Operation Verification Results of ExpressCluster on IBM Power Systems

January 15, 2010
2nd Edition (revised July 12, 2010)

1. Introduction

The reduction of TCO is required in an IT environment, and server integration, which uses server virtualization technology, is one way to reduce power consumption and server setting spaces, and also to cut down costs by making effective use of hardware resources. IBM Power Systems enables to integrate multiple business applications into a few servers, by equipping a server virtualization function (PowerVM), which allows multiple operating systems to run on one server.
However, the same sign can be seen in the OS field, and many systems are opting to use Linux to save costs. As many companies and divisions are being integrated and consolidated to be optimized, Linux is making it easy to reuse the know-hows and to ensure engineers. Also, Linux supports various platforms, so it enables each system to choose the most suitable platform.

The main use of Linux used to be WEB front-ends or mail servers, but currently Linux plays the roll of infrastructures that can stably operate relational databases (RDBMS), which are the backbones of systems.

Even though the usable range of Linux on the Power System is growing, it did have a few issues. IBM Power Systems offers functions to maintain the high availability of each component, such as processors, memories and PCI adapters, but it does not support the failure of business applications running on each logical partition (LPAR). Also, IBM Power Systems cannot support failure when the power of the package completely stops. This issue has to be solved in order to use RDBMS as an infrastructure. RDBMS is the key of data maintenance, so it is extremely important to improve the availability of it.

To solve this issue, it is common to use HA clustering software, which switches the business to a different server when detecting a system error, but there were no HA clustering software that could fully bring out the functions of IBM Power Systems.

Virtual I/O Server (VIOS) is one of the virtualization functions (PowerVM) that IBM Power Systems offers. Using the VIOS makes it possible to share PCI devices between multiple LPARs, which enable the share of networks or external disks with the minimum necessary number of PCI devices. But usual HA clustering software depends on physical disk environments, so it could not control the external disks in VIOS configured environment.

One of NEC ExpressCluster's features is that it can work independently from the hardware environment, which can also be used in VIOS configured environment.

This document will report the results of a verification combining VIOS environment, Linux, RDBMS (DB2) and ExpressCluster over the IBM Power Systems.

2. Configuration Settings

2.1. Hardware Configuration

We have verified in a cluster configuration by using two physical servers. One LPAR have been configured on each server, the one is configured as an active node, and the other is as a standby node.

It is possible to configure physical and virtual adapters for networks or disk I/O that are used by an operating system and applications on each LPAR, and physical and virtual adapters can be on the same LPAR. In this case, we used virtual adapters which enable flexible configuration.
As for the shared disk connected to the servers configuring the cluster, we used a Storage Area Network (SAN) disk, and connected via SAN switch.

2.1.1. Virtual Adapter and Virtual I/O Server

Virtual adapter is one of the virtualization functions (PowerVM) that IBM Power Systems offers. A virtual adapter provides an interface for networks or disk I/O on the LPAR, and it enables you to create LPARs without any physical adapters. However, connection through a physical adapter is necessary for communicating to the external networks or accessing to physical disks such as SAN disks. The VIOS connects the virtual adapters configured on each LPAR with external networks or disks. Namely, it works as a bridge for virtual and physical, as you can see in figure1.

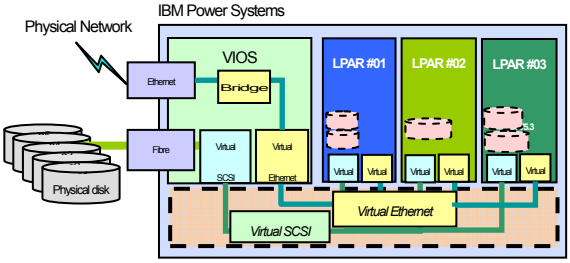

Figure 1: Virtual I/O Server Overview

Physical adapters connected to the networks or disks are configured in each VIOS partition, and VIOS works as a bridge for those physical adapters and the virtual adapters configured on each LPAR. It is possible to configure redundant physical adapters on the VIOS considering the availability and maintainability. Moreover, it is possible to multiplex the VIOS partition itself inside one package.

2.1.2. Disk Configuration

The disk configuration used in this verification is shown in figure2.

We configured a LPAR on each of the two Power Systems, and allocated disks for OS area and data area to each LPAR. RAID5 is configured on the external SAN disk, and both the OS area and the data area use the LUN configured on RAID5.

For the OS area, a disk is allocated for each LPAR individually. This OS area will not be the failover target of ExpressCluster. On the other hand, the data area will be selected as a failover target, so it will be configured as a shared disk between the LPARs configuring the cluster. In this verification, the data used by the database will be stored in this shared storage, so that the database can access to the same data after failing over the business application from the active server to the standby server. As for the details of database configuration, please refer to "Database configuration" in chapters 2.3 and 2.4.

Next is the disk allocation to LPAR in virtualization environments. Considering availability and maintainability, the VIOS in each package is configured redundantly. The LPARs configuring the cluster configure the disks (LUN) as multipath disks via the path supplied by each of these redundant VIOS Servers. Therefore, availability can be improved by configuring two VIO Servers in one package and configuring multipath which supplies paths from each VIOS to the same LUN, and allowing the LPAR to access to the disk via the other VIOS even if one of the VIO Servers has stopped.

(Ex ;) Even if "VIOS1-1" is in figure2 has been shuted down,, it is possible to access to RHEL5.4 and the data area via "VIOS1-2".

We configured redundant physical adapters which are used to access from each VIOS to the external SAN disk (in this verification test, it is a Fibre Channel (FC) adapter shown as "F" in figure2), and built a configuration which is able to tolerate physical adapter errors.

In this configuration, there are four paths for the LPAR routed through two VIO Servers to access to the same LUN, and it can provide a higher level of availability than configuring two physical FC ports on the LPAR. LPAR is able to keep on operating until all of four physical connections on the dual VIO Servers are disconnected.

