NetApp
Go further, faster

Technical Report

Uninterrupted Data Availability with NetApp MetroCluster and DB2 9.7 High-Availability Feature

Jawahar Lal, NetApp
November 2009 | TR-3807

EXECUTIVE SUMMARY

Whether you have a single data center, a campus, or a metropolitan-wide environment, NetApp® MetroCluster is a cost-effective solution that can provide continuous data availability for your mission-critical DB2 business environment. MetroCluster is an industry-leading solution that combines storage array–based clustering with synchronous mirroring to help deliver continuous availability and minimal data loss at a lower cost. MetroCluster combined with the DB2 high-availability (HA) feature offers transparent recovery from failures so mission-critical applications can continue uninterrupted.

TABLE OF CONTENTS

1 INTRODUCTION

For today's enterprises 24x7 data availability is not an option but a necessity to succeed in an ever-increasing competitive environment. With globalization, data growth, and business reliance on data, enterprises today operate in an extremely complex environment and are more susceptible to interruptions than in the past. Organizations recognize the importance of having a bulletproof business continuance plan and architecture in place to deal with a disaster. The costs of not having one—lost productivity, revenue, and customer loyalty and possibly even business failure—make it mandatory to have a plan that makes sure of an absolute minimum of downtime and rapid recovery from a disaster or failure, with no loss of data. However, such plans, often with added high-availability requirements, can be costly and difficult to implement and administer and can be complicated by a storage infrastructure that includes data centers at sites located miles apart. An additional challenge is making sure that the data at the remote site is fully up to date so that it can serve as the principal data store in case of a disaster at the primary site.

NetApp MetroCluster is an integrated high-availability and business continuance solution that leverages proven technologies from NetApp. It expands the capabilities of the comprehensive NetApp portfolio of high-availability and disaster recovery solutions—a portfolio that includes failover, data replication, and backup solutions. A simple-to-administer solution, MetroCluster extends failover capability from within a data center to a site located many miles away. It also replicates data from the primary site to the remote site to help keep the data completely current. The combination of failover and data replication helps you recover from disaster—with minimal loss of data—in minutes rather than hours or days. The built-in simplicity of MetroCluster allows you to quickly fail over to a remote site and continue operations while turning your attention back to critical business decisions. NetApp MetroCluster combined with IBM DB2 takes high availability and business continuance to the next level and offers customers a solution that is reliable and easy to deploy, maintain, and administer.

2 PURPOSE AND SCOPE

The purpose of this technical report is to serve as a proof of concept for a high-availability database solution for DB2 9.7 running on a NetApp storage system with MetroCluster. MetroCluster can be used to simultaneously protect any mission-critical application and improve availability and is ideal for campus and metropolitan environments where the distance between primary and remote data centers permits synchronous replication without undue latency delay.

In the event of an outage, whether due to an isolated hardware problem or an overall site disaster, MetroCluster extends the benefits of clustered server technology to sites located miles apart. MetroCluster instantly accesses the replicated data on the remote server without any manual intervention or disruption to client application availability. Business continuance can go on for your application data, you can avoid costly downtime, and you can find and fix the source of the outage with minimal to your operation.

MetroCluster design and configuration used in this document are based on NetApp technical report TR-3548: MetroCluster Design and Implementation Guide.[1]

2.1 INTENDED AUDIENCE

This technical report is intended for information technology professionals, storage professionals, DB2 DBAs, and business continuity professionals responsible for the database management infrastructure. For methods and procedures in this technical report, it is assumed that the reader has reasonable knowledge of the following:

IBM DB2 database system architecture and workload generator:

- DB2 storage architecture and database administration
- Benchmark Factory, a benchmark workload generator for databases developed by Quest Software Inc[2]
- Cluster manager software
- Tivoli System Automation for Multiplatform (SA MP)

Working knowledge on NetApp solutions, including the following:

- Data ONTAP®
- NetApp MetroCluster

[1] www.netapp.com/us/library/technical-reports/tr-3548.html.
[2] www.quest.com/benchmark-factory.

2.2 ASSUMPTIONS

Throughout this document it is assumed that we have two physical sites, “Site-A (primary)” and “Site-B (DR).” These sites are separated by 10.5 km. Figure 1 illustrates the architecture components used at both sites. The architecture components are clearly named based on their physical location.

