Windows Server 2012 and Windows Server 2012 R2 NIC Optimization and Best Practices with Dell PS Series

Dell EMC Engineering
January 2017

A Dell EMC Best Practices Guide

Revisions

Date	Description
July 2013	Initial release
September 2013	Updated configuration recommendations
January 2017	Include Windows 2012 R2 Delayed ACK configuration using PowerShell

Acknowledgements

Updated by: Chuck Farah

The information in this publication is provided “as is.” Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

Copyright © 2017 Dell Inc. or its subsidiaries. All Rights Reserved. Dell, EMC, and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be the property of their respective owners. Published in the USA. [1/4/2017] [Best Practices Guide] [BP1057]

Dell EMC believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Table of contents

Executive summary

This document explores the configuration options available for improving Dell™ PS Series SAN performance using the Broadcom® BCM57810 or Intel® X520 10 GbE network adapters and Microsoft® Windows Server® 2012 and Windows Server 2012 R2 on a Dell EMC™ PowerEdge™ 12th-generation server. Recommended OS and NIC configurations are given based on the results of SAN performance testing.

1 Introduction

Dell PS Series arrays provide a storage solution that delivers the benefits of consolidated networked storage in a self-managing iSCSI storage area network (SAN) that is affordable and easy to use, regardless of scale.

In every iSCSI SAN environment, there are numerous configuration options at the storage host which can have an effect on overall SAN performance. These effects can vary based on the size and available bandwidth of the SAN, the host/storage port ratio, the amount of network congestion, the I/O workload profile and the overall utilization of system resources at the storage host. Furthermore, a particular setting might greatly improve SAN performance for a large block sequential workload while having an insignificant or slightly negative effect on a small block random workload. Another setting might improve SAN performance at the expense of host processor utilization.

Keeping these ideas in mind, this technical paper quantifies the effect on iSCSI performance, as measured in throughput and IOPS, of several configuration options within the Broadcom and Intel 10 GbE adapter properties and the Windows Server 2012 TCP stack during three common SAN workloads. It also takes into account the value of certain settings in congested network environments and when host processor resources are at premium. From the results, recommended configurations for a Dell PS Series SAN are given for each tested NIC type.

In order to focus on the pure SAN performance benefits of the tested configuration options, Data Center Bridging (DCB) and Broadcom Switch Independent Partitioning, also known as NIC Partitioning (NPAR) were excluded from testing.

Note: The performance data in this paper is presented relative to baseline configurations and is not intended to express maximum performance or benchmark results. Actual workload, host-to-array port ratios, and other factors may also affect performance.

1.1 Audience

This technical white paper is for storage administrators, SAN system designers, storage consultants, or anyone who is tasked with configuring a host server as an iSCSI initiator to Dell 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, 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).

2 Technical overview

iSCSI SAN traffic takes place over an Ethernet network and consists of communication between Dell PS Series array member network interfaces and the iSCSI initiator of storage hosts. The Broadcom BCM57810 NetXtreme® II and the Intel X520 10GbE network adapters were used as the iSCSI initiators during this project.

The Broadcom BCM57810 network adapter features iSCSI Offload Engine (iSOE) technology which offloads processing of the iSCSI stack to the network adapter. When using iSOE mode, the network adapter becomes a host bus adapter (HBA) and a host-based, software iSCSI initiator is not utilized. This is as opposed to non-iSOE mode in which the network adapter functions as a traditional NIC and works with a software iSCSI initiator. This paper refers to the non-iSOE mode of operation as NDIS mode. In Windows, NDIS refers to the Network Driver Interface Specification, a standard application programming interface (API) for NICs.

The following three initiator modes of operation were tested:

- Broadcom BCM57810 NDIS mode
- Broadcom BCM57810 iSOE mode
- Intel X520

Appendix A provides a detailed list of the tested configuration options and default values for each NIC type as well as for the Windows Server 2012 TCP stack.

3 Test configurations and methodology

The following section addresses the reasoning behind SAN design decisions and details the SAN configurations. Performance testing methodology, test case sequence, and results analysis are also explained.

3.1 Simplified SAN

Every effort was made to simplify and optimize the test configurations so that the performance effects of each option could be isolated. The following configuration and design elements helped to achieve this goal.

- Host Integration Toolkit for Microsoft (HIT/Microsoft) with default settings
- All unused protocols disabled for each network adapter
- Eight volumes within a single storage pool, evenly distributed across array members
- An isolated SAN with no LAN traffic
- Load balancing (volume page movement) disabled on the array members
- DCB was not used
- The NIC bandwidth was not partitioned

Note: Windows Server 2012 natively supports NIC Load Balancing and Failover (LBFO). Dell recommends using the HIT/Microsoft and MPIO for any NIC that is connected to Dell PS Series iSCSI storage. The use of LBFO is not recommended for NICs dedicated to SAN connectivity since it does not add any benefit over MPIO. Microsoft LBFO or other NIC vendor specific teaming can still be used for non-SAN connected interfaces. If DCB is enabled in a converged network, LBFO can also be used for NIC partitions that are not SAN connected.

Load balancing is recommended for production environments because it can improve SAN performance over time by optimizing volume data location based on I/O patterns. While it is enabled by default, it was disabled for performance testing to maintain consistent test results.

The following two SAN designs were chosen. See appendix A for more detail about the hardware and software infrastructure.

3.1.1 Base SAN configuration

The first SAN design chosen was a basic SAN with a redundant SAN fabric and an equal number of host and storage ports. Having a 1:1 host/storage port ratio is ideal from a bandwidth perspective. This helped to ensure that optimal I/O rates were achieved during lab testing. Figure 1 shows only the active ports of the Dell PS Series array members.

The following configuration was used:

- Two switches in a LAG configuration
- Two array members each with a single port connected
- A single host with two 10 GbE NIC ports
- A 1:1 storage/host port ratio

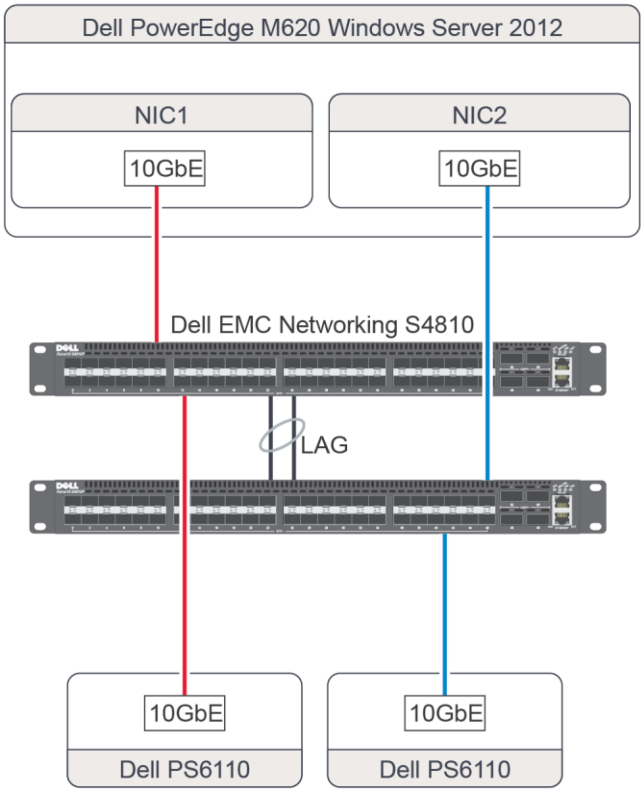

Dell PowerEdge M620 Windows Server 2012
NIC1
10GbE
NIC2
10GbE
Dell EMC Networking S4810
LAG
10GbE
Dell PS6110
10GbE
Dell PS6110

Figure 1 Physical diagram of the base SAN configuration

3.1.2 Congested SAN configuration

The second SAN design was constructed to mimic a host port experiencing network congestion. In this SAN design, a single host port was oversubscribed by four storage ports for a 4:1 storage/host port ratio. Since only one host port existed, the SAN fabric was reduced to a single switch. Figure 2 shows only the active ports of the Dell PS Series array members.

Note: A non-redundant SAN fabric is not recommended for a production SAN environment.