2.1.3. Network Configuration

The network configuration used in this verification is shown in figure3. Likewise the disk configuration, we will use two Power Systems, configure a LPAR on each of them, and configure two paths of LANs (Public and Interconnect) on each LPAR. The Public LAN is used for services, and the Interconnect LAN is used for heartbeat which monitors whether the LPARs configuring the cluster are working normally or not.

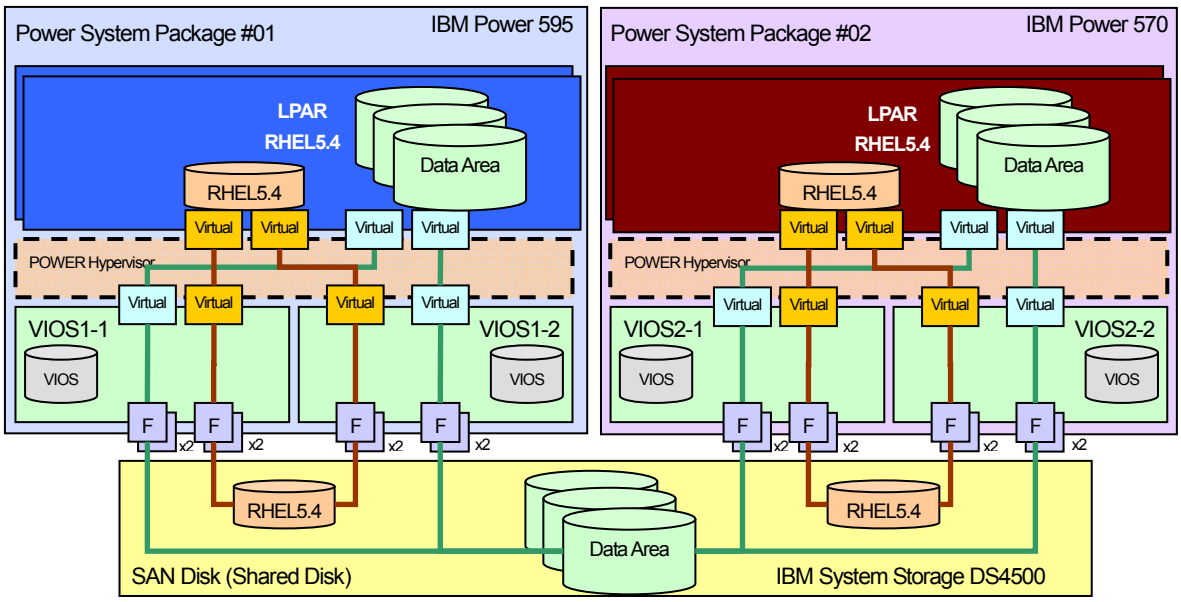

Figure 2:
Verification Environment Disk Configuration

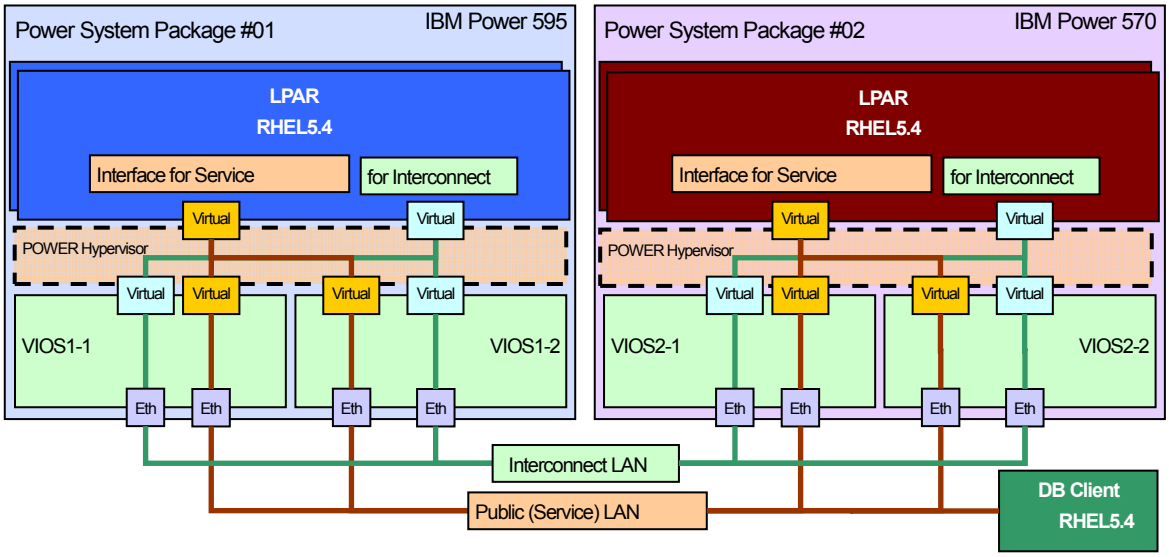

Figure 3:
Verification Environment Network Configuration

As for the network configuration on LPARs in virtualization environments, multiple routes will be provided via the dual VIO Servers in each package. Therefore, even if one VIOS stops, the virtual adapter configured on the LPAR can access to the external network via the other VIOS. It is also possible to multiplex the adapters allocated to the VIOS and configure EtherChannel or EtherChannel Backup considering the physical adapter's fault tolerance, but in this verification we did not configure either Ether Channel or EtherChannel Backup.

In this VIOS redundant configuration, the function to provide two routes from the LPAR is implemented in SEA Failover. You set the VIOS as "Primary" or "Backup". If the "Primary" VIOS stops, the network communication path will be switched to "Backup" automatically.. The SEA Failover function is provided by the VIOS, and the VIOS will monitor whether the other VIOS (Backup VIOS) is in normal status or not.

2.2. OS Configuration

We will install Red Hat Enterprise Linux 5.4 (RHEL5.4) for IBM POWER in LPAR, and configure drivers to access to the LUN (which is allocated by the shared disk) from multiple paths, make partitions, and define RAW devices.

