XCAT 2 High Availability Management Node Setup

10/13/2010, 08:56:54 PM

Table of Contents

1.0 Overview....1
2.0Configuration Requirements....2
3.0Setup Primary Management Node....3
4.0Setup Standby Management Node....3
5.0 Setup Database Replication....4
5.1DB2 High Availability Disaster Recovery (HADR) Setup 4
5.1.1 Disconnect all the DB2 Clients....5
5.1.2 Setup configuration parameters for xcatdb....5
5.1.3 Backup xcatdb on Primary Management Node....5
5.1.4 Restore xcatdb on Standby Management Node....6
5.1.5 Configure HADR services ports....6
5.1.6 Configure the HADR Parameters....6
5.1.7 Start HADR....7
5.1.8 Verify HADR Status....8
5.1.9 Test Database Synchronization....9
5.1.10 Some useful HADR commands....10
5.2Database Replication for Postgresql....10
5.3Database Replication for Other Database Systems....11
6.0Files Synchronization....12
6.1SSL Credentials and SSH Keys....12
6.2Node Deployment Packages....12
6.3Network Services Configuration Files....13
6.4Additional Customization Files and Production files....13
7.0Cluster Maintenance Considerations....14
8.0Failover....15
9.0Failback....17
10.0 References....18

Note: If DB2 is used in your cluster, this documentation only works for xCAT 2.5 or beyond.

1.0 Overview

This documentation illustrates how to setup the second management node, or standby management node, in your cluster to provide high availability management capability.

With the primary management node fails, the administrator can easily have the standby management node take over the management node role, thus avoid long duration of a bad state that the cluster does not have management node available.

The xCAT high availability management node(HAMN) feature is not designed for automatic setup or automatic failover in this version, this documentation will describe how to synchronize various data between the primary management node and standby management node automatically, and describe how to perform some manual steps to have the standby management node takeover the management node role when fail occurs on the primary management node.

The primary management node will be taken down during the failover process, so any NFS mount or other network connections from the compute nodes to the management node should be temporary disconnected during the failover process, if the network connection is required for the compute nodes running, you should consider some other ways to provide high availability for the network services unless the compute nodes can also be taken down during the failover process. It also implies:

1. This HAMN approach is primarily intended for clusters in which the management node manages diskful nodes or linux stateless nodes, which also includes hierarchical clusters in which the management node only directly manages the service nodes, and they are diskful or linux stateless, the compute nodes in hierarchical can be anything.

2. This documentation is **not** primarily intended for clusters in which the nodes directly managed by the management node are linux statelite or aix diskless nodes, because the nodes depend on the management node being up to run its operating system over NFS. But if the nodes use only readonly nfs mounts from the MN management node, then you can use this doc as long as you recognize that your nodes will go down while you are failing over to the standby management node.

2.0 Configuration Requirements

xCAT HAMN requires the operating system version, xCAT version and database version are identical on the two management nodes.

The hardware type/model are not required to be the same on the two management nodes, but it is recommended to have similar hardware capability on the two management nodes to support the same operating system and have similar management capability.

Since the management node needs to provide IP services through broadcast such as DHCP to the compute nodes, the primary management node and standby management node should be in the same subnet to make the network services could work after failover.

The HAMN setup can be performed at any time during the life of the cluster. This documentation assumes the HAMN setup is performed from the very beginning of the cluster setup. You can skip the corresponding steps in case part of the setup has

already been done in your cluster.

The twin-tailed disks are not required, different methods are used to ensure the data synchronization between the primary management node and standby management node, however, if you have the twin-tailed disks in cluster, then the data synchronization will be easier, you can put the related directories and files mentioned in section **Setup Database Replication** and section **Files Synchronization** onto the twin-tailed disks, re-mount the twin-tailed disks to the standby management node during the failover, the corresponding steps to keep the data synchronized can be skipped.

The steps in this documentation are based on a real cluster environment:

Primary Management Node: aixmn1(9.114.47.103) running AIX 6.1L and DB2 9.7
Standby Management Node: aixmn2(9.114.47.104) running AIX 6.1L and DB2 9.7

You need to substitute the hostnames and ip address with yours when setting up your own HAMN environment.

