Best Practices for Deploying Dell EqualLogic Synchronous Replication

A Dell EqualLogic Reference Architecture

Dell Storage Engineering
October 2013

Revisions

Date	Description
October 2013	Initial release

Table of contents

Feedback

We encourage readers of this publication to provide feedback on the quality and usefulness of this information by sending an email to SISfeedback@Dell.com.

SISfeedback@Dell.com

Executive summary

EqualLogic Synchronous Replication (SyncRep) is a datacenter protection solution for groups where members remain within a close proximity of one another. With this arrangement in mind, tests with different I/O block sizes, a varying number of SyncRep enabled volumes and a varying distance between the SyncActive and SyncAlternate pool locations were run to determine the most efficient settings for maximizing the benefits of SyncRep. Results of testing performed for this white paper show that the best results when utilizing the SyncRep feature are obtained by:

- Maintaining the same number of arrays in both the SyncActive and SyncAlternate pools
- Using the same drive types in all arrays, and
- Configuring the arrays in both pools with the same RAID type.
- Separating the pools by a maximum distance of 1Km

1   Introduction

Synchronous Replication (SyncRep), included in Dell™ EqualLogic™ firmware 6.0 and later, simultaneously writes data to two hardware independent copies of a volume, contained in two different storage pools that are part of the same PS Series group. This feature is designed to provide volume data protection in the event of a site power outage or complete hardware failure. It is intended as a datacenter protection solution where group members are within close proximity to each other. The distance between the pools may be between separate racks in the same datacenter, or across different floors in the same building, or even between buildings on the same campus. However, SyncRep is not designed to be a disaster recovery or a long distance replication solution.

1.1   Audience

This technical white paper is intended for storage administrators, SAN system designers, storage consultants, or anyone tasked with configuring a host server as an iSCSI initiator to EqualLogic PS Series storage for use in a production SAN. It is assumed that readers have experience in designing and/or administering a shared storage solution. Also, there are some assumptions made in terms of familiarity with all current Ethernet standards as defined by the Institute of Electrical and Electronic Engineers (IEEE) as well as TCP/IP and iSCSI standards as defined by the Internet Engineering Task Force (IETF).

1.2   Terminology

The following terms will be used throughout this document:

Group: Consists of one to 16 EqualLogic PS Series arrays connected to an IP network.

In sync: A term used to indicate that both the SyncActive and SyncAlternate pools contain the same volume data.

Out of sync: A term used to indicate that the SyncActive pool and SyncAlternate pool may not contain the same data.

Paused: When activated, data is only written to the SyncActive pool. The group tracks all the changes and will write them to the SyncAlternate pool when replication is resumed.

Pool: A pool is a container that each member (array) is assigned to after being added to a group. A pool can have up to eight members.

SyncActive: The active copy of the volume to which iSCSI initiators connect when reading and writing data.

SyncAlternate: When data is written to the SyncActive volume, the group simultaneously writes the same data to this copy of the volume in the alternate pool.

Note: SyncActive and SyncAlternate are always from the perspective of the volume and not the pool. For example, if volume 1 is in Pool A, then Pool A is initially the SyncActive pool for Volume 1 and the hosts

will be connected to Volume 1. If Volume 2 is in Pool B, then Pool B will initially be the SyncActive pool for Volume 2 and the hosts will be connected to Volume 2.

1.3 Objective

This document provides best practice guidance and design considerations for deploying SyncRep based on the following parameters:

- Network topology and overall design
- Distance between pools
- Pool design
- RAID level configuration

Pausing or switching to the SyncAlternate volume and failover to a SyncAlternate volume are not within the scope of this paper. For more information and details on performing these operations, please see the technical report (number TR1085) titled:

Dell EqualLogic PS Series Arrays: Understanding Synchronous Replication at

http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/20175375/download.aspx

2 Replication technologies

The EqualLogic SyncRep feature is intended to be used as a real time data protection solution for groups where arrays are in close proximity to each other – such as within the same datacenter, same building or for a cross-campus solution. SyncRep volumes are always synchronized and will always contain the same data at any point in time. If the SyncActive volume were to fail, hosts could access the data in the SyncAlternate volume. Typical use cases for SyncRep include:

- Datacenter protection when single member pools are in the same rack or on the same floor.
- Extended datacenter protection of pools located on different floors or in separate rooms of the same floor.
- Continued availability during planned maintenance by switching from SyncActive to SyncAlternate volumes.