The following configuration was used:

- One switch
- Four array members each with one port connected
- A single host with one 10 GbE NIC
- A 4:1 storage/host port ratio

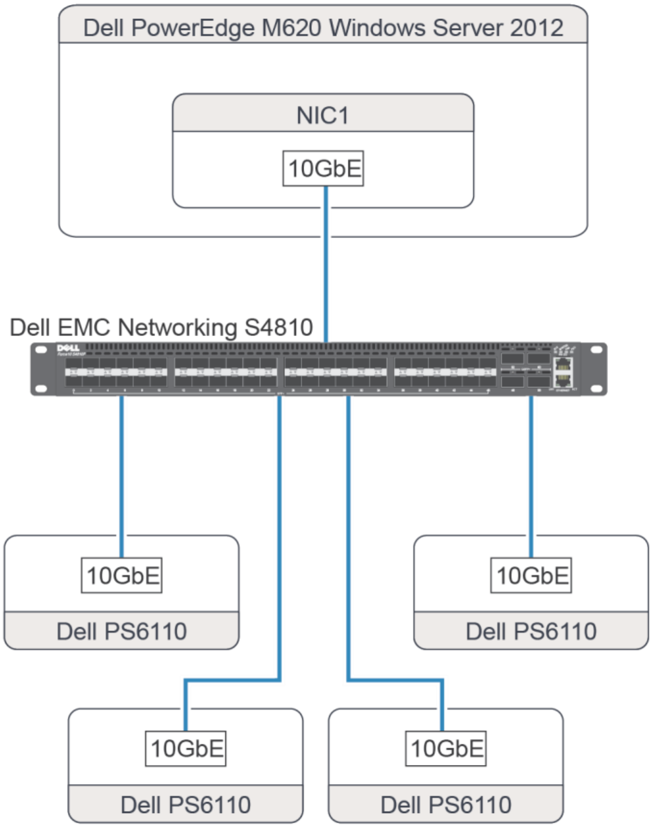


Figure 2 Physical diagram of the congested SAN configuration

3.2 I/O execution and evaluation

Prior to each test run, the host was restarted to confirm configuration changes were in effect. After boot, the even distribution of iSCSI connections across host and storage ports and of active array member ports across SAN switches was confirmed.

The following three vdbench workloads were run:

- 8 KB transfer size, random I/O, 67% read
- 256 KB transfer size, sequential I/O, 100% read
- 256 KB transfer size, sequential I/O, 100% write

For every test case, each vdbench workload was run three times for twenty minute durations and the results were averaged.

Vdbench IOPS results were used to evaluate 8K random workload performance. Vdbench throughput results were used to evaluate 256K sequential workload performance. Host and array member retransmission rates and processor utilization were also examined.

See appendix C for a list of vdbench parameters.

3.3 Statistical analysis

In order to ensure stable results, the relative standard deviation among the three performance results for each test case workload was calculated. When the relative standard deviation was greater than one percent for a given test case workload, that particular workload performance test was re-run.

3.4 Test case sequence

The performance and effect on system resources of adapter and Windows Server 2012 TCP stack options were evaluated using the test cases listed in this section.

Initially, tests were run to compare a totally default configuration to the baseline configuration chosen for testing. The baseline configuration included the following non-default settings:

- Jumbo frames enabled for the NIC (see the best practice recommendations in section 5); the Dell EMC Networking S4810 switch should set Jumbo frames to 12K on all ports.
- Flow control enabled for the NIC (if not already by default); the Dell EMC Networking S4810 switch should be configured with flow control on for receive and off for transfer.
- Maximum receive and transmit buffers for the NIC (if applicable)

Each subsequent test case consisted of a single option being toggled from the default setting to show its effect relative to the baseline configuration defined previously. In test cases where a technology option had an adapter and a corresponding OS setting (for example, Receive Side Scaling), both settings were changed simultaneously prior to the test case execution.

3.4.1 Broadcom BCM57810 NDIS mode test case sequence

The following tables show the test case sequence used to evaluate the effect of tested configuration options for the Broadcom BCM57810 in NDIS mode. **Bold text** indicates the changed value for each test scenario.

Table 1 Baseline test case sequence for Broadcom BCM57810 NDIS mode

Test case	Frame size	Flow Control	Rx/Tx buffers	Other adapter settings	Windows Server TCP stack setting	Comments
1	Standard	On	Default	Default	Default	Default configuration
2	**Jumbo**	On	Default	Default	Default	Jumbo performance effect
3	Jumbo	On	**Maximum**	Default	Default	Baseline configuration to evaluate all subsequent settings

Table 2 Test case sequence for Broadcom BCM57810 NDIS mode to evaluate other configuration options

Test case	Frame size	Flow Control	Rx/Tx buffers	Other adapter settings	Windows Server TCP stack setting	Comments
4	Jumbo	On	Maximum	**Interrupt moderation disabled**	Default	
5	Jumbo	On	Maximum	**Receive Side Scaling (RSS) disabled**	**RSS disabled**	
6	Jumbo	On	Maximum	**RSS queues of 16**	Default	RSS enabled by default at adapter and in Windows Server TCP stack
7	Jumbo	On	Maximum	**Receive Side Coalescing (RSC) disabled**	**RSC disabled**	
8	Jumbo	On	Maximum	**TCP Connection Offload enabled**	**Chimney enabled**	Both settings required to activate TCP Offload Engine (TOE)
9	Jumbo	On	Maximum	Default	**Receive Window Auto-tuning disabled**	
10	Jumbo	On	Maximum	Default	**Delayed ACK algorithm disabled**	

Test case	Frame size	Flow Control	Rx/Tx buffers	Other adapter settings	Windows Server TCP stack setting	Comments
11	Jumbo	On	Maximum	Default	**Nagle's algorithm disabled**	
12	Jumbo	On	Maximum	**Large Send Offload (LSO) disabled**	Default	

3.4.2 Broadcom BCM57810 iSOE mode test case sequence

The following table shows the test case sequence used to evaluate the effect of tested configuration options for the Broadcom BCM57810 in iSOE mode. **Bold text** indicates the changed value for each test scenario.

As seen in Table 3, when in iSOE mode the Broadcom 57810 has a much more limited set of adapter options. Also, in iSOE mode the Windows Server 2012 TCP stack options have no effect since the entire iSCSI and TCP stack are offloaded to the adapter by design. **Bold** text indicates the changed value for each test scenario.

Table 3 Test case sequence for Broadcom BCM57810 iSOE mode

Test case	Frame size	Flow Control	Rx/Tx buffers	Other adapter settings	Windows Server TCP stack	Comments
1	Standard	Auto	N/A	N/A	N/A	Default configuration
2	**Jumbo**	Auto	N/A	N/A	N/A	Jumbo performance effect
3	Jumbo	**On**	N/A	N/A	N/A	Recommended settings

3.4.3 Intel X520 test case sequence

The following table shows the test case sequence used to evaluate the effect of tested configuration options for the Intel X520. **Bold text** indicates the changed value for each test scenario.

Table 4 Baseline test case sequence for Intel X520

Test case	Frame size	Flow Control	Rx/Tx buffers	Other adapter settings	Windows Server TCP stack	Comments
1	Standard	On	Default	Default	Default	Default configuration
2	**Jumbo**	On	Default	Default	Default	Jumbo performance effect
3	Jumbo	On	**Maximum**	Default	Default	Baseline configuration to evaluate all subsequent settings

Table 5 Test case sequence for Intel X520 to evaluate other configuration options

Test case	Frame size	Flow Control	Rx/Tx buffers	Other adapter settings	Windows Server TCP stack	Comments
4	Jumbo	On	Maximum	**Interrupt moderation disabled**	Default	
5	Jumbo	On	Maximum	**Receive Side Scaling (RSS) disabled**	**RSS disabled**	
6	Jumbo	On	Maximum	**RSS queues of 16**	Default	RSS enabled by default at adapter and in Windows Server TCP stack
7	Jumbo	On	Maximum	**RSS profile of NUMA scaling**	Default	RSS enabled by default at adapter and in Windows Server TCP stack
8	Jumbo	On	Maximum	**Receive Side Coalescing (RSC) disabled**	**RSC disabled**	
9	Jumbo	On	Maximum	Default	**Receive Window Auto-tuning disabled**	
10	Jumbo	On	Maximum	Default	**Delayed ACK algorithm disabled**	
11	Jumbo	On	Maximum	Default	**Nagle’s algorithm disabled**	
12	Jumbo	On	Maximum	**Large Send Offload (LSO) disabled**	Default	

4 Results and analysis

All test case performance results for each NIC mode and workload combination are presented in this section. For the sake of analysis, a five percent margin of error is assumed and only performance differences greater than this were acknowledged as significant.

Based on the results, recommended configurations for each NIC mode are given in section 5.

4.1 Baseline testing

The following non-default settings were used as a baseline configuration for further testing and evaluation.