2.2.1. Linux Installation

We will install RHEL5.4 in the LUN on the external disk storage, using the storage connection path multiplexed by the two VIO Servers. The following two points mainly differ from the normal RHEL5.4 installation process.

- We will use the Device Mapper Multipath (DMMP), bundled in RHEL5.4, when installing.
- We will use the VNC Viewer from a PC on the same network segment to install, controlling the GUI installer via network.

As for the VIOS, we will configure the followings.

- As for the two VIO Servers, we will configure the attribution "reserve policy" of the LUN's hdisk (where RHEL5.4 will be installed) as "no reserve".
- After installing RHEL5.4, we will assign the two vSCSIs to the boot list from LPAR's SMS.

Installation will be conducted in the following process
Installation will be conducted in the following procedure

1) Media Preparation

Connect the DVD drive to the install target LPAR, and set the RHEL5.4 install media.

2) LPAR activation

Activate the LPAR, select the DVD drive, and conduct "Normal Mode Boot" from the SMS menu.

3) Installer start

The installer's startup message will appear on the console, so enter the following options after "boot:" prompt. Assign "mpath" to use multipath driver DMMP. We will also use VNC, so we need to assign VNC related options. As for the IP address and password, enter the actual ones.

Example:

linux vnc vncpassword=xxx ip=10.7.10.123 netmask=255.255.0.0 mpath

4) VNC connection

A message requiring a connection from the VNC viewer will appear, so access to the GUI install screen using a VNC viewer from a PC on the same network. The login screen will appear, and if you enter the password you specified before as the vncpassword, the Anaconda (the GUI installer) screen will appear.

The following process is the same as the normal RHEL process, but be sure that the install target device is a multipath device.

- Display the confirmation screen when configuring the disk partition layout, and make sure that the device

name is “/dev/mapper/mpath0”.

- If the device name is different, such as “/dev/sda”, the multipath environment is not configured correctly.

Next, we will add “Service and productivity tools” provided by IBM, in order to use Power VM virtualization functions such as Dynamic LPAR and other functions such as reliability, availability, and serviceability (RAS) on RHEL5.4.

1) rpm package preparation

Choose the appropriate system’s link from the following website, and download the rpm package file from the RHEL 5 tab.
http://www14.software.ibm.com/webapp/set2/sas/f/lopdiags/home.html
In this configuration, we will choose “on HMC-or IVM-managed servers” from “Red Hat” category.

2) rpm package installation

Transfer the downloaded file to the install target LPAR, and install with the rpm command.
Example:
rpm -Uvh*.rpm

That is all for the installation process, but please refer to the following points to operate the RHEL5.4 environment installed with multipath driver DMMP.

A wwid number, which is a storage LUN’s unique ID, will be written automatically in to the following three files by the installer. If there is need to install to a new LUN in cases such as system backup and restore, it is necessary to rewrite this wwid to the new wwid.

1) /var/lib/multipath/bindings file

The association of multipath device mpathX and wwid is defined. In case of restoring to a different LUN environment, please change to the new wwid immediately.

2) /etc/multipath.conf file

The wwid of a LUN managed by DMMP is defined. In case of restoring to a different LUN environment, please change the “blacklist_exceptions” wwid to the new one.

3) /boot/initrd-XXX file

The wwid of a LUN which the DMMP configures as a multipath drive when starting the Linux, is defined in the start script file. In cases such as restoring in a different LUN environment, please reconfigure this file with the mkinitrd command, after rewriting the two files above.

2.2.2. Database Area Multipath Driver Configuration

When configuring database area’s multipath in a Linux environment, we will use multipath driver DMMP like we did when installing RHEL5.4, since we will use storage connecting paths multiplexed by the two VIO Servers.

Configuration will be conducted in the following procedure.

1) Confirm wwid

Find out the wwid of each LUN by using “scsi_id” command.
Example:
/sbin/scsi_id -g -u -s /block/sdc

2) Edit DMMP configuration file

Describe wwid in DMMP configuration file.
Add wwid to “blacklist_exceptions” in configuration file “/etc/multipath.conf”.

3) Create device file

Reflect the changes of the configuration file.
Create a device file (ex; /dev/mapper/mpath1) by executing “multipath” command and make DMMP recognize the LUN.

We will access to the LUN from database software and applications running on the OS (such as fdisk), by specifying the device file we created right now.

However, when not using the VIOS, you need to use multipath driver RDAC instead of DMMP. Please refer to the site below for details to get RDAC.
http://www.lsi.com/rdac/

In case of using RDAC in a cluster environment, a path switching movement like noted below will repeatedly occur when an error occurs in the main path.

1. The controller’s Ownership will switch on every LUN mapped to the node where main path error occurred. This includes LUNs used by other nodes.
2. Another node will detect the connection to the appropriate LUN via Preferred Path and return Ownership to Preferred Path.
3. The node where the main path error occurred will switch the Ownership again.

2. and 3. will keep on repeating, so a certain amount of error messages will be read out to the system log every minute. The switching itself will take only a few minutes so there is hardly any affect on performance, but it is possible to change the configuration to stop the repeating movement.

The following is the procedure to configure RDAC’s configuration item, “Disable LUN Rebalance”, and prevent Ownership to switch back when the primary path recovers. In an environment not using the VIOS, please configure as necessary.

1) Open “/etc/mpp.conf” from the editor, and change the value of DisableLUNRebalance to 3.

2) Use “mppUpdate” command and remake initrd image.

3) Reboot the system and start with the new initrd image.

2.2.3. Shared Disk Partition Configuration

In a HA clustering environment using ExpressCluster, a cluster partition, which is used to monitor between the clustered servers whether the other server is in normal status

or not, is necessary.

As for these partitions, it is possible to either configure the amount of LUNs required for partitions, map, and configure multipath, or divide one LUN into plural partitions. This time, we will adopt the latter method.