EqualLogic Auto-Replication provides a disaster-recovery option in case the original volume (or the entire EqualLogic group) becomes unavailable or if a site-wide disaster occurs. Auto-Replication is a point-in-time replication solution that offers extended-distance replication. Point-in-time copies can be made manually or on a schedule. The frequency in which replication occurs determines how old the replica becomes relative to the current state of the source volume. This also determines the potential maximum amount of data that may be lost if the primary site experienced a complete failure. This is also known as the Recovery Point Objective or RPO. Typical use cases for Auto-Replication include:

- Disaster recovery over a Storage Area Network (same datacenter or across campus)
- Disaster recovery over an extended distance (i.e., over a WAN link to a remote location)
- Data consolidation when sending replicas to a data center for backup

2.1 How SyncRep works

Enabling SyncRep will result in the creation of two hardware independent copies of volume data. SyncRep will only acknowledge write completion when both the source and target volume write have completed. A key benefit of SyncRep is that if one pool is not available, volume data from the other pool can be retrieved.

SyncRep is enabled on a per volume basis and supports up to 32 SyncRep enabled volumes per group for a PS6xxx array. PS4xx0 arrays support up to four SyncRep enabled volumes per group. A SyncRep solution must consist of two pools with at least one member in each pool. The volumes connected to the hosts are called the SyncActive volume. A copy of the data exists simultaneously in both pools, but hosts will only directly access the SyncActive volume of the SyncRep enabled volumes. The host is unaware of the secondary copy. When SyncRep is enabled, any new data or updated data that is written to the active volume is also written to the alternate volume before the active volume will acknowledge success of the write transaction back to the host.

A SyncRep enabled volume acknowledges the writes on the SyncActive volume only after the writes are also confirmed on the SyncAlternate volume. If a failure on the SyncActive volume occurs, SyncRep allows host connections to move to the SyncAlternate volume (which then becomes the SyncActive volume) so

that no data loss occurs. Data is only read from the SyncActive volumes. If the replication is paused or out of sync, then SyncRep keeps track of the changes and the group creates a snapshot of the volume to reflect its contents. This snapshot resides in the SyncActive pool. When synchronous replication is resumed, the tracked changes are applied to the alternate volume.

3 Test configurations and methodology

The test cases in this paper were designed to assess the ability and performance of EqualLogic SyncRep across different scenarios. Identical workloads were run in all configurations to ensure that the data could be compared directly without introducing variables beyond those specified in the test plan.

In order to determine the relative performance of each SAN design, the performance tool vdbench was used to capture throughput values at four distinct I/O workloads. Vdbench is "a disk and tape I/O workload generator for verifying data integrity and measuring performance of direct attached and network connected storage." (http://sourceforge.net/projects/vdbench/)

> **Note**: All EqualLogic SAN best practices, such as enabling flow control and Jumbo frames, were implemented.
> See Appendix A for details about specific configuration of the hardware and software infrastructure.

3.1 Test approach

The test topologies for this paper included three scenarios: pools located in the same rack (1m), pools located in different racks on the same floor (<100m), and pools located in different racks on different floors (<1km). SyncRep was tested by varying the number of volumes (up to the maximum of 32), the block size (4K-256K), and the distance between pools (at 1m, 100m, and 1km). The following technique was used to complete the tests.

- A baseline was first established by connecting all hosts and all arrays through a single pair of redundantly configured switches separated by 1m.
- A second pair of switches was added and separated by distance initially of 1m to establish a baseline. The pairs of switches were then separated by a distance of 100m using SR optics on either pair of switches. The test was conducted with a redundant pair of cables configured as a LAG to increase throughput and provide redundancy.
- The same test was then repeated with LR optics and a 1km optical cable between pairs of switches.
- All tests had the SyncActive volumes in Pool A and all the SyncAlternate volumes in Pool B. Although mixing SyncActive and SyncAlternate volumes in the same pool is possible it was not configured that way in this paper in order to simplify the presentation of the test reults.

