Advanced Performance Tuning for Dell EqualLogic Auto-Replication

Dell Storage Engineering

July 2014

A Dell Technical White Paper

Revisions

Date	Description
July 2014	Initial release

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

PRODUCT WARRANTIES APPLICABLE TO THE DELL PRODUCTS DESCRIBED IN THIS DOCUMENT MAY BE FOUND AT: http://www.dell.com/learn/us/en/19/terms-of-sale-commercial-and-public-sector Performance of network reference architectures discussed in this document may vary with differing deployment conditions, network loads, and the like. Third party products may be included in reference architectures for the convenience of the reader. Inclusion of such third party products does not necessarily constitute Dell's recommendation of those products. Please consult your Dell representative for additional information.

Trademarks used in this text:

Dell™, the Dell logo, Dell Boomi™, Dell Precision™ ,OptiPlex™, Latitude™, PowerEdge™, PowerVault™, PowerConnect™, OpenManage™, EqualLogic™, Compellent™, KACE™, FlexAddress™, Force10™ and Vostro™ are trademarks of Dell Inc. Other Dell trademarks may be used in this document. Cisco Nexus®, Cisco MDS®, Cisco NX-0S®, and other Cisco Catalyst® are registered trademarks of Cisco System Inc. EMC VNX®, and EMC Unisphere® are registered trademarks of EMC Corporation. Intel®, Pentium®, Xeon®, Core® and Celeron® are registered trademarks of Intel Corporation in the U.S. and other countries. AMD® is a registered trademark and AMD Opteron™, AMD Phenom™ and AMD Sempron™ are trademarks of Advanced Micro Devices, Inc. Microsoft®, Windows®, Windows Server®, Internet Explorer®, MS-DOS®, Windows Vista® and Active Directory® are either trademarks or registered trademarks of Microsoft Corporation in the United States and/or other countries. Red Hat® and Red Hat® Enterprise Linux® are registered trademarks of Red Hat, Inc. in the United States and/or other countries. Novell® and SUSE® are registered trademarks of Novell Inc. in the United States and other countries. Oracle® is a registered trademark of Oracle Corporation and/or its affiliates. Citrix®, Xen®, XenServer® and XenMotion® are either registered trademarks or trademarks of Citrix Systems, Inc. in the United States and/or other countries. VMware®, Virtual SMP®, vMotion®, vCenter® and vSphere® are registered trademarks or trademarks of VMware, Inc. in the United States or other countries. IBM® is a registered trademark of International Business Machines Corporation. Broadcom® and NetXtreme® are registered trademarks of Broadcom Corporation. Qlogic is a registered trademark of QLogic Corporation. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and/or names or their products and are the property of their respective owners. Dell disclaims proprietary interest in the marks and names of others.

Table of contents

Executive summary

This white paper studies the effects of network latency, packet loss, TCP window size and Jumbo frames when implementing Dell EqualLogic auto-replication. Additionally, multi-volume replication and the impact of host I/O performance on the storage system are analyzed.

Fine tuning a few parameters on the storage arrays may help improve replication and I/O performance for certain specific use cases. This white paper provides a detailed, lab-tested data analysis for the improvement of replication performance by fine-tuning EqualLogic storage configuration parameters.

The major observations are:

- Host storage I/O performance can be improved while replication is occurring by reducing the concurrent number of volumes replicated.
- Increasing TCP window size can help increase the maximum throughput over high-bandwidth, high-latency links.
- Using Jumbo frames for replication can improve replication performance significantly for high speed links.

1 Introduction

Auto-replication is an enterprise feature included with each Dell EqualLogic PS Series storage array. This white paper focuses on Dell EqualLogic auto-replication and advanced EqualLogic auto-replication tuning for a few specific scenarios.

EqualLogic auto-replication is a point-in-time replication designed to be a Disaster Recovery (DR) solution that offers extended-distance replication. It provides asynchronous, incremental data synchronization from a primary to a secondary (DR) site. Every occurrence of replication updates the remote copy of the data with any changes that occurred on the primary copy since the last replication event. This provides protection from local failures and site disasters. In case of a disaster or failure, the replica can be used to recover data. Auto-replication occurs without incurring any down time for the primary volume and in most cases has minimal effect on performance of the primary volumes.

EqualLogic firmware 7.x has several enhancements that can improve the replication performance in certain specific scenarios. This white paper discusses link bandwidth, latency and packet loss, and the effects on replication performance. There are also support-only commands incorporated to fine tune replication parameters and improve performance. This white paper presents lab validated data and analysis to help IT and SAN administrators design and plan EqualLogic replication for the best performance.

Note: For an in-depth understanding of Dell EqualLogic auto-replication best practices, see the *Dell EqualLogic Auto-Replication: Best Practices and Sizing Guide* white paper at: http://en.community.dell.com/techcenter/storage/w/wiki/2641.dell-equallogic-auto-replication-best-practices-and-sizing-guide-by-sis.aspx

1.1 Audience

This technical white paper is for storage administrators, SAN system designers, storage consultants, or anyone who is tasked with configuring and administering EqualLogic PS Series storage for use in a production SAN. It is assumed that all readers have experience in designing or administering a shared storage solution. Also, 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) is assumed.

1.2 Terminology

Auto-replication: The built-in point-in-time replication feature included with every Dell EqualLogic PS Series array. The terms replication and auto-replication are used interchangeably.

Concurrent replications: Two or more volumes replicated simultaneously from a primary site to a secondary site.

Group: An EqualLogic PS group consists of one or more PS Series arrays connected to an IP network that work together to provide SAN resources to servers. A group may contain up to 16 arrays and is managed as a single storage entity.

Network latency: Time taken by a data packet to travel from one point to another over an Ethernet network.

Packet loss: Data packets sent over an Ethernet link may not reach the receiving end due to factors like congestion on the network, faulty network hardware, signal attenuation over long distance, and others. This is referred to as packet loss on the Ethernet network.

Primary Group: A group containing the source volume(s) to be copied or replicated.

Replica: A point-in-time copy of an EqualLogic volume stored in a Secondary Group.

Secondary Group: A group containing the replica or copy of the source volume(s).

WAN Emulator: A device used to simulate distance and impairments that may be encountered over a Wide Area Network (WAN).

2 Objectives

Replication performance between the primary and secondary storage array groups is dependent on the Ethernet network between the replication partners. There are many factors on the Ethernet network, like link speed, error rate, network congestion, distance and latency of the link that can affect replication performance. Similarly, on the storage arrays the hardware resources, pool configurations, host workloads and other factors also determine the replication performance.

EqualLogic arrays are configured to give the best performance with default settings under most scenarios. A few tunable settings were introduced on the EqualLogic storage arrays to enhance replication performance in certain configurations. These settings can be changed by using support-only commands which are reserved for EqualLogic support. Support-only commands are not documented and should not be tried without consulting EqualLogic support. The white paper focuses on how tuning these parameters can create an improvement in replication and host I/O performance.

Note: Support commands are restricted for the use of EqualLogic support or used under EqualLogic support supervision to make changes to EqualLogic storage array parameters. This paper discusses the use of these commands and their potential effect, however the exact syntax for each command is not documented publicly and can only be provided by engaging Dell support.