We will use RHEL5.4's fdisk command to configure partitions, but in a multipath environment, the fdisk command will be performed to the multipath device.
Example:
fdisk /dev/mapper/mpath1

Details of the use of each partition are described in "2.5.3 Disk Configuration".

2.2.4. RAW Device Definition

The cluster partition, which is crucial for ExpressCluster configuration, is needed to be defined as a RAW device from RHEL5.4.

We will use "rawdevices" service function, which is bundled with RHEL5.4, to define RAW devices.
The configuration procedure is the following.

1) Device definition
Define the relation of the actual device name and the RAW device name in the "rawdevices" configuration file.
Specify such as "/dev/raw/raw1 /dev/mapper/mpath1" in configuration file "/etc/sysconfig/rawdevices". Define for the number of all required RAW devices.

2) Device file generation
Generate a device file (ex:" /dev/raw/raw1") by executing "service rawdevices restart" command reporting the change of the configuration file to the kernel. This file will disappear when the OS is rebooted.

3) Auto generating configuration
Execute "chkconfig rawdevices on" command, and enable to generate the device file automatically after rebooting the OS too.

Other than using "rawdevices" service function, it is possible to define by using "udev" function. Please refer to the following website for details.
http://kbase.redhat.com/faq/docs/DOC-10164

As for ExpressCluster, we will specify the device files made here.

2.3. Database Configuration (DB2)

This chapter organizes information of when installing IBM DB2 (DB2) in ExpressCluster.

2.3.1. Installed Software

The installed software is the following.

- DB2 Package
  - DB2 v9.7 Enterprise Server Edition for POWER Linux (64bit)
- Components needed for DB2
  - Runtime for XL C/C++ Advanced Edition for Linux, V9.0 (C++ Runtime)
  - libaio, pdksh (Installed from OS CD-ROM)

Additional software needed to make DB2 work properly is organized in the following URL.
http://www.ibm.com/db2/linux/validate

2.3.2. Installation Plan (Physical Arrangements)

DB2 verification environment is shown in figure 4.

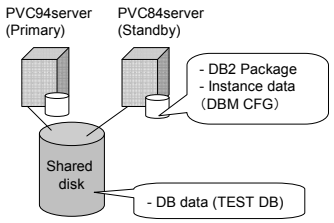

Figure4: DB2 Verification Environment

First of all, when installing DB2 in a shared disk type cluster environment, you need to decide whether to install the following components into the local disk or the shared disk.

DB2 package and instance data can be stored in either the local disk or the shared disk. DB data must be stored in the shared disk.

If you install both DB2 package and instance data in the local disk, it will enable to activate DB2 instance or apply a patch without mounting the shared disk.

This makes it possible to apply a patch (Fix Pack) to the standby server while running business applications on the active server. Generally, for a cluster system requirement, it is more likable to shorten downtime as much as possible, so this time we stored instance data in the local disk as shown in table1.

Table 1: Directory Configuration / Arrangement

Component	Disk Device	Path
DB2 Package	Local	/opt/ibm/db2/V9.7/
Instance Data	Local	/home/user name/
DB Data	Shared	/db2data/sharedb/

In this configuration, there are two reminders.

1) The directory to keep the DB data has to be completely empty. (No files existing.)
   In ia Linux environment, there are many cases that a LOST+FOUND directory is automatically configured in the device's top directory (this time it is the directory mounted as /db2data), so you can't directly store data there. This time, we dealt by configuring sharedb/ directory under it.

2) In case of changing the parameter at instance level, you need to update on both servers.
   This time, instance data is stored separately in both local disks, so if you changed the parameter at instance level, you need to make the same change on the other server and equalize the parameters.

Changing DBM CFG (Database Manager Config) and configuring or deleting database is equivalent to instance level parameter change.

2.3.3. Installation Plan (DB2 user ID)

You need to decide DB2 user ID before installing DB2. DB2 generally uses three user IDs.

- Instance user (Instance owner)
- Fenced user
- Administration server user (Admin)

Instance user is a user to manage the instance (DB2 program itself). Fenced user is used when executing stored procedure or user definition functions.

Management server user is necessary for starting and shutting down the management server, but it is not crucial for the DB2, so it is not configured this time. Therefore we designed the instance user and separate user as shown below.

	User (ID Number)	Password	Group(ID Number)
Instance User	db2inst1 (500) ※Port Number:50000	password	db2group (500)
Fenced User	db2fenc1 (501)	password	db2group (500)

The value is optional, but it is important to decide the ID number at this time. You can configure disk permission properly by equalizing the ID number on both servers.

Also, instance users will need one TCP/IP port number. Please choose a port number that is not used by other daemons (services), and bigger than 1,024. This time we assigned the default value 50,000.

2.3.4. Installation Procedure to the Primary

In this configuration, we need to install twice, to the primary and to the secondary, since we will install the DB2 package to the local disk. First, we will install to the primary (PVC94).

1) Prerequisite Software Installation
Install C++ Runtime and libaio,pdksh by executing rpm command.

2) DB2 package Installation
DB2 installer can be chosen from two types; db2setup which is in GUI format, or db2_install which is in command line format. This time we installed using db2_install, which can be used in any environment.

Extract the DB2 file (DB2_ESE_97_Linux_ipSeries.tar.gz) as a root user, and execute db2_install.
This time we assigned /opt/ibm/db2/V9.7 (default value) as the install directory, and selected ESE (Enterprise Server Edition) as the install product.