Workloads of 100% sequential read, 100% sequential write and random with a mix of 67% random reads and 33% random writes were applied using vdbench. The block size was varied and the number of SyncRep enabled volumes was increased from one to 32 (the maximum allowed with PS6xx0 Series). IOPS, bandwidth, retransmits (an indicator of network congestion) and other statistics were carefully observed and gathered to analyze the system performance and ensure that iSCSI sessions were not adversely affected. Each test was run once for 20 minutes.

Varying combinations and types of arrays (for example, three arrays in one pool, and a single array in the second pool) were also studied. These tests were designed to establish best practice limitations imposed by mixing the drive types and number of arrays.

3.2 Configuration

This section describes the configuration illustrated in Figure 1.

- Two PowerEdge™ R620 servers
  - Microsoft® Windows Server® 2008 R2 SP1
  - One Broadcom 57810 NIC per 10 Gb host (two active 10 GbE ports on the SAN)
- Six EqualLogic PS6110XV (10 Gb) arrays
  - One active storage controller port (one active 10 GbE port on the SAN)
- Four Force10™ S4810 switches
  - VLT within active or alternate sites
  - LACP between pairs across sites
  - Two pools of three arrays each
    - A SyncActive pool with three arrays
    - A SyncAlternate pool with three arrays
  - The LAG is initially 1 meter long and will be increased up to 1km

Each pair of S4810 switches was in a VLT configuration and switch pairs were connected via LACP. The entire configuration was cabled using EqualLogic best practices. Hardware specifications of the servers and arrays are in Appendix A titled, "Solution infrastructure details".
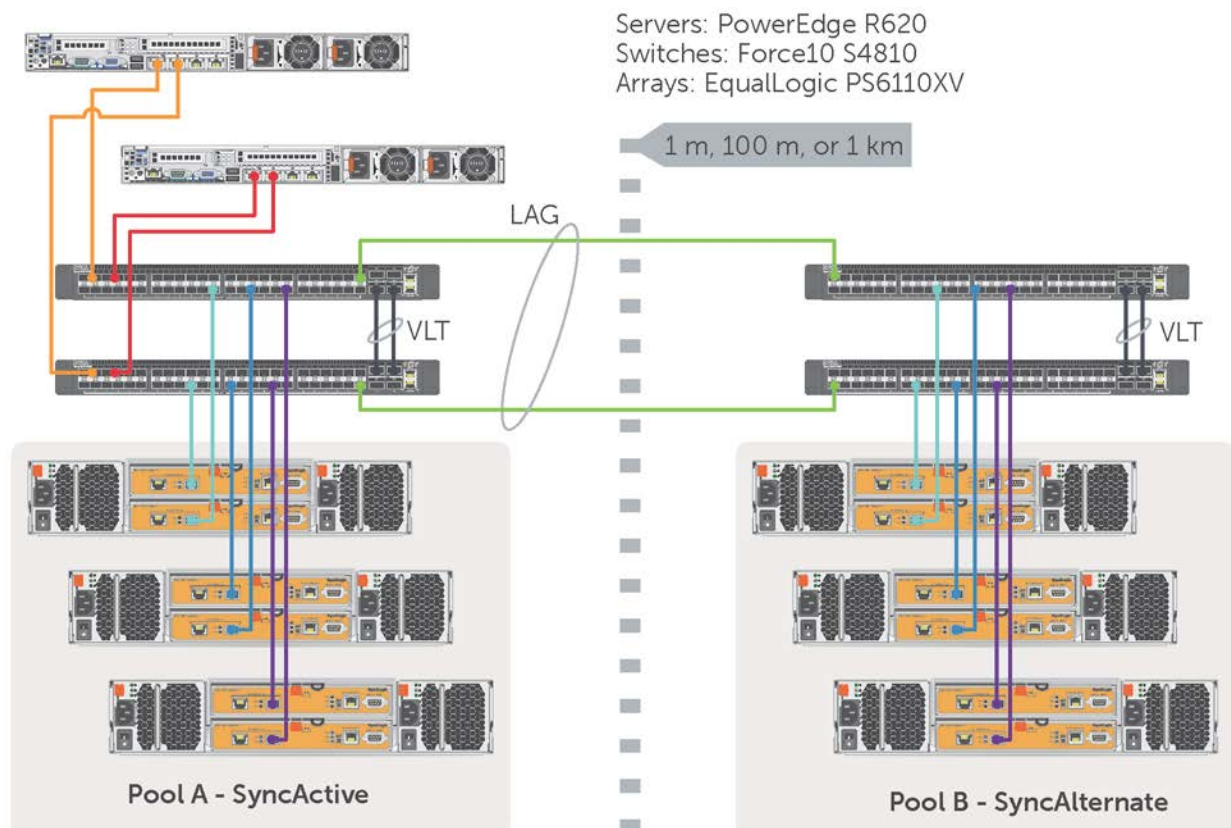