The paper focuses on the following scenarios and factors on the replication Ethernet link:

- Multiple volume replication
- Effects of reducing the number of concurrent volumes replicated
- Effects on I/O performance due to replication
- Replication bandwidth
- Network latency
- TCP window size
- Packet loss
- Ethernet MTU size

3 Technology overview

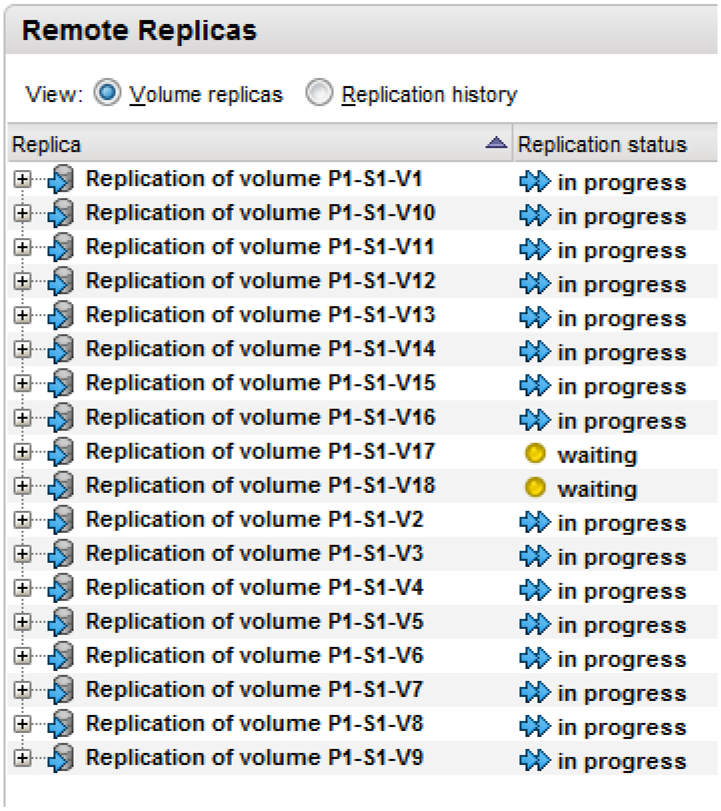
3.1 Concurrent volume replication on EqualLogic storage arrays

Auto-replication technology in EqualLogic supports enabling replication for up to 256 volumes simultaneously, but the default settings for concurrent replication replicates a round-robin set of up to 16 volumes. The number of volumes that can be concurrently replicated is called **Replication transfer size**. The default Replication transfer size is 16 volumes and can be changed to a value from 1 to 16 using support-only commands.

Note: The replication transfer size (number of volumes concurrently replicated) can be set to a whole number between 1 and 16 using support-only commands. The default value is 16.

When more volumes than the set value of the replication transfer size are configured for replication, the maximum number of volumes actively replicated will be limited to the replication transfer size. Volumes with ongoing replication are displayed as **in-progress** within the EqualLogic group manager GUI. If the number of volumes to be replicated is greater than the replication transfer size value, some volumes will be in a **waiting** state. Although the actual replication for the volumes in **waiting** state is delayed, the replica snapshot itself will still be from the point-in-time that the replica was created or initiated.

As shown in Figure 1, all the in-progress volumes show replication progress over time. As a volume completes replicating data, it is moved out of the queue, another volume in **waiting** state is added to the queue, and the volume state changes to **in-progress**.



Remote Replicas

View: ● Volume replicas ○ Replication history

Replica	Replication status
Replication of volume P1-S1-V1	in progress
Replication of volume P1-S1-V10	in progress
Replication of volume P1-S1-V11	in progress
Replication of volume P1-S1-V12	in progress
Replication of volume P1-S1-V13	in progress
Replication of volume P1-S1-V14	in progress
Replication of volume P1-S1-V15	in progress
Replication of volume P1-S1-V16	in progress
Replication of volume P1-S1-V17	waiting
Replication of volume P1-S1-V18	waiting
Replication of volume P1-S1-V2	in progress
Replication of volume P1-S1-V3	in progress
Replication of volume P1-S1-V4	in progress
Replication of volume P1-S1-V5	in progress
Replication of volume P1-S1-V6	in progress
Replication of volume P1-S1-V7	in progress
Replication of volume P1-S1-V8	in progress
Replication of volume P1-S1-V9	in progress

Figure 1 Volumes in in-progress and waiting state during replication

3.2 Replication and concurrent I/O activity

Auto-replication is designed as a background process and shares hardware resources with the processing of host I/O on the storage arrays. The static resources on the storage arrays must serve the incoming host I/O along with the background replication process of all of the **in-progress** volumes, and therefore it is possible that some host workloads may cause an impact on the replication performance or that replication performance may impact host I/O. However, an appropriately sized storage solution with sufficient disk spindles should show little performance impact when both types of I/O are in use.

In a production environment, volume replication is typically scheduled during non-work hours, for example at night or on weekends when I/O activity on the storage is low. Some businesses like to replicate their data every few hours or the data to be replicated is large enough that the time required would overlap with work hours. For such a situation, where volume replication and host I/O simultaneously demand resources from the storage arrays, the static resources of the storage arrays may be strained and may cause an impact on host I/O performance.

The incoming I/O from the servers is the highest priority, and the impact on the server I/O performance can be mitigated by adding additional array members to a pool, which increases the cumulative processing power of the pool and distributes the workload over more disk spindles. As a temporary alternative to increasing the number of arrays in a pool, an administrator can reduce the impact on host I/O performance by reducing the replication transfer size (number of concurrently replicated volumes) and releasing some of the resources used by the background replication process for the host I/O on the storage arrays.

3.3 Network latency and replication

A data packet moves at a fixed speed over the Ethernet network which causes the round trip time for the data packet to increase with an increase in distance traveled by the data packet. The maximum possible theoretical speed of data over the network is the speed of light in a fiber optic cable (124,188 miles per second). Table 1 shows the approximate theoretical distance a data packet can travel in a given time across fiber optic cables (values in the table should be doubled for round-trip calculations). Additionally the data traveling on the Ethernet network will have added latency from the networking equipment such as switches and routers, so the numbers in Table 1 are purely theoretical calculations using the following formula:

```
Distance travelled = 124188 miles per second x Time in seconds
```

Table 1 Theoretical one-way network latency due to distance.

Time in ms	Approximate distance traveled through fiber optic cable in Miles
1	124
20	2480
50	6200

The time taken for data to be replicated between replication partners depends of several factors which can be combined under Service time. Service time is the time for data to be retrieved from a storage system and delivered to an end device. It depends on disk time, storage processing time, propagation delay over the network, and other factors.

3.4 TCP window size over the replication link

The TCP window size is the number of bytes a sender can transmit before receiving an acknowledgment for the sent data over a TCP connection. The window size determines the amount of unacknowledged data. TCP window size is a 16-bit field in the TCP header and can have a maximum value of 64 Kb. To improve performance on high-bandwidth, high-delay networks, the TCP window scale option was introduced (RFC 1072 and RFC 1323) as a TCP header option. This supports scalable windows by allowing TCP to negotiate a scaling factor for the window size at connection establishment and allows a maximum TCP window of up to 1 GB.