4.1.1 Jumbo frames

Jumbo frames enable Ethernet frames with payloads greater than 1,500 bytes. In situations where large packets make up the majority of traffic and additional latency can be tolerated, jumbo packets can reduce processor utilization and improve wire efficiency.

Figure 3 illustrates the dramatic effect that enabling Jumbo frames can have on large block workloads. Significant throughput increases were observed for both read and write large block workloads on both the Broadcom and Intel network adapters in all supported operating modes (such as NDIS and iSOE).

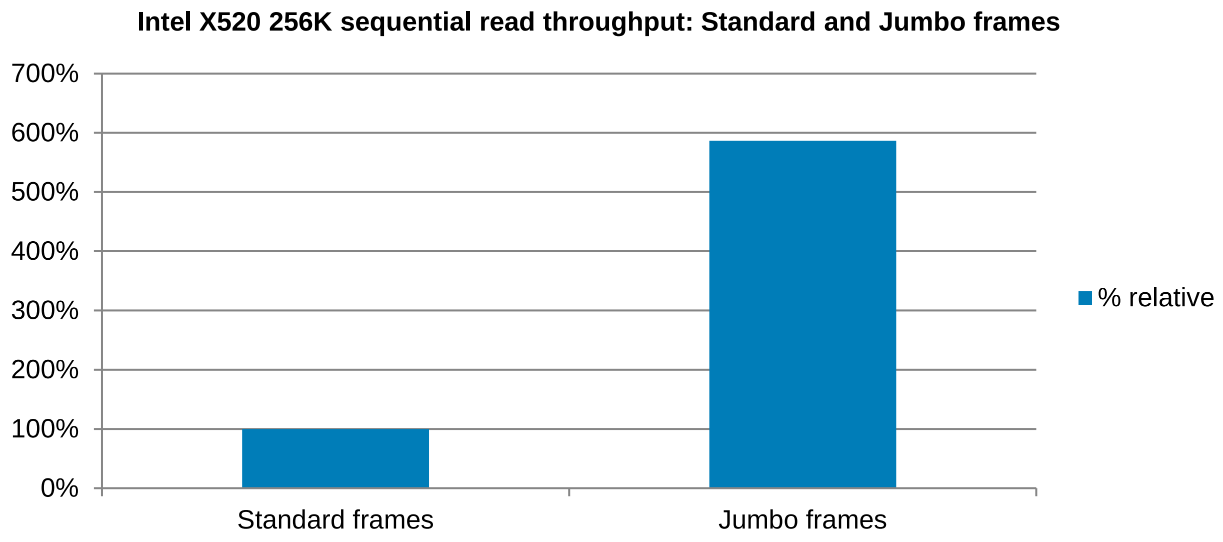


Figure 3 Performance effect on 256K sequential read workload of enabling Jumbo frames on an Intel X520

4.1.2 Flow control

Flow control is a link-level mechanism that enables the adapter to respond to or to generate flow control (PAUSE) frames, helping to regulate network traffic. Flow control is enabled by default and is well-known to be of benefit in a congested network environment.

4.1.3 Receive and Transmit buffers

Rx/Tx buffers are used by the adapter when copying data to memory. Increasing this value can enhance performance with only a slight increase in system memory utilization. Maximizing buffer allocation is particularly important on a server with heavy processor utilization and can also be beneficial during times of network congestion.

4.2 Other available performance tuning options

For both the network adapter and the Windows Server 2012 TCP stack, there are other options available which can have an effect on SAN performance under the right circumstances. This section defines these options and discusses the results of the performance testing. It is important to understand that the performance results described below may not translate to all Dell PS Series SAN environments. The material presented in this section identifies setting and workload combinations that have a clearly positive or negative impact on iSCSI SAN performance. It is recommended that each potential configuration change be evaluated in the environment prior to implementation.

For more information on performance tuning options for the networking subsystem see *Performance Tuning Guidelines for Windows Server 2012* at
http://msdn.microsoft.com/en-us/library/windows/hardware/jj248719.aspx

Note: The tuning guidelines and recommendations in *Performance Tuning Guidelines for Windows Server 2012* are for TCP/IP network interfaces in general and are not particular to iSCSI networking.

4.2.1 Interrupt moderation

With interrupt moderation, a network adapter attempts to reduce the number of interrupts required by sending a single interrupt for multiple events rather than an interrupt for each event. While this can help lower processor utilization, it can also increase SAN latency. It is enabled by default on both the Broadcom and Intel adapters. Neither a significant performance enhancement nor degradation was observed when it was disabled.

4.2.2 Receive Side Scaling (RSS)

RSS balances incoming traffic across multiple processor cores, up to one logical process per processor core. This can improve the processing of receive-intensive workloads when the number of available logical processors outnumbers network adapters. It is enabled by default on both the Broadcom and Intel adapters and globally in Windows Server 2012.

In addition to disabling RSS to test against the baseline (RSS enabled), two primary customizations to RSS were also tested:

- Increasing the number of RSS queues to the maximum: A larger number of queues increases network throughput at the expense of processor utilization.
- Changing the RSS load balancing profile from Closest Processor to NUMA Scaling: The Closest Processor setting dynamically balances the RSS load across the available processors while the NUMA Scaling dynamically assigns processors in a round robin across the available NUMA nodes.

Outright disabling of RSS reduced 256K sequential read throughput by 30 percent on the Intel adapter. No other significant performance effects were observed on other workloads or on the Broadcom adapter.

Increasing the number of RSS queues or changing the RSS load balancing profile did not have a significant effect on any workload performance for either the Broadcom or Intel adapter.

4.2.3 Receive Segment Coalescing (RSC)

RSC enables the adapter to collect packets from a TCP/IP stream and combine them into larger packets, thus reducing the number of IP headers that must be processed. Like RSS, RSC is meant to benefit receive-intensive workloads. It is enabled by default on both the Broadcom and Intel adapters and globally in Windows Server 2012.

When RSC was disabled, no significant performance effects were observed for any workload on either the Broadcom or Intel adapters.

4.2.4 TCP Offload Engine (TOE)

TOE offloads the processing of the entire TCP/IP stack to the network adapter and is available on the Broadcom 57810 network adapter. For TOE to be active, TCP connection offload must be enabled on the network adapter and chimney support must be enabled in Windows Server 2012, both of which are disabled by default.

While TOE did not significantly improve performance in any of the three workloads, it did have a positive effect on storage array member retransmission and processor utilization. In the congested SAN configuration during the 256K sequential read workload, the array member retransmission rate was reduced from .11 percent when disabled to zero when enabled. The processor utilization decreased from ~6 percent to ~4 percent. The TOE effects are illustrated in Figure 4.

4.2.5 Large send offload

LSO enables the adapter to offload the task of segmenting TCP messages into valid Ethernet frames. Because the adapter hardware is able to complete data segmentation much faster than operating system software, this feature may improve transmission performance. In addition, the adapter uses fewer processor resources. It is enabled by default on both Broadcom and Intel adapters. Disabling LSO resulted in decreased performance in almost every workload.

4.2.6 TCP checksum offload

TCP checksum offload enables the adapter to verify received packet checksums and compute transmitted packet checksums. This can improve TCP performance and reduce processor utilization and is enabled by default. With TCP checksum offload disabled, iSCSI connection instability and increased array packet retransmission were observed during testing.

4.2.7 TCP receive window auto-tuning

TCP window auto tuning enables Windows Server 2012 to monitor TCP connection transmission rates and increase the size of the TCP receive window if necessary. It is enabled by default. Disabling auto-tuning reduced 256K sequential read workload throughput by 30 percent on both Broadcom and Intel adapters.

4.2.8 Delayed ACK algorithm

The delayed ACK algorithm is a technique to improve TCP performance by combining multiple acknowledgment (ACK) responses into a single response. This algorithm has the potential to interact negatively with a TCP sender utilizing Nagle's algorithm, since Nagle's algorithm delays data transmission until a TCP ACK is received. Though disabling this algorithm had no effect on the performance of any tested workload, there are cases where disabling TCP delayed ACK for an iSCSI interface may improve performance. For example, poor read performance and unreliable failover have been observed during periods of network congestion on Microsoft iSCSI cluster nodes. In certain cases, disabling TCP delayed ACK on iSCSI interfaces might be recommended by Dell Support.

4.2.9 Nagle's algorithm

Nagle's algorithm is a technique to improve application TCP performance by buffering output in the absence of an ACK response until a packet's worth of output has been reached. This algorithm has the potential to interact negatively with a TCP receiver utilizing the delayed ACK algorithm, since the delayed ACK algorithm may delay sending an ACK under certain conditions up to 500 milliseconds.