Figure 1   Test configuration

3.3 One meter tests

These tests determined a performance baseline and were used as a starting point for the distance tests. Pool A refers to the SyncActive pool and Pool B refers to the SyncAlternate pool. All tests were configured with RAID 10 and the link aggregated switch pairs were separated by one meter of multimode fiber optic cabling. Each test was run for 20 minutes and the variables were changed one at a time, beginning with the block size and the same workloads were run again. After all the block sizes were tested, the number of volumes per host was changed and the tests were repeated. The following test cases were run:

Test case 1: SyncRep not enabled

- One pool of three PS6110XV arrays
- Used to determine a performance baseline without SyncRep enabled
- Results can be found in Section 4.1

Test case 2: 3:3 member pool design
This test was repeated with two pools of three PS6110XV arrays each for a comparison with test case 5.

- Two pools of three PS6110XV arrays each
- SyncRep was enabled
- SyncActive volumes were in Pool A
- SyncAlternate volumes were in Pool B
- Results can be found in Section 4.2 and 4.3

Test case 3: 3:1 member pool design

- One pool of three PS6110XV arrays (Pool A) and one pool of one PS6110XV array (Pool B).
- SyncRep was enabled
- SyncActive volumes were in Pool A
- SyncAlternate volumes were in Pool B
- Results can be found in Section 4.2

Test case 4: Mixed drive speed and array type pool design

- One pool of three PS6110XV arrays (Pool A) and one pool of one PS4110E array (Pool B).
- SyncRep was enabled
- SyncActive volumes were in Pool A
- SyncAlternate volumes were in Pool B
- Results can be found in Section 4.3

Test case 5: Mixed RAID type design

- Two pools of one PS6110XV array each. Pool A was RAID 10 and Pool B was RAID 6.
- SyncRep was enabled
- SyncActive volumes were in Pool A
- SyncAlternate volumes were in Pool B
- Results can be found in [Section 4.4](#)

Table 1 lists all the permutations for test cases 1-5. Each test case was run by sweeping the number of volumes per host with each specified block size. Each permutation had 100% sequential writes, 100% sequential reads and a mix of random R/W (67%/33%) using vdbench.

Table 1   1 meter test cases

1 meter test cases		
Volumes (per host)	**Block Sizes**	**Workloads**
1	4k, 8k, 16k, 32k, 64k, 128k, 256k	Sequential writes, sequential reads, random R/W
4	4k, 8k, 16k, 32k, 64k, 128k, 256k	Sequential writes, sequential reads, random R/W
8	4k, 8k, 16k, 32k, 64k, 128k, 256k	Sequential writes, sequential reads, random R/W
16	4k, 8k, 16k, 32k, 64k, 128k, 256k	Sequential writes, sequential reads, random R/W

3.4   Distance testing

The even pool design test case was repeated with added distance to the LAG between the switch pairs (and pools) to observe effects of the additional latency with longer cables. These tests used the following configurations:

- Distance of 100m
    - SR optics and multi-mode optical cable spool
    - Pool A had three PS6110XV arrays
    - Pool B had three PS6110XV arrays
- Distance of 1km
    - LR optics and single-mode optical cable spool
    - Pool A had three PS6110XV arrays
    - Pool B had three PS6110XV arrays

Each test case was run by sweeping the number of volumes per host with each specified block size. Each permutation had 100% sequential writes, 100% sequential reads and a mix of random R/W (67%/33%) using vdbench.