3.0 Setup Primary Management Node

The procedure described in xCAT2 Top Doc can be used for the primary management node setup. If DB2 will be used as the xCAT database system, please refer to the doc Setup DB2 as the xCAT Database.

4.0 Setup Standby Management Node

The procedure described in xCAT2 Top Doc can also be used for the standby management node setup. The database system on the standby management node should be the same as the one running on the primary management node.

After the standby management node setup is done, perform the following additional configuration steps:

1. Make sure the primary management node can resolve the hostname of the standby management node, and vice versa.
2. Setup ssh authentication between the primary management node and standby management node. It should be setup as “passwordless ssh authentication" and it should work in both directions. The summary of procedure is:
   i. Cat key from /.ssh/id_rsa.pub on the primary management node and add to /.ssh/authorized_keys on the standby management node. Remove the standby management node entry from /.ssh/known_hosts on the primary management node prior to issuing ssh to the standby management node.
   ii. Cat key from /.ssh/id_rsa.pub on the standby management node and add to /.ssh/authorized_keys on the primary management node. Remove the

primary management node entry from /.ssh/known_hosts on the standby management node prior to issuing ssh to the primary management node.

3. Make sure the time on the primary management node and standby management node is synchronized. Some tips on setting up the timezone and time:
   i. Command echo $TZ returns the current timezone setting
   ii. Command date and chtz can be used to adjust the time and timezone.
   iii. To setup ntp on the management nodes on AIX:
      1. Update the /etc/ntp.conf file with a valid ntp server.
      2. stopsrc -s xntpd
      3. startsrc -s xntpd
      4. Use ntpq -p to show the peer status of the ntp server, should see * to left of server after successful association with server is established
4. Stop the xcatd daemon and DHCP service, for example, using commands stopsrc -s xcatd and stoprsc -s dhcpsd on AIX.
5. (Optional) Backup the xCAT database tables for the current configuration on primary management node, using command dumpxCATdb -p <backupdir>

5.0 Setup Database Replication

The most important data that needs to be kept synchronized on the primary management node and standby management node is the xCAT database. Most of the commercial database systems and some free database systems such as Postgresql and MySQL provide database replication feature, the database replication feature can be used for high availability capability. The configuration for database replication is quite different with various database systems, so this documentation can not cover all of the configuration scenarios. This documentation will focus on database replication configuration for DB2, and will also provide some documentation link for the replication setup for some other database systems. You can refer to the "Setup DB2 as the xCAT Database" document link at http://xcat.svn.sourceforge.net/viewvc/xcat/xcat-core/trunk/xCAT-client/share/doc/xCAT2SetupDB2.pdf for more details on how to setup DB2 as the xCAT database.

5.1 DB2 High Availability Disaster Recovery (HADR) Setup

DB2 High Availability Disaster Recovery (HADR) is a database replication feature that provides a high availability solution. HADR transmits the log records from the primary database server to the standby server. The HADR standby replays all the log records to its copy of the database, keeping it synchronized with the primary database server. Applications can only access the primary database and have no access to the standby database.

HADR communication between the primary and the standby is through TCP/IP, so the primary database server and standby database server do not need to be in the same subnet.

This documentation will only describe some basic configuration steps for HADR setup, there might be some configuration deviations in different cluster environment, please refer to the following links for more details:

1. Redbook "High Availability and Disaster Recovery Options for DB2 on Linux UNIX and Windows"
   http://www.redbooks.ibm.com/abstracts/sg247363.html
2. DB2 Information Center
   http://publib.boulder.ibm.com/infocenter/db2luw/v9r5/index.jsp?topic=/com.ibm.db2.luw.admin.ha.doc/doc/c0011748.html

Please be aware that all the DB2 commands in this section should be run as xcatdb unless otherwise noted.

5.1.1 Disconnect all the DB2 Clients

Before proceeding with the DB2 HADR setup, all the DB2 clients should be disconnected from the DB2 database sever. For xCAT environment, the only DB2 clients should be xcatd, so the xcatd on both management node and service nodes need to be stopped using command stopsrc -s xcatd.

