xCAT 2 InfiniBand Support

Date: 02/19/2009

Contents

Table of Contents

1. Setup IB support in xCAT

1.1. IB Configuration

XCAT provides one sample postscript configiba to config IB secondary adapter. This script can run on AIX and Linux managed nodes both.

The script configiba is stored in /opt/xcat/share/xcat/ib/scripts, so the user needs to manually copy configiba to /install/postscript. You can configure IB adapters either during node installation use the updatenode command after node installation.

To use this script, setup DNS for the new adapters:

1)The IP address entries for IB interfaces in /etc/hosts on xCAT managed nodes should have the node short hostname and the unique IB interface name in them. The format should be <ip_address_for_this_ib_interface node_short_hostname-ib_interfacename>.

For example:

```
c890f11ec01 is the node short hostname, c890f11ec01-ib0, c890f11ec01-
ib1, c890f11ec01-ib2, etc. are the IP names for the IB interfaces on
c890f11ec01.
```

2)Update networks table with IB sub-network

For example:

```
chtab net=172.16.0.0 networks.netname=ib0 networks.mask=255.255.0.0
networks.mgtifname=ib
```

Note: Attributes gateway, dhcpserver, tftpserver, and nameservers in networks table are not necessary to assign, since the xCAT management work is still running on ethernet.

3)On AIX, change the default connection between management nodes and compute nodes from ssh to rsh:

```
chtab key=useSSHonAIX site.key=no
```

4)If the computer node have already been installed and are runnning, make sure /etc/resolv.conf is available on the compute node before running updatenode, since configiba script will connect to name server to resolve IP address for the IB interfaces. If not, define /etc/resolv.conf on compute node or use rcp to copy resolv.conf from management node to the compute node. For example:

```
domain ppd.pok.ibm.com
search ppd.pok.ibm.com
nameserver  172.16.0.1
```

Note: 172.16.0.1 is the name server address which provide the IP addresses for IB interfaces on compute nodes.

5)Add the entries in the /etc/hosts into DNS and restart the DNS

Following is an example of /etc/hosts

```
192.168.0.10        c890f11ec01-ib0
192.168.0.11        c890f11ec01-ib1
```

For Linux Managed Nodes:

```
makedns
service named restart
```

For AIX Managed Nodes:

```
makedns
stopsrc -s named
startsrc -s named
lssrc -s named
```

Node: Make sure the state of named is active

6)Check if DNS for the IB network has been setup successfully

```
nslookup c890f11ec01-ib0
nslookup c890f11ec01-ib1
```

7)For RHEL and SLES, prepare the IB drivers/libraries for compute nodes. The packages have been listed in the below table.

Put IB drivers/libraries rpms under /install/post/otherpkgs/<os>/<arch> directory where <os> and <arch> can be found in the nodetype table.

Add rpm names (without version number) into /install/custom/install/<ostype>/profile.otherpkgs.pkglist, where <profile> is defined in the nodetype table. <ostype> is the operating system name without the version number. The following os types are recognized by xCAT.

```
centos
fedora
rh
sles
windows
```

IB Drivers and Libraries

On RHEL:

Driver/Library	Corresponding rpms in RHEL5.3

openib	*openib-*. el5.noarch.rpm*	
libib	**32bit**	*libibcm-*.el5.ppc.rpm* *libibcm-devel-*.el5.ppc.rpm* *libibcm-static-*.el5.ppc.rpm* *libibcommon-*.el5.ppc.rpm* *libibcommon-devel-*.el5.ppc.rpm* *libibcommon-static-*.el5.ppc.rpm* *libibmad-*.el5.ppc.rpm* *libibmad-devel-*.el5.ppc.rpm* *libibmad-static-*.el5.ppc.rpm* *libibumad-*.el5.ppc.rpm* *libibumad-devel-*.el5.ppc.rpm* *libibumad-static-*.el5.ppc.rpm* *libibverbs-*.el5.ppc.rpm* *libibverbs-devel-*.el5.ppc.rpm* *libibverbs-static-*.el5.ppc.rpm* *libibverbs-utils-*.el5.ppc.rpm*
	64bit	*libibcm-*.el5.ppc64.rpm* *libibcm-devel-*.el5.ppc64.rpm* *libibcm-static-*.el5.ppc64.rpm* *libibcommon-*.el5.ppc64.rpm* *libibcommon-devel-*.el5.ppc64.rpm* *libibcommon-static-*.el5.ppc64.rpm* *libibmad-*.el5.ppc64.rpm* *libibmad-devel-*.el5.ppc64.rpm* *libibmad-static-*.el5.ppc64.rpm* *libibumad-*.el5.ppc64.rpm* *libibumad-devel-*.el5.ppc64.rpm* *libibumad-static-*.el5.ppc64.rpm* *libibverbs-*.el5.ppc64.rpm* *libibverbs-devel-*.el5.ppc64.rpm* *libibverbs-static-*.el5.ppc64.rpm* *libibverbs-utils(it is used to ship ibv_* commands and depends on 32bit IB libraries) 64bit rpm is not available in RedHatEL5.3. Please install 32bit IB libraries also if user needs both ibv_* commands and the 64bit*

