Deploying DB2 pureScale Feature 10.5.0.3 on AIX with RDMA over Converged Ethernet

Aslam Nomani
DB2 Quality Assurance Manager
IBM Toronto

Gopi Varadarajulu (gvaradar@in.ibm.com)
Architect, DB2 Up and Running Development
IBM

Jimmy Stamkos (jstamkos@ca.ibm.com)
DB2 Software Developer
IBM

Ronen Grosman (roneng@ca.ibm.com)
Staff software developer
IBM

This tutorial is intended to show users how they can deploy DB2® pureScale® on AIX using Remote Direct Memory Access (RDMA) over Converged Ethernet. The step-by-step guide will provide details of a sample deployment that could be replicated to allow for other successful pureScale deployments.

Executive summary

In today's highly competitive marketplace, it's important to deploy a data processing architecture that not only meets your immediate tactical needs but that also provides the flexibility to grow and change to adapt to your future strategic requirements. Today, companies around the world are using the DB2 pureScale Feature to manage mission-critical data. IBM has brought the proven industry-leading technology and reliability of DB2 for z/OS® to open systems. The DB2 pureScale Feature is intended to meet the needs of many customers by providing:

- **Virtually unlimited capacity**—The ability to scale out your system by adding additional machines to your database system with ease.
- **Application transparency**—The ability to leverage your existing applications without changes.
• **Continuous availability**— By providing an active-active architecture for your database system with inherent redundancy.

• **Reduced Total Cost of Ownership (TCO)**— DB2 pureScale can help reduce TCO through providing the ability for a simplified deployment and management of advanced technology.

The DB2 pureScale Feature can be acquired through a simplified appliance model called IBM PureData® System for Transactions. This offering provides built-in database management along with integration of a factory-optimized hardware/software offering. Learn more about PureData System for Transactions.

For customers who have customized needs and prefer not to go with an appliance offering, this tutorial describes how to deploy the DB2 pureScale Feature. The scenario is based on a configuration using two P7 740 servers connected to a DS5300 storage controller.

For more information about supported servers and storage, please refer to Installing a DB2 pureScale environment in the DB2 V10.5 Knowledge Center.

**Introduction**

The DB2 pureScale Feature leverages an active-active shared-disk database implementation based on the DB2 for z/OS data sharing architecture. It leverages proven technology from the DB2 database software on the mainframe to bring the active-active shared-disk technology to open systems. Using DB2 pureScale offers:

• **Practically unlimited capacity**— The DB2 pureScale Feature provides practically unlimited capacity by allowing for the addition and removal of compute nodes (also called members on demand). DB2 pureScale can scale up to 128 members and has a highly efficient centralized management facility that allows for efficient scale-out. The DB2 pureScale Feature also leverages a technology called Remote Direct Memory Access (RDMA), which provides a highly efficient inter-node communication mechanism that also facilitates its scaling capabilities.

• **Application transparency**— An application that runs in a DB2 pureScale environment does not need to have any knowledge of the different members in the cluster or be concerned about partitioning data. The DB2 pureScale Feature will automatically route applications to the members deemed most appropriate. DB2 pureScale provides native support for a great deal of syntax used by other database vendors, allowing those applications to run in a DB2 pureScale environment with minimal or no changes. The benefits of DB2 pureScale can be leveraged in many cases without having to modify the application.

• **Continuous availability**— The DB2 pureScale Feature provides a fully active-active configuration such that if one member goes down, processing can continue at the remaining active members. During a failure, only data being modified on the failing member is temporarily unavailable until database recovery completes for that set of data, which is quick. This is in direct contrast to other competing solutions, where an entire system freeze may occur as part of the database recovery stage.

• **Reduced TCO**— The DB2 pureScale Feature can help reduce TCO through its integrated and simplified deployment and maintenance capabilities. The DB2 pureScale interfaces
handle the deployment and maintenance of components integrated within the DB2 pureScale Feature. This helps reduce what might amount to steep learning curves that would be associated with some of the competing technologies.

Some of the pureScale features in DB2 10.5 that allow it to truly deliver on the value propositions stated above include:

- **Online Fix-pack updates** — The ability to apply a DB2 fix pack while the database is still available for access.
- **High Availability Disaster Recovery (HADR)** — The ability for DB2 to automatically maintain a secondary copy of the database at another location that could be thousands of miles away.
- **Workload Balancing (WLB) and Client Affinity** — WLB offers the ability to automatically balance clients across all members or a specific set of members through a feature called Member Subsetting. Alternatively, Client Affinity allows users to prioritize which members they prefer their application to connect.
- **Automatic Client Reroute (ACR)** — ACR allows clients to automatically be reconnected to another member after a failure. In some cases, this failover can be completely seamless to the application.
- **Add a member online** — The ability to dynamically add a member to a cluster while the existing members remain online and available to process work.
- **Mobility of backup images** — Backup images of DB2 Enterprise Server Edition can be restored into a DB2 pureScale instance and vice versa.

To understand better how DB2 pureScale offers the above benefits, it is important to understand a little more of the architecture. The image below depicts the components that are part of a DB2 pureScale configuration. Even though there are multiple advanced components, a significant portion of this is transparent to the end user because DB2 pureScale internally deploys and manages these components.
Clients can initially connect to any member. DB2 pureScale can automatically load balance the clients across members based on machine utilization of each machine. If any host in the DB2 pureScale cluster fails, the DB2 pureScale Feature will redirect clients among the active members on the remaining hosts.

Each DB2 member represents a DB2 processing engine. Up to 128 members can be deployed in a single DB2 pureScale configuration. The members cooperate with each other and the Cluster Caching Facility (CF) to provide coherent access to the database from any member. Members can be added online and removed as processing demands change.

Integrated with DB2 pureScale is a cluster services layer that provides failure detection, recovery automation, and a clustered file system. These technologies are integrated within the DB2 pureScale Feature and leverage IBM technologies optimized for DB2 software. They include IBM Tivoli® Systems Automation for Multiplatforms, Reliable Scalable Cluster Technology, and General Parallel File System (GPFS™). DB2 pureScale automatically deploys and configures these technologies in accordance with a best-practice pre-defined configuration. A customer does not need to determine how to configure the clustering technology that comes with DB2 pureScale since that is kept transparent from the end user.

In the DB2 pureScale configuration, there is a need for communication between the members and the CFs. To make this communication as efficient as possible, DB2 pureScale Feature leverages the RDMA technology. RDMA allows one machine to read or write to the memory of another machine without requiring any processor cycles on the target machine. This mechanism in conjunction with high-speed networks, such as RDMA over Converged Ethernet (RoCE) and
InfiniBand, allows for an extremely efficient transport layer, which allows DB2 pureScale to scale efficiently and is a key differentiator from competing technologies.

The CFs provide a scalable and centralized locking mechanism to ensure data coherency. They also act as a fast cache for DB2 pages, leveraging RDMA technology to provide increased performance in situations where an expensive physical disk operation may otherwise have been required. The CF, along with the efficient transport layer, are two key components that allow DB2 pureScale to scale so well that each member does not have to negotiate with all other members when performing a task.

As discussed, DB2 pureScale leverages a shared-disk technology. Basically, any member can read or write to any portion of the database on the shared-disk system. If any member fails, the full set of data is still accessible from the other active members.

**Understanding redundancy of the DB2 pureScale Feature**

The DB2 pureScale Feature is designed to provide redundancy at the software and hardware levels. Redundancy is critical for high availability, so typical pureScale deployments tend to have redundant servers, networks, and storage. For example, the following identifies typical characteristics of a pureScale configuration:

- Two or more members across at two or more physical machines to protect against a member failure or a server failure
- Two CFs across two physical machines to protect against a CF failure or a server failure
- Fully redundant Ethernet network with multiple paths defined at each host leveraging multiple switches
- Fully redundant RoCE network with multiple paths defined at each host leveraging multiple switches
- Fully redundant storage network with multiple paths defined at each host leveraging multiple switches
- Disks defined within the storage leveraging the appropriate RAID settings

It should be noted that further levels of redundancy can be implemented through leveraging HADR which can allow for availability even after an entire site failure.

**Planning and sizing for DB2 pureScale Feature**

In preparing for the sizing of your environment, the first step would be to contact your IBM sales representative to work through a base sizing, and he will then leverage the IBM Techline process. After that, some of the basic guidelines from below can be applied to the configuration.

In addition to normal database deployment best practices common with Extensible Storage Engine (ESE), there are a few pureScale considerations.

**Storage**

For ultra-fast recovery, it is recommended that a validated pureScale storage controller that provides fast I/O fencing be used. Fast I/O fencing is used for quick fencing of failed members.
from storage for faster recovery times. Learn more about I/O fencing and pureScale storage considerations, check out Storage Considerations for pureScale in the IBM Knowledge Center.

File systems

The general guideline is to try to limit the number of file systems since each file system introduces a 1- to 2-second additional delay in recovery after any potential failures. A best practice is to have one GPFS file system for the instance home directory, one for the database table spaces, and one for the database logs.

By default, creating file systems using the \texttt{db2cluster} command implements a best practice configuration of the file system.

Capacity planning

CPU capacity

CF capacity is proportional to member capacity being used. In particular:

- \texttt{Write-heavy workload}: One CF physical core for every six member physical cores
- \texttt{Read-intensive workload}: 1 CF physical core for every 10 member physical cores

Thus, you could expect to have one CF physical core for every six to 10 member physical cores.

CF network capacity

\textbf{One CF RoCE adapter for every six to eight CF physical cores}

It should be noted that a minimum of two RoCE adapters is recommended for best availability.

CF memory

CF Memory = 40 percent of the sum of all memory on the members

Member network planning

\textbf{32 member physical cores per RoCE adapter}

It should be noted that a minimum of two RoCE adapters is recommended for best availability.

LPAR configuration

For best isolation, the CFs and member are placed on separate LPARs

For low-latency, high-performance communications, the CF employs a spinning model for communicating with the members where a thread continually spins in a loop waiting for requests from the members. To ensure appropriate CF response times, it is recommended to ensure that at least one physical core is allocated to each CF. One means of guaranteeing this LPAR gets priority to this CPU processing is by providing dedicated cores to the CF.
Deploying DB2 pureScale Feature

Configuration overview

In this scenario, we will deploy the DB2 pureScale Feature on a pair of P7 740 machines with each physical machine having two LPARs. Each LPAR will have the following characteristics:

- It will be on a public network that will allow for client connectivity.
- It will also be on a private RoCE network to allow for high-speed low-latency communication between members and CFs.
- It will have shared connectivity to a common set of disks.

Figure 2. Sample DB2 pureScale Feature hardware configuration

Table 1 details the high-level configurations of each LPAR.

Table 1. Configuration overview

<table>
<thead>
<tr>
<th>Hostname</th>
<th>ServerA</th>
<th>ServerC</th>
<th>ServerB</th>
<th>ServerD</th>
</tr>
</thead>
<tbody>
<tr>
<td>OS Level</td>
<td>AIX® 7.1 TL2 SP2</td>
<td>AIX 7.1 TL2 SP2</td>
<td>AIX 7.1 TL2 SP2</td>
<td>AIX 7.1 TL2 SP2</td>
</tr>
<tr>
<td>uDAPL Level</td>
<td>7.1.2.15</td>
<td>7.1.2.15</td>
<td>7.1.2.15</td>
<td>7.1.2.15</td>
</tr>
<tr>
<td>Server Type</td>
<td>Member 0</td>
<td>Primary CF</td>
<td>Member 1</td>
<td>Secondary CF</td>
</tr>
<tr>
<td>Cores</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>RAM</td>
<td>20 GB</td>
<td>16 GB</td>
<td>20 GB</td>
<td>16 GB</td>
</tr>
<tr>
<td>Shared Disks</td>
<td>hdisk5 – 25 MB Disk leveraged by DB2 Cluster Services layer</td>
<td>hdisk5 – 100GB disk used to create a shared file system to hold shared instance files across hosts</td>
<td>hdisk4 – 200GB disk used for DB2 data</td>
<td>hdisk3 – 20</td>
</tr>
</tbody>
</table>
DB2 pureScale Feature pre-installation steps

Prerequisites

Note: All commands listed during the installation steps will be executed as a user with root privileges unless specified otherwise.

1. Ensure that password-less ssh is set up within all the hosts participating in a DB2 pureScale cluster. You can validate the SSH configuration by issuing the following command from each host to every other host in the cluster and ensure that it returns the valid hostname without any prompting:

   # ssh <target host> hostname

2. You should have at least 5GB of free space in /tmp and /var file systems of each host. You will also require 6GB of free space in the installation path of each host.

3. Choose the two disks with sizes as described in the Shared Disks section in Table 1 that will be used for deploying the DB2 pureScale Feature and ensure that each is tagged with a physical volume identifier (PVID). These disks will be used in Step 14 below in the DB2 pureScale Feature Installation Steps section. One will be used for the Shared Instance Directory, with the second disk being leveraged by the Cluster Services Layer for internal operations.

   Tip: To validate the size of a disk you may run the bootinfo command.

   # bootinfo -s hdisk3

   a. The lspv command can list all physical volumes available on a host:

   # lspv
   hdisk0   00cc14e2ad7c570e   rootvg   active
   hdisk1   00cc14e2ba1a1dcd   homevg   active
   hdisk2   00cc14e2ebd1bce   othervg   active
   hdisk3   None               None
   hdisk4   None               None
   hdisk5   None               None
   hdisk6   None               None

   hdisk3 to hdisk6 do not have a PVID identified in the second column.
b. To add a PVID, run the `chdev` command from one host in the cluster:

```bash
# chdev -l hdisk3 -a pv=yes
# chdev -l hdisk4 -a pv=yes
# chdev -l hdisk5 -a pv=yes
# chdev -l hdisk6 -a pv=yes
```

On each of the other hosts in the cluster, run the following commands to remove the disk names.

```bash
# rmdev -dl hdisk3
# rmdev -dl hdisk4
# rmdev -dl hdisk5
# rmdev -dl hdisk6
```

Run the `cfgmgr` command from each host in the cluster: `# cfgmgr`.

c. Run the `lspv` command to display each hdisk's PVID that can uniquely identify the disk from any host within the cluster.

```bash
# lspv
hdisk0          00cc14e2ad7c570e        rootvg         active
hdisk1          00cc14e2bala1dcc        homevg         active
hdisk2          00cc14e2a1bicce          othervg        active
hdisk3          00cc14e23bb31033e        None
hdisk4          00cc14e23bb3b99ee        None
hdisk5          00cc14e23bb3e9dd        None
hdisk6          00cc14e23bb3e9df        None
```

4. Configure the RoCE network adapters and software required for DB2 pureScale Feature (see Appendix A).

5. Configure the 10G RoCE switch(es) required for DB2 pureScale Feature (see Appendix C).

6. Run the `db2prereqcheck` command in the installation media to verify that all prerequisites are met.

```bash
db2prereqcheck -p -v 10.5.0.3 -dl   hdisk2, hdisk3, hdisk4, hdisk5  -hl ServerA, ServerB, ServerC, ServerD
```

**DB2 pureScale Feature installation steps**

1. Identify one of the LPARs as the installation initiating host (IIH). This host will run the installation program for DB2 pureScale. Ensure that the display is set accordingly to display graphical tools. In this case, we will use ServerA as the IIH.

2. Navigate to the server folder on the downloaded product image, or the root directory of the product installation DVD, and as root user, run `db2setup`, which will launch the DB2 graphical installer: `#./db2setup`.

3. The execution of `db2setup` will take you to a welcome screen, and an information panel will be displayed, as shown below. A mouse click on the various topics on the left pane will provide more information. To install
the DB2 pureScale Feature, click **Install a Product** on the left pane.
4. On the next panel, click **Install New** under **DB2 Version 10.5 Fix Pack 3 Advanced Editions with DB2 pureScale**, and a progress indicator bar will appear at the bottom of the window.

5. The setup wizard will then appear and provide a set of intuitive instructions to navigate through the remainder of the deployment. Click **Next** to continue the install.
Optionally, you can click **View Features** to see what features will be installed.

6. The next panel will display the software licensing associated with DB2 pureScale. Read through the license, select the **Accept**
radio button if you accept the licensing terms, then click **Next**.
7. Select one of the install methods. In this example, we will install the DB2 pureScale Feature directly from the graphical interface by selecting the first option. Click **Next**.
8. Provide information for the installation directory of the DB2 binaries. In most cases, the default path will be fine. Click **Next** after determining the installation path.
9. Select the languages to install, then click **Next**.
10. Specify the location of the DB2 information center. In this case, we use the **On the IBM Web site** option. Click **Next**.
11. The next panel will ask whether to create a DB2 instance. In this case, we will select **Create a DB2 instance**, then click **Next**.

12. The next panel prompts for information regarding the DB2 instance owner in a similar manner to prior DB2 releases. Enter the information of the instance owner and click **Next**. **Note:** If an existing user is selected for the DB2 instance owner or the DB2 fenced user, the user must exist on all the hosts with the same UID, GID, group name, and $HOME path. The hosts should not share the instance owner's $HOME directory because that will be local to each host. If a new user creation is selected, the defined new user must not exist in any of the hosts.
13. The next panel will prompt for information regarding the fenced user ID. Enter the information and click **Next**.
14. The next steps use one of the predefined disks to create a shared file system used by the DB2 pureScale instance environment for instance files shared across all machines. The file system will be mounted as `/db2sd<timestamp>`, and the directory `/db2sd<timestamp>/<instance_name>` will be the default database path (as defined by the `DFTDBPATH` configuration parameter). Another small disk will be leveraged for automatic
internal cluster recovery purposes. Provide the full path to the disks and click **Next.**

**Set up a DB2 Cluster File System**

Setting up a IBM DB2 pureScale environment requires access to a shared file system. The embedded DB2 cluster file system will be used for this purpose.

A shared disk partition is required to store shared database data. This partition must be accessible by all hosts. The partition will be used exclusively by this DB2 copy and must not have a file system created.

Example of a device path is `/dev/hdisk10`.

**Shared disk partition device path**: `/dev/hdisk6`

**Mount point**: `/data/purescale/db2cluster`

**DB2 Cluster-Services Tiebreaker**

A shared disk partition is required for cluster recovery purposes. This partition should be at least 25 megabytes in size and accessible by all hosts. This partition must be different from the shared disk partition specified above.

**Device path**: `/dev/hdisk5`
15. On the next panel, add the remaining hosts that should be part of the DB2 pureScale cluster. By default, the IIH will already be included. For each host you need to add, click **Add**.

16. After clicking **Add**, you will be prompted for a host name. Enter the host name and click **OK**. Repeat the adding of a host for each host you need to add. Note that after each host is entered, the DB2 installation program will do validate to ensure that everything is configured appropriately for the new host. This validation might take a
few minutes, and it is normal as this validates all prerequisites, storage, and adapters.

17. After adding all the hosts, the host list window will appear as below with a check mark beside each host to confirm that it has been validated. When satisfied with the configuration identified in the bottom portion of the window, click **Next**. At this time, the DB2 installation
program will do one additional validation of the password-less SSH configuration.

If there is more than one physical machine in the configuration, the installation wizard will automatically assign each CF to a different machine. You can confirm the assignments in the Instance Settings section of the panel. To modify which hosts serve which function in terms of members or CFs, you can change it by clicking Advanced in the above window. It is strongly recommended to have two CFs identified on two distinct physical machines to ensure it does not create a single point of failure.

18. Click Verify Prerequisites to check that each host meets the prerequisites for installation. If there are any issues on the hosts, correct them at this point on the host side and
click **Verify Prerequisites** again. Once this is successful for all hosts, click **Next**.

19. After the DB2 installation program has completed all input validations, it will show a summary of the inputs before proceeding with the installation.
If you are satisfied with the inputs, click **Finish** to start the install.

20. During the actual installation process, a progress monitor bar will appear that will allow you to see progress of the installation. This will take several minutes as the DB2 pureScale Feature is being deployed to all hosts in the configuration, and all the components defined in the earlier sections are being deployed.

21. Upon successful completion of installing the DB2 pureScale Feature, the following window will appear and prompt you to click **Finish**. At this point, the installation of the
DB2 pureScale across all four hosts is complete, with a DB2 instance ready for use.

Adding additional ports to members and CFs

After the installation has completed, add cluster interconnect net names to the members and CFs to achieve the desired redundancy on network adapter ports. It is recommended that two or four ports be used for every member and CF. In this cluster, there are two network adapter ports that can be used for each host, so an additional one is added for each. The following commands are run.

From the db2sdin1 user:
> db2stop
10/02/2013 13:00:51  0   0   SQL1064N  DB2STOP processing was successful.
10/02/2013 13:00:52  1   0   SQL1064N  DB2STOP processing was successful.

From the root user:
#<DB2 Install Path>/instance/db2iupdt -update -cf serverD -cfnet serverD-ro0,serverD-ro1 -u db2sdfe1 db2sdin1
#<DB2 Install Path>/instance/db2iupdt -update -cf serverC -cfnet serverC-ro0,serverC-ro1 -u db2sdfe1 db2sdin1
#<DB2 Install Path>/instance/db2iupdt -update -m serverB -mnet serverB-ro0,serverB-ro1 -u db2sdfe1 db2sdin1
#<DB2 Install Path>/instance/db2iupdt -update -m serverA -mnet serverA-ro0,serverA-ro1 -u db2sdfe1 db2sdin1

DB2 pureScale Feature post-installation steps

With a DB2 pureScale Feature for Enterprise Server Edition instance ready for use, it is recommended to create additional file systems to use for the data and the logs. This file system creation can be done using db2cluster. The following commands are run from the root user:

1. Create one file system for data and one for logs.
   #<DB2 Install Path>/bin/db2cluster -cfs -create -filesytem data -disk /dev/hdisk4
   #<DB2 Install Path>/bin/db2cluster -cfs -create -filesytem log -disk /dev/hdisk3

   Note: The DB2 install path in this deployment would be /opt/IBM/db2/V10.5.
**Note:** The data and log file systems will by default be created under /db2fs and will be accessible on all hosts in the DB2 pureScale instance.

2. Modify the owner of the file system to be the DB2 instance owner so the DB2 instance owner will have full access to this file system. In this case, db2sdin1 is the instance owner's name and db2iadm1 is the instance owner's group name.

```bash
#chown db2sdin1:db2iadm1 /db2fs/data
#chown db2sdin1:db2iadm1 /db2fs/log
```

3. Start the DB2 instance by issuing `db2start`. You can see the state of the DB2 pureScale instance at any point by using the `db2instance` command.

```
> db2start
10/02/2013 13:50:00 1 0 SQL1063N  DB2START processing was successful.
10/02/2013 13:50:00 0 0 SQL1063N  DB2START processing was successful.
```

**Tip:** You can view the state of a DB2 pureScale cluster using the `db2instance -list` command.

4. Create the database and move the logs to the log file system. The following commands must be run from member hosts, not CF hosts.

```
> db2 create db testdb on /db2fs/data
> db2 update db cfg for testdb using newlogpath /db2fs/log
```

5. Catalog client connections to any active pureScale members and proceed with database requests.

**Understanding DB2 pureScale Feature usage**

There are many advantages that the DB2 pureScale Feature adds as was discussed in the Introduction. Below is more detail of some use cases that demonstrate the added value that DB2 pureScale provides. The steps documented in the introduction already demonstrate that the DB2 pureScale Feature can help to reduce TCO by the simplicity of deployment.

**Adding and removing members**

The DB2 pureScale Feature offers the ability to add members to the configuration quickly and without data redistribution requirements. The DB2 installation binaries are automatically stored on the IIH, not requiring access to the original install media when members are being added. A member can be added online by simply running the following command from any DB2 member:

```
db2iupdt -d -add -m ServerX -mnet ServerX-ro0,ServerX-ro1 db2sdin1
```

Members can be started or quiesced transparently to the application such that the application is unaware a change has occurred.

**Automatic workload balancing**

The DB2 pureScale Feature provides the ability to dynamically distribute a workload across all active members based on the machine utilization characteristics of the different machines. Multi-threaded CLI applications will by default have connection-level workload balancing without any changes. This workload balancing can be modified such that it applies at the transaction level as opposed to the connection level. For multi-threaded Java™ applications, the following can...
be changed in the connection string to take advantage of transaction-level workload balancing:

```
enableSysplexWLB=true
```

As additional members are started, clients will automatically route to the new member without any interruption of service. Also, members can be quiesced without the application knowing this operation has even occurred.

Clients can also be configured to have a preference to which member it should connect. This feature is referred to as *client affinity* and can be beneficial if a partitioned workload already exists.

**Note:** To take advantage of DB2 pureScale features such as transaction-level workload balancing or client affinity, the minimum client level should be 9.7 Fix pack 1 or the correlating JCC level.

**Online fix-pack updates**

In many cases, it is critical to apply maintenance to a system, but you don't want any negative impact to the client applications. Online or rolling fix-pack updates allows the user to apply DB2 fix-pack, machine, or OS-level updates to a given machine while the rest of the cluster is still available to process database requests. Allowing these maintenance activities to be performed while the database is still accessible further enhances the continuous availability characteristics of DB2 pureScale.

**High availability**

One of the significant value propositions of the DB2 pureScale Feature is the high-availability characteristics integrated into the architecture. All necessary resources are automatically monitored by the DB2 pureScale cluster services and restarted as needed. Applications connected to a failing member will automatically be rerouted to an active member, where the application can reissue any failed transactions. Applications connected to a non-failing component will not be impacted.

A differentiating factor of the DB2 pureScale Feature compared to other competing technologies is that upon a member failure, no cluster-wide freeze occurs. In fact, only data in the process of being updated on the failing member is temporarily unavailable until recovery is complete. Applications on active members trying to access the locked data on the failing member will be briefly in a lock-wait state, and by default will not receive any errors. The recovery will be completed quickly so that data availability through a member failure will look similar to the representation in the following figure.
Figure 4. Typical data availability pattern during member recovery

Disaster Recovery (HADR)

HADR is a tightly integrated data replication solution that provides a disaster recovery solution if a catastrophe happens on the primary site. HADR automatically maintains a secondary copy of the database by sending changes across the TCP/IP network from the primary site to the secondary site. The HADR feature compliments the integrated high availability inherent within pureScale to provide a continuously available solution necessary for mission-critical applications.

Conclusion

The DB2 pureScale Feature for Advanced Enterprise Server Edition provides a database solution that meets the needs of the most demanding customers. It is designed to leverage the CF and RDMA technologies, allowing it to scale effectively to meet the growing and dynamic needs of different organizations. Additional members can be added to the DB2 pureScale environment without any impact to existing applications to meet the demands of peak processing times. The DB2 pureScale Feature automatically balances the workload across all DB2 members in the cluster without any changes at the application side, taking full advantage of the additional processing capacity. If a DB2 member fails, applications will be automatically routed among the other active members. When the failed member host comes back online, applications will be transparently routed to the restarted member.

The DB2 pureScale Feature’s design and capabilities can help reduce total cost of ownership compared to other solutions. This is because the DB2 pureScale Feature allows for a simplified deployment and maintenance model. The installation of the DB2 pureScale Feature manages the deployment and configuration of all the bundled software components to all the hosts in the DB2 pureScale environment. Once the DB2 pureScale Feature environment is up and running, its operating status is monitored and maintained easily from any of the active members.

Contributors

Serge Boivin

DB2 LUW information development
Appendix A: Configuring 10GE RoCE

As discussed, the DB2 pureScale Feature leverages a 10GE RDMA over Ethernet (RoCE) network to allow for optimal communication between members and CFs. This appendix describes the high level steps of deploying RoCE.

1. Deploy the 10GE switch(es) and perform the necessary configurations. See Appendix C for the required switch configuration for RoCE with an example for a IBM System Networking RackSwitch switch.

2. From the hardware management console's graphical interface, assign the 10GE card so that each LPAR has a dedicated assignment. For each LPAR:
   a. Select the checkbox beside the LPAR.
   b. From the pop-up menu, select Configuration.
   c. Select Manage Profiles.
   d. Select the default profile.
   e. Select the I/O tab.
   f. Find the 10GE cards with the description "Ethernet controller."
   g. Ensure that the card that is to be assigned is in the correct Pci Gen 2 Slot.
   h. Click the checkbox beside the card and click Add as required to dedicate it to the LPAR.
      Note that RoCE does not have virtualization capabilities, and the 10GE card must be assigned as required and be dedicated to a particular LPAR. Note also that this card should not be assigned to the VIOs for RoCE, and must be assigned directly to the LPAR.

3. Shut down and then reactivate each of the LPARs from the HMC.

4. Download the required uDAPL and RoCE packages as documented in the DB2 pureScale Feature documentation, install these on all LPARs, and then reboot each LPAR.

5. Enable RoCE on each LPAR by running the smitty icm command and select Add an InfiniBand Communication Manager.

6. Set aside pseudo-IP addresses that are to be used for the RDMA (private) network. These may be any IP address that are not in any other reachable private or public network. These IP addresses will not be pingable once configured. They are not to be associated with regular Ethernet cards, and not to be configured with a network interface. In this tutorial, we assigned 10.1.1.1 on ServerA, 10.1.1.2 on ServerB, 10.1.1.3 on ServerC, and 10.1.1.4 on ServerD for the internet addresses, so for the first IP subnet, we use ServerA-ro0 for 10.1.1.1, ServerB-ro0 for 10.1.1.2, ServerC-ro0 for 10.1.1.3, and ServerD-ro0 for 10.1.1.4. For the second IP subnet, we use ServerA-ro1 for 10.1.2.1, ServerB-ro1 for 10.1.2.2, ServerC-ro1 for 10.1.2.3, and ServerD-ro1 for 10.1.2.4.

7. On each LPAR modify the /etc/dat.conf file to include the device, port and IP address corresponding to the RoCE hostname, similar to the following:
   ```
   hca0 u2.0 nonthreadsafe default /usr/lib/libdapl/libdapl2.a(shr_64.o) IBM.1.1 ""/dev/roce0 1 10.1.1.1"
   hca1 u2.0 nonthreadsafe default /usr/lib/libdapl/libdapl2.a(shr_64.o) IBM.1.1 ""/dev/roce0 2 10.1.2.1"
   ```

   In the above configuration, there are two host channel adapters. The first (hca0) is using /dev/roce0 on port 1, with IP address 10.1.1.1. The second (hca1) is using /dev/roce0, but on port 2, with IP address 10.1.2.1. Other hosts will be the same, except that the IP addresses
will match the RoCE IP addresses corresponding to what was set aside for on each host in Step 6.

8. Edit the `/etc/hosts` file on each LPAR similar to the following for routing purposes:

```
10.1.1.1    ServerA-ro0.<domain_name>    ServerA-ro0
10.1.1.2    ServerB-ro0.<domain_name>    ServerB-ro0
10.1.1.3    ServerC-ro0.<domain_name>    ServerC-ro0
10.1.1.4    ServerD-ro0.<domain_name>    ServerD-ro0
10.1.2.1    ServerA-ro1.<domain_name>    ServerA-ro1
10.1.2.2    ServerB-ro1.<domain_name>    ServerB-ro1
10.1.2.3    ServerC-ro1.<domain_name>    ServerC-ro1
10.1.2.4    ServerD-ro1.<domain_name>    ServerD-ro1
```

You can validate the state of the RoCE on each LPAR by running the `ibstat -v` command. Verify that port 1 and port 2 are active and the links are up.

**Appendix B: Configuring QDR InfiniBand**

As discussed, the DB2 pureScale Feature leverages an InfiniBand network to allow for optimal communication between members and CF servers. This appendix describes the high level steps of deploying QDR InfiniBand.

1. Deploy the InfiniBand switch(es), each with an active subnet manager.
2. From the hardware management console's graphical interface, assign the InfiniBand card so that each LPAR has a dedicated assignment. For each LPAR:
   a. Select the checkbox beside the LPAR.
   b. From the pop-up menu, select **Configuration**.
   c. Select **Manage Profiles**.
   d. Select the default profile.
   e. Select the **I/O** tab.
   f. Find the InfiniBand cards with the description "InfiniBand controller."
   g. Ensure that the card that is to be assigned is in the correct Pci Gen 2 Slot.
   h. Click the check box beside the card and click **Add as required** to dedicate it to the LPAR. Note that QDR InfiniBand does not have virtualization capabilities, and the InfiniBand card must be assigned as required and be dedicated to a particular LPAR. Note also that this card should not be assigned to the VIOs for InfiniBand, and must be assigned directly to the LPAR.
3. Shut down and reactivate each of the LPARs from the HMC.
4. Download the required uDAPL and InfiniBand packages as documented in the DB2 pureScale Feature documentation, install these on all LPARs, then reboot each LPAR.
5. Enable InfiniBand on each LPAR by running the `smitty icm` command and then selected "Add an InfiniBand Communication Manager."
6. Set up the InfiniBand network interfaces on each LPAR:
   a. Run the `smitty inet` command.
   b. Select **Change/Modify Characteristics**.
   c. Select `ib0`.
   d. Change the internet address and netmask, then select the appropriate HCA adapter (iba0) and the adapter's port number (1).
   e. Run the `smitty inet` command again for the second adapter port.
Deploying DB2 pureScale Feature 10.5.0.3 on AIX with RDMA over Converged Ethernet

f. Select **Add a Network Interface**.
g. Select **Add an IB Network Interface**.
h. Change the network interface name (ib1), change the Internet address and netmask, then select the appropriate HCA adapter (iba0) and adapter's port number (2).
i. Reboot all LPARs.

In this tutorial, we assigned 10.1.1.1 on ServerA, 10.1.1.2 on ServerB, 10.1.1.3 on ServerC, and 10.1.1.4 on ServerD for the internet addresses, so for the first IP subnet, we use ServerA-ib0 for 10.1.1.1, ServerB-ib0 for 10.1.1.2, ServerC-ib0 for 10.1.1.3, and ServerD-ib0 for 10.1.1.4. For the second IP subnet, we assigned ServerA-ib1 for 10.1.2.1, ServerB-ib1 for 10.1.2.2, ServerC-ib1 for 10.1.2.3, and ServerD-ib1 for 10.1.2.4. We used 255.255.255.0 for the netmask on each, and used the defaults for all other settings.

7. On each LPAR, ensure that the `/etc/dat.conf` file has a format similar to the following:
   
   ```
   hca0 u2.0 nonthreadsafe default /usr/lib/libdapl/libdapl2.a(shr_64.o) IBM.1.1 "/dev/iba0 1 ib0" " 
   hca1 u2.0 nonthreadsafe default /usr/lib/libdapl/libdapl2.a(shr_64.o) IBM.1.1 "/dev/iba0 2 ib1" " 
   ```

   In the above configuration, the first host channel adapter (hca0) is using `/dev/iba0` on port 1, and interface ib0. The second host channel adapter (hca1) is using `/dev/iba0` on port 2, and interface ib1.

8. Edit the `/etc/hosts` file on each LPAR similar to the following for routing purposes:

   ```
   10.1.1.1    ServerA-ib0.<domain_name> ServerA-ib0
   10.1.1.2    ServerB-ib0.<domain_name> ServerB-ib0
   10.1.1.3    ServerC-ib0.<domain_name> ServerC-ib0
   10.1.1.4    ServerD-ib0.<domain_name> ServerD-ib0
   10.1.2.1    ServerA-ib1.<domain_name> ServerA-ib1
   10.1.2.2    ServerB-ib1.<domain_name> ServerB-ib1
   10.1.2.3    ServerC-ib1.<domain_name> ServerC-ib1
   10.1.2.4    ServerD-ib1.<domain_name> ServerD-ib1
   ```

   You can validate the state of the InfiniBand on each LPAR by running the `ibstat -v` command. Verify that port 1 is active and the link is up. Also perform a ping test using the addresses and hostnames defined in the `/etc/hosts` file.

### Appendix C: RoCE switch configuration

Prior to configuration of the switch(es) for use with DB2 pureScale, please note that the switch used should support the following:

- Disablement of the Converged Enhance Ethernet (CEE) feature
- Enablement of Global Pause (IEEE 802.3x)
- Disablement of Spanning Tree Protocol (STP)
- Enablement Link Aggregate Control Protocol (LACP) for configurations with two switches
- Activation of all configured Inter-Switch Links (on both switches) at the same time for configurations with two switches

The following example shows how to configure an IBM System Networking RackSwitch G8124 switch from its default configuration to the DB2 required configuration.

1. First run the following to start configuration:
2. Connect to the switch and assign the switch an IP address, default gateway, and netmask to the switch. Commands to configure:

   interface ip-mgmt address <your new switch IP address>
   interface ip-mgmt netmask <mask>
   interface ip-mgmt enable
   interface ip-mgmt gateway address <your public gateway IP address>
   interface ip-mgmt gateway enable

Note that this step is optional but required if the switch is being connected on the public network.

3. Plug the cables into the ports on the switches and into the cards on the machines.

4. Disable spanning tree and enable flow control for:
   - Each port used by DB2 pureScale
   - Each port used for the inter-switch links, if there are two switches in the configuration

Command sequence to disable spanning tree:

RS G8124(config)#interface port <port number>
RS G8124(config-if)#no spanning-tree stp <port number> enable
RS G8124(config-if)#flowcontrol both

For example, if port 1 was to be configured:

RS G8124(config)#interface port 1
RS G8124(config-if)#no spanning-tree stp 1 enable
RS G8124(config-if)#flowcontrol both

The command to show the port config is like this:

RS G8124(config-if)#show interface port <port number>

For example:

RS G8124(config)#interface port 2
RS G8124(config-if)#no spanning-tree stp 2 enable
RS G8124(config-if)#flowcontrol both

Afterward, interface details will appear as such:
RS G8124(config-if)#show interface port 2
Current port 2 configuration: enabled, PVID 1
  ErrDisable recovery enabled
  STP: non-edge, auto link-type, no guard
  The DLF rate control currently turned off
  The Multicast rate control currently turned off
  The Broadcast rate control currently turned off
  802.1p priority: 0
  DSCP remarking for port: disabled
  BPDU guard: disabled
  Flood blocking: disabled
  MAC address notification: disabled
  L2 Learning: enabled
  ACL Port config is empty
  UDLD: disabled, mode normal
  OAM: disabled, mode active
  DHCP Snooping trust disable, limit rate: none
  VLANs: 1
Current Port 2 Gig link configuration:
  speed 10000, mode full duplex, fctl both, auto off

**fctl both** will appear to indicate that flow control both is enabled.

Another way to verify flow control is to show interface links:

```
RS G8124>show interface link
----------------------------------------------------------------------------
Alias Port Speed Duplex Flow Ctrl Link
------- ---- ----- -------- --TX-----RX-- -----
1      1    10000 full yes      yes up
2      2    10000 full yes      yes up
```

TX and RX under flow control will show yes for each configured port.

5. Disable spanning tree globally on each switch. Command to configure:

```
RS G8124(config)#spanning-tree mode disable
```

6. Configure LACP on inter switch links (ISLs) on each switch. Commands to configure:

```
RS G8124(config)#interface port <port number>
RS G8124(config-if)#lacp mode active
RS G8124(config-if)#lacp key 1
RS G8124(config-if)#exit
RS G8124(config)#interface port <second port number>
RS G8124(config-if)#lacp mode active
RS G8124(config-if)#lacp key 1
RS G8124(config-if)#exit
```

Where `<port number>` is one port number, and `<second port number>` is a second port.

Repeat the commands in this step above for additional ISLs.

Note that all ISLs that are connecting the same pair of switches used for DB2 should have the same key. If total VLAN separation is required between DB2 clusters on the same set of switches, use different ISLs, different IP subnets, and a different (single) VLAN for each of those clusters. Additional VLAN configuration is required for such a case.

Command to check if lACP is enabled:

```
RS G8124(config-if)#show lacp
```

For example, if ports 7 and 12 are designated for ISLs:
RS G8124(config)#interface port 7
RS G8124(config-if)#lacp mode active
RS G8124(config-if)#lacp key 1
RS G8124(config-if)#exit
RS G8124(config)#interface port 12
RS G8124(config-if)#lacp mode active
RS G8124(config-if)#lacp key 1
RS G8124(config-if)#exit

Then show LACP will show ports 7 and 12 active here with admin key 1:
RS G8124(config)#show lacp
Current LACP system ID: fc:cf:62:17:2e:00
Current LACP system Priority: 32768
Current LACP timeout scale: long
Current LACP params for 7:  active, Priority 32768, Admin Key 1, Min-Links 1
Current LACP params for 12:  active, Priority 32768, Admin Key 1, Min-Links 1

7. Save the configuration. Copy the running config to the startup config so that it is re-enabled on switch reboot. Command to save the configuration:
RS G8124(config-if)#copy running-config startup-config

Be sure to answer 'y' to any questions asked by this command.

Appendix D: Frequently asked questions

The DB2 pureScale Feature supports a variety of IBM storage, as well as storage provided by others vendors. To leverage the DB2 pureScale Feature with rapid fencing required for ultra-fast failover times, the disk storage should be validated with SCSI-3 persistent reserve with the underlying clustered file system. The storage subsystems validated with DB2 pureScale can be found at the following location: IBM Knowledge Center: Shared storage support for DB2 pureScale environments.
Resources

- The Information Management area on developerWorks provides resources for architects, developers, and engineers.
- Stay current with developer technical events and webcasts focused on a variety of IBM products and IT industry topics.
- Follow developerWorks on Twitter.
- Watch developerWorks demos ranging from product installation and setup demos for beginners, to advanced functionality for experienced developers.
- Get involved in the developerWorks Community. Connect with other developerWorks users while you explore developer-driven blogs, forums, groups, and wikis.
About the authors

Aslam Nomani

Aslam Nomani is a senior technical staff member within the IBM Canada Toronto Laboratory. He has worked with IBM for more than 15 years, with a focus on delivering best-of-breed quality. He is currently a quality assurance architect for IBM DB2 and cloud offerings.

Gopi Varadarajulu

Gopi Varadarajulu is a senior software engineer and architect for DB2 for LUW Deployment and Upgrade Development, and is responsible for DB2 First user experience. He had various design and engineering responsibilities for the past 10 years and has three patents in database management systems and operating systems.

Jimmy Stamkos

Jimmy has been working at IBM for 4 years in the DB2 Quality Assurance team. Over the past few years, Jimmy has focused on the quality of the DB2 pureScale Feature and validating high availability and AIX integration capabilities.

Ronen Grosman

Ronen Grosman is a member of the DB2 development team. He is currently working on DPF performance. Prior to his role as the lead developer of SACR, he worked on such features as Range Clustered Tables.

© Copyright IBM Corporation 2015
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)