5.1.2 Setup configuration parameters for xcatdb

Several configuration parameters need to be updated for HADR on both the primary management node and standby management node.

```
db2 UPDATE DB CFG FOR XCATDB USING LOGRETAIN ON
db2 UPDATE DB CFG FOR XCATDB USING TRACKMOD ON
db2 UPDATE DB CFG FOR XCATDB USING LOGINDEXBUILD ON
db2 UPDATE DB CFG FOR XCATDB USING INDEXREC RESTART
```

5.1.3 Backup xcatdb on Primary Management Node

The xcatdb on the primary management node and standby management node should be synchronized before setting up the HADR, otherwise, we will run into errors when trying to start HADR.

as root

```
mkdir /var/lib/db2/backup
chown xcatdb:xcatdb backup
```

as xcatdb

```
db2 BACKUP DB XCATDB TO /var/lib/db2/backup/
```

The command output will be something like:

```
Backup successful. The timestamp for this backup image
is : 20100805161232
```

Record the timestamp for later use, this timestamp is also part of the filename saved in /var/lib/db2/backup

Note: if you get an error, like "SQL1035N The database is currently in use. SQLSTATE=57019", make sure your xcatd daemons on management node and service nodes are not running, deactivating the xcatdb using command "db2 DEACTIVATE DB XCATDB" may also be helpful.

5.1.4 Restore xcatdb on Standby Management Node

Copy the xcatdb backup from the primary management node to standby management node:

```
scp -rp /var/lib/db2/backup xcatdb@aixmn2:/var/lib/db2/
```

Restore the xcatdb database:

```
db2 RESTORE DATABASE XCATDB FROM "/var/lib/db2/backup"
TAKEN AT 20100805161232 REPLACE HISTORY FILE
```

You will be prompted with the following question:

```
SQL2539W  Warning!  Restoring to an existing database that
is the same as the
backup image database.  The database files will be deleted.
Do you want to continue ? (y/n)
```

Answer: y

5.1.5 Configure HADR services ports

Add the following lines into /etc/services on both the primary management node and standby management node, you need to run as root to edit /etc/services.

```
DB2_HADR_1      55001/tcp
DB2_HADR_2      55002/tcp
```

5.1.6 Configure the HADR Parameters

Use the following commands to configure the HADR parameters:

On primary management node:

```
db2 UPDATE ALTERNATE SERVER FOR DATABASE XCATDB USING
HOSTNAME 9.114.47.104 PORT 60000
db2 UPDATE DB CFG FOR XCATDB USING HADR_LOCAL_HOST
9.114.47.103
```

```
db2 UPDATE DB CFG FOR XCATDB USING HADR_LOCAL_SVC
DB2_HADR_1
db2 UPDATE DB CFG FOR XCATDB USING HADR_REMOTE_HOST
9.114.47.104
db2 UPDATE DB CFG FOR XCATDB USING HADR_REMOTE_SVC
DB2_HADR_2
db2 UPDATE DB CFG FOR XCATDB USING HADR_REMOTE_INST xcatdb
db2 UPDATE DB CFG FOR XCATDB USING HADR_SYNCMODE NEARSYNC
db2 UPDATE DB CFG FOR XCATDB USING HADR_TIMEOUT 3
db2 UPDATE DB CFG FOR XCATDB USING HADR_PEER_WINDOW 120
db2 CONNECT TO XCATDB
db2 QUIESCE DATABASE IMMEDIATE FORCE CONNECTIONS
db2 UNQUIESCE DATABASE
db2 CONNECT RESET
```

On Standby management node:

```
db2 UPDATE ALTERNATE SERVER FOR DATABASE XCATDB USING
HOSTNAME 9.114.47.103 PORT 60000
db2 UPDATE DB CFG FOR XCATDB USING HADR_LOCAL_HOST
9.114.47.104
db2 UPDATE DB CFG FOR XCATDB USING HADR_LOCAL_SVC
DB2_HADR_2
db2 UPDATE DB CFG FOR XCATDB USING HADR_REMOTE_HOST
9.114.47.103
db2 UPDATE DB CFG FOR XCATDB USING HADR_REMOTE_SVC
DB2_HADR_1
db2 UPDATE DB CFG FOR XCATDB USING HADR_REMOTE_INST xcatdb
db2 UPDATE DB CFG FOR XCATDB USING HADR_SYNCMODE NEARSYNC
db2 UPDATE DB CFG FOR XCATDB USING HADR_TIMEOUT 3
db2 UPDATE DB CFG FOR XCATDB USING HADR_PEER_WINDOW 120
```

Substitute the IP addresses in the example with yours.