While disabling Nagle's algorithm did not demonstrate an effect on the performance results of the sustained workloads, there may be cases where disabling Nagle's algorithm on the storage host improves iSCSI SAN performance. Bursty SAN I/O or a high frequency of iSCSI commands can trigger ACK delays and increase write latency. As with disabling TCP delayed ACK, Dell Support might recommend disabling Nagle's algorithm on iSCSI interfaces.

4.2.10 iSCSI Offload Engine (iSOE)

iSOE offloads the processing of the entire iSCSI stack to the network adapter and is available on the Broadcom 57810 network adapter. For iSOE to be active it must be enabled in the Broadcom Advanced Control Suite (BACS). By default it is disabled.

Once the iSOE function is enabled for the adapter (and the NDIS function disabled), it disappears from the native Windows Server networking administration and monitoring tools and appears as a storage controller in Windows device manager. It must be configured and managed through BACS.

In the base SAN configuration, iSOE increased the IOPS of the 8K random read/write workload by over 30 percent while increasing the throughput of the 256K sequential read workload by 10 percent relative to the baseline NDIS configuration. 256K sequential write throughput was unchanged.

Like TOE, iSOE had a positive effect on storage array member retransmission and processor utilization. In the congested SAN configuration during the 256K sequential read workload, the array member retransmission rate was reduced from .11 percent when disabled to zero when enabled. The processor utilization decreased significantly from ~6 percent to ~1.5 percent. See Figure 4 for an illustration.

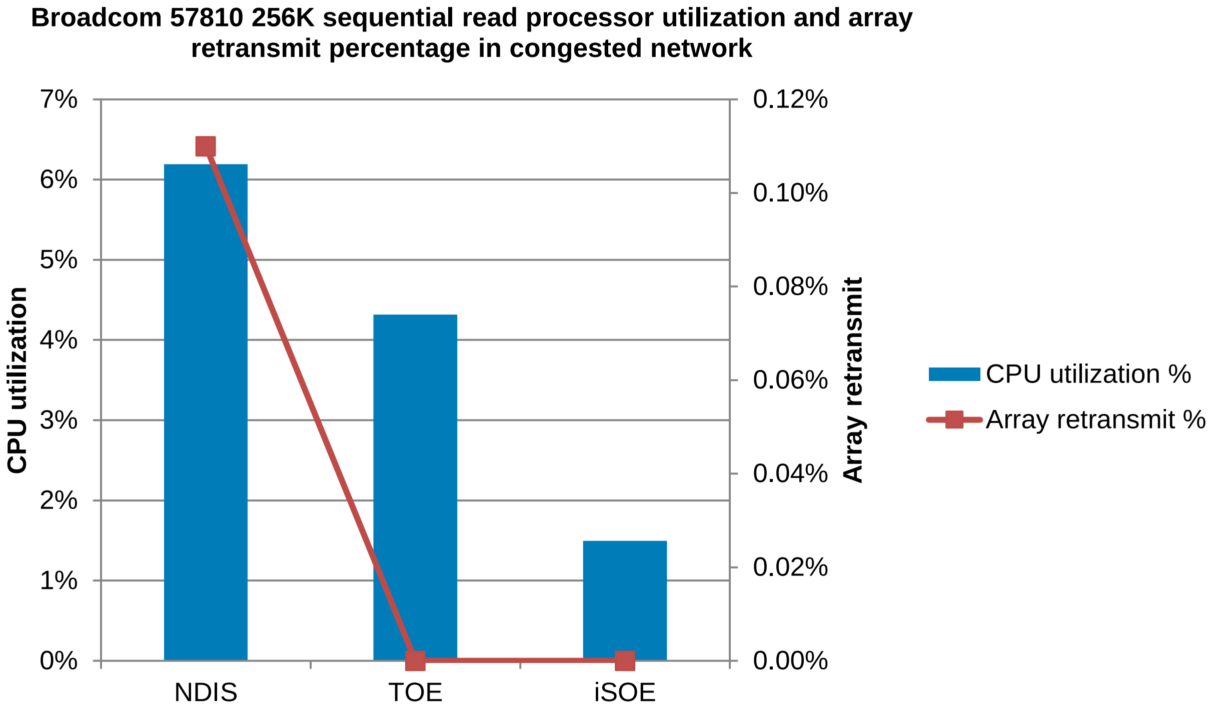


Figure 4 Processor utilization and array retransmit percentage of a Broadcom 57810 in NDIS, TOE, and iSOE mode while running the 256K sequential read workload in the congested SAN configuration

4.3 Broadcom BCM57810 NDIS mode performance results

This section shows the performance results when using the different configuration options during the three tested workloads using the Broadcom BCM57810 network adapter in NDIS mode.

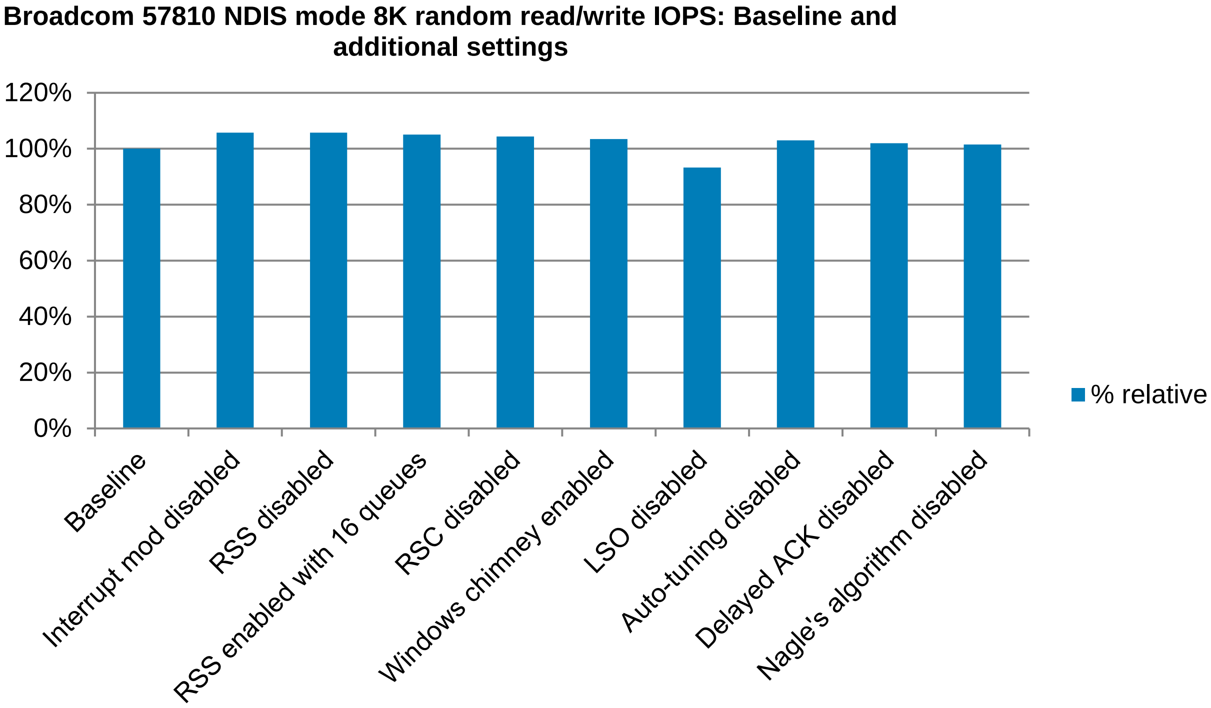


Figure 5 Broadcom 57810 NDIS mode, 8K random 67% read workload, IOPS relative to the baseline configuration.

