile>.exlist
```

Add the packages name that need to be installed on the stateless node into the pkglist file

```
/install/custom/install/<os>/<profile>.pkglist
```

[RH]:

Add following packages name into the <profile>.pkglist

```
bash
nfs-utils
stunnel
dhclient
kernel
```

```
openssh-server
openssh-clients
busybox-anaconda
wget
vim-minimal
ntp
```

You can add any other packages that you want to install on your compute node. For example, if you want to have userids with passwords you should add the following:

```
cracklib
libuser
passwd
```

[SLES11]:

Add following packages name into the <profile>.pkglist

```
aaa_base
bash
nfs-utils
dhcpcd
kernel-ppc64
openssh
psmisc
wget
sysconfig
syslog-ng
klogd
vim
```

Add following package name into the <profile>.otherpkgs.pkglist

```
stunnel
```

Since SLES11 does not ship the stunnle package, you need to download it and copy it to /install/post/otherpkgs/sles11/ppc64.

You can get the stunnle package for SLES11 from here:

```
http://xcat.sourceforge.net/yum/xcat-dep/sles11/ppc64

mkdir -p /install/post/otherpkgs/sles11/ppc64
copy the stunnel package to the /install/post/otherpkgs/sles11/ppc64
```

4.3.1.2. Run image generation

[RHEL]:

```
cd /opt/xcat/share/xcat/netboot/rh
./genimage -i eth0 -n ibmveth -o rhels5.2 -p compute
```

[SLES11]:

```
cd /opt/xcat/share/xcat/netboot/sles
./genimage -i eth0 -n ibmveth -o sles11 -p compute
```

4.3.1.3. Pack the image

[RHEL]:

```
packimage -o rhels5.2 -p compute -a ppc64
```

[SLES]:

```
packimage -o sles11 -p compute -a ppc64
```

4.3.2. Set the node status ready for network boot

```
nodeset pnode2 netboot
```

4.3.3. Use network boot to start the installation

```
rnetboot pnode2
```

4.3.4. Check the installation result

1. Check that ssh service on the node is working and you can login without password
2. If ssh is working but cannot login without password, force exchange the ssh key to the compute node using xdsh:

```
xdsh pnode1 -K
```

After exchanging ssh key, following command should work.

```
xdsh pnode1 date
```

5. Firmware upgrade

5.1. Requirements

POWER5 and POWER6 Licensed Internal Code updates must meet the following prerequisites:

5.1.1. Enable the HMC to allow remote ssh connections.

[AIX]

Ensure that ssh is installed on the AIX xCAT management node. If you are using an AIX management node, make sure the value of "useSSHonAIX" is “yes” in the site table.

chtab key="useSSHonAIX" site.value=yes

5.1.2. Define the necessary attributes

The Lpar , CEC, or BPA has been defined in the nodelist, nodehm, nodetype, vpd, ppc tables.

5.1.3. Define the HMC as a node

Define the HMC as a node on the management node. For example,

```
nodeadd  hmc01.clusters.com groups=hmc
```

5.1.4. Setup SSH connection to HMC

Run the rspconfig command to set up and generate the ssh keys on the xCAT management node and transfer the public key to the HMC. You must also manually configure the HMC to allow remote ssh connections. For example:

```
rspconfig hmc01.clusters.com sshcfg=enable
```

5.1.5. Get the Microcode update package and associated XML file

Download the Microcode update package and associated XML file from the IBM Web site: http://www14.software.ibm.com/webapp/set2/firmware/gjsn.

5.2. Perform Firmware upgrade for CEC on P5/P6

5.2.1. Define the CEC as a node on the management node

Update the xCAT required xCAT tables:

Modify the nodelist table

```
nodeadd  Server-m_tmp-SNs_tmp groups=hmc,all
```

Modify the table nodehm

```
chtab node="Server-m_tmp-SNs_tmp" nodehm.mgt="hmc"
```

Modify the table nodetype:

```
chtab node="Server-m_tmp-SNs_tmp" nodetype.nodetype="fsp"
```

Modify the table ppc:

```
chtab node="Server-m_tmp-SNs_tmp" ppc.hcp=
hmc01.clusters.com
```

Modify the tab vpd:

```
chtab node=Server-m_tmp-SNs_tmp vpd.serial=s_tmp
vpd.mtm=m_tmp
```

Set the account of the HMC(Modify the ppchcp):

```
chtab hcp=hmc01.clusters.com  ppchcp.username=hscroot
ppchcp.password=abc123
```

5.2.2. Setup SSH connection to HMC

Generate the ssh keys on the xCAT management node and transfer the public key to the HMC to configure the HMC to allow remote ssh connections.