		libraries.
libehca (for Galaxy1/ Galaxy2 support)	**32bit**	*libehca-*.el5.ppc.rpm* libehca-static-*.el5.ppc.rpm
	64bit	*libehca-*.el5.ppc64.rpm* *libehca-static-*.el5.ppc64.rpm*
libmthca (for Mellanox InfiniHost support)	**32bit**	*libmthca-*.el5.ppc.rpm* *libmthca-static-*.el5.ppc.rpm*
	64bit	*libmthca-*.el5.ppc64.rpm* *libmthca-static-*.el5.ppc64.rpm*
libmlx4 (for Mellanox ConnectX support)	**32bit**	*libmlx4-*.el5.ppc.rpm* *libmlx4-static-*.el5.ppc.rpm*
	64bit	*libmlx4-*.el5.ppc64.rpm* *libmlx4-static-*.el5.ppc64.rpm*

On SLES:

Platforms	**Driver/Library**
SLES11	*ofed-1.4.0-3.4.ppc64.rpm*
SLES10	*libcxgb3-64bit-*.ppc.rpm* *libcxgb3-devel-*.ppc.rpm* *libcxgb3-devel-64bit-*.ppc.rpm* *libehca-*.ppc.rpm* *libehca-64bit-*.ppc.rpm* *libehca-devel-*.ppc.rpm* *libehca-devel-64bit-*.ppc.rpm* *libibcm-*.ppc.rpm* *libibcm-64bit-*.ppc.rpm* *libibcm-devel-*.ppc.rpm* *libibcm-devel-64bit-*.ppc.rpm* *libibcommon-*.ppc.rpm* *libibcommon-64bit-*.ppc.rpm* *libibcommon-devel-*.ppc.rpm* *libibcommon-devel-64bit-*.ppc.rpm*

	libibmad-.ppc.rpm* *libibmad-64bit-*.ppc.rpm* *libibmad-devel-*.ppc.rpm* *libibmad-devel-64bit-*.ppc.rpm* *libibumad-*.ppc.rpm* *libibumad-64bit-*.ppc.rpm* *libibumad-devel-*.ppc.rpm* *libibumad-devel-64bit-*.ppc.rpm* *libibverbs-*.ppc.rpm* *libibverbs-64bit-*.ppc.rpm* *libibverbs-devel-*.ppc.rpm* *libibverbs-devel-64bit-*.ppc.rpm* *libipathverbs-*.ppc.rpm* *libipathverbs-64bit-*.ppc.rpm* *libipathverbs-devel-*.ppc.rpm* *libipathverbs-devel-64bit-*.ppc.rpm* *libmlx4-*.ppc.rpm* *libmlx4-64bit-*.ppc.rpm* *libmlx4-devel-*.ppc.rpm* *libmlx4-devel-64bit-*.ppc.rpm* *libmthca-*.ppc.rpm* *libmthca-64bit-*.ppc.rpm* *libmthca-devel-*.ppc.rpm* *libmthca-devel-64bit-*.ppc.rpm* *librdmacm-1.0.6-*.ppc.rpm* *librdmacm-64bit-*.ppc.rpm* *librdmacm-devel-*.ppc.rpm* *librdmacm-devel-64bit-*.ppc.rpm* *libsdp-*.ppc.rpm* *libsdp-64bit-*.ppc.rpm* *libsdp-devel-*.ppc.rpm* *libsdp-devel-64bit-*.ppc.rpm* *mpi-selector-*.ppc.rpm* *mstflint-*.ppc.rpm* *mvapich2-*.ppc.rpm* *mvapich2-64bit-*.ppc.rpm*