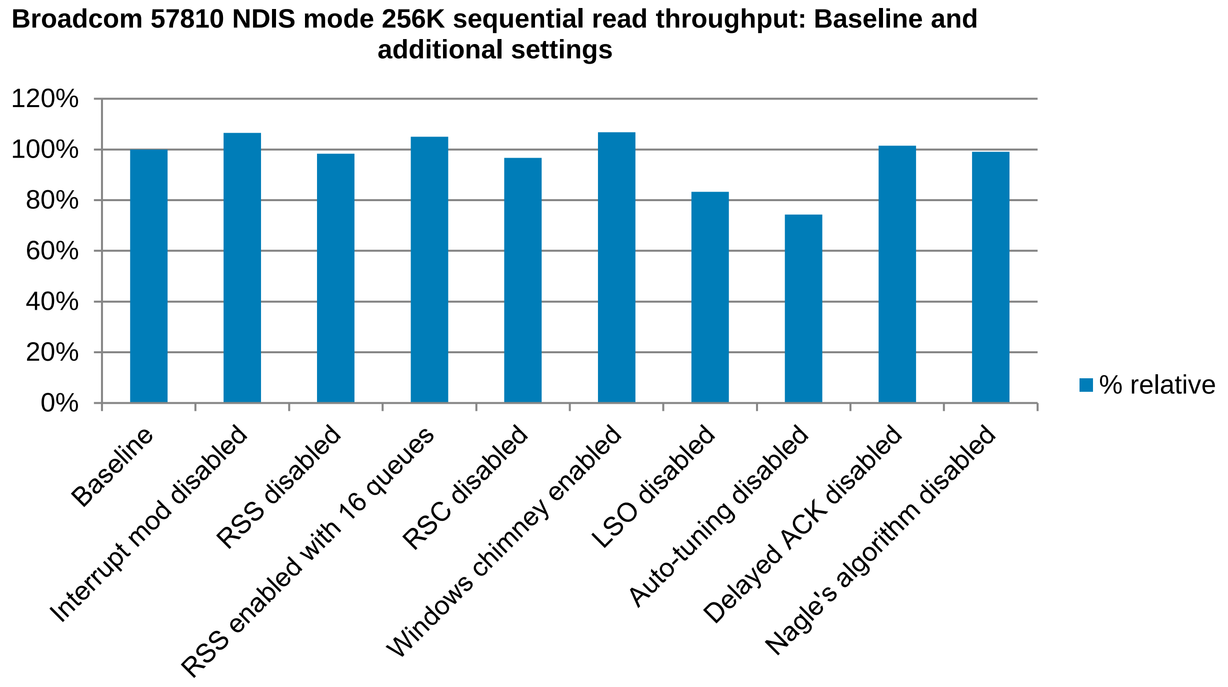


Figure 6 Broadcom 57810 NDIS mode, 256K sequential read workload, throughput relative to the baseline configuration

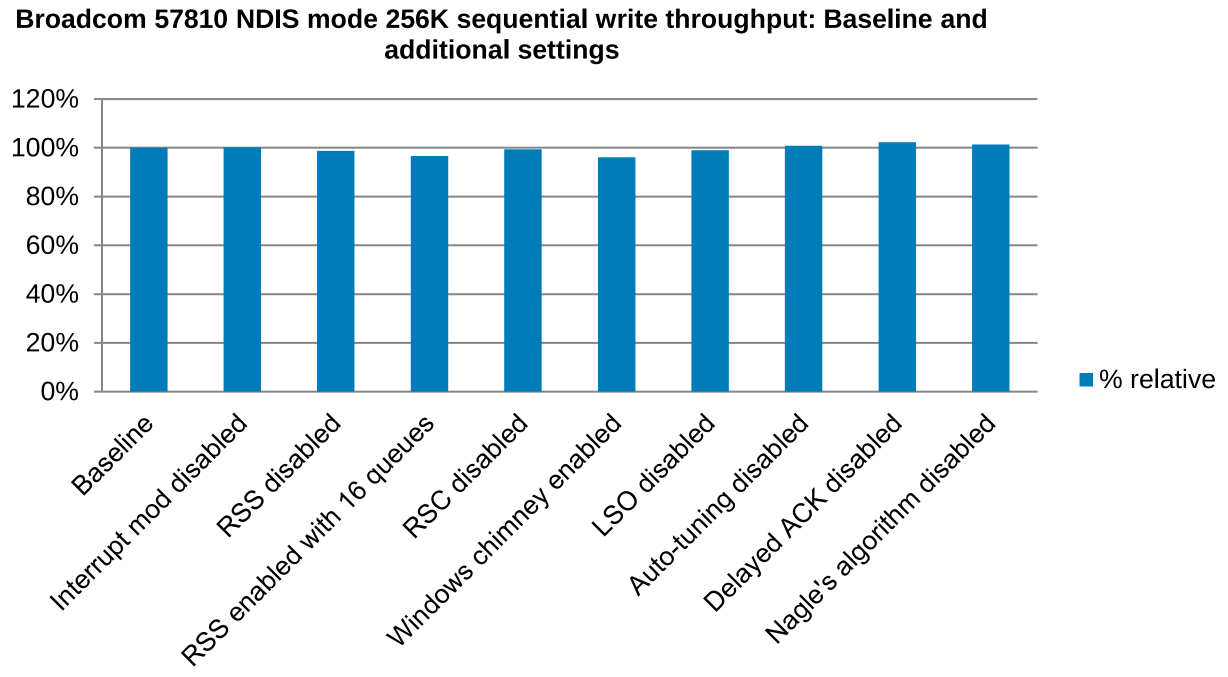


Figure 7 Broadcom 57810 NDIS mode, 256K sequential write workload, throughput relative to the baseline configuration

4.4 Broadcom BCM57810 iSOE mode performance results

This section compares the performance results with iSOE mode to those of NDIS mode and NDIS mode with TOE during the three tested workloads using the Broadcom BCM57810 network adapter.

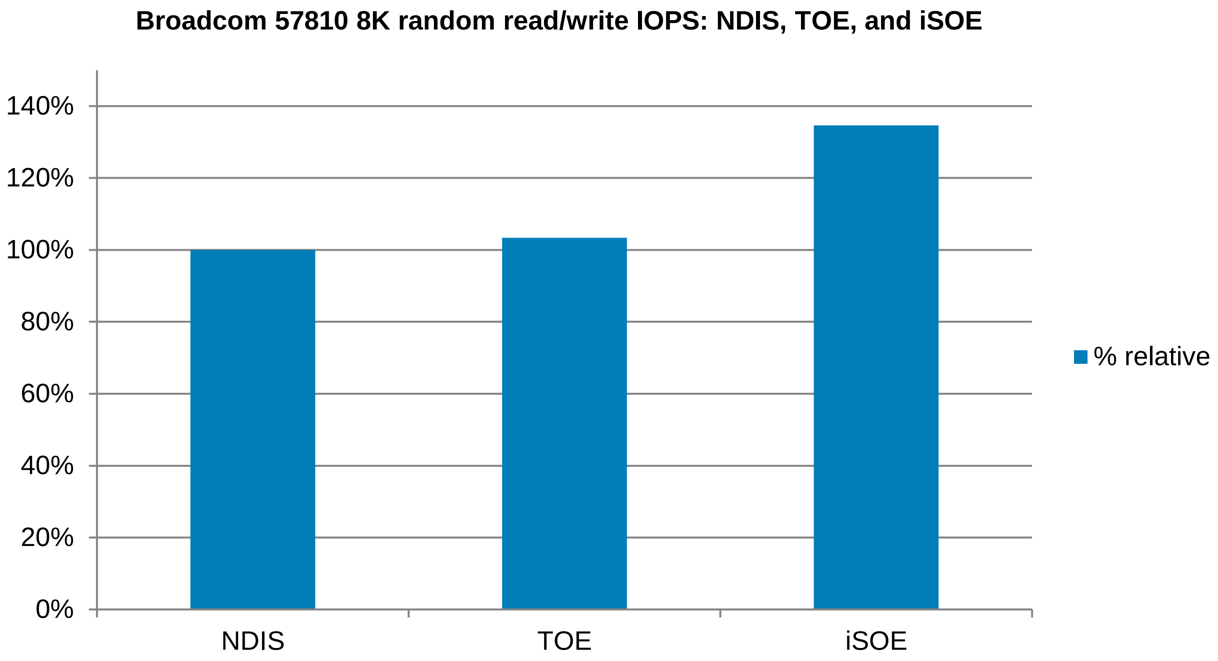


Figure 8 Broadcom 57810 iSOE mode, 8K random 67% read workload, IOPS relative to the baseline NDIS configuration and to the baseline NDIS configuration with TOE enabled