5.1.7 Start HADR

On standby management node, start HADR as the standby database:

```
db2 DEACTIVATE DATABASE XCATDB
db2 START HADR ON DATABASE XCATDB AS STANDBY
```

On primary management node, start HADR as the primary database:

```
db2 DEACTIVATE DATABASE XCATDB
db2 START HADR ON DATABASE XCATDB AS PRIMARY
```

If you get any message other than "*DB20000I The START HADR ON DATABASE command completed successfully*", make sure all the steps described above have been done correctly, or refer to the DB2 information center for troubleshooting.

5.1.8 Verify HADR Status

HADR can be in wrong state even if the "START HADR" command returns successfully, command "db2 GET SNAPSHOT FOR DB ON XCATDB" or "db2pd -d xcatdb -hadr" can be used to verify HADR status, the HADR status output is quite similar between these two commands, here is an example:

db2 GET SNAPSHOT FOR DB ON XCATDB

```
HADR Status
 Role                    = Primary
 State                   = Peer
 Synchronization mode    = Nearsync
 Connection status       = Connected, 08/05/2010
20:33:00.412948
 Peer window end         = 08/05/2010 21:03:07.000000
(1281013387)
 Peer window (seconds)   = 120
 Heartbeats missed       = 0
 Local host              = 9.114.47.103
 Local service           = DB2_HADR_1
 Remote host             = 9.114.47.104
 Remote service          = DB2_HADR_2
 Remote instance         = xcatdb
 timeout(seconds)        = 3
 Primary log position(file, page, LSN) = S0000002.LOG, 18,
000000000FA18D7C
 Standby log position(file, page, LSN) = S0000002.LOG, 18,
000000000FA18D7C
 Log gap running average(bytes) = 0
```

db2pd -d xcatdb -hadr

```
Database Partition 0 -- Database XCATDB -- Active -- Up 0
days 01:17:11

HADR Information:
Role    State                SyncMode HeartBeatsMissed
LogGapRunAvg (bytes)
Primary Peer                 Nearsync 0                 0

ConnectStatus ConnectTime                            Timeout
Connected     Thu Aug  5 20:33:00 2010 (1281011580) 3

PeerWindowEnd                          PeerWindow
Thu Aug  5 21:52:07 2010 (1281016327) 120

LocalHost                                 LocalService
9.114.47.103                              DB2_HADR_1
```

```
RemoteHost                                  RemoteService
RemoteInstance
9.114.47.104                                DB2_HADR_2
xcatdb

PrimaryFile  PrimaryPg  PrimaryLSN
S0000002.LOG 66         0x000000000FA4869D

StandByFile  StandByPg  StandByLSN
S0000002.LOG 66         0x000000000FA4869D
```

The attributes “Role”, “State” and “ConnectStatus” need to be checked, for an operating HADR environment, the “Role” should be “Primary” or “Standby”; the “State” should be “Peer” and the “ConnectStatus” should be “Connected”. If any of the attribute is not correct, you need to go back to check the HADR settings and try to restart the HADR, if the problem persists, refer to DB2 documentation or contact DB2 service team.

5.1.9 Test Database Synchronization

After the HADR setup is done, we should verify the database synchronization between the primary management node and standby management node. Here are the recommended steps:

On primary management node:

1) start xcatd, for example, using startsrc -s xcatd on AIX

2) Add a new testnode

3) stop xcatd, for example, using stopsrc -s xcatd on AIX

On standby management node:

1) Takeover as the HADR primary using command “db2 TAKEOVER HADR ON DATABASE XCATDB USER xcatdb USING cluster”

2) start xcatd, for example, using startsrc -s xcatd on AIX

3) Verify the testnode is in database and the node attributes are correct

4) Delete the testnode from database

5) stop xcatd, for example, using stopsrc -s xcatd on AIX

On primary management node:

1) Takeover as the HADR primary using command “db2 TAKEOVER HADR ON DATABASE XCATDB USER xcatdb USING cluster”

2) start xcatd, for example, using startsrc -s xcatd on AIX

3) Verify the testnode is not in the database

5.1.10 Some useful HADR commands

Besides the HADR related commands described above, there are still some other HADR commands that are useful for administration and debugging. When debugging errors, a good resource is the DB2 Information Center at http://publib.boulder.ibm.com/infocenter/db2luw/v9r7/index.jsp For example, the message SQL1117N can be found in Database reference > Messages > SQL Messages > SQL1000 - SQL1499

1. Stop HADR

   ```
   db2 STOP HADR ON DATABASE XCATDB
   ```

   Note: On the HADR standby database server, after the HADR is stopped, the database is in "ROLL-FORWARD PENDING" state and the xcatdb can not be activated, this error is returned: "SQL1117N A connection to or activation of database "XCATDB" cannot be made because of ROLL-FORWARD PENDING. SQLSTATE=57019", for the SQLSTATE=57019 , use the command "db2 ROLLFORWARD DATABASE XCATDB TO END OF LOGS AND COMPLETE" to fix this problem.

2. Check xcatdb configuration

   ```
   db2 CONNECT TO XCATDB
   db2 GET DB CFG
   ```

3. Takeover HADR role

   ```
   db2 TAKEOVER HADR ON DATABASE XCATDB USER xcatdb USING
   cluster
   ```

   OR

   ```
   db2 TAKEOVER HADR ON DATABASE XCATDB USER xcatdb USING
   cluster BY FORCE
   ```

The "BY FORCE" option should be used only if the primary database server is not functional.

5.2 Database Replication for Postgresql

Postgresql does provide feature "Continuous Archiving and Point-In-Time Recovery (PITR)" that can be used to provide high availability cluster configuration, see http://www.postgresql.org/docs/8.4/interactive/warm-standby.html and http://www.postgresql.org/docs/8.4/interactive/continuous-archiving.html for more details.

But this feature actually uses the "backup on the primary database server" and "restore on the standby database server", PITR is not real-time replication, the backup interval is configured manually in the postgresql.conf file, the recovery interval is configured in recovery.conf, it will save a lot of database logging files and the logging files take big amount of disk space, each logging file uses about 16MB disk space. Based on the considerations above, the database backup command

pg_dump and restore command pg_restore seem to be a better solution for the xCAT postgresql database replication.

On the primary management node
add crontab entries to:

1. dump the xcatdb into a file
2. scp the xcatdb backup file to the standby management node

Here is an example of the crontab entries for user postgres:

```
0 3 * * * /var/lib/pgsql/bin/pg_dump -f /tmp/xcatdb -F t
xcatdb
```

Here is an example of the crontab entries for user root:

```
0 4 * * * scp /tmp/xcatdb aixmn2:/tmp/
```

On the standby management node:
stop the xcatd and Postgresql.
AIX:

```
stopsrc -s xcatd
su - postgres
/var/lib/pgsql/bin/pg_ctl -D /var/lib/pgsql/data  stop
```

Linux:

```
service xcatd stop
su – postgres
service postgresql stop
```

Add crontab entry to restore the database, here is an example of the crontab entries for user postgres:

```
0 5 * * * /var/lib/pgsql/bin/pg_restore -d xcatdb -c
/tmp/xcatdb
```

5.3 Database Replication for Other Database Systems

This documentation will not cover the details for setting up replication for the database systems other than DB2, here are some useful links for setting up database replication for various database systems supported by xCAT.

MySQL: http://dev.mysql.com/doc/refman/5.5/en/replication.html

sqlite: sqlite does not provide replication feature, but since the sqlite is file-based database, so you can use file copy or synchronization mechanism on Unix/Linux to achieve the database synchronization.