Table 2   Distance tests cases

Distance Test Cases		
SyncRep enabled volumes (per host)	Block Sizes	Workloads
1	4k, 8k, 16k, 32k, 64k, 128k, 256k	Sequential writes, sequential reads, random R/W
4	4k, 8k, 16k, 32k, 64k, 128k, 256k	Sequential writes, sequential reads, random R/W
8	4k, 8k, 16k, 32k, 64k, 128k, 256k	Sequential writes, sequential reads, random R/W
16	4k, 8k, 16k, 32k, 64k, 128k, 256k	Sequential writes, sequential reads, random R/W

4 Results and analysis

> **Note:** The results provided in this paper are intended for the purpose of comparing the specific configurations used in our lab environment. The results do not portray the maximum capabilities of any system, software, or storage.

4.1 3:1 member pool design results

This section describes the test results where a different number of arrays in the SyncActive pool were used versus the number of arrays in the SyncAlternate pool.

The baseline case used in this section had the following setup:

- Three PS6110XV arrays in the SyncActive pool
- Three PS6110XV arrays in the SyncAlternate pool
- Two hosts
- Eight SyncRep volumes per host (16 total SyncRep enabled volume)
- 100% sequential writes, 100% sequential reads and random R/W (67%/33%) workloads using vdbench
- Block sizes: 4k, 8k, 16k, 32k, 64k, 128k, 256k

The results were then compared to the case in which the SyncActive pool and the SyncAlternate pool had a different number of arrays. The setup used for the uneven pool case was:

- Three PS6110XV arrays in the SyncActive pool
- One PS6110XV arrays in the SyncAlternate pool
- Eight SyncRep volumes per host (16 total SyncRep enabled volume)
- Two hosts
- 100% sequential writes, 100% sequential reads and random R/W (67%/33%) workloads using vdbench
- Block sizes: 4k, 8k, 16k, 32k, 64k, 128k, 256k

The results of our comparison showed that overall performance was reduced when the pools did not have the same number of arrays. The effect on performance was greater at larger block sizes because there was an inherent I/O oversubscription when three arrays in the SyncActive pool were writing data to only one array in the SyncAlternate pool.

This shows that the SyncAlternate pool should not be thought of as a backup pool. Instead Dell's best practice recommendation is that pools should always be configured with an equal number of the same type of arrays to ensure that performance is not adversely affected.

4.2 Mixed drive type results

This section discusses the results of the tests with different drive types in the arrays of the SyncActive pool and the SyncAlternate pool.

The baseline case used in this section had the following setup:

- Three PS6110XV arrays in the SyncActive pool
- One PS6110XV array in the SyncAlternate pool
- Eight SyncRep volumes per host (16 total SyncRep enabled volume)
- Two hosts
- 100% sequential writes, 100% sequential reads and random R/W (67%/33%) workloads using vdbench
- Block sizes: 4k, 8k, 16k, 32k, 64k, 128k, 256k

The baseline results were compared to the case in which the SyncActive pool and the SyncAlternate pool had different drive types in the arrays. The setup used for the mixed drive type case was:

- Three PS6110XV arrays in the SyncActive pool
- One PS4110E array in the SyncAlternate pool
- Eight SyncRep volumes per host (16 total SyncRep enabled volume)
- 100% sequential writes, 100% sequential reads and random R/W (67%/33%) workloads using vdbench
- Block sizes: 4k, 8k, 16k, 32k, 64k, 128k, 256k

The results of our comparison showed that overall performance was reduced when the arrays in the SyncAlternate pool had a drive type with a slower rotational speed than the arrays in the SyncActive pool. Therefore, Dell's best practice recommendation is to have the same drive type in both the SyncActive and SyncAlternate pools to ensure optimal performance.