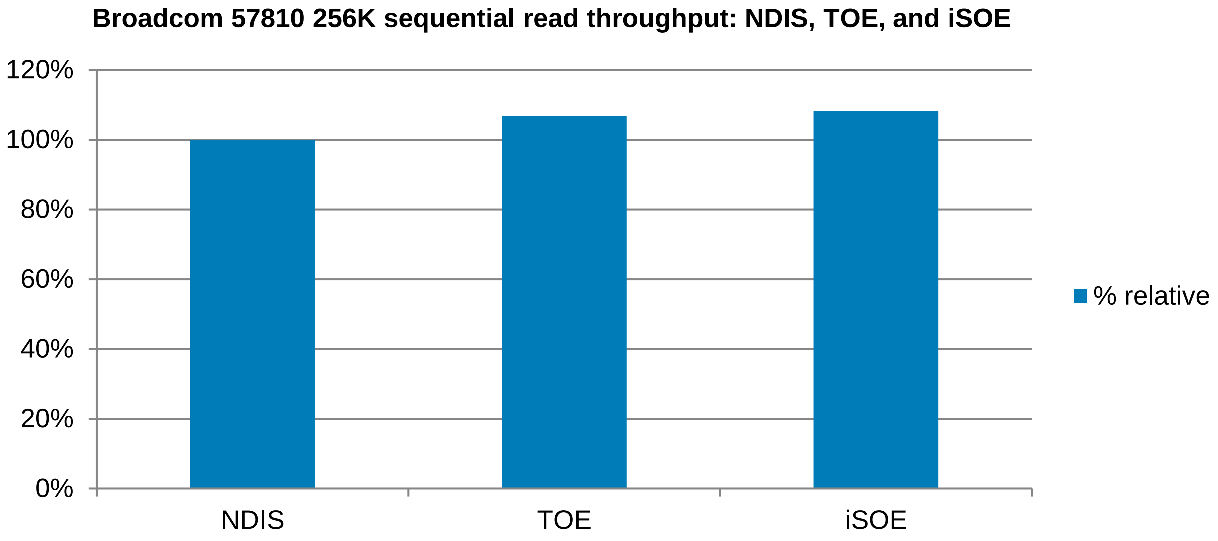


Figure 9 Broadcom 57810 iSOE mode, 256K sequential read workload, throughput to the baseline NDIS configuration and to the baseline NDIS configuration with TOE enabled

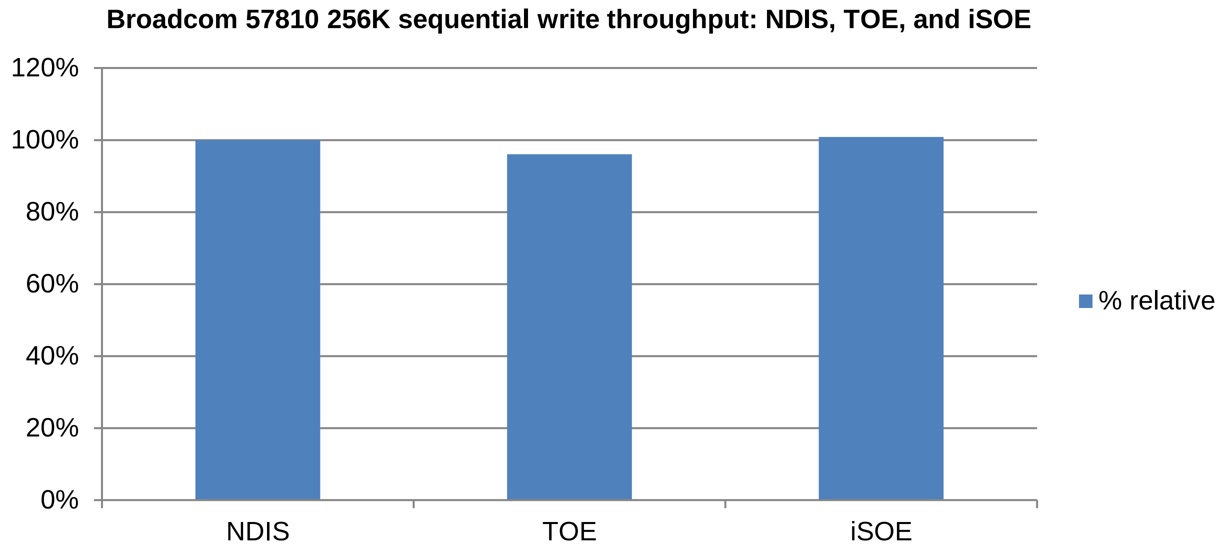


Figure 10 Broadcom 57810 iSOE mode, 256K sequential write workload, throughput relative to the baseline NDIS configuration and to the baseline NDIS configuration with TOE enabled

4.5 Intel X520 performance results

This section shows the performance results with the different configuration options during the three tested workloads using the Intel X520 network adapter.

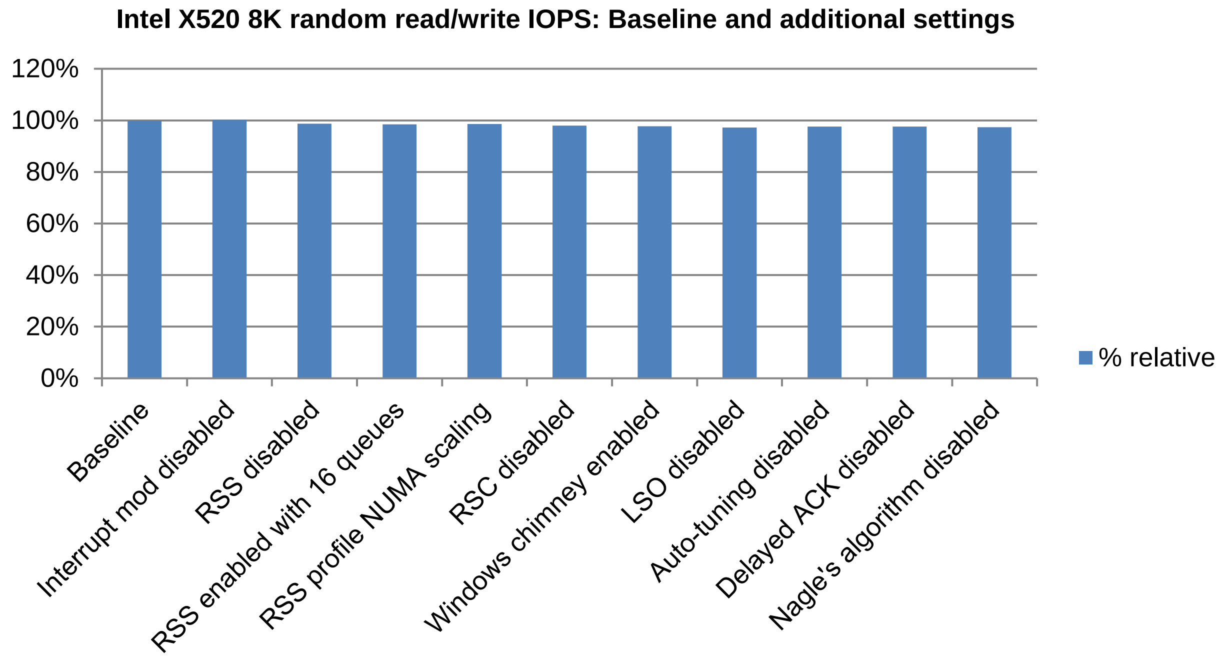


Figure 11 Intel X520, 8K random 67% read workload, IOPS relative to the baseline configuration