A high-bandwidth link, though capable of transferring a higher amount of data, can be fundamentally limited if there is high latency on the network. This means that the maximum replication speed it can support depends on the latency of the network along with the bandwidth of the link. For a high-latency link, the maximum bandwidth possible from a TCP/IP session is equal to the TCP window size divided by the latency of the link. The maximum possible bandwidth is lower than the network bandwidth on a high-latency link. This is shown in Table 2 where the approximate theoretical maximum bandwidth for a given TCP window size and round trip time is calculated according to the formula below:

Theoretical latency-limited bandwidth = TCP window size / Network latency in sec

Table 2 Window size and theoretical latency-limited bandwidth (approximate)

Window size	1 ms	20 ms	50 ms
72 KB	600 Mbps	30 Mbps	11 Mbps
1 MB	8 Gbps	400 Mbps	160 Mbps
2 MB	16 Gbps	800 Mbps	330 Mbps

Increasing the TCP window size can sometimes improve performance by allowing more data to be in-flight. Increasing the TCP window size can improve replication performance for high-speed, high-latency links or when there is a long distance between the replication partners. The default maximum window size on EqualLogic arrays for a replication TCP connection is 72 Kb but can be changed and increased up to 2 MB using a support-only command. The actual TCP connection window size in use depends on the CPU load of the storage array and other factors affecting the network.

3.5 Jumbo frame enabled network for replication

A standard TCP/IP packet uses a MTU of 1500 bytes, which includes 20 bytes of IP header and 20 or more bytes of TCP header depending on the TCP options used. TCP/IP protocol carries an overhead of TCP and IP headers leaving only 1460 bytes Maximum Segment Size (MSS) or less of data to be transferred per packet. To reduce the overhead of TCP/IP headers and increase the data transfer rate over the TCP/IP connection, Dell recommends using Jumbo frames, which are TCP/IP packets of 9000+ bytes, with EqualLogic storage arrays. A data payload of 8960 bytes can be transferred within each Jumbo frame.

Jumbo frames along with reducing the overhead for every data packet sent helps improve performance over links with significant packet loss. Ethernet links can have packet loss due to factors like congestion on the network, faulty network hardware, signal attenuation over long distances, or other factors. Packet loss on a TCP connection activates the TCP congestion avoidance algorithm. The TCP congestion avoidance algorithm treats packet loss as feedback to discover congestion, and when it is detected, it throttles back the amount of packets being sent over the connection. Replication performance is affected by packet loss on the network, so this needs to be taken into consideration as a network variable.

EqualLogic storage arrays by default attempt to use Jumbo frames for replication between the two replication partners to increase the replication performance. In case of a failure to establish a Jumbo frame connection between replication partners, it will revert back to a replication connection with standard frames.

Note: Jumbo frames for a replication connection between Primary and Secondary sites are negotiated separately from the iSCSI connection between the storage arrays and the host.

4 Test topology and architecture

The test to study the effects of the configuration parameters stated in Section 2 required an environment where two replication partners, a **Primary Group** and **Secondary Group**, were interlinked using a WAN emulator to create different scenarios on the Ethernet network and storage arrays. Servers were connected to the SAN of the Primary Group to generate a workload for some test cases.

A lab environment configured with two groups, a **Primary Group** and **Secondary Group,** each consisting of three Dell EqualLogic PS6110X 10 Gb arrays was used for testing. Each group member was connected with two Dell Force10 S4810 TOR switches interconnected with a Link Aggregation Group (LAG) creating a separate SAN network for the primary site and another for the secondary site. The two SAN networks for the Primary and Secondary groups were independent of each other except for the WAN connection between the two sites. The Primary Group replicated its data to the Secondary Group for DR.

To simulate a production environment, Dell PowerEdge R620 servers running Windows Server 2008 R2 were connected to the SAN of the Primary site for generating I/O workload on the Primary Group. This allowed for the study of the I/O performance during replication activities. This configuration was used for all testing done in this paper, unless mentioned otherwise.

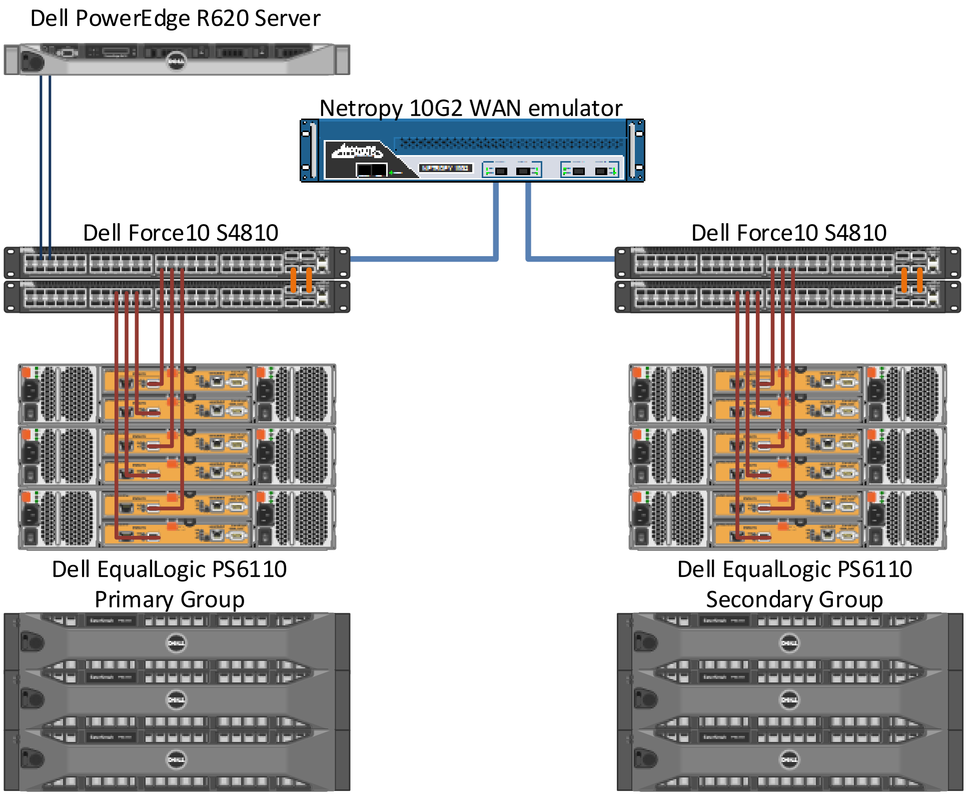


Figure 2 Test topology and architecture

Depending on the distance and networking hardware between the sites, replication performance can vary. To simulate a WAN connection for a variety of scenarios, a WAN emulator was used to connect the Primary site with the Secondary site. Apposite's Netropy 10G2 WAN emulator was used to test the replication with different scenarios on the WAN link.

5 Test methodology

To test one-way replication from the Primary Group to the Secondary Group, multiple 10 GB volumes were created on the Primary Group and replicated to the Secondary Group. Data change of 10% between replications is very typical, so 10 GB volumes were selected to represent 10% data change for a 100 GB volume. Similar results would be expected for higher volume sizes and larger data change between replications. The replication test was carried out on three to 18 volumes depending on the test performed.

The WAN emulator was configured to emulate link speed, error rate, latency and other network scenarios between the Primary and Secondary replication partners. It was configured initially before the start of each test and no changes were made on the WAN link during the test run, which provided a consistent network for replication during each test.