4.3 Mixed RAID type results

This section discusses the results of the testing with different RAID types in the arrays of the SyncActive pool and the SyncAlternate pool.

The baseline case used in this section had the following setup:

- One PS6110XV array in the SyncActive pool
- One PS6110XV array in the SyncAlternate pool
- 16 SyncRep volumes per host (32 total SyncRep enabled volume)
- 100% sequential writes, 100% sequential reads and random R/W (67%/33%) workloads using vdbench
- Block sizes: 4k, 8k, 16k, 32k, 64k, 128k, 256k
- RAID 10 on both arrays

The baseline results were compared to the case in which the SyncActive pool and the SyncAlternate pool had different RAID types configured on the arrays. The setup used for the mixed RAID type case was:

- One PS6110XV array in the SyncActive pool
- One PS6110XV array in the SyncAlternate pool
- 16 SyncRep volumes per host (32 total SyncRep enabled volume)
- 100% sequential writes, 100% sequential reads and random R/W (67%/33%) workloads using vdbench
- Block sizes: 4k, 8k, 16k, 32k, 64k, 128k, 256k
- RAID 10 in the SyncActive pool
- RAID 6 in the SyncAlternate pool

The results of our comparison showed that overall performance was reduced when the arrays in the SyncActive pool had a different RAID type than the arrays in the SyncAlternate pool. Therefore, Dell's best practice recommendation is to have the same RAID type configured on all of the arrays to ensure optimal performance.

4.4 Distance results

This section explains the results of the tests with varying distances between the SyncActive pool and the SyncAlternate pool.

The baseline case used in this section had the following setup:

- Three PS6110XV arrays in the SyncActive pool
- Three PS6110XV arrays in the SyncAlternate pool
- Eight SyncRep volumes per host (16 total SyncRep enabled volume)
- Two Hosts
- 100% sequential writes, 100% sequential reads and random R/W (67%/33%) workloads using vdbench
- Block sizes: 4k, 8k, 16k, 32k, 64k, 128k, 256k
- Distance between pools: one meter

The baseline results were compared to the cases in which the SyncActive pool and the SyncAlternate pool had different distances between the pools. The setup used for the distance cases was:

- Three PS6110XV arrays in the SyncActive pool
- Three PS6110XV arrays in the SyncAlternate pool
- Eight SyncRep volumes per host (16 total SyncRep enabled volume)
- Two Hosts
- 100% sequential writes, 100% sequential reads and random R/W (67%/33%) workloads using vdbench
- Block sizes: 4k, 8k, 16k, 32k, 64k, 128k, 256k
- Distance between pools: 100 meters, 1 kilometer

Table 3 shows the baseline performance (one meter between switch pools) compared to the tests cases where the distance between the pools was increased to 100 meters over multimode fiber optic and 1 kilometer over single-mode fiber optic.   In all cases, the impact on performance due to the increased distance between the SyncActive (Pool A) and SyncAlternate (Pool B) pools was less than 5% and therefore considered statistically insignificant.

Table 3   Comparison of baseline (1m) with distances of 100m and 1Km

Workload	100m	1Km
Random R/W Mix	No significant differences found	No significant differences found
Sequential Read	No significant differences found	No significant differences found
Sequential Write	No significant differences found	No significant differences found

Dell therefore recommends configuring SyncRep with a distance of 1km or less between the SyncActive (Pool A) and SyncAlternate (Pool B) pools.

5 Best practice recommendations and conclusions

SyncRep, included in Dell EqualLogic firmware 6.0 and later, simultaneously writes data to two separate copies of a volume in two different storage pools that are in the same PS Series group. This feature provides a datacenter protection solution for group members that are within close proximity to each other. SyncRep acknowledges the writes on the SyncActive volume only when the writes are also made on the SyncAlternate volume. If a failure on the SyncActive volume does occur, SyncRep allows a move to the SyncAlternate volume so that no data loss occurs.

When enabling SyncRep for EqualLogic volumes, there are several Dell best practice recommendations to keep in mind.