3 HIGH-LEVEL TOPOLOGY DIAGRAM

The solution uses NetApp MetroCluster as a back end for storage availability and DB2 9.7 Enterprise Server Edition with high-availability (HA) feature as the database management system. Tivoli SA MP was used as the cluster manager software. Two IBM p520 servers running AIX 5.3.1, one on each site, were configured in an SA cluster domain to support the front-end database application availability. Each node hosted an instance with one database per instance and accessed storage using NFS.

IBM Tivoli SA MP is bundled with IBM Data Server on AIX and Linux® as part of the DB2 HA feature and is integrated with the DB2 installer. You can install, upgrade, or uninstall SA MP using either the DB2 installer or the install/uninstall scripts that are included in the IBM Data Server install media. For further detail on configuration and installation of SA MP, refer to NetApp technical report TR-3492: DB2 9 for Linux: High Availability Using Tivoli System Automation and NetApp FAS or IBM N Series Storage System.[3]

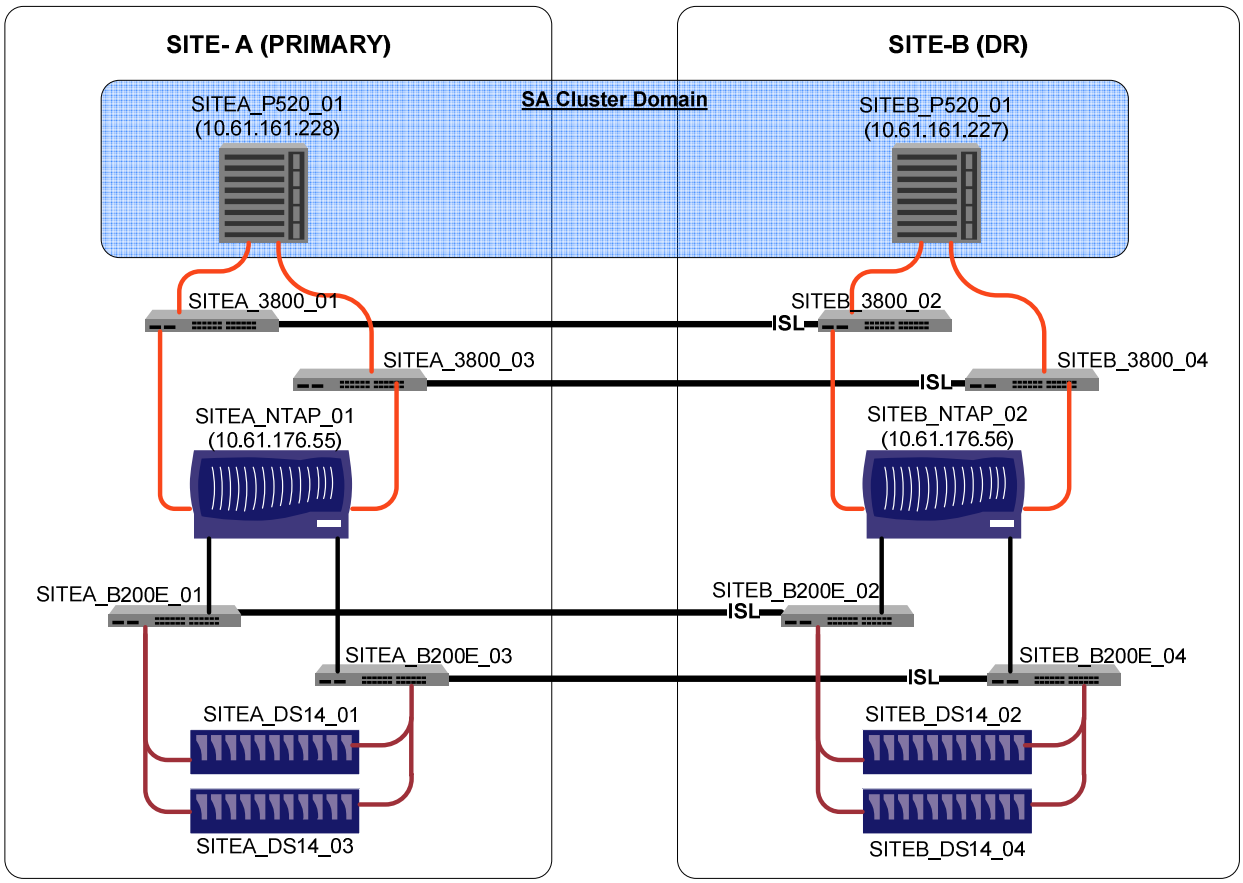

Figure 1) Overview of DB2 9.7 and MetroCluster HA architecture.