If the parameters on the EqualLogic storage arrays were changed for the test, then, using the support-only commands, these parameters were configured on both the Primary and Secondary sites before the start of the test.

Each test was run three times and the average of the results was taken to get consistent and reproducible test results. Replication performance depends on many variables such as CPU availability on the storage arrays, memory availability on the arrays, and others, which were not controlled on this test environment. The data generated may vary for different environments depending on the variables described above.

5.1 Concurrent volume replication test

This test studied the effects of replicating multiple volumes simultaneously and is described further in Section 6.1. For this test case, the replication transfer size value was varied in increments from 1 to 16 and its effects were studied on the replication activity for three to 18 volumes.

Eighteen volumes of 10 GB each were created on the Primary Group and were replicated on the Secondary Group. The actual number of volumes replicated varied depending on the test variable as shown below. For this test case, the WAN emulator was configured to simulate different Ethernet link speeds between the Primary and Secondary sites. Host I/O was generated from the Dell PowerEdge server on the Primary Group storage arrays to simulate an application workload and to test effects of replication on host I/O. Using the VDbench I/O generation tool, 30 threads of 8K block size, random I/O with 70% read and 30% write were generated. No host I/O was generated on the Secondary site.

Test variables for concurrent volume replication test:

- Number of Volumes: 3, 9, 12, 18
- Replication Transfer Size: 1, 3, 9, 12, 16 (default)
- Link Speed: 10Gb, 1Gb, OC3, T3

All 80 unique combinations of the test variables (number of volumes, replication transfer size, and link speed) were tested. Host IOPS, Host I/O latency and replication performance, measured as time taken for replication, were used to determine the effects of concurrent EqualLogic volume replication.

5.2 Network latency and TCP window size test

This test scenario used the WAN emulator to simulate added network latency and then increased the TCP windows size over the default setting. Testing was carried out using three volumes of 10 GB each. The WAN emulator emulated 10 Gb, 1 Gb, OC3, and T3 Ethernet link speeds. For this test, no host I/O was generated on the Primary or Secondary site.

Test variables for Network latency and TCP window size test:

- TCP window size: 72 KB (default), 1024 KB
- Network latency: 0 ms, 1 ms, 10 ms, 20 ms, 50 ms
- Link Speed: 10 Gb, 1 Gb, OC3, T3

The test variables as listed above are used for the Network Latency and TCP Window Size tests. In all, 40 unique combinations of the TCP window size, network latency, and link speed were tested. Replication performance, measured as time taken for replication, was measured and used to determine the effects of network latency and TCP window size on EqualLogic volume replication.

5.3 Jumbo frames versus Standard frames test

This test determined the effects of Jumbo versus Standard frames and used the WAN emulator to simulate an appropriate network. First, Jumbo frames were allowed on the WAN emulator (default setting) to simulate a network capable of carrying Jumbo Ethernet frames. Then, during the Standard frame test, the WAN emulator was limited to carrying standard frames. Different link speeds and packet loss scenarios were emulated on the WAN emulator. Three volumes of 10 GB each were replicated from the primary to the secondary site and their replication times were measured. No host I/O was generated during this test.

Test variables for Jumbo frames versus Standard frames test:

- Ethernet frame size: Jumbo frame (MTU 9000+) (default), Standard frame (MTU 1500)
- Packet loss: 0%, 0.1%, 1%
- Link Speed: 10 Gb, 1 Gb, OC3, T3

All 24 unique combinations of Ethernet frame size, packet loss, and link speed, as shown above, were tested. Replication performance, measured as time taken for replication, was measured and used to determine the effects of Jumbo frames versus Standard frames on EqualLogic volume replication.

6 Test results and analysis

6.1 Concurrent volume replication on EqualLogic storage arrays

6.1.1 Efficiency of concurrent replication

Figure 3 demonstrates the replication time for multiple volumes simultaneously replicated over a 10 Gb link. Here, the time taken for replicating 18 volumes is three times longer than the time taken to replicate three volumes. If it took 100 minutes to replicate three volumes, then replicating 18 total volumes, in groups of three volumes at a time, would take 600 minutes total (6 x 100 minutes). On the contrary, if we concurrently replicate all 18 volumes, the time it takes over a 10 Gb link is only 300 minutes total. This clearly shows the advantage of replicating multiple volumes simultaneously because the storage arrays are able to use the entire 10 Gb link available for replication between the primary and the secondary site. The performance of replicating multiple volumes depends on the available bandwidth and other network and storage factors. As the available bandwidth between the primary and secondary sites decreases, the difference between replication time for three volumes and 18 volumes increases. For OC3 bandwidth (data not shown in graph), the replication time required for replicating 18 volumes simultaneously is six times that of replication for three volumes. This means that replicating 18 volumes simultaneously is equal to replicating 18 volumes as three volumes at a time.

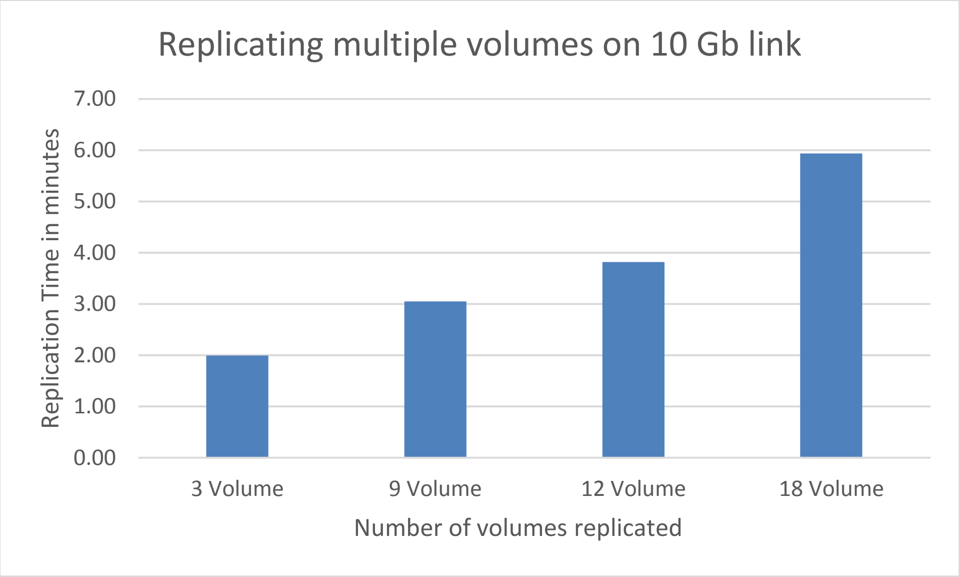


Figure 3 Comparing replication time for three to 18 volumes. (Link speed = 10 Gb, replication transfer size = 16)

6.1.2 Effects of replication transfer size on host I/O

The number of IOPS from a storage array is limited by the number of disk spindles available. I/O from the host workload and from replication activity share this maximum I/O capability of the storage system. Reducing the replication transfer size value reduces the replication I/O on the storage arrays and I/O from the host workload increases to a higher value and lower I/O latency. The effects of replication on the incoming storage I/O for changing replication transfer size is illustrated in Figure 4. The replication time increased by 3.5 times going from the replication transfer size of 16 to one and the IO rate and latency improved by 1.5 times. This clearly shows how, at the cost of replication performance, the performance of host I/O can be improved by reducing the number of volumes concurrently replicated. An ideally designed storage system will never encounter this issue. However, when a situation arises where the IOPS from the storage system need to be increased, the storage administrator can choose to temporarily change the replication transfer size to improve IOPS at the cost of increased replication time.