6.0 Files Synchronization

To make the standby management node be ready for an easy take over, there are a lot files should be kept synchronized between the primary management node and standby management node.
A straightforward way to keep files synchronized is to use rsync, rsync is shipped with xCAT as part of the xcat-dep on AIX and also shipped with Linux distribution, you can see more details on the official rsync website http://samba.org/rsync/. The crontab can make the synchronization be automatic. This documentation will use the rsync and crontab as the files synchronization solution, you can use your own files synchronization solution as long as it could keep the corresponding files be synchronized between the primary management node and standby management node.

6.1 SSL Credentials and SSH Keys

The SSL credentials need to be identical on the primary management node and standby management node, the xcatd request submit from service nodes and compute nodes depend on the SSL credentials.
To setup the ssh authentication between the primary management node, standby management node, service nodes and compute nodes, the ssh keys should be kept synchronized between the primary management node and standby management node.
The SSL credentials reside in directory /etc/xcat/ca, /etc/xcat/cert and $HOME/.xcat/, the ssh keys are in directory /etc/xcat/hostkeys.

Here is an example of the crontab entries for synchronizing the SSL credentials and SSH keys:

```
0 1 * * * /usr/bin/rsync -az /etc/xcat/ca /etc/xcat/cert
/etc/xcat/hostkeys aixmn2:/etc/xcat
0 1 * * * /usr/bin/rsync -az $HOME/.xcat aixmn2:$HOME/
```

Note: you can backup the $HOME/.ssh directory in case some information from the $HOME/.ssh on the primary management node is needed after failover. This is an optional step:

```
0 1 * * * /usr/bin/rsync -az $HOME/.ssh
aixmn2:$HOME/sshbackup/
```

6.2 Node Deployment Packages

The node deployment packages are under the directory specified by the “installdir” in site table, the default location is /install directory. The node deployment packages need to be synchronized to the standby management node.
For Linux, it will be easy to achieve this by copying the whole /install directory from the primary management node to the standby management node; however, copying the whole /install directory is not enough for AIX, we will have to create the NIM

resources on the standby management node. Some manual steps are required to create the NIM resources on the backup management node.

Here is an example of the crontab entries for synchronizing the node deployment packages:

```
0 2 * * * /usr/bin/rsync -az /install aixmn2:/
```

If you do not want to do the manual steps on the standby management node to re-create the NIM resources, the AIX feature High Availability Network Installation Manager(HANIM) can be used for keeping the NIM resources synchronized between the primary management node and standby management node Please refer to AIX redbook "NIM from A to Z in AIX 5L" at http://www.redbooks.ibm.com/redbooks/pdfs/sg247296.pdf for more details about HANIM.

6.3 Network Services Configuration Files

A lot of network services are configured on management node, such as DNS, DHCP and HTTP, the network services are mainly controlled by the configuration files, however, some of the network services configuration files contain the local hostname/ipaddresses related information, so simply copying these network services configuration files to the standby management node may not work, considering that generating these network services configuration files are very easy and quick by running xCAT commands such as makedhcp, makedns or nimnodeset, as long as the xCAT database contains correct information. It will be easier to configure the network services on the standby management node by running xCAT commands when failing over to the standby management node. An exception is the /etc/hosts and /etc/resolve files, the /etc/hosts and /etc/resolv.conf may be modified as cluster maintenance activities, the /etc/hosts and /etc/resolv.conf are very important for xCAT commands, so the /etc/hosts and /etc/resolv.conf will be synchronized between the primary management node and standby management node. Here is an example the crontab entries for synchronizing the /etc/hosts and /etc/resolv.conf:

```
0 2 * * * /usr/bin/rsync -az /etc/hosts /etc/resolv.conf
aixmn2:/etc/
```

6.4 Additional Customization Files and Production files

Besides the files mentioned above, there might be some additional customization files and production files need to be copied over to the standby management node, depending on the users unique requirements. In a word, we should try to make the standby management node be a clone of the primary management node. Here are some example files that can consider:

/.profile
/.rhosts

/etc/auto_master
/etc/auto/maps/auto.u
/etc/hosts
/etc/motd
/etc/security/limits
/etc/resolv.conf
/etc/netscvc.conf
/etc/ntp.conf
/etc/inetd.conf
/etc/passwd
/etc/security/passwd
/etc/group
/etc/security/group
/etc/exports
/etc/dhcpsd.cnf
/etc/sevices
/etc/inittab
(and more)

7.0 Cluster Maintenance Considerations

The standby management node should be taken into account when doing any maintenance work in the xCAT cluster with HAMN setup.