[3] www.netapp.com/us/library/technical-reports/tr-3492.html.

3.1 ARCHITECTURE COMPONENTS

Hardware	Vendor	Name	Count	Version	Description
Storage	NetApp	FAS series	2	NA	Storage controller
Host	IBM p520	IBM p520 server on each site; eServer (1.65 Ghz/8GB RAM)	2	NA	Database server
Server: storage network	Brocade	3800	4	3.2.1	16-port FC switch
Storage: disk network (MetroCluster)	Brocade	200E	4	5.1.0	16-port FC switch
Storage shelf	NetApp	DS14 MK2 15K rpm 144GB	4	NA	Storage shelves

Software	Vendor	Name	Version	Description
Storage	NetApp	SyncMirror®	7.2.5	Replication
	NetApp	Data ONTAP	7.2.5	Operating system
	NetApp	Cluster_Remote	7.2.5	Failover
Host	IBM	AIX 5.3	5.3.1	Operating system
	IBM	DB2 Enterprise Server Edition	9.7	Database

3.2 PLATFORM SPECIFICATION

a) FAS Series Storage Controller

The controller and back-end Fibre Channel switches were configured using the instructions described in the Data ONTAP 7.2.5. Active/Active Configuration Guide, which can be found on the NOW™ (NetApp on the Web) site. NetApp MetroCluster supports two types of configurations; stretch and fabric. The stretch MetroCluster is ideal for short distances (500 meters or less) and is ideal for campus DR scenarios. For longer distances greater than 500 meters, the fabric MetroCluster is ideal. The fabric MetroCluster provides protection for the systems that are located at 100 kilometers or less. We used a fabric MetroCluster configuration to produce this document.

Two NetApp FAS series controllers (each with two DS14mk2-HA shelves full of 144GB 15K drives) connected with the VI-MC (Virtual Interface MetroCluster) interconnect and four Brocade 200E switches were used in this test. The controllers were named SITEA_NTAP_01 and SITEB_NTAP_02, and the switches were named SITEA_B200E_01, SITEB_ B200E_02, SITEA_ B200E_03, and SITEB_ B200E_04. The process for cabling disk shelves to the Fibre Channel switches in a MetroCluster configuration depends on whether you have hardware-based disk ownership or software-based disk ownership. The test environment we used to produce this document was configured using hardware-based ownership.

b) Slot Assignments

The storage controllers are configured identically in terms of hardware with the following cards/slot assignments:

Slot Number	Card	Purpose
1	X3300A: Remote management card	Remote monitoring/management
5	X2050A: Dual optical Fibre Channel for mirroring	Disk shelf connection
6	X1922A: VI-MC cluster adapter	Cluster interconnect
7	X3140A: NVRAM4	NVRAM card
8	X2050A: Dual optical Fibre Channel for mirroring	Disk shelf connection
11	X2050A: Dual optical Fibre Channel for target interconnect	Fibre Channel target card

c) Network Settings

Site	System	Interface	IP	Purpose
SITE-A	SITEA_NTAP_01	E0	10.61.176.55	Management network
	SITEA_P520_01	En1	10.61.161.228	Management network
	SITEA_3800_01		10.61.176.51	
	SITEA_3800_03		10.61.176.53	
SITE-B	SITEB_NTAP_02	E0	10.61.176.56	Management network
	SITEB_P520_02	En1	10.61.161.227	Management network
	SITEB_3800_02		10.61.176.52	
	SITEB_3800_04		10.61.176.54	

d) Storage Layout: Aggregate

Controller	Aggregate Name	Number of Disks	Options	Purpose
SITE-A	db2aggr_a	8	RAID_DP®, aggr mirrored	Database
SITE-B	db2aggr_b	8	RAID_DP, aggr mirrored	Database