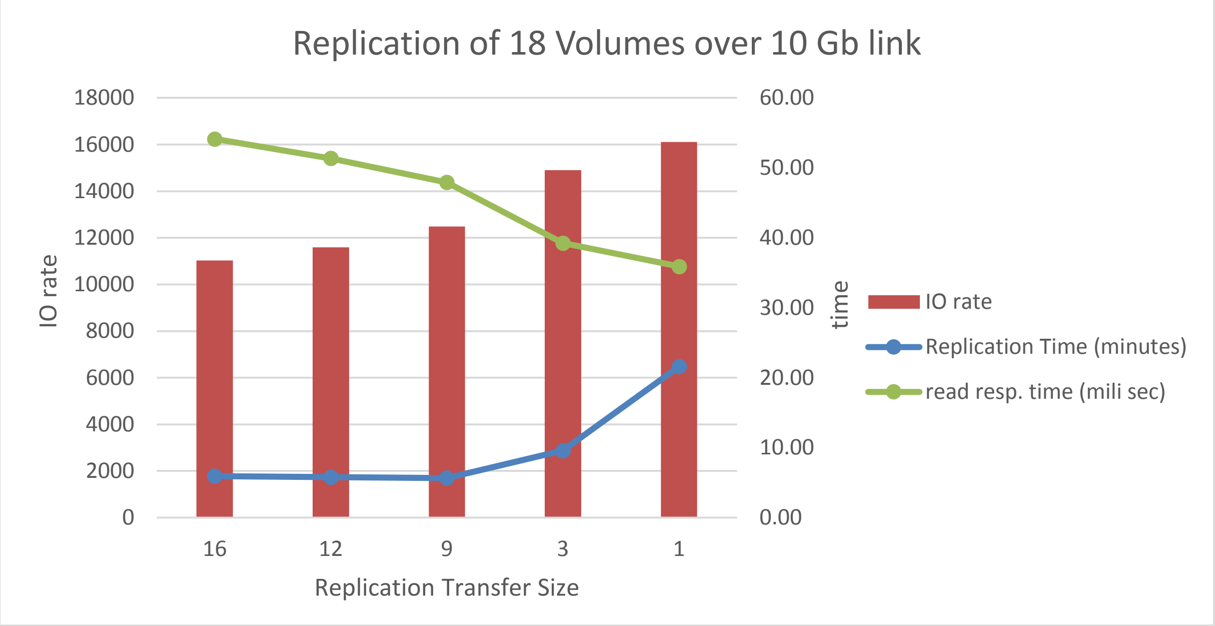


Figure 4 Replication and I/O performance for replication of 18 volumes. (Number of Volumes = 18, Link speed = 10 Gb)

The performance of random 8K I/O was tested with concurrent replication of 3, 9, 12 and 18 volumes, and Figure 4 plots results for 18 volume replication. Figure 5 plots the replication time and incoming host IOPS for 3, 9, 12 and 18 volumes replicated over the 10 Gb link while changing the replication transfer size from 16 to 1. The graph depicts that in a situation where the storage arrays are overloaded with a high number of volumes replicated concurrently, the performance of ongoing IOPS might be lower while a portion of array resources are attending to background replication activity. This is not likely to occur with a well-designed and properly-sized storage solution.

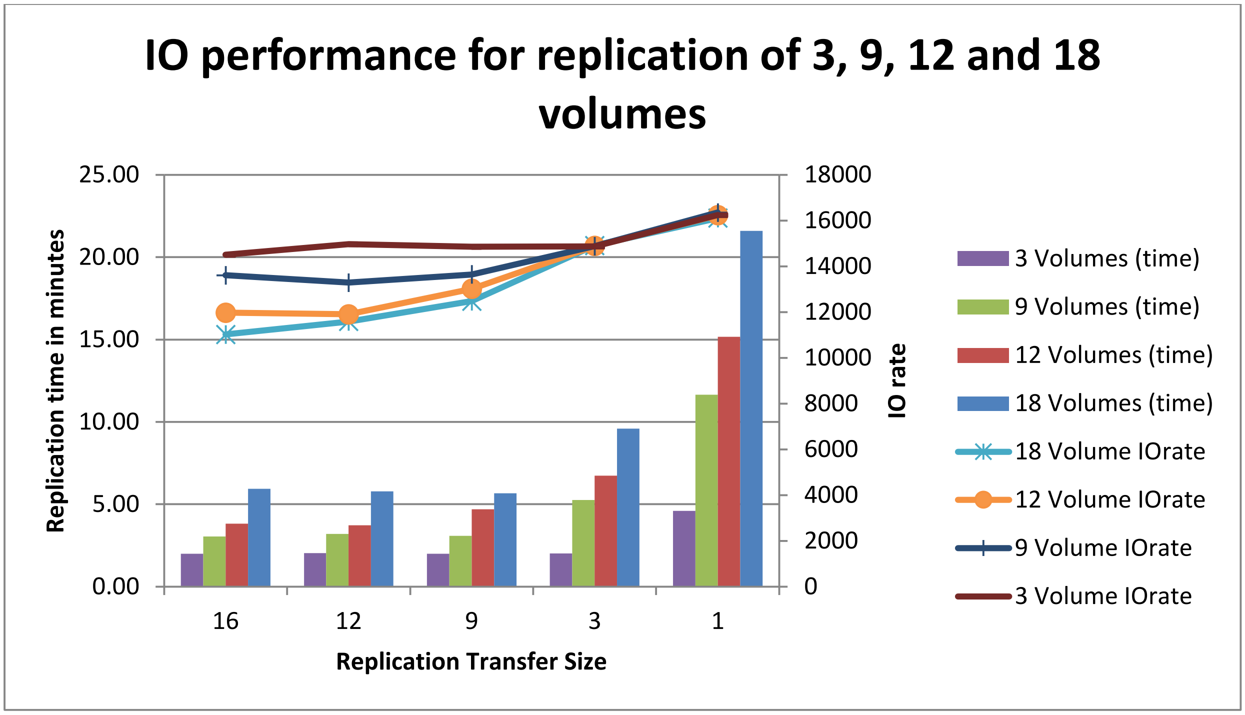


Figure 5 Replication and I/O performance over 10 Gb replication link (Link speed = 10 Gb)

Note: EqualLogic customer support must run the support-only command to change the value of the replication transfer size. Please contact support to make any support-only changes on the EqualLogic arrays.

The above results applies to 10 Gb replication links between the replication groups. Although as the bandwidth decreases, the effects of reducing the number of volumes simultaneously replicated also decreases. As the available bandwidth for replication is reduced, the storage array does not have enough replication link bandwidth to maintain its full capability for replication. This means the storage array is already using much less of its resources for replication, and reducing the number of concurrent volume transfers will not significantly impact host I/O performance.