1. Software Maintenance
   Any software update on the primary management node should also be done on the standby management node.
2. Files Synchronization
   Although we have setup crontab to synchronize the related files between the primary management node and standby management node, but the crontab entries are only run in some specific time slots, the synchronization delay brings in potential problems with HAMN, so it is recommended to manually synchronize the files mentioned in the section above whenever the files are modified.
3. Reboot management nodes
   In case the primary management node needs to be rebooted, the HADR will failover to the standby management node, to avoid the unnecessary failover, it is recommended to power off the standby management node before rebooting the primary management node. Rebooting the standby management node does not require additional steps.

At this point, the HA MN Setup is complete, and customer workloads and system administration can continue on the primary management node until a failure occurs. The xcatdb and files on the standby management node will continue to be synchronized until such a failure occurs.

8.0 Failover

When the primary management node fails for whatever reason, the administrator should start the failover process, some manual steps are involved in the failover process. The following procedure should be followed in the event of a failure on the primary management node.

1. Failover the database replication, use the description in the section “setup database replication” to failover the database replication to the standby management node if necessary. Take the DB2 HADR configuration as an example, there are two scenarios that require different procedure, if the outage is a known outage, this is where the standby management node takes over before the primary management node goes down, in this scenario, command "db2 TAKEOVER HADR ON DATABASE XCATDB USER xcatdb USING cluster" can be used to failover the HADR; if the outage is a unknown outage, this is where the primary management node remains in control until the primary management goes down, in this scenario, the command "db2 TAKEOVER HADR ON DATABASE XCATDB USER xcatdb USING cluster BY FORCE" can be used to failover the HADR, the “BY FORCE” option is required when the DB2 database on the primary management node is not functional.
2. Shutdown the primary management node
   If the primary management node is not totally dead, shutdown the primary management node. The standby management node could not take over the management role if the primary management node is still up, because the standby management node will be configured with the hostname and ip address that the primary management was configured. When the primary management node is shutdown, the nodes may no longer function, depending on the type of node installation. If xCAT is still active on the primary management node at this time, rpower and xdsh can be used to achieve shutting down the nodes if needed.
3. Stop the database system on the standby management node. Take DB2 as an example, the following commands can be used to stop the DB2:

   ```
   db2 STOP HADR ON DATABASE XCATDB
   ```

   Note: If you get error message SQL1769N Stop HADR cannot complete. Reason code = "2", try to run command

   ```
   db2 DEACTIVATE DATABASE XCATDB USER XCATDB USING cluster
   ```

   and then rerun the

   ```
   db2 STOP HADR ON DATABASE XCATDB

   db2 connect reset
   db2 force applications all
   db2 terminate
   db2stop
   ```

4. On the standby management node, change the ip address and hostname configured on cluster-facing adapters to the ones that were configured on the primary management node. Here is an example:
   ```
   /usr/sbin/mktcpip -h'aixmn1' -a'9.114.47.103'
   -m'255.255.255.192' -i'en1' -g'9.114.47.126' -t'N/A'
   ```
   Note: the mktcpip command will update /etc/hosts also, if it is not desired, you can use chdev command instead.
   It is recommended to open a console to Standby MN" prior to making any ethernet interface changes. Also, keep the console open, to observe any errors while issuing commands in the remainder of this section.
5. Update database configuration to use the new ip address and new hostname. For DB2, use the following command:
   ```
   re-login as xcatdb
   db2gcf -u -p 0 -i xcatdb
   ```
   This command will update the DB2 database configuration file "*/var/lib/db2/sqllib/db2nodes.cfg*" and start DB2.
   For Postgres, update the line "host all all x.x.x.x/32 md5" in file /var/lib/pgsql/postgresql.conf and update the line "listen_addresses = 'x.x.x.x'" in file /var/lib/pgsql/pg_hba.conf.
6. Start xcatd on the standby management node. If you get error "SQL1117N A connection to or activation of database "XCATDB" cannot be made because of ROLL-FORWARD PENDING. SQLSTATE=57019", use the following steps to workaround:
   i. Roll forward DB2 database, using:

      db2 ROLLFORWARD DB XCATDB TO END OF LOG
      db2 ROLLFORWARD DB XCATDB COMPLETE

   ii. Verify xcatdb is usable, via db2

      db2 CONNECT TO XCATDB USER XCATDB USING cluster

      db2 LIST TABLES
7. Setup the network services and conserver:
   i. DNS: run makedns. Verify dns services working for node resolution.
   ii. DHCP(Linux only): run makedhcp. Verify dhcp operational for hardware management.
   iii. conserver: makeconservercf
   iv. Verify that bootp is operational for booting the nodes.
8. Setup os deployment environment
   i. Create operating system images:
      1. For Linux: the operating system images definitions are ready in xCAT database, and the operating system images files are already in /install directory.

2. For AIX: If the HANIM is being used for keeping the NIM resources synchronized, then no manual steps are needed to create the NIM resources on the standby management node; otherwise, the operating systems images files are in /install directory, but have to create the NIM resources manually. Refer to the xCAT AIX documents listed at http://xcat.svn.sourceforge.net/viewvc/xcat/xcat-core/trunk/xCAT-client/share/doc/index.html for more details on how to create the NIM resources. Here are some manual steps that can be referred for re-creating the NIM resources:
    1. If the nim master is not initialized, run command "nim_master_setup -a mk_resource=no -a device=<*source directory*>" to initialize the NIM master, where the <*source directory*> is the directory that contains the AIX installation image files.
    2. Run "lsdef -t osimage -l" to list all the AIX operating system images.
    3. For each osimage:
        1. Create the lpp_source resource: /usr/sbin/nim -Fo define -t lpp_source -a server=master -a location=/install/nim/lpp_source/<*osimagename*>_lpp_source <*osimagename*>_lpp_source
        2. Create the spot resource: /usr/lpp/bos.sysmgt/nim/methods/m_mkspot -o -a server=master -a location=/install/nim/spot/ -a source=no <*osimage*>
        3. Check if the osimage has any kind of the following resources: "installp_bundle", "script", "root", "shared_root", "tmp", "home", "shared_home", "dump" and "paging". If yes, use commands "/usr/sbin/nim -Fo define -t <*type*> -a server=master -a location=<*location*> <*osimagename*>_<*type*>" to create all the necessary nim resources, where the <*location*> is the resource location returned by "lsdef -t osimage -l" command.

Note: If the NIM master was already up and running on the standby management node prior to failover, the NIM master hostname needs to be changed, you can use smit nim to perform the NIM master hostname change.

If you are seeing ssh problems when trying to ssh the compute nodes or any other nodes, the hostname in ssh keys under directory $HOME/.ssh needs to be updated.

ii. Run nodeset, nimnodeset or mkdsklsnode

9. Performing management operations
   After finished the step 8, the standby management node is ready for managing the cluster, you can run any xCAT command to manage the cluster. For example, if the diskless nodes need to be rebooted to boot from

network, you can run rpower <*noderange*> reset or rneboot <*noderange*> to initialize the network boot.

9.0 Failback

When the previous primary management node is back up and running, you may want to failback to the primary management node, since the xCAT database and related files are not up to date on the previous primary management node when it is down. So failing back to the the previous primary management node is not an easy action, you can go through all the steps described in this documentation to setup the previous standby management node as the new primary management node and setup the previous primary management node as the new standby management node, and then do a failover from the new primary management node to the new standby management node.

10.0 References

- Redbook "High Availability and Disaster Recovery Options for DB2 on Linux UNIX and Windows" http://www.redbooks.ibm.com/abstracts/sg247363.html
- DB2 Information Center http://publib.boulder.ibm.com/infocenter/db2luw/v9r5/index.jsp?topic=/com.ibm.db2.luw.admin.ha.doc/doc/c0011748.html
- Setup DB2 as the xCAT Database
- http://www.postgresql.org/docs/8.4/interactive/warm-standby.html
- http://wiki.postgresql.org/wiki/Replication,_Clustering,_and_Connection_Pooling