Controller	Aggregate Name	Volume Name	Size	Options	Purpose
SITE-A	db2aggr_a	db2data	50GB	RAID_DP, flex mirrored, create_ucode=on,convert_ucode=on	DB2 user data
	db2aggr_a	db2logs	20GB	RAID_DP, flex mirrored, create_ucode=on,convert_ucode=on	DB2 transaction logs
	db2aggr_a	db2sys	5GB	RAID_DP, flex mirrored, create_ucode=on,convert_ucode=on	DB2 home and system files
SITE-B	db2aggr_b	db2data	50GB	RAID_DP, flex mirrored, create_ucode=on,convert_ucode=on	DB2 user data
	db2aggr_b	db2logs	20GB	RAID_DP, flex mirrored, create_ucode=on,convert_ucode=on	DB2 transaction logs
	db2aggr_b	db2sys	5GB	RAID_DP, flex mirrored, create_ucode=on,convert_ucode=on	DB2 home and system files

4 METROCLUSTER CONFIGURATION

4.1 SWITCH CONFIGURATION

The back-end FC switches in a MetroCluster environment must be set up in a specific manner for the solution to function properly. In the following sections, the switch and port connections are detailed and should be implemented exactly as documented.

SITEA-B200E-01 – (10.61.176.47)			
Port Number	Bank/Pool	Connected With	Purpose
0	1/0	SITEA_NTAP_01, FC port 4a	SITEA_ NTAP_01 FC HBA
1	1/0	SITEA_NTAP_01, FC port 10a	SITEA_ NTAP_01 FC HBA
2	1/0		
3	1/0		
4	1/1		
5	1/1	SITEA_NTAP_01, Pool1 SITEA_DS14_03 B	Disk shelf
6	1/1		

7	1/1		
8	2/0		
9	2/0	SITEA_NTAP_01, Pool0, SITEA_DS14_01 B	Disk shelf
10	2/0		
11	2/0		
12	2/1	SITEA_ NTAP_01, FC-VI, CI (cluster interface) 1	Cluster Interconnect
13	2/1	SITEB_B200E_03, FC port 13	ISL (interswitch link)
14	2/1		
15	2/1		

SITEB-B200E-02 – (10.61.176.48)			
Port Number	**Bank/Pool**	**Connected With**	**Purpose**
0	1/0	SITEB_ NTAP_02, Pool0 SITEB_DS14_02 B	Disk Shelf
1	1/0		
2	1/0		
3	1/0		
4	1/1		
5	1/1		
6	1/1		
7	1/1		
8	2/0	SITEB_ NTAP_02, FC port 4a	SITEB_ NTAP_02 FC HBA
9	2/0	SITEB_ NTAP_02, FC port 10a	SITEB_ NTAP_02 FC HBA
10	2/0		
11	2/0		
12	2/1	SITEB_ NTAP_02, FC-VI, CI 1	Cluster interconnect
13	2/1	SITEA_B200E_01, FC port 13	ISL
14	2/1	SITEB_ NTAP_02, Pool1 SITEB_DS14_02 B	Disk shelf
15	2/1		

SITEA-B200E-03 – (10.61.176.49)			
Port Number	**Bank/Pool**	**Connected With**	**Purpose**
0	1/0	SITEA_ NTAP_01, FC port 4b	SITEA_ NTAP_01 FC HBA
1	1/0	SITEA_ NTAP_01, FC port 10b	SITEA_ NTAP_01 FC HBA
2	1/0		
3	1/0		
4	1/1		
5	1/1	SITEA_ NTAP_01, Pool1 SITEA_DS14_01A	Disk shelf
6	1/1		
7	1/1		
8	2/0		
9	2/0	SITEA_ NTAP_01, Pool0, SITEA_DS14_03 A	Disk shelf

10	2/0		
11	2/0		
12	2/1	SITEA_ NTAP_01, FC-VI, CI 2	Cluster interconnect
13	2/1	SITEB_B200E_04, FC port 13	ISL
14	2/1		
15	2/1		