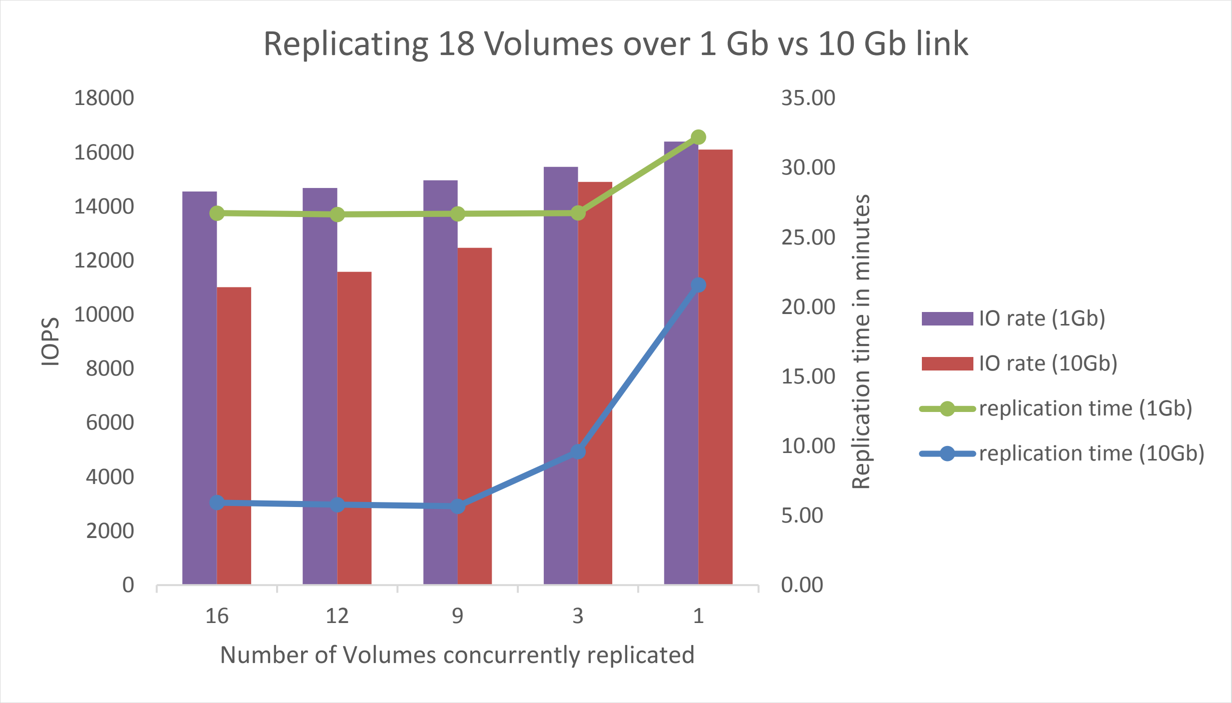


Figure 6 Replication and I/O performance for replicating 18 Volumes over 10 Gb and 1 Gb links (Number of volumes = 18, Link speed 1Gb and 10Gb)

The chart above shows the performance comparison between the 1 Gb and 10 Gb replication link and the behavior of the host I/O on the storage system. The number of simultaneous replications was changed from 16 to one for a total of 18 volumes replicated.

Figure 6 shows a significant increase in the IO rate and the time taken to finish replication at three concurrent volumes being replicated. This shows how hardware resources released by the replication process are used by host I/O on the storage arrays increasing host I/O performance by 150% in the case of the 10 Gb bandwidth and 10-15% with 1 Gb bandwidth. The improvement diminishes as the available replication bandwidth decreases and the storage array is not able to send more data due to the bandwidth limitations with lower bandwidth links. At OC3 link speed (not shown in the chart), there is no significant improvement seen from reducing the number of simultaneous transfers. Therefore, change in replication transfer size is beneficial to host IOPS for high-bandwidth links (1 Gb and higher), while for link speeds of OC3 and below, the default value of the replication transfer size (16) should be used.

6.2 Effects of network latency and TCP window size on replication

6.2.1 Network latency in replication

This test for replication over a 10 Gb link with latency induced using a WAN emulator shows how latency on the network can affect the replication performance. With higher latency values, the effective bandwidth of the replication link is reduced.

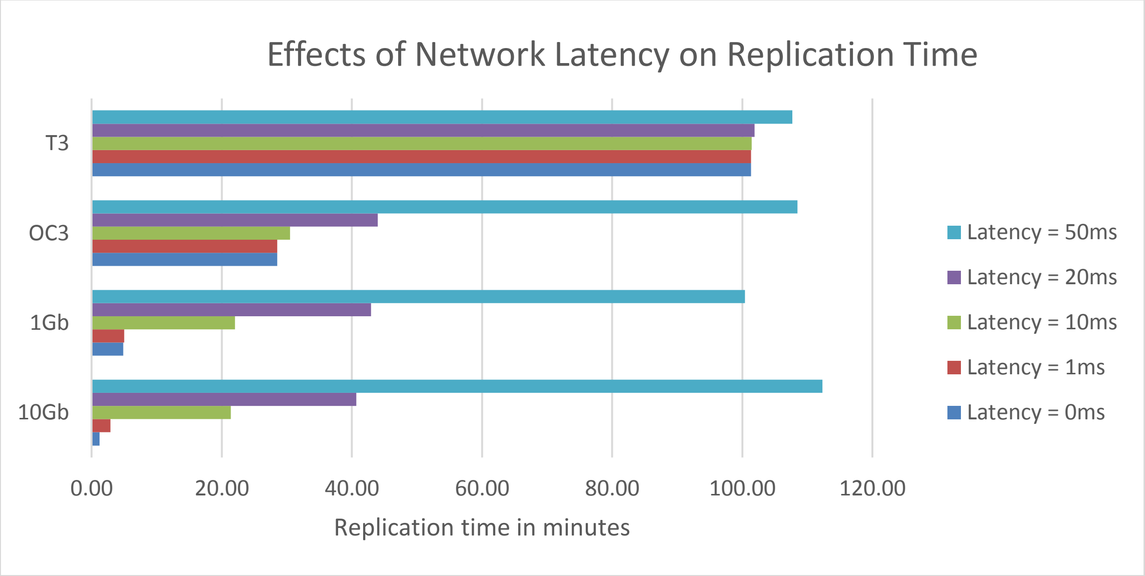


Figure 7 Network latency effect on Replication (TCP window size = 72 KB default)

Replication behavior for three volumes being replicated on a network with varied latencies and bandwidths is presented in the graph above in Figure 7. No host I/O operations were performed on the storage system during this test. For (one way) latencies of 0 ms, 1 ms, 10 ms, 20 ms and 50 ms (round trip time is twice that of the one-way latency), the replication time for 10 Gb, 1 Gb, OC3 and T3 bandwidth indicates that the effective bandwidth is reduced for high-latency links. The graph demonstrates using higher bandwidth and a low latency link can reduce the replication time between the Primary and the Secondary sites. As the link latency grows, the effective link bandwidth is reduced. The time for replication over a 50 ms one-way latency link (100 ms round trip) is similar for 10 Gb, 1 Gb, OC3 and T3 link speeds.

Note: Network latency is inherently added by the physical length of the Ethernet link, routers, switches and other network hardware. Network latency of 0 ms represents no additional latency added by the WAN emulator.