- Having the same number of arrays in the SyncActive and SyncAlternate pools will yield better performance than if the number or arrays in the SyncActive and SyncAlternate pools differentiate. When the number of arrays differentiates, the write latency can increase which results in a write throughput decreases as compared to having the same number of arrays. When this happens it can affect application response time and performance.
- Having the drives in both the SyncActive and SyncAlternate pools the same type and speed will yield better performance than if the drive speeds in the SyncAlternate pool are of a slower type. When the speed of the drives in the pools differentiate, the write latency increases and the write throughput decreases as compared to having the same speed drives in all the arrays. This will affect application response time and performance.
- Having the same RAID type in both the SyncActive and SyncAlternate pools will yield better performance than if the RAID types in the SyncActive and SyncAlternate pools differentiate. When the RAID types of the pools differentiate, the write latency increases and the write throughput decreases as compared to having the RAID type configured on all the arrays. When this occurs, the application response time and performance are affected.
- Keep the distance between the SyncActive (Pool A) and SyncAlternate (Pool B) pools at 1km or less.
- Volumes with SyncRep enabled and running large sequential workloads resulted in a decrease in performance when compared to the case where SyncRep was not enabled. If an application will have large sequential workloads ensure that the performance overhead is taken into consideration when enabling SyncRep.

A   Solution infrastructure details

Switch Configuration Guides (SCG) are also published to provide step-by-step instructions for configuring Ethernet switches for use with EqualLogic PS Series storage using Dell best practices. These SCGs are located on Dell Tech Center at http://en.community.dell.com/techcenter/storage/w/wiki/4250.switch-configuration-guides-by-sis.aspx.

Switches	
Switch Model	Force10 S4810
Firmware	Firmware version: 9.1

Server	
Server Model	PowerEdge R620
BIOS	Dell Inc. 6.2.3, 4/26/12
Processor	Two - Intel Xeon CPU E5-2620 @ 2GHz, 2261 Mhz, 6 Core(s), 12 Logical Processor(s)
OS Name	Windows Server 2008 R2 Enterprise
OS Version	6.1.7601 SP1 Build 7601

Array (10G)	
Model	PS6110XV
Firmware	6.0.4 (R322829)
Enable performance load balancing in pools	Disabled*
Raid Type	RAID 10
Control module	model70-0477 (TYPE 14)

*This option was disabled to prevent page movement during performance tests. For normal production environments, it is recommended to leave this option enabled at all times.

Network Adapter (10G) (All settings default except where noted)	
Model	Broadcom BCM57810
Driver Version	7.4.23.0
Family Firmware version	FFV7.4.8
Bootcode version	MFW1 7.4.19
iSCSI boot version	V7.4.2
Jumbo Packet	9014 Bytes

Network Adapter (10G) (All settings default except where noted)	
Flow Control	RX/TX enabled
RX buffers	3000
TX Buffers	5000

MPIO Configuration	
HITKit version	4.5.0.6492
MPIO Device Specific Module	Maximum sessions per slice:2
	Maximum sessions per volume:6
Default Load Balancing Policy	Least Queue Depth

Additional resources

Support.dell.com is focused on meeting your needs with proven services and support.

DellTechCenter.com is an IT Community where you can connect with Dell Customers and Dell employees for the purpose of sharing knowledge, best practices, and information about Dell products and your installations.

Referenced or recommended Dell publications:

- Dell EqualLogic Configuration Guide:
  http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/19852516/download.aspx
- Dell EqualLogic Compatibility Matrix:
  http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/19856862/download.aspx
- Dell EqualLogic Group Manager Online Help
  http://psonlinehelp.equallogic.com/en/V6.0/groupmanager.htm

For EqualLogic best practices white papers, reference architectures, and sizing guidelines for enterprise applications and SANs, refer to Storage Infrastructure and Solutions Team Publications at:

- http://dell.to/sM4hJT

This Technical Report describes how EqualLogic Synchronous Replication works to protect your data and offers configuration recommendations when planning to use EqualLogic SyncRep.

- http://www.dellstorage.com/WorkArea/DownloadAsset.aspx?id=3050