	mvapich2-devel-.ppc.rpm* *mvapich2-devel-64bit-*.ppc.rpm* *ofed-1.3-*.ppc.rpm* *ofed-cxgb3-NIC-kmp-ppc64-*.ppc.rpm* *ofed-doc-*.ppc.rpm* *ofed-kmp-ppc64-*.ppc.rpm* *open-iscsi-*.ppc.rpm* *opensm-*.ppc.rpm* *opensm-64bit-*.ppc.rpm* *opensm-devel-*.ppc.rpm* *opensm-devel-64bit-*.ppc.rpm* *perftest-*.ppc.rpm* *qlvnictools-*.ppc.rpm* *rds-tools-*.ppc.rpm* *release-notes-as-*.ppc.rpm* *ruby-*.ppc.rpm* *sdpnetstat-*.ppc.rpm* *srptools-*.ppc.rpm* *tvflash-*.ppc.rpm*

8)Add this script to the xCAT postscripts table

```
chtab node=c890f11ec01 postscripts.postscripts=otherpkgs,configiba
```

```
Note: postscript otherpkgs is used to install IB libraries/drivers.
       please include other postscripts that you need.
```

9)Now all the preparation work for IB configuration has been done, user can use the updatenode command to update the nodes, or if the nodes have not been installed continue with the installation process.

Note: In the sample postscript, the netmask is set to default value: 255.255.0.0 and gateway is set to "X.X.255.254". If the IB interface name is not a simple combination of short hostname and ibX or netmask and gateway does not meet the user's requirement, then modify the sample script , like in the example below:

The node short hostname is 890f11ec01-en, and the IB interface name is 890f11ec01-ib0, c890f11ec01-ib1, etc. The user should modify as follows:

```
   my $hostname = "$ENV{NODE}-$nic";
```

to

```
   my $fullname = `echo $ENV{NODE} | cut -c 1-11`;
   chomp($fullname);
   my $hostname = "$fullname-$nic";
```

It is assumed every node has two IB adapters, if only one adapter is available on each node, modify script as following:

```
my @nums = (0..3);
```

to

```
my @nums = (0..1);
```

1.2. xdsh support for IB Devices

A new device configuration file on management node is introduced to allow the xdsh command to setup ssh, that is transfer the ssh keys to the IB device. The device configuration file is located in /var/opt/xcat/<DevicePath>/config.

The path to the device configuration file is parsed by xdsh from the attribute value of the "--devicetype" flag or the environment variable "DEVICETYPE" which is input to the xdsh call.

For example:

If the devicetype for Qlogic switch is "IBSwitch::Qlogic" then the device configuration file must be be found in the following directory:

```
/var/opt/xcat/IBSwitch/Qlogic/config
```

The following is an example of a device configuration file:

```
# Qlogic switch device configuration
[main]
ssh-setup-command=sshKey add
[xdsh]
pre-command=NULL
post-command=showLastRetcode -brief
```

Below is the explanation of the file attributes:

- ssh-setup-comand

Specify the ssh key appending command supported by device specified. If this entry is not provided, xCAT uses default ways for HMC and IVM-managed devices to write ssh keys of Management Nodes.

- pre-command

Specify the pre-execution commands before remote command. For example,

users might want to export some environment variables before executing real commands.

If the value of this entry is assigned "NULL", it means no pre-execution commands are needed.
For example, the Qlogic Switch does not support environment variable, the 'pre-command' is assigned with "NULL" to disable environment variables usage.

If no entry is provided, the default behavior is to export the environment variables that are normally exported by xdsh when running remote commands.

・post-command

Specify the built-in command provided by device specified to show the last command execution result. For example, the Qlogic Switch provides “showLastRetcode -brief” to display a numeric return code of last command execution.