SITEB-B200E-04 – (10.61.176.50)			
Port Number	**Bank/Pool**	**Connected With**	**Purpose**
0	1/0	SITEB_ NTAP_02, Pool0 SITEB_DS14_02 A	Disk shelf
1	1/0		
2	1/0		
3	1/0		
4	1/1		
5	1/1		
6	1/1		
7	1/1		
8	2/0	SITEB_ NTAP_02, FC port 4b	SITEB_ NTAP_02 FC HBA
9	2/0	SITEB_ NTAP_02, FC port 10b	SITEB_ NTAP_02 FC HBA
10	2/0		
11	2/0		
12	2/1	SITEB_ NTAP_02, FC-VI, CI 1	Cluster interconnect
13	2/1	SITEA_B200E_03, FC port 13	ISL
14	2/1	SITEB_ NTAP_02, Pool1 SITEB_DS14_04 A	Disk shelf
15	2/1		

4.2 HOST SERVERS

HOST OPERATING SYSTEM

Two single-node database servers were set up, one on each site, running on an IBM eServer with four POWER5 1.65GHz processors and 8GB RAM. The hosts were named SITEA_P520_01 and SITEB_ P520_02.

DATABASE SOFTWARE

IBM DB 9.7 with high-availability feature was installed on each host according to procedures described in the DB2 9.7 installation guide available at the IBM DB2 Database for Linux, UNIX, and Windows Information Center page[4].

NETWORK CONFIGURATION

The following table details the network settings for the DB2 database host servers.

Host Name	IP Address	Purpose
SITEA_P520_01	10.61.161.228	Database server for site A
SITEA_P520_02	10.61.161.227	Database server for site B

[4] http://publib.boulder.ibm.com/infocenter/db2luw/v9r7/index.jsp.

4.3 DB2 DATABASE CONFIGURATION

Our test environment consisted of a DB2 9.7 Enterprise Server Edition with HA feature running an online transaction processing (OLTP) workload. We followed best practices for DB2 and disabled file system caching by creating table spaces with the NO FILE SYSTEM CACHING clause.

AIX FCP-RELATED PARAMETER CHANGES

The default queue depth for each disk needs to set a correct value to get proper throughput. In this configuration, we changed queue_depth to 256 for each hdisk by executing the following command (for example, for hdisk1):

```
chdev -l hdisk1 -a queue_depth=256
```

Additionally, `num_cmd_elems`, `vmo`, `maxservers`, and `maxreqs` were modified:

```
chdev -l fcs0 -a num_cmd_elems=2048
aioo -o maxservers=20

aioo -o maxreqs=32768
vmo -o lru_file_repage=0
```

DATABASE WORKLOAD

To simulate database transactions during the test scenarios, Benchmark Factory was used to generate TPCC workload for the DB2 database.

5 FUNCTIONAL TEST SCENARIOS

The following subsections describe the various test scenarios that were executed upon successful build of the solution discussed earlier in this document. The test scenarios include various component failures, including server hardware, network, storage system, and so on. Prior to the execution of each test, the environment was reset to the normal running state. In the normal running state, a DB2 instance and a database are up and running on the host SITEA_P520_01 in site-A as well as the host SITEB_DB2_01 in site-B. Additionally, Benchmark Factory was configured to perform typical user transactions on both sites.

5.1 COMPLETE LOSS OF POWER TO DISK SHELF

In order to achieve high availability, there shouldn't be a single point of failure in the entire architecture. This scenario tested a loss of an entire disk shelf. The test was simulated by turning off both disk shelf power supplies at the same time while the database workload was running.

Scenario 1	Expected Result	(Pass/Fail)
Power off the shelf "SITEA_DS14_01"	Relevant disks went offline	Pass
	Plex mirror changed to broken state	Pass
	Service to clients (availability and performance) is unaffected	Pass
Power the disk shelf back ON	The disks were detected	Pass
	A resync of the plexes occurred without any manual action	Pass
DB2 and application	No error message in db2diaglog	Pass
	DB2 service was uninterrupted; workload continued to run without any error	Pass

5.2 LOSS OF ONE LINK ON ONE DISK LOOP

This test was simulated by unplugging a fiber cable connecting one of the disk shelves.

Scenario 2	Expected Result	(Pass/Fail)
Remove fiber entering SITEA_DS14_02 Pool0, ESH A	Relevant disks went offline	Pass
	Plex mirror changed to broken state	Pass
	Service to clients (availability and performance) is unaffected	Pass
Reconnect the fiber	The disks were detected	Pass

	A resync of the plexes occurred without any manual action	Pass
DB2 and application	No error message in db2diaglog	Pass
	DB2 service was uninterrupted; workload continued to run without any error	Pass

5.3 LOSS OF A BROCADE SWITCH ON THE STORAGE SIDE

This test was simulated by turning off power for the switch while the database workload was running.