```
rspconfig hmc01.clusters.com sshcfg=enable
```

5.2.3. Check firmware level

```
rinv Server-m_tmp-SNs_tmp firm
```

5.2.4. Update the firmware

Download the Microcode update package and associated XML file from the IBM Web site: http://www14.software.ibm.com/webapp/set2/firmware/gjsn. Create the /tmp/fw directory, if necessary, and copy the downloaded files to the /tmp/fw directory.

Run the rflash command with the --activate flag to specify the update mode to perform the updates.（Please see the "rflash" manpage for more information）

```
rflash Server-m_tmp-SNs_tmp  -p /tmp/fw --activate
disruptive
```

NOTE:You Need check your update is concurrent or disruptive here!! other commands sample:

```
rflash Server-m_tmp-SNs_tmp  -p /tmp/fw --activate
concurrent
```

Notes:
1)If the noderange is the group lpar, the upgrade steps are the same as the CEC's.
2)System p5 and p6 updates can require time to complete and there is no visual indication that the command is proceeding.

5.3. Perform Firmware upgrades for BPA on P5/P6

5.3.1. Define the BPA as a node on the management node

Update the xCAT tables:

Modify the nodelist table. Define the BPA as a node

```
nodeadd  Server-m_tmps_tmp groups=hmc,all
```

Modify the table nodehm

```
chtab node="Server-m_tmps_tmp"  nodehm.mgt="hmc"
```

Modify the table nodetype:

```
chtab node="Server-m_tmps_tmp"  nodetype.nodetype="fsp"
```

Modify the table ppc:

```
chtab node="Server-m_tmps_tmp"  ppc.hcp= hmc01.clusters.com
ppc.id=x
```

Modify the tab vpd:

```
chtab node=Server-m_tmps_tmp  vpd.serial=s_tmp vpd.mtm=m_tmp
```

Set the account of the HMC(Modify the ppchcp):

```
chtab hcp=hmc01.clusters.com  ppchcp.username=hscroot
ppchcp.password=abc123
```

5.3.2. Setup SSH connection to HMC

Generate the ssh keys on the xCAT management node and transfer the public key to the HMC to configure the HMC to allow remote ssh connections.

```
rspconfig hmc01.clusters.com sshcfg=enable
```

5.3.3. User rinv to check the firmware level (see rinv manpage)

```
rinv  Server-m_tmps_tmp  firm
```

5.3.4. Update the firmware

Download he Microcode update package and associated XML file from the IBM Web site:

http://www14.software.ibm.com/webapp/set2/firmware/gjsn

Create the /tmp/fw directory, if necessary, and copy the downloaded files to the /tmp/fw directory.

Run the rflash command with the --activate flag to specify the update mode to perform the updates.

```
rflash Server-m_tmps_tmp  -p /tmp/fw --activate disruptive
```

NOTE:You Need check your update is concurrent or disruptive here!! other commands sample:

```
rflash Server-m_tmps_tmp  -p /tmp/fw --activate concurrent
```

5.4. Commit currently activated LIC update(copy T to P) for a CEC/BPA on p5/p6

5.4.1. Check firmware level

Refer to the environment setup in the section 'Firmware upgrade for CEC on P5/P6' to make sure the firmware version is correct.

5.4.2. Commit the firmware LIC

Run the rflash command with the –commit flag.

```
rflash  Server-m_tmp-SNs_tmp --commit
```

Notes:
(1)If the noderange is Lpar, the commit steps are the same as the CEC's.
(2) When the –commit or --recover two flags is used, the noderange cannot be BPA . It only can be CEC or LPAR ,and will take effect for both managed systems and power subsystems.