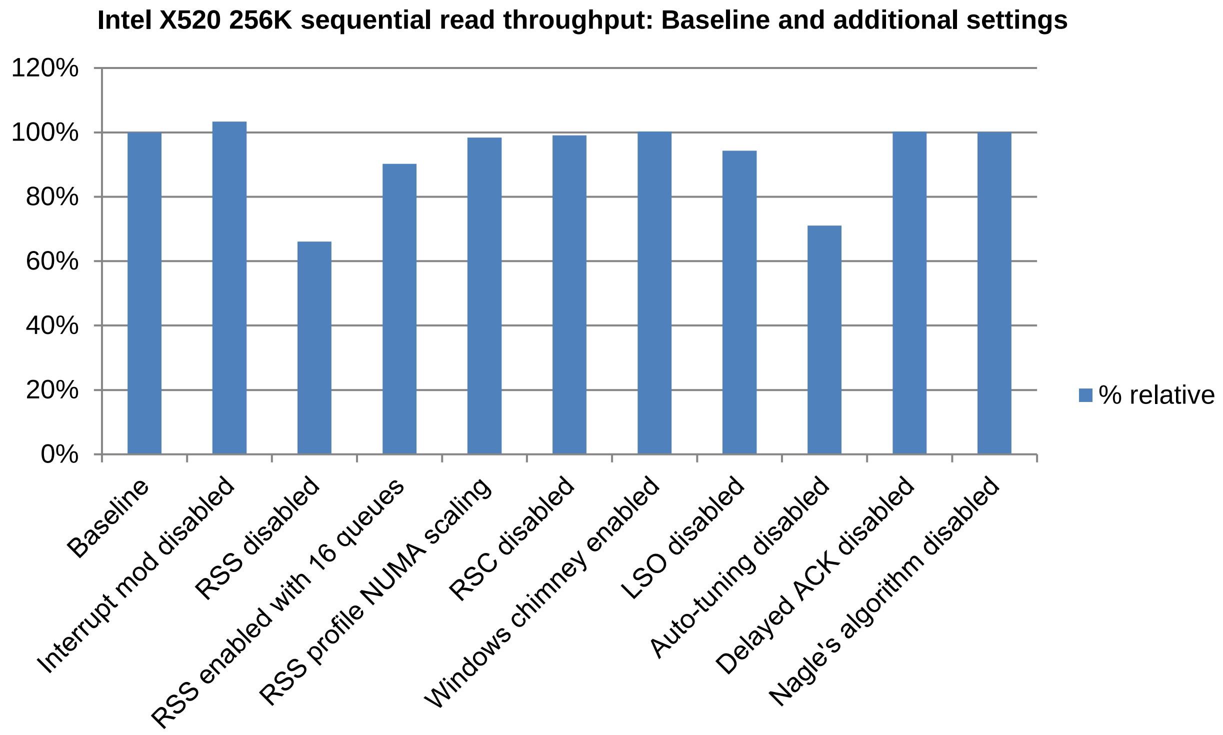


Figure 12 Intel X520, 256K sequential read workload, throughput relative to the baseline configuration

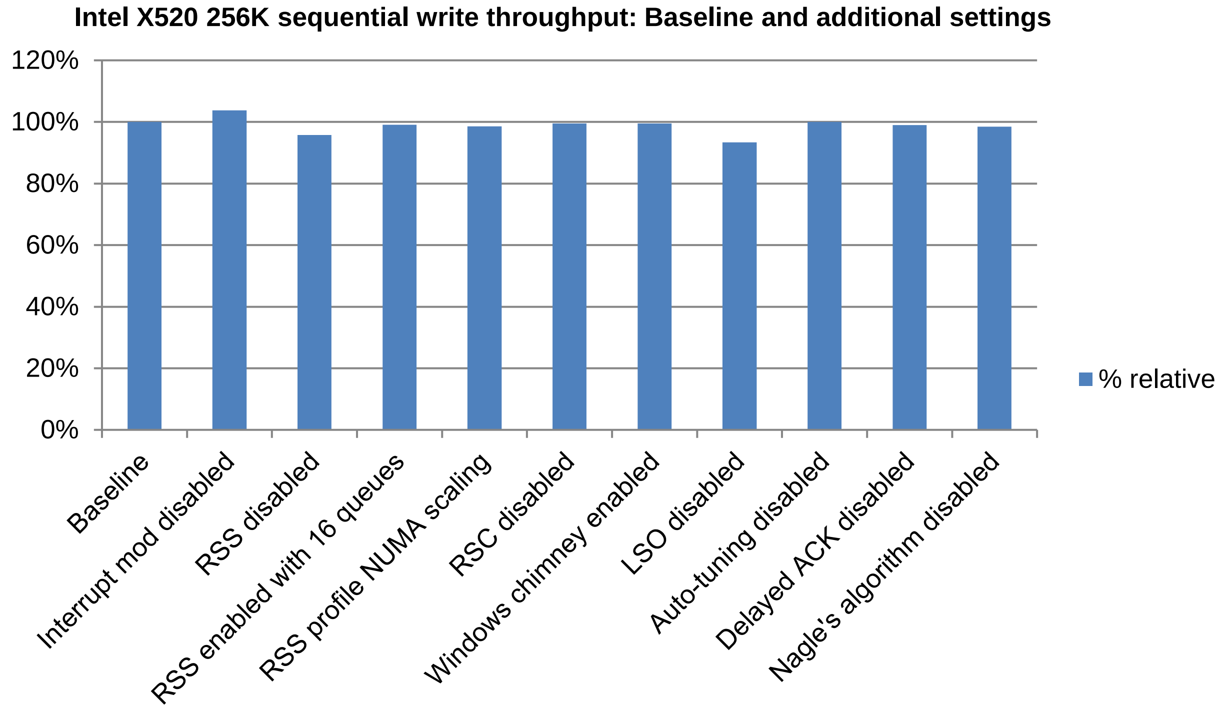


Figure 13 Intel X520, 256K sequential write workload, throughput relative to the baseline configuration

5 Best practice recommendations

In this section, recommended configurations are given based on the performance results and analysis detailed in section 4. Non-default settings are recommended only when a compelling difference in performance for one or more workloads is observed, or when a setting is known to provide benefit during network congestion or heavy processor utilization. Only the non-default settings are listed.

For a complete list of tested options and default values as well as instructions on making configuration changes to the storage host, see appendix A.

5.1 Broadcom BCM57810 NDIS mode recommended configuration

Based on the performance results and analysis for each workload, the following NIC and OS configuration changes are recommended.

Table 6 Broadcom BCM57810 NDIS mode recommended adapter configuration

Setting	Default value	Recommended value
Jumbo packet	1514	9614
Receive buffers	0=Auto	3000
TCP Connection Offload	Disabled	Enabled
Transmit Buffers	0=Auto	5000

Table 7 Broadcom BCM57810 NDIS mode recommended Windows Server 2012 TCP configuration

Setting	Default value	Recommended value
Chimney Offload State	Disabled	Enabled

5.2 Broadcom BCM57810 iSOE mode recommended configuration

Based on the performance results and analysis for each workload, the following NIC configuration changes are recommended.

Since all TCP functions for the SAN interfaces are offloaded in iSOE mode, the Windows Server 2012 TCP configuration has no effect and no changes are required.

Table 8 Broadcom BCM57810 iSOE mode recommended adapter configuration

Setting	Default value	Recommended value
Flow control	Auto	Rx and Tx enabled
MTU	1500	9600

5.3 Intel X520 recommended configuration

Based on the performance results and analysis for each workload, the following NIC configuration changes are recommended.

Table 9 Intel X520 recommended adapter configuration

Setting	Default value	Recommended value
Jumbo packet	Disabled	Enabled
Receive Buffers	512	4096
Transmit Buffers	512	16384

6 Conclusion

For Dell PS Series SANs, it is recommended that Jumbo frames and flow control be enabled for both the Broadcom 57810 and the Intel X520 adapters. If not using the Broadcom iSOE, receive and transmit buffers should also be maximized.

When using the Broadcom BCM57810, TOE should be considered for its ability to decrease processor utilization and also to lower retransmission rates in congested networks.

The Broadcom BCM57810 iSOE is another compelling option. Not only did it decrease retransmission rates in a congested network and lower processor utilization by an even greater amount than TOE, it also exhibited some performance benefits, particularly during the 8K random read/write workload.

One thing to consider when using iSOE is the difference in administration procedures. Once iSOE is enabled, the network adapter disappears from the native Windows Server network management and monitoring tools and appears as a storage controller in Windows Server device manager. It must be configured and monitored in the Broadcom Advanced Control Suite.

A Network adapter and TCP stack configuration details

This section provides more detail about the configuration options and default settings of the network adapter properties and the Windows Server 2012 TCP stack.

A.1 Broadcom BCM57810 NDIS mode adapter options

Table 10 lists the tested adapter options for the Broadcom BCM57810 NetXtreme II 10 GigE NIC in NDIS mode along with the default value.