If the value of this entry is assigned “NULL”, it means no post-command is used.

If no entry is provided, the default behavior to run “echo $?” used to dump return code of last command execution.

The default remote shell on the AIX management node is rsh. Changing the site useSSHonAIX attribute=yes will change the default to ssh for xdsh.

```
chtab key=useSSHonAIX site.key=yes
```

Define IB switches as a node, this is required by xdsh which only support the input as a node.

```
mkdef –t node –o c890f11ec01 groups=all nodetype=switch
```

You can use xdsh to configure ssh login to the IB device by running the following. Note you must use the correct userid for your device. After this configuration is complete, you will be able to login to the device without a password.

```
xdsh c890f11ec01 -K -l admin --devicetype IBSwitch::Qlogic

Enter the password for the userid on the node where the ssh keys will
be updated.

/usr/bin/ssh setup is complete.
return code = 0
```

After setup of the ssh keys for the login, the admin can run the commands on IB switches from the management node using xdsh.

Below is an example of using xdsh to list the current directory on the device.

```
/opt/xcat/bin/xdsh c890f11ec01 -l admin --devicetype
IBSwitch::Qlogic ls

export DEVICETYPE=IBSwitch::Qlogic /opt/xcat/bin/xdsh
c890f11ec01 -l admin ls
```

2. Sample Scripts

2.1. Annotatelog

annotatelog is a sample script to parse the QLogic log entries in file /var/log/xCAT/errorlog/[xCAT management nodes] on the xCAT Management Node by subnet manager, IB node, chassis, FRU(Field-Replaceable Unit) or a particular node. This script is supported on AIX and Linux management nodes.

From xCAT's point of view, the log to analyze must be xCAT consolidated log, which means this log file must come from xCAT syslog/errorlog monitoring mechanism, such as /var/log/xCAT/errorlog/[xCAT Management Nodes] file.

Since the log format is varies, xCAT does not support other log files.

The syntax of the annotatelog command will be:

annotatelog -f log_file [-s start_time] [-e end_time]
{ [-i -g guid_file -l link_file] [-S] [-c] [-u]| [-A -g guid_file -l link_file]}

{[-n node_list -g guid_file] [-E]}

[-h]

-f log_file

Specifies a log file fullpath name to analyze; must be xCAT consolidated log got from Qlogic HSM or ESM.

-s start_time

Specifies the start time for analysis, where the **start_time** variable has the format ddmmyyhh:mm:ss (day, month, year, hour, minute, and second), if it is not specified, annotatelog will parse the log file from the beginning.

-e end_time

Specifies the end time for analysis, where the **end_time** variable has the format ddmmyyhh:mm:ss (day, month, year, hour, minute, and second), if it is not specified, annotatelog will parse the log file to the end.

-l link_file

Specifies a link file fullpath name, which concatenates all '/var/opt/iba/analysis/baseline/fabric*links' files from all fabric management nodes.

-g guid_file

Specifies a guid file fullpath name, which has a list of GUIDs as obtained from the "getGuids" script.

-E

Annotate with node ERRLOG_ON and ERRLOG_OFF information. This can help determine if a disappearance was caused by a node disappearing. It is for AIX nodes only and should be used with -n or -i flag

-S

Sort the log entries by subnet manager only.

-i

Sort the log entries by IB node only.

-c

Sort the log entries by chassis only.

-u

Sort the log entries by FRU only.

-A

Output the combination of -i, -S, -c and -u. It should be used with -g and -l flags.

-n node_list

Specifies a comma-separated list of xCAT Managed Node host names, IP addresses to look up in log entries, it should be used with -g flag.

-h

Display usage information.

2.2. getGuids

getGuids is a sample script to get GUIDs for Infiniband Galaxy HCAs (Host Channel Adapter) and their ports from xCAT Management Nodes. It needs to be run on the xCAT Management Node. It will use a xdsh call to all the xCAT Managed Nodes to get the information about the IB devices. It uses the ibstat command on AIX system or ibv_devinfo command on Linux system to get the information about the IB devices.

The syntax of the getGuids command will be:

getGuids [-h] [-f output_file]

-f output_file

Specifies a file full path name that is used to save the GUIDs output.

-h

Display usage information.