3) Instance User/Separate User Generation
First, register the TCP/IP port that the instance user will use, as we planned ahead. Add the following entry to /etc/services file.

```
db2c_db2inst1     50000/tcp
```
(db2c_db2inst1 is an arbitrary name)

Next, configure an instance user and a separate user on the OS.

```
# groupadd -g 500 db2group
# useradd -g db2group -u 500 db2inst1
# passwd db2inst1

# useradd   -g db2group -u 501 db2fenc1
# passwd db2fenc1
```

Configure DB2 instance using a user on the OS

#chmod 777 /tmp (Warning will appear if db2inst1 doesn't have access authority to /tmp)
#/opt/ibm/db2/V9.7/instance/db2icrt -p 50000 -u db2fenc1 db2inst1

4) Instance Configuration
We will configure the minimum required configuration to the instance.

```
# /opt/ibm/db2/V9.7/instance/db2iauto -off db2inst1
(Suspend auto-activation)
# su - db2inst1
> db2start
> db2set DB2COMM=TCPIP
> db2 update dbm cfg using SVCENAME db2c_db2inst1
(Configure to enable TCP/IP communication)
> db2stop
> db2start
```

(Reboot instance and reflect the parameter)
> exit

5) Mount Shared Disk to the Primary

We mounted as /db2data, using a root user account. Please don't describe the mount information in /etc/fstab. If you do, the OS will mount automatically and ExpressCluster will not be able to control the movement.

6) Configure a Directory to Store Data

Make a directory using a root user account, and change the owner of the directory to instance user (db2inst1:db2group). Now it is enabled to read and write data from DB2.

mkdir /db2data/sharedb/
chown db2inst1:db2group /db2data/sharedb/

7) Test Database Configuration

Make a database inside the directory we configured on the shared disk and check if it is connectable.

su - db2inst1
> db2 CREATE DB TEST1 ON /db2data/sharedb/
> db2 CONNECT TO TEST1
> db2 TERMINATE

Now we have finished DB2 installation to the primary. We will stop the instance.

> db2stop
> exit

2.3.5. Installation Procedure to the Standby

The installation procedure to the standby is the same as that to the primary. Perform primary's procedure 1) to 5). Do not perform 6) or after, since we already stored data in the shared disk.

We configured database on the primary, but not on the standby, so we need to make the standby recognize the existence of database, by executing the CATALOG command.

> db2 CATALOG DATABASE TEST1 ON /db2data/sharedb/

Now we are able to connect to the database from the standby too.

> db2 CONNETC TO TEST1
> db2 TERMINATE
> db2stop
> exit

2.3.6. DB2 Client Configuration

From DB2 client, we will execute the CATALOG command and make it recognize the database at the remote site. When using ExpressCluster, a floating IP address will be generally prepared at the server side, so we will assign the floating IP in CATALOG too.

1) Client Instance Configuration

Configure a client instance on the client machine.

(After installing DB2 client)
chmod 777 /tmp
groupadd -g 500 db2group
useradd -g db2group -u 500 db2inst1
passwd db2inst1
/opt/ibm/db2/V9.7/instance/db2icrt -s client db2inst1

2) CATALOG Remote DB

Specify IP address and port number of the remote DB by executing the CATALOG TCP/IP command. This time we used 172.16.11.4 for floating IP address.

su - db2inst1
> db2 CATALOG TCPIP NODE DB2SVR REMOTE 172.16.11.4 SERVER 50000
> db2 CATALOG DB TEST1 AT NODE DB2SVR
> db2 CONNECT TO TEST1 USER db2inst1 USING password
> db2 TERMINATE

2.3.7. Command to Control DB2 from the Cluster

Other than general commands that users normally use, DB2 has db2gcf commands, which are used to control DB2 from cluster software.

These commands are able to be used directly without assigning environment variables, and also able to specify timeout values, so you should use these commands when customizing the cluster software's control script.

Instance Start	db2gcf -u -i <Instance Name> -t <Timeout seconds> -L
Instance Stop	db2gcf -d -i <Instance Name> -t < Timeout seconds> -L
Instance Forced Stop	db2gcf -k -i <Instance Name> -t < Timeout seconds> -L
Instance Status Display	db2gcf -s -i <Instance Name> -t < Timeout seconds> -L

The DB2 manual includes details of db2gcf commands. (http://publib.boulder.ibm.com/infocenter/db2luw/v9r7/index.jsp?topic=/com.ibm.db2.luw.admin.cmd.doc/doc/r0010986.html)

2.3.8. Monitoring DB2 from ExpressCluster

There are two ways to monitor DB2 from ExpressCluster.

1) Trial database connection

ExpressCluster's DB2 monitor resource will automatically check if the database is in normal status or not, by actually

connecting to the database, and routinely executing the SQL (by making lists, for example).

2) Instance status monitoring

There are two ways to monitor the entire instance status; monitor the existence of db2sysc (db2syscs.exe in a Windows environment), which is the parent process of the whole instance, by executing ps command, or monitor the status from the db2gcf command. You can embed instance status monitoring in ExpressCluster by configuring a shell script to carry out the monitoring noted above, and to return a 0 if normal and something else if abnormal, and registering the script to ExpressCluster's custom monitor resource.

This time we set 1). As for DB2 module configuration, we need to set DB2 install directory (path to the library) specification and database character code.

In case of installing with the default setting, the directory will be /opt/ibm/db2/V9.7/ and the library will be directly under it, so assign either /opt/ibm/db2/V9.7/lib32/ or /opt/ibm/db2/V9.7/lib64/ according to the module. The ExpressCluster module for POWER Linux is 64bit binary, so we assigned /opt/ibm/db2/V9.7/lib64/.

As for character code, we will assign the same one as we assigned when creating database with DB2 (CREATE DATABASE). If you execute CREATE DATABASE without assigning anything, the default value is Unicode (UTF-8). We are using UTF-8 in this verification.

2.4. Cluster Configuration

In this configuration, we installed ExpressCluster in each LPAR configured on each Power Systems, and configured a Active-Active configuration (except for database, which is configured Active-Passive).

2.4.1. Installed Software

In this verification, we used the following products for clustering software.

- ExpressCluster Package
  - ExpressCluster X 2.1 for Linux
    - Update (CPRO-XL050-05) applied
- Database monitoring option
  - ExpressCluster X Database Agent 2.1 for Linux [1]

2.4.2. Network Configuration

In order to increase redundancy of the heartbeat path, ExpressCluster recommends using two or more network paths, including at least one interconnect LAN noted below.