6.2.2 Effects of TCP window size over high latency replication links

The graph in Figure 8 displays the observations made during the lab test when the WAN emulator was used to induce latency between the Primary and Secondary replication groups, and link speeds of 10 Gbps, 1 Gbps, OC3 and T3 were emulated. The storage arrays were configured to use the default of 72 KB TCP window size and a 1024 KB TCP window size for replication.

Note: The default TCP window size of 72 KB for EqualLogic replication can be changed by customer support only through a support-only command. This change in window size is a persistent change. Please contact support to make any changes using support-only commands on the EqualLogic arrays.

Results of the test showed improvement in replication performance by increasing the TCP window size for replication connections. The results may vary for each specific environment. On a 10 Gbps Ethernet link, using 1 MB window size for 20 ms and 50 ms (one way) latency links helps improve replication performance over the default 72 KB window size. The replication time for a 50 ms latency link reduces by 15% when using 1 MB window size. Similarly for 1 Gbps and OC3 link, using 1 MB window size reduces the replication time by 8% and 12% respectively compared to using the 72 KB TCP window. The observed data shows how replication time can be reduced by increasing the TCP window size for replication connections.

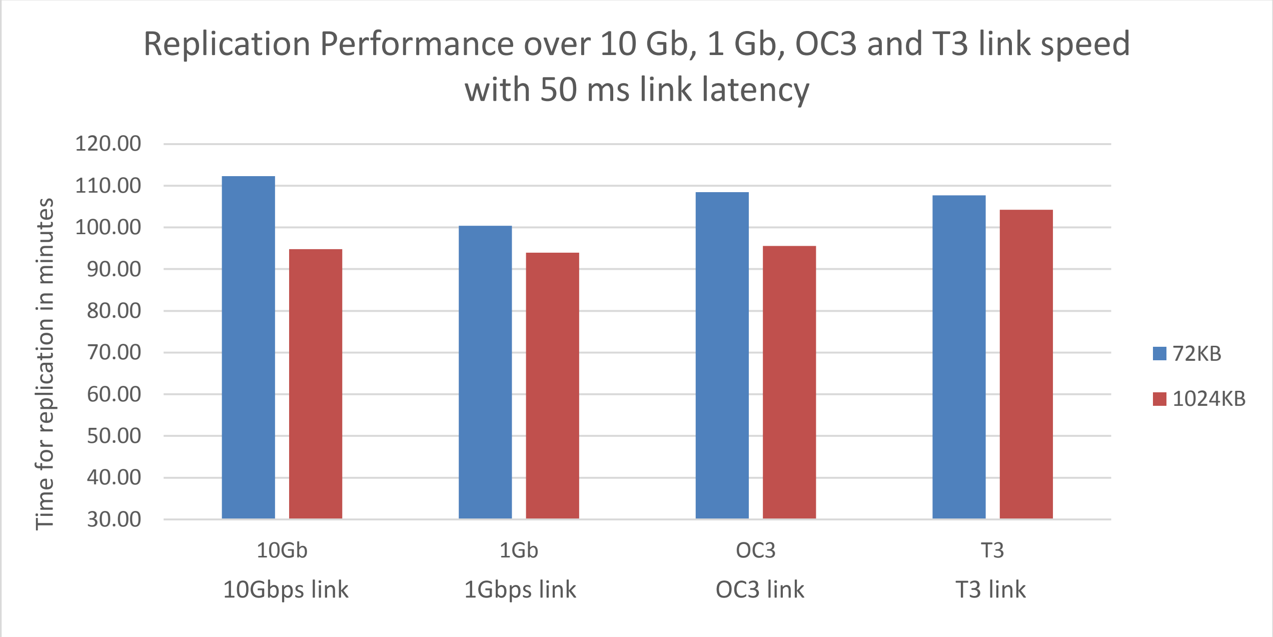


Figure 8 Improvements in replication time by increasing TCP window size for high-latency link (Link Latency = 50 ms)

The advantage seen by increasing the window size at 10 Gb, 1 Gb and OC3 link speeds for a high-latency link reduces as the bandwidth between the replication partners goes lower. Limited bandwidth causes limited capacity of the link to carry data, in turn reducing the effects of increasing the window size.

6.3 Effects of using a Jumbo frame enabled network for replication

6.3.1 Advantages of Jumbo frames over a lossless link

This section focuses on the performance advantage gained by using Jumbo frames compared to Standard Frames for replication over a lossless link. The objective of the analysis is to demonstrate when a Jumbo frame capable network is needed.

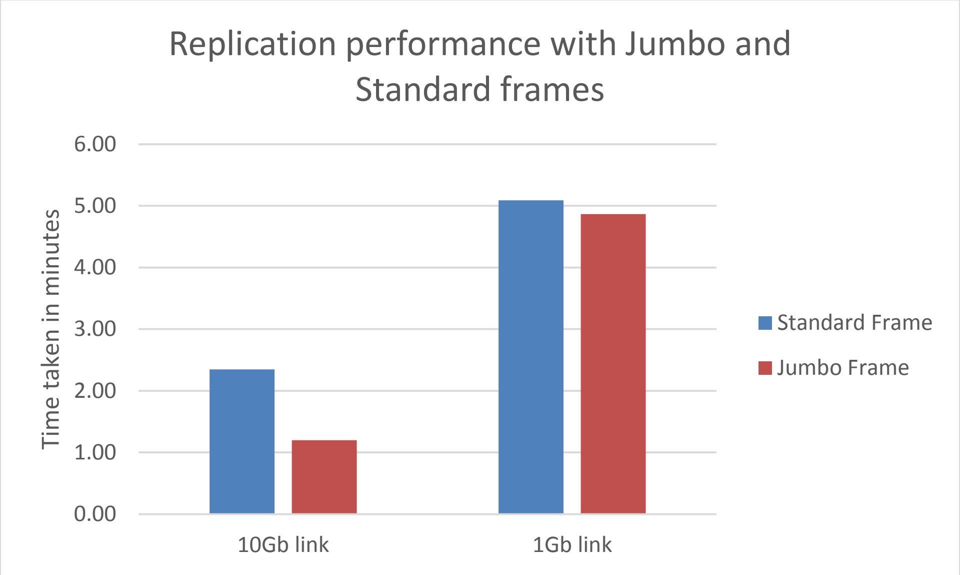


Figure 9 Effects of Ethernet Jumbo frames vs Standard frame size on replication (Link speed is 10Gb and 1Gb, Packet loss = 0%)

As shown in Figure 9, when replication partners are connected using a 10 GB link, using Jumbo frames helps reduce the replication time by almost half. Using Jumbo frames can be highly advantageous for high bandwidth links. The 1 Gb link shows the replication time is 10%-15% lower when using Jumbo frames instead of Standard frames. Tests with lower speeds of OC3 and T3 showed no significant improvement while using Jumbo frames compared to Standard frames.

6.3.2 Packet loss and advantage of using Jumbo frames

This section analyzes the effects of using Jumbo frames on a network which has a loss of 0.1% and 1%.

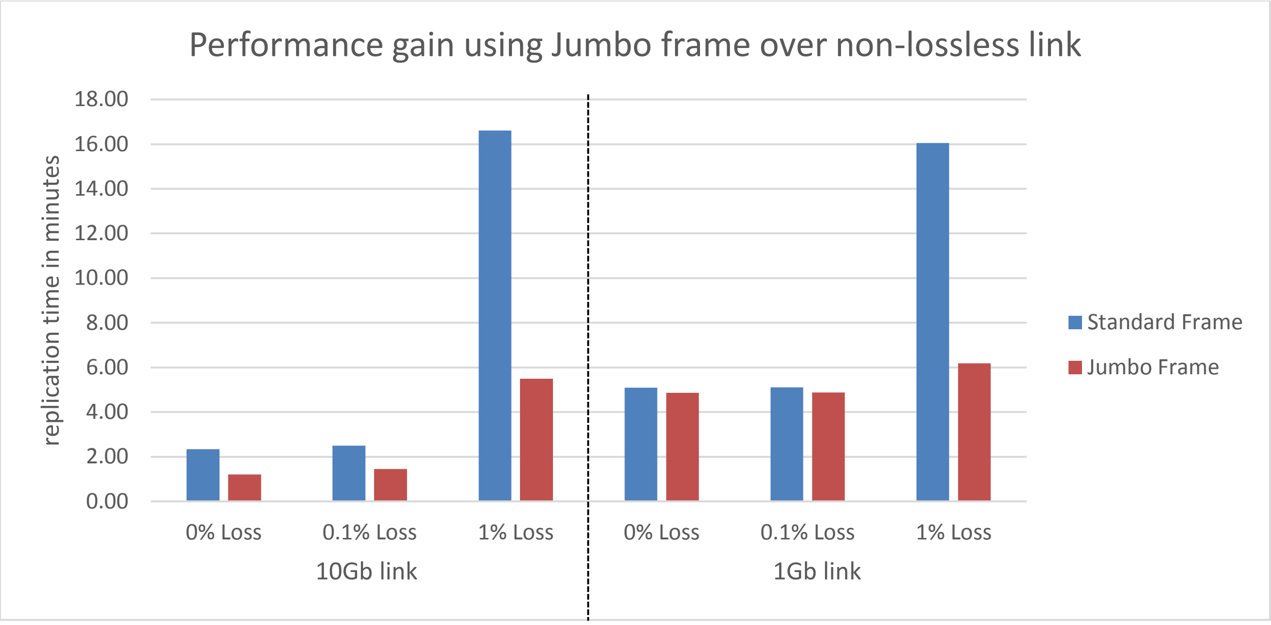


Figure 10 Replication performance over a link with packet loss (Link speed = 10 Gb and 1 Gb, Packet loss = 0%, 0.1% and 1%)

The data in Figure 10 shows that using Jumbo frames on a 10 Gb link is more effective than using Jumbo frames on a 1 Gb link and lower speed links with packet loss. The advantage of Jumbo frames over Standard frames diminishes as the link bandwidth is lowered to OC3 and T3. For a link with packet loss, the TCP congestion avoidance throttles the traffic, but using Jumbo frames helps it to recover more quickly compared to using standard frames. The replication test with Jumbo frames uses higher bandwidth even after the packet loss of 0.1% to 1% is induced by the WAN emulator. This is also shown in the graph in Figure 10 where the 10 Gb link with 0.1% and 1% packet loss shows Jumbo frames performing 40% and 67% better compared to Standard frames. Similarly for the 1 Gb link, Jumbo frames performs 60% better over standard frames for 1% packet loss.

Using Jumbo frames may be helpful in improving performance of replication over a link with significant packet loss or over a WAN network which is congested.

7 Recommendations for tuning auto-replication

EqualLogic storage arrays give the best performance with default settings under most scenarios. Most often there are no changes needed on the storage array to get good replication performance. In a few scenarios the replication time and host I/O performance may be enhanced by making a few changes on the EqualLogic storage arrays. Tuning of these settings for auto-replication scenarios is summarized below.

7.1 Heavy host I/O on the storage array during auto-replication

Reducing the replication transfer size helps release hardware resources used by the replication process. These hardware resources can be used by storage arrays for other processes like serving incoming host I/O.

In scenarios where storage arrays are subjected to host I/O and auto-replication simultaneously, the constraint on array hardware resources can be decreased by reducing replication transfer size temporarily. It improves host I/O performance at the cost of replication performance. A balance between host I/O workload and replication performance must be determined.

When replicating within the same datacenter or replicating over a shorter distance where high bandwidth links are available, it is recommended to replicate multiple volumes simultaneously to efficiently use the available bandwidth.

7.2 High speed, high-latency Ethernet links

Increasing TCP window size helps to efficiently use link bandwidth and improve replication performance over long distances and high-latency links. For efficient use of the available bandwidth on a high-latency link, the TCP window size may be increased up to 2 Mb. Remote office or branch office environments where data is replicated over long distances can take advantage of this increase in the TCP window size.

In general any environment where the network latency between replication partners is greater than the ratio of TCP window size to available bandwidth can benefit from increasing TCP window size.

7.3 Data packet loss over Ethernet link

Jumbo frames can help replication performance overall and helps over a high bandwidth link with packet loss. It can be helpful to mitigate the effects of high loss on the network for 1 Gb and higher link speeds. Jumbo frames might not be available over a long distance WAN network, but for replication between a single datacenter and environments with a dedicated link between Primary and Secondary sites, you should enable Jumbo frames (on switches and routers) because it gives a huge advantage in replication performance over Standard frames.

Note: A WAN optimizer can help optimize the replication performance over low speed and/or high-latency links, or when large amounts of data must be replicated. The advantages of a WAN optimizer are not explored in this white paper but should be considered when higher replication performance is required.

7.4 Additional auto-replication best practices

Designing and planning of replication for disaster recovery of a storage system requires careful consideration of Recovery Time Objective, Recovery Point Objective, replication space allocation, network connectivity of the Primary and Secondary site, the WAN network connecting the replication sites and other factors.

Detailed auto-replication best practices can be found in the following white paper:

Dell EqualLogic Auto-Replication: Best Practices and Sizing Guide
http://en.community.dell.com/techcenter/storage/w/wiki/2641.dell-equallogic-auto-replication-best-practices-and-sizing-guide-by-sis.aspx

A Configuration details

Table 3 Component details

Component	Description
EqualLogic firmware	Storage Array Firmware V7.0.2
Switch firmware	FTOS 9.2
Netropy firmware	Firmware version 2.0
Vdbench (Load generation tool)	Version 5.04.01
Cabling	Fiber optic cable used for connection.
Server	Dell R620 server
OS	Windows Server 2008 R2 Enterprise

B 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 installations.

Referenced or recommended Dell publications:

- Dell EqualLogic Auto-Replication: Best Practices and Sizing Guide
http://en.community.dell.com/techcenter/storage/w/wiki/2641.dell-equallogic-auto-replication-best-practices-and-sizing-guide-by-sis.aspx
- Understanding Data Replication Between Dell EqualLogic PS Series Groups
http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/19861448.aspx
- EqualLogic Compatibility Matrix
http://en.community.dell.com/techcenter/storage/w/wiki/2661.equallogic-compatibility-matrix.aspx
- EqualLogic Configuration Guide
http://en.community.dell.com/techcenter/storage/w/wiki/2639.equallogic-configuration-guide.aspx