Scenario 3	Expected Result	(Pass/Fail)
Power off the Fibre Channel switch "SITA-B200E_02"	Controller messages: some disks are connected to only one switch	Pass
	Controller messages: one of the cluster interconnects is down	Pass
	Service to clients (availability and performance) is unaffected	Pass
Power it back on	Switch completes its boot process	Pass
	Controller messages: second cluster interconnects is active	Pass
DB2 and application	No error message in db2diaglog	Pass
	DB2 service was uninterrupted; workload continued to run without any error	Pass

5.4 LOSS OF ISL ON THE STORAGE SIDE

Redundancy at ISLs is also required for high availability. This test was simulated by simply removing the fiber connection between two of the fiber switches at storage side while a load was applied.

Scenario 4	Expected Result	(Pass/Fail)
Remove the fiber between SITEA_B200E_01 and SITEB_B200E _02	Controller messages: some disks are connected to only one switch	Pass
	Controller messages: one of the cluster interconnects is down	Pass
	Service to clients (availability and performance) is unaffected	Pass
Reconnect ISL	Controller messages: the disks are now connected to two switches	Pass
	Controller messages: second cluster interconnects is active again	Pass
DB2 and application	No error message in db2diaglog	Pass
	DB2 service was uninterrupted; workload continued to run without any error	Pass

5.5 LOSS OF ONE STORAGE CONTROLLER

This test consisted of two parts; first, we tested the failover scenario and second, we tested failback. To simulate failover we turned off both power supplies on the SITE A storage controller at the same time and observed the results. After failover simulation, we turned on power and simulated a failback scenario by issuing the give back command on the surviving controller to request that processing be returned to the previously failed controller.

Scenario 5	Expected Result	(Pass/Fail)
Power off the controller SITEA_NTAP_01	Disk ownership is assumed by the second controller without any manual intervention	Pass
	No or minimal host interruption	Pass
	No interruption on client service (availability and performance)	Pass
Power on the controller SITEA_NTAP_01	Controller powers ON without any error	Pass
On controller SITEB_NTAP_02	No or minimal host interruption	Pass

Issue a “cf giveback” command	Disk ownership is assumed by the SITE A storage controller	Pass
DB2 and application	No error message in db2diaglog	Pass
	DB2 service was uninterrupted; workload continued to run without any error	Pass

5.6 LOSS OF ONE DATABASE SERVER

To test availability of the overall solution we simulated the loss of one database server by shutting it down and restarting the server after results were observed.

Scenario 6	Expected Result	(Pass/Fail)
Shut down the server SITEA_P520_01 and observe SA status on the SITEB_P520_02 by executing `/opt/ibm/db2/V9.7/ha/tsa/get status`	SA resources that were active on the SITE A server node are started on the SITE B server node in the SA cluster	Pass
	The database instance becomes available on the SITE B server node within a few seconds	Pass
	Status of the resources that were on the SITE A server node becomes `Failed_Offline`	Pass
	No or minimal interruption on client service (availability and performance)	Pass
Restart the server SITEA_P520_01 and observe SA status on the SITEB_P520_02	Status of the resources from `Failed_offline` to `offline`	Pass
Add node SITEA_P520_01 back to SA cluster by executing the following command: **`startrpnode sitea_p520_01`**	SA resources are started on the SITE A server, and status of resources becomes online	Pass
	DB2 instance is started, and application connection moved back to the SITE A server SITEA_P520_01	Pass
	No or minimal interruption on client service (availability and performance)	Pass

6 CONCLUSION

An effective high-availability (HA) solution must address both unplanned and planned causes of downtime to achieve a truly fault-tolerant and resilient IT infrastructure. To make sure of high availability of application data, the underlying architecture design and components must support high availability. By combining NetApp MetroCluster and IBM DB2 9.7 HA features, customers can achieve true high availability with minimum acquisition cost and operational complexity.

7 ACKNOWLEDGEMENTS

I want to acknowledge contribution and great feedback from Lee Dorrier, Bill Heffelfinger, Chattur Narayankumar, and Gary Zasman to make this writeup a finished product.

www.netapp.com