- Interconnect LAN
  Used to monitor whether the other server is in normal status or not, or to exchange cluster information between the servers configuring a cluster. It is recommended not to be used for communication of other use. (Also referred to as heartbeat LAN)

- Public LAN
  Service LAN used to communicate with the client.
  Also used as a backup LAN for when an error occurs on the interconnect LAN

In this verification, we used total two LANs, one for interconnect LAN and one for public LAN.

2.4.3. Disk Configuration

We partitioned plural partitions on a same volume on the storage, and used each partition for the following use.

However, it is necessary to do some setting before installing ExpressCluster, so that these partitions will not be mounted automatically from the OS.

- Cluster partition
  Used for alive monitoring (disk heartbeat) between the servers configuring the cluster.
  In case of using a shared disk, it is recommended to configure more than one on each LUN, including the switchable partition.
  There is no need to configure a file system for this partition, since ExpressCluster's disk heartbeat resource will perform Read/Write access by RAW.
- Switchable partition (ext3 format)
  Used for storing business data.
  This area will be able to access only from the active server, by controlling access and mounting from ExpressCluster's disk resource.
  Supports major file systems such as ext3 and xfs. (But recommendation is a journaling file system.)
  In this verification, this partitionis used to store DB2 database.
- Switchable partition (RAW format)
  Used for storing business data accessed in RAW format.
  This partition will be able to access only from the active server, by being bound to the OS RAW device from ExpressCluster's RAW resource.
  This time, there were no applications using this RAW partition, so this was used as a dummy area for verification.

2.4.4. Heartbeat and Network Partition Solution

ExpressCluster performs heartbeat for keepalive monitoring between the servers.

If heartbeat can't be performed, the server will become a split brain situation and business suspension or at worst, data will be destroyed by a double mount of the shared disk.

[1] As for ppc version database monitoring, we used a module developed and supported to match this verification. To get this module, please refer to the end of this document for contact information. (DB2 and Oracle monitorings are supported)

Therefore, we employ the following structures to avoid these situations.

- Using multiple paths
  In this configuration, heartbeat will be able to continue operating even if a part of it becomes disconnected by hardware error, by using the paths below.
  - LAN heartbeat (using the two paths described in "2.5.2 Network Configuration")
  - Disk heartbeat (using the cluster partition described in "2.5.3 Disk Configuration")

- Using Kernel LAN heartbeat
  To perform LAN heartbeat surely in a high-load environment, each server performs heartbeat not only in user space but also in Kernel space too.

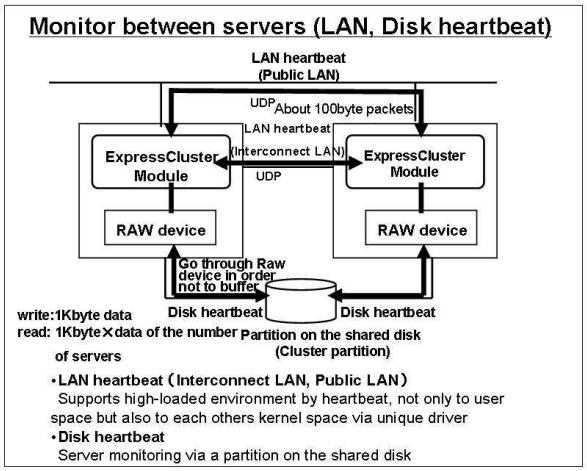

Figure 5: Heartbeat Path

Also, we employed the following structures to avoid data destruction even when an error occurs on all heartbeat paths.

- Ping network partition solution resource
  Register an external network equipment that is always activated and can be accessed from each server as a target, and if the server cannot communicate with that network equipment when the heartbeat from the other server disrupts, the server will forcibly shutdown its own server.
  Therefore, even if all of the heartbeat paths breakdown and the server gets into a split brain situation, the server which cannot communicate with the network equipment registered will spontaneously shutdown itself, so you can avoid double activation of the business and the data destruction that could have occurred from that.

2.4.5. Group resources

ExpressCluster enables to register each group resource as a failover target unit.

This time we registered the following group resources.

- Disk resource
  This provides the structure to mount the device existing on the shared disk from the active server.
  In this verification, we stored the DB2 database in the disk partition managed by this resource.
- Exec resource
  This provides the structure to execute scripts to activate/shutdown the services such as DB2.
  Also, the services to be activated/shutdown by this resource must be set as not to be automatically activated by the OS.
- Floating IP resource
  This provides virtual IP addresses. Therefore, clients can access to the server without considering which server is active.
- RAW resource
  Used for storing business data accessed in RAW format.
  This partition will be able to access only from the active server, by being bound to the OS RAW device from ExpressCluster's RAW resource.
  This time, although there were no applications using this RAW partition, this resource was provided for verification purpose.

2.4.6. Monitor resources

ExpressCluster enables to register each monitor resource corresponding to the monitoring target.

In this verification, we registered the following monitor resources.

- ARP monitor resource
  This provides the structure to deliver ARP packets for floating IP resource.
- DB2 monitor resource
  This provides the monitoring structure for DB2 database. Monitoring in perspective of services will be performed by executing SQL statements (create/drop/insert/update/select) to read and write dummy data toward the DB2 instance, and checking if the database responds properly.
- Disk monitor resource
  This provides the monitoring structure for the local disk and the shared disk.
- IP monitor resource
  This provides the monitoring structure for the conduction to external network equipments.
- RAW monitor resource
  This provides RAW device monitoring structure.
- User space monitor resource
  This provides user space stall monitoring structure.

3. Test Items and Results

3.1. Normal Status Check

We performed the following items, and confirmed that each performance had no problem in normal status.

- WebManager Connection

Items	Check Contents
WebManager Connection	Confirmed that the access to the WebManager (the manager to manage ExpressCluster) is enabled.