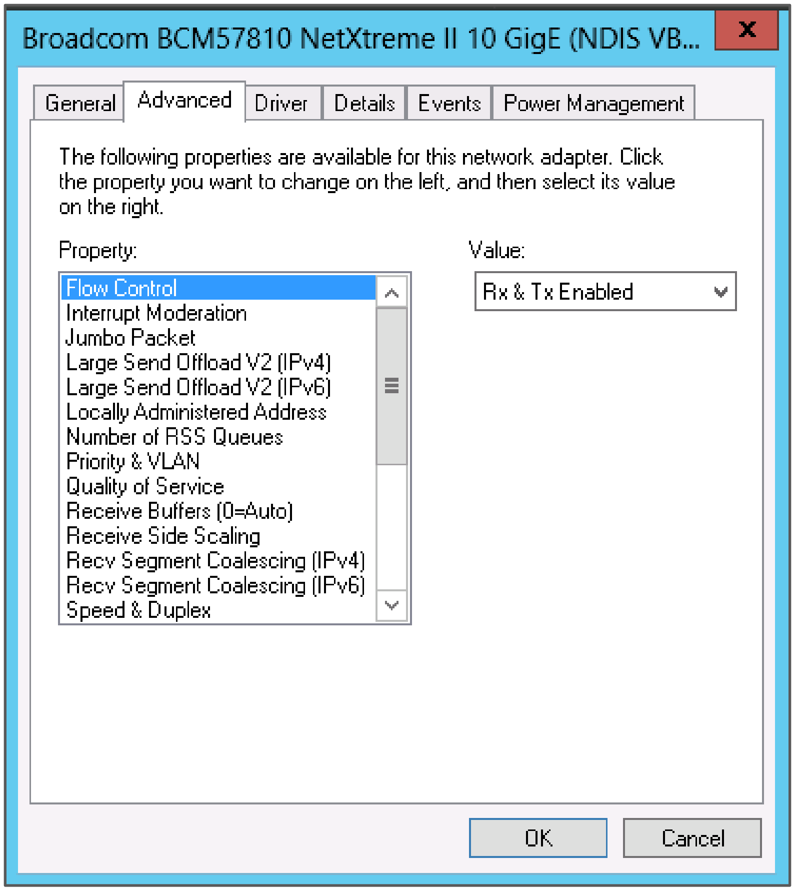
Table 10 Broadcom BCM57810 NDIS mode adapter options

Setting	Default value
Flow control	Rx and Tx Enabled
Interrupt moderation	Enabled
Jumbo packet	1514
Large Send Offload V2	Enabled
Number of RSS queues	4
Receive buffers	0=Auto
Receive Side Scaling	Enabled
Recv Segment Coalescing	Enabled
TCP Connection Offload*	Disabled
TCP/UDP Checksum Offload	Rx and Tx Enabled
Transmit buffers	0=Auto

*To be enabled, this option must be enabled in the NIC adapter settings and Windows Server TCP Chimney offload must be enabled.

A.2 Configuring Broadcom BCM57810 adapter properties in NDIS mode

Adapter properties for the Broadcom BCM57810 NDIS adapter can be set in the traditional Windows Server adapter properties dialog box in the **Advanced** tab or with the Broadcom Advanced Control Suite (BACS), a separate application from the native Windows management tools.



Broadcom BCM57810 NetXtreme II 10 GigE (NDIS VB...

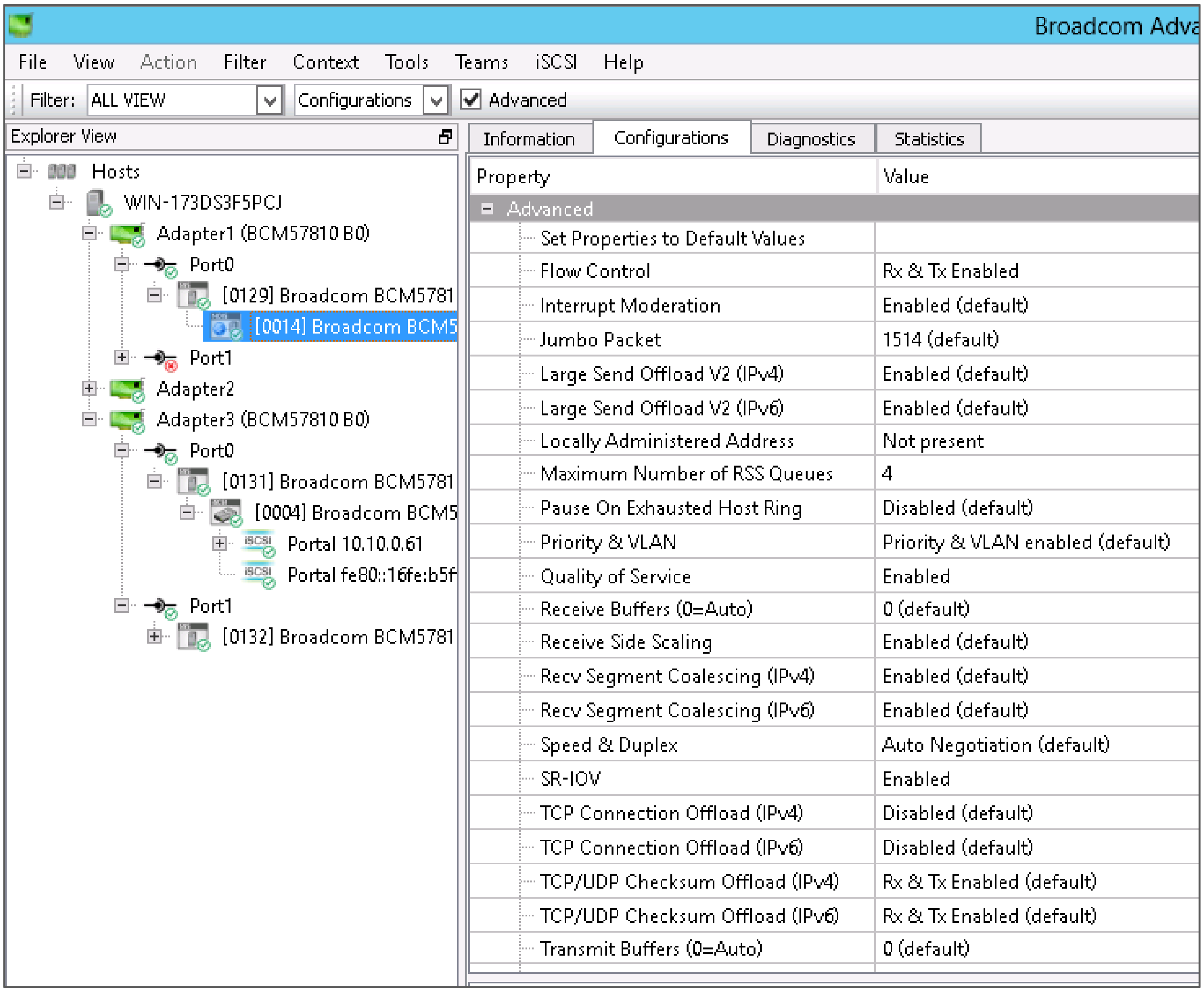
The following properties are available for this network adapter. Click the property you want to change on the left, and then select its value on the right.

Property:

- Flow Control
- Interrupt Moderation
- Jumbo Packet
- Large Send Offload V2 (IPv4)
- Large Send Offload V2 (IPv6)
- Locally Administered Address
- Number of RSS Queues
- Priority & VLAN
- Quality of Service
- Receive Buffers (0=Auto)
- Receive Side Scaling
- Recv Segment Coalescing (IPv4)
- Recv Segment Coalescing (IPv6)
- Speed & Duplex

Value: Rx & Tx Enabled

Figure 14 Windows 2012 adapter properties window for the Broadcom NDIS adapter



Filter: ALL VIEW | Configurations | ☑ Advanced

Property	Value
⊟ Advanced	
Set Properties to Default Values	
Flow Control	Rx & Tx Enabled
Interrupt Moderation	Enabled (default)
Jumbo Packet	1514 (default)
Large Send Offload V2 (IPv4)	Enabled (default)
Large Send Offload V2 (IPv6)	Enabled (default)
Locally Administered Address	Not present
Maximum Number of RSS Queues	4
Pause On Exhausted Host Ring	Disabled (default)
Priority & VLAN	Priority & VLAN enabled (default)
Quality of Service	Enabled
Receive Buffers (0=Auto)	0 (default)
Receive Side Scaling	Enabled (default)
Recv Segment Coalescing (IPv4)	Enabled (default)
Recv Segment Coalescing (IPv6)	Enabled (default)
Speed & Duplex	Auto Negotiation (default)
SR-IOV	Enabled
TCP Connection Offload (IPv4)	Disabled (default)
TCP Connection Offload (IPv6)	Disabled (default)
TCP/UDP Checksum Offload (IPv4)	Rx & Tx Enabled (default)
TCP/UDP Checksum Offload (IPv6)	Rx & Tx Enabled (default)
Transmit Buffers (0=Auto)	0 (default)

Figure 15 Configuring properties for the Broadcom NDIS adapter using the Broadcom Advanced Control Suite

A.3 Broadcom BCM57810 iSOE mode adapter options

Table 11 lists the tested adapter options for the Broadcom BCM57810 NetXtreme II 10 GigE NIC in iSOE mode along with the default value.

Table 11 Broadcom BCM57810 iSOE mode adapter options

