Changing Physical Disk Storage in IBM DB2 pureScale: A Practical Example

IBM® DB2® pureScale® is a feature that provides continuous availability to multiple active servers that share a common set of disk storage subsystems. This feature includes multiple components, such as IBM Tivoli® System Automation (TSA) for high availability (HA) and General Parallel File System (IBM GPFS™) for clustered file systems.

Like any component in an IT infrastructure, DB2 pureScale eventually might require a migration to meet the demands of upgraded hardware and software.

Database administrators, operating system administrators, and system administrators working with DB2 pureScale in different capacities must understand how to upgrade the system and manage the changes that are required for the upgrade effectively by performing the correct steps.

This IBM Redbooks® Analytics Support web doc introduces a reference procedure and related considerations for storage migration that is based on a real-life test example. Database and system administrators can use this document as a reference when performing a storage migration in DB2 pureScale. This document applies to DB2 pureScale V9.8, V10.1, V10.5, and V11.

The DB2 pureScale system diagram that is shown in Figure 1 describes the migration of disks that include logs and data in to a new physical storage hardware.
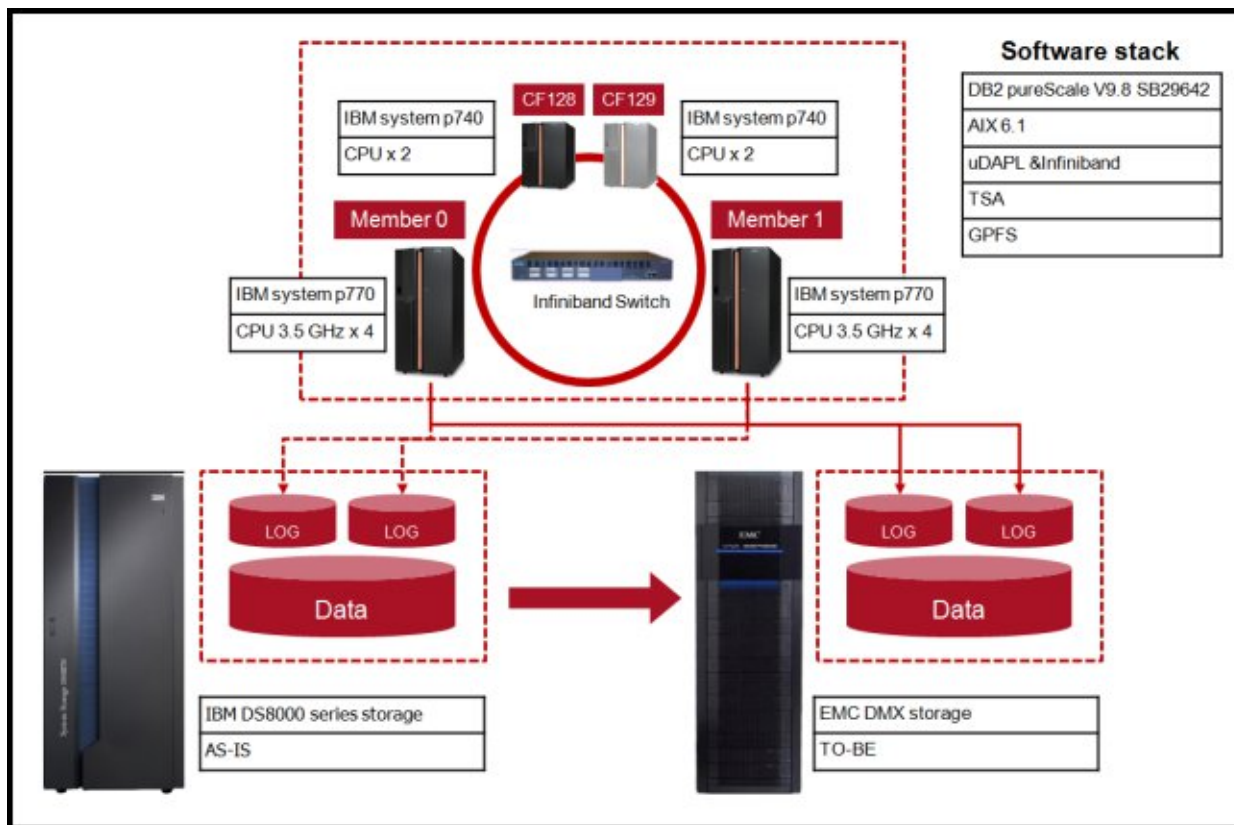

Figure 1. Physical disk storage change scenario diagram

The basic steps are based on the IBM Technote "Changing physical disk storage of GPFS file system" (found at http://www.ibm.com/support/docview.wss?uid=swg21633330). This web doc provides more details that help to answer the following questions:

- Can you add or remove multiple disk physical volumes by using one command?
- What is the potential impact of adding disks during an online transaction?
- Does GPFS file system auto-rebalancing happen automatically?
- How do you monitor the rebalancing status of a GPFS file system?
- What are the expected outputs of each step and the corresponding estimated elapsed time?

    Note: Elapsed times vary depending on the system resources and configuration. All the elapsed time information that is provided in this web doc is for reference only.

- How do you change GPFS and the TSA tiebreaker disk? Is it possible to implement this task while the system is online?

In this scenario example, we use an IBM DS8000® storage system as the source storage system and EMC DMX as the destination storage system.

The following sections describe the steps to take after you attach new EMC storage to DB2 pureScale cluster hosts.

Adding EMC disks to an existing GPFS file system that is configured with IBM Storage

Assume that you have a file system that is named db2data4 that has three 100 GB IBM disks, which are named gpfs229nsd, gpfs230nsd, and gpfs231nsd.

Now, you add three new EMC disks that should use the naming convention gpfs*XXX*nsd. (In this example, the disks are named gpfs237nsd, gpfs238 nsd, and gpfs239nsd). If you run the `db2cluster -cfs -add` command, you can add multiple physical volumes at once.

Run this command as root. You can find this command in `{DB2 installation path}/bin`.

```
# db2cluster -cfs -add -filesystem db2data4 -disk
/dev/hdisk210,/dev/hdisk211,/dev/hdisk212
The specified disks have been successfully added to file system 'db2data4'.
```

The elapsed time for this command to run in our scenario was 2 minutes and 7 seconds.

You can monitor the status of adding the disks by running the GPFS command `/usr/lpp/mmfs/bin/mmdf db2data4`. Figure 2 shows that new three disks are being allocated to the file system.

```
disk          driver   sector failure holds    holds                          storage
name          type     size   group metadata data  status       availability pool
------------- -------- ------- ------- -------- ----- ------------- ------------- ------------
gpfs229nsd    nsd        512     -1 yes      yes   ready        up            system
gpfs230nsd    nsd        512     -1 yes      yes   ready        up            system
gpfs231nsd    nsd        512     -1 yes      yes   ready        up            system
gpfs237nsd    nsd        512     -1 yes      yes   allocmap add up            system
gpfs238nsd    nsd        512     -1 yes      yes   allocmap add up            system
gpfs239nsd    nsd        512     -1 yes      yes   allocmap add up            system
disk                   disk size failure holds    holds                 free KB              free KB
name                   in KB     group metadata data          in full blocks       in fragments
------------- --------------- --------- -------- ----- -------------------- --------------------
Disks in storage pool: system (Maximum disk size allowed is 1.6 TB)
gpfs229nsd          104857600       -1 yes      yes        62303232 ( 59%)        30208 ( 0%)
gpfs230nsd          104857600       -1 yes      yes        62567424 ( 60%)        36000 ( 0%)
gpfs231nsd          104857600       -1 yes      yes        62228480 ( 59%)        29728 ( 0%)
gpfs237nsd          100669440       -1 yes      yes       100666368 (100%)          992 ( 0%)
gpfs238nsd          100669440       -1 yes      yes       100666368 (100%)          992 ( 0%)
gpfs239nsd          100669440       -1 yes      yes       100666368 (100%)          992 ( 0%)
                    ---------------                        --------------------  --------------------
(pool total)        616581120                             489098240 ( 79%)        98912 ( 0%)

                    ===============                        ====================  ====================
(total)             616581120                             489098240 ( 79%)        98912 ( 0%)
```

Figure 2. Monitor the addition of disks by running the mmdf command

In similar fashion, you can add new disks to the same or other file systems by running the same command. During the disk map information change, SQL applications running against this file system might be delayed for 5 - 20 seconds temporarily.

GPFS file system rebalancing check

After the disks are added to a file system, the data on existing disks should be rebalanced to the new disks.

Automatic GPFS rebalancing might happen, depending on the storage disk vendors that are added. For example, there were four findings during this case scenario test:

- When you add new EMC DMX disks on a file system that consists of IBM DS8000 storage disks, GPFS rebalancing does not start automatically.
- When you add IBM DS8000 storage disks on a file system that consists of IBM DS8000 storage disks, GPFS rebalancing starts automatically.
- When you add new EMC DMX disks and IBM DS8000 storage disks, after you add the new EMC DMX disks to a file system, GPFS rebalancing does not start automatically even if you add IBM DS8000 storage disks afterward.
- When you add new EMC DMX disks on a file system that consists of EMC DMX storage disks, GPFS rebalancing does not start automatically.

No matter what the storage product is, the necessary actions are clear:

- Check whether the rebalancing is in progress after you add the new disks.
- If the GPFS rebalancing is activated, let it run. Otherwise, start the GPFS rebalancing manually.

After you add the new EMC disks, the new disk status changes to `ready` and the free size on the disks is 100%, which means these disks are empty and that the GPFS rebalancing did not happen in this case (see Figure 3).

Figure 3. Check whether the GPFS rebalancing started

To start the GPFS rebalancing manually, run the following command as root:

```
# db2cluster –cfs –rebalance –filesystem db2data4
```

You can monitor the rebalancing progress by running `mmdf <file system name>` and checking the free space on disks.

Removing old IBM storage disks

After you add new EMC disks for GPFS file systems, the next step is removing old IBM disks. When you create a DB2 pureScale cluster, you configure the GPFS tiebreaker disk in the db2fs1 file system unless you change the location manually after you create it.

So, regarding the db2fs1 file system, remove the disk after you change the GPFS tiebreaker disk to another new disk on the db2fs1 file system or another independent disk. Changing the GPFS tiebreaker is described next.

In this section, the example commands show the removal of three disks from db2data4. All commands should be run as the root user. Run the following commands:

```
# db2cluster –cfs –remove –filesystem db2data4 –disk /dev/hdisk12
The specified disks have been successfully removed from file system 'db2data4'
```

The elapsed time for this command to run in our scenario was 6 minutes and 24 seconds.

```
# db2cluster –cfs –remove –filesystem db2data4 –disk /dev/hdisk13
The specified disks have been successfully removed from file system 'db2data4'
```

The elapsed time for this command to run in our scenario was 4 minutes and 56 seconds.

```
# db2cluster –cfs –remove –filesystem db2data4 –disk /dev/hdisk14
The specified disks have been successfully removed from file system 'db2data4'
```

The elapsed time for this command to run in our scenario was 5 minutes and 46 seconds.

Unlike when you add disk commands, only one disk can be specified with the `–remove` option, which means that disks can be removed only one at a time in a GPFS file system. To save time, multiple commands can be run against multiple file systems. For example, here are some sample commands running in parallel:

- To remove a disk from the db2data1 file system on the DB2 pureScale member 0 host, run the following command:

  ```
  # db2cluster –cfs –remove –filesystem db2data1 disk /dev/hdisk74
  ```

- To remove a disk from the db2data2 file system on the DB2 pureScale member 1 host, run the following command:

  ```
  # db2cluster –cfs –remove –filesystem db2data2 disk /dev/hdisk100
  ```

While a disk is being removed, it is shown as `being emptied` by the output of the `mmdf <file system>` command. GPFS rebalancing always happens when disks are removed because the data on the disk should be moved on to the remaining disks on the file system (see Figure 4).

```
disk            driver    sector failure holds   holds                              storage
name            type        size   group metadata data  status         availability pool
---------       --------  ------- ------- -------- ----- -------------- ------------ -----------
gpfs229nsd      nsd          512      -1 yes      yes   being emptied  up           system
gpfs230nsd      nsd          512      -1 yes      yes   ready          up           system
gpfs231nsd      nsd          512      -1 yes      yes   ready          up           system
gpfs237nsd      nsd          512      -1 yes      yes   ready          up           system
gpfs238nsd      nsd          512      -1 yes      yes   ready          up           system
gpfs239nsd      nsd          512      -1 yes      yes   ready          up           system
disk                  disk size   failure holds    holds                 free KB              free KB
name                     in KB     group metadata data          in full blocks          in fragments
---------            ---------   --------- -------- ----- --------------------- ---------------------
Disks in storage pool: system (Maximum disk size allowed is 1.6 TB)
gpfs229nsd           104857600        -1 yes      yes       96391168 ( 92%)       19584 ( 0%) *
gpfs230nsd           104857600        -1 yes      yes       56864768 ( 54%)       37760 ( 0%)
gpfs231nsd           104857600        -1 yes      yes       56649728 ( 54%)       30720 ( 0%)
gpfs237nsd           100669440        -1 yes      yes       92023808 ( 91%)        2944 ( 0%)
gpfs238nsd           100669440        -1 yes      yes       92020736 ( 91%)        2944 ( 0%)
gpfs239nsd           100669440        -1 yes      yes       92014592 ( 91%)        3904 ( 0%)
                     -------------                          ---------------------   ---------------------
(pool total)         511723520                              389573632 ( 76%)       78272 ( 0%)

                     =============                          =====================   =====================
(total)              511723520                              389573632 ( 76%)       78272 ( 0%)
```
Figure 4. Monitor the progress of removing a disk

Changing the GPFS tiebreaker disk

When you change the GPFS tiebreaker disk, DB2 pureScale must be in maintenance mode. In the scenario, a new disk that is named hdisk151 from EMC storage is set as the new GPFS tiebreaker disk.

Complete the following steps:

● Check the current GPFS tiebreaker disk by running the following command as the DB2 instance user or the root user:

```
$ db2cluster -cfs -list -tiebreaker
 The current quorum device is of type Disk with the following specifics:
/dev/hdisk1
```

● Stop the DB2 instance and cluster services by running the following commands as the DB2 instance user:

```
$ db2stop
$ db2stop instance on hostname1
$ db2stop instance on hostname2
```

● Put TSA and GPFS in to maintenance mode by running the following commands as the root user:

```
# db2cluster -cm -enter -maintenance -all
Domain 'db2domain_20130718132645' has entered maintenance mode.
# db2cluster -cfs -enter -maintenance -all
The shared file system has successfully entered maintenance mode.
```

- Change the GPFS tiebreaker disk by running the following commands as the root user:

```
# db2cluster -cfs -set -tiebreaker -majority
The quorum type has been successfully change to 'majority'.
# db2cluster -cfs -set -tiebreaker -disk /dev/hdisk151/
The quorum has been successfully changed to 'disk'.
```

- Exit maintenance mode by running the following commands as the root user:

```
# db2cluster -cm -exit -maintenance -all
Host 'hostname1' has exited maintenance mode. Domain
'db2domain_20130718132645' has been started.
# db2cluster -cfs -exit -maintenance -all
The shared file system has successfully exited maintenance mode.
```

- Start the DB2 instance by running the following command as the DB2 instance user:

```
$ db2start instance on hostname1
$ db2start instance on hostname2
$ db2start
```

Changing the TSA tiebreaker disk

Finally, the TSA tiebreaker disk should be changed to a disk from the new EMC storage. This work can be done while the DB2 pureScale cluster is online. In the scenario, a new disk that is named hdisk200 from EMC storage is set as new TSA tiebreaker.

Complete the following steps:

- Check the current TSA tiebreaker disk by running the following command as the DB2 instance user or the root user:

```
$ db2cluster -cm -list -tiebreaker
 The current quorum device is of type Disk with the following specifics:
DEVICE=/dev/hdisk2.
```

- Change the TSA tiebreaker disk by running the following command as the root user:

```
# db2cluster -cm -set -tiebreaker -disk /dev/hdisk200
Configuring quorum device for domain 'db2domain_20130718132645' …
Configuring quorum device for domain 'db2domain_20130718132645' was
successful.
```

Summary

Consider the following points:

- Multiple disks can be added to file systems by running one command, and multiple commands against multiple file systems can be run in parallel.
- In the process of adding disks to a GPFS file system, there might be a delay when the applications access the file system during the allocation of disk maps.
- Depending on the storage vendor product, GPFS rebalancing might happen when you add disks. You must confirm whether the GPFS rebalancing was done to the new disks; if not, the GPFS rebalancing must be done manually.

- A GPFS tiebreaker disk can be changed only when the DB2 pureScale cluster is in maintenance mode.

- To monitor the progress of adding and removing disks and GPFS rebalancing, run the GPFS `mmdf` `<file system>` command.

Related information

- Changing the physical disk storage of GPFS file systems:
  http://www.ibm.com/support/docview.wss?uid=swg21633330

- GPFS mmdf command reference:
  https://ibm.biz/BdrLeX

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

This document was created or updated on September 23, 2016.

Send us your comments in one of the following ways:
- Use the online **Contact us** review form found at:
  ibm.com/redbooks
- Send your comments in an e-mail to:
  redbooks@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

This document is available online at http://www.ibm.com/redbooks/abstracts/tips1346.html .

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml.

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

DB2®
DS8000®
GPFS™
IBM®
pureScale®
Redbooks®
Redbooks (logo)®
Tivoli®

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.
Other company, product, or service names may be trademarks or service marks of others.