- Server Monitoring

Items	Check Contents
Server	Confirmed from the WebManager that the server status is in “normally operating”

- Heartbeat, Network Partition Resolution

Items	Check Contents
LAN heartbeat resource	Confirmed from the WebManager that the status of all the LAN heartbeat resources are in “normally operating” (both the heartbeat dedicated LAN and the public LAN)
Disk heartbeat resource	Confirmed from the WebManager that the status of disk heartbeat resources are in “normally operating”
Ping network partition resolution resource	Confirmed from the WebManager that the status of Ping network partition resolution resource is in “normally operating”

- Group, Group Resource

Items	Check Contents
Group	Confirmed from the WebManager that the status of all groups are in “normally operating”
Disk resource	Confirmed from the WebManager that the status of disk resource is in “normally operating”, and that it is mounted in the target mount point.
EXEC resource	Confirmed from the WebManager that the status of EXEC resource is in “normally operating”, and that the start script / stop script is executed when the resource starts/stops
Floating IP resource	Confirmed from the WebManager that the status of floating IP resource is in “normally operating”, and that the client can access to the active server with the floating IP address
RAW resource	Confirmed from the WebManager that the status of all RAW resources are in “normally operating”

- Monitor Resource

Items	Check Contents
Disk monitor resource	Confirmed from the WebManager that the status of disk monitor resource is in “normally operating”, and the monitor is not detecting any errors
IP monitor resource	Confirmed from the WebManager that the status of IP monitor resource is in “normally operating”, and the monitor is not detecting any errors
NIC Link Up/Down monitor resource	Confirmed from the WebManager that the status of NIC Link Up/Down monitor resource is in “normally operating”, and the monitor is not detecting any errors
PID monitor resource	Confirmed from the WebManager that the status of PID monitor resource is in “normally operating”, and the monitor is not detecting any errors
RAW monitor resource	Confirmed from the WebManager that the status of RAW monitor resource is in “normally operating”, and the monitor is not detecting any errors
User space monitor resource	Confirmed from the WebManager that the status of user space monitor resource is in “normally operating”, and the monitor is not detecting any errors
ARP monitor resource	Confirmed from the WebManager that the status of ARP monitor resource is in “normally operating”, and the monitor is not detecting any errors
DB2 monitor resource	Confirmed from the WebManager that the status of DB2 monitor resource is in “normally operating”, and the monitor is not detecting any errors

3.2. Error Status Check

We made a pseudo error occur to the following components assuming hardware or OS errors, and confirmed that there are no problems to either reaction.

- Shared Disk Device

Items	Action	Check Contents
Shared disk device SCSI/FC path	Unplug the FC cable from the server	- The disk monitor resource will detect the error - Disk heartbeat resource will stop - The group will failover to the standby server

- Network Path

Interconnect LAN	Unplug the interconnect LAN	- The heartbeat resource corresponding to the interconnect LAN will deactivate - The group will not failover, and continue running on the active server (Heartbeat will continue by using the public LAN as a backup heartbeat path)
Public LAN	Unplug the public LAN	- The heartbeat resource corresponding to the public LAN will deactivate - IP monitor resource will detect an error and the group will failover to the standby server

- VIOS

VIOS	Stop one of the two VIO Servers existing on the same server	- I/O will continue properly from the other VIOS - ExpressCluster will not detect an error and the group will continue activating on the active server

- Group Resource, Monitor Resource

Items	Action	Check Contents
Disk resource	Activate the group with the disk mounted	- Disk resource activation will fail and the group will failover to the standby server
EXEC resource	Configure a script ending with “exit 1”and activate the group	- EXEC resource activation will fail and the group will failover to the standby server
Floating IP resource	Assign the same address to another server in the same network, and activate the group	- Floating IP resource activation will fail and the group will failover to the standby server
RAW resource	Specify a RAW device that does not exist, and activate the group	- RAW resource activation will fail and the group will failover to the standby server
PID monitor resource	Crash one of EXEC resource’s resident process externally, using the kill command	- PID monitor error will occur, and the group will failover to the standby server

3.3. Failover Group Transition Check

We performed the following status transitions, and confirmed that the status of each server and each group will transit properly.

sequence	Items	Check Contents
1	Activate server 1 and server 2	The status of server 1 and server 2 will become “activated” The group will activate on server 1
2	Failover the group to server 2 manually	The group will failover to server 2
3	Failover the group to server 1 manually	The group will fail over to server 1
4	Execute a command to shutdown server 1	The group will failover to server 2
5	Perform cluster shutdown from the WebManager	The group will deactivate and server 2 will shutdown
6	Activate server 1 and server 2	The status of server 1 and server 2 will become “activated” The group will activate on server 1
7	Execute a command to shutdown server 1	The status of server1will become “deactivated” The group will failover to server 2
8	Execute a command to shutdown server 2	The group will deactivate
9	Activate server 1 and server 2	The status of server 1 and server 2 will become “activated” The group will activate on server 1
10	Perform cluster shutdown from the WebManager	The group will deactivate and server 1 and server 2 will shutdown
11	Activate server 2 [After the activation wait time of the other server (five minutes) passes]	The status of server 2 will become “activated” The group will activate on server 2
12	Activate server 1	The status of server 1 will become “activated”
13	Execute a command to shutdown server 2	The status of server 2 will become “deactivated” The group will failover to server 1
14	Activate server 2	The status of server 2 will become “activated”
15	Perform cluster shutdown from the WebManager	The group will deactivate and server 1 and server 2 will shutdown
16	Activate server 1 [After the activation wait time of the other server (five minutes) passes]	The status of server 1 will become “activated” The group will activate on server 1
17	Activate server 2	The status of server 2 will become “activated”
18	Execute a command to shutdown server 1	The status of server1will become “deactivated” The group will failover to server 2
19	Activate server 1	The status of server 1 will become “activated”
20	Perform cluster shutdown from the WebManager	The group will deactivate and server 1 and server 2 will shutdown
21	Activate server 1 and server 2	The status of server 1 and server 2 will become “activated” The group will activate on server 1

3.4. Database Action Check

We performed the following items to the DB2 cluster set, and confirmed that there are no problems to either performance

(In this chapter, DB2 Agent will be noted as “Database Agent”)

Items	Action	Check Contents
Activation	Activate the group from the WebManager manually	- The Database will activate, and the group status will become “activated” - Able to access to the database from a client terminal and process transactions properly
Normal Action	Same as above	- The status of Database Agent will become ready to monitor, and it will not detect an error
Deactivation	Deactivate the group manually from the WebManager	- The database will deactivate and the group status will change to “Deactivated”
DB Process Error	Crash the DB process externally, using the kill command	- The Database Agent will detect an error - The group will failover to the standby server, and can continue to process transactions from the client
Server Error	Shutdown the OS by an external command from the active server	- The group will failover to the standby server, and can continue to process transactions from the client

3.5. Test Result Summary

As shown above, we confirmed that the client can process transaction properly, by clustering DB2 on the Power Systems.

Also, as a result of making pseudo errors occur in perspective of hardware, OS, and application (DB2), we confirmed that failover will occur in each situation, and that the transaction continues to be provided to the client.

4. Conclusion

In this document, we verified DB2 redundant configuration on IBM Power Systems by installing ExpressCluster.

Therefore, the operation was performed as assumed in both normal status and abnormal status, and we were able to confirm that the availability of services can be improved by performing failover in cases of hardware, OS, or application errors.

Also, we verified that by adopting this configuration, you can get the following advantages, which are hard to realize when combining with other cluster software.

- Able to configure a cluster in a VIOS configuration with external disk control function
- By installing Database Agent, you can monitor the database action by not only a simple process existence monitoring, but also at service level.

5. References

- ExpressCluster
  http://www.nec.com/global/prod/expresscluster/
- IBM Power Systems
  http://www.ibm.com/systems/power/
- IBM PowerVM
  http://www.ibm.com/systems/power/software/virtualization/
- IBM Linux for Power Systems
  http://www.ibm.com/systems/power/software/linux/
- IBM DB2
  http://www.ibm.com/software/data/db2/

IBM Japan,Ltd.
Power System Division

NEC Corporation
IT Network Global Solutions Division
Software Group
info@expresscluster.jp.nec.com

Appendix

ExpressCluster Configuration Information

- DB2 Activation Script(start.sh)

```
#! /bin/sh
#****************************************
#*                start.sh                *
#****************************************

ulimit -s unlimited

if [ "$CLP_EVENT" = "START" ]
then
    if [ "$CLP_DISK" = "SUCCESS" ]
    then
        echo "NORMAL1"

# The value specified by "-t" is the timeout time to wait for instance
#process to start. Please specify the time that surpasses the
#largest amount of time required to finish activation
        /opt/ibm/db2/V9.7/bin/db2gcf -i db2inst1 -u -t 600
        if [ $? -ne 0 ]; then
            exit 1
        fi

        if [ "$CLP_SERVER" = "HOME" ]
        then
            echo "NORMAL2"
        else
            echo "ON_OTHER1"
        fi
    else
        echo "ERROR_DISK from START"
    fi
elif [ "$CLP_EVENT" = "FAILOVER" ]
then
    if [ "$CLP_DISK" = "SUCCESS" ]
    then
        echo "FAILOVER1"

# The value specified by "-t" is the timeout time to wait for instance
#process to start. Please specify the time that surpasses the
#largest amount of time required to finish activation
        /opt/ibm/db2/V9.7/bin/db2gcf -i db2inst1 -u -t 600
        if [ $? -ne 0 ]; then
            exit 1
        fi

        if [ "$CLP_SERVER" = "HOME" ]
        then
            echo "FAILOVER2"
        else
            echo "ON_OTHER2"
        fi
    else
        echo "ERROR_DISK from FAILOVER"
    fi
else
    echo "NO_CLP"
fi
echo "EXIT"
exit 0
```

- DB2 Deactivation Script(stop.sh)

```
#! /bin/sh
#****************************************
#*                stop.sh                 *
#****************************************

ulimit -s unlimited

if [ "$CLP_EVENT" = "START" ]
then
    if [ "$CLP_DISK" = "SUCCESS" ]
    then
        echo "NORMAL1"

# The value specified by "-t" is the timeout time to wait for instance
#process to start. Please specify the time that surpasses the
#largest amount of time required to finish deactivation
        /opt/ibm/db2/V9.7/bin/db2gcf -i db2inst1 -d -t 600
        if [ $? -ne 0 ]; then
            /opt/ibm/db2/V9.7/bin/db2gcf -i db2inst1 -k -t 600
            exit 1
        fi

        if [ "$CLP_SERVER" = "HOME" ]
        then
            echo "NORMAL2"
        else
            echo "ON_OTHER1"
        fi

    else
        echo "ERROR_DISK from START"
    fi
elif [ "$CLP_EVENT" = "FAILOVER" ]
then
    if [ "$CLP_DISK" = "SUCCESS" ]
    then
        echo "FAILOVER1"

# The value specified by "-t" is the timeout time to wait for instance
#process to start. Please specify the time that surpasses the
#largest amount of time required to finish deactivation
        /opt/ibm/db2/V9.7/bin/db2gcf -i db2inst1 -d -t 600
        if [ $? -ne 0 ]; then
            /opt/ibm/db2/V9.7/bin/db2gcf -i db2inst1 -k -t 600
            exit 1
        fi

        if [ "$CLP_SERVER" = "HOME" ]
        then
            echo "FAILOVER2"
```

```
            else
                echo "ON_OTHER2"
            fi

        else
            echo "ERROR_DISK from FAILOVER"
        fi
else
        echo "NO_CLP"
fi
echo "EXIT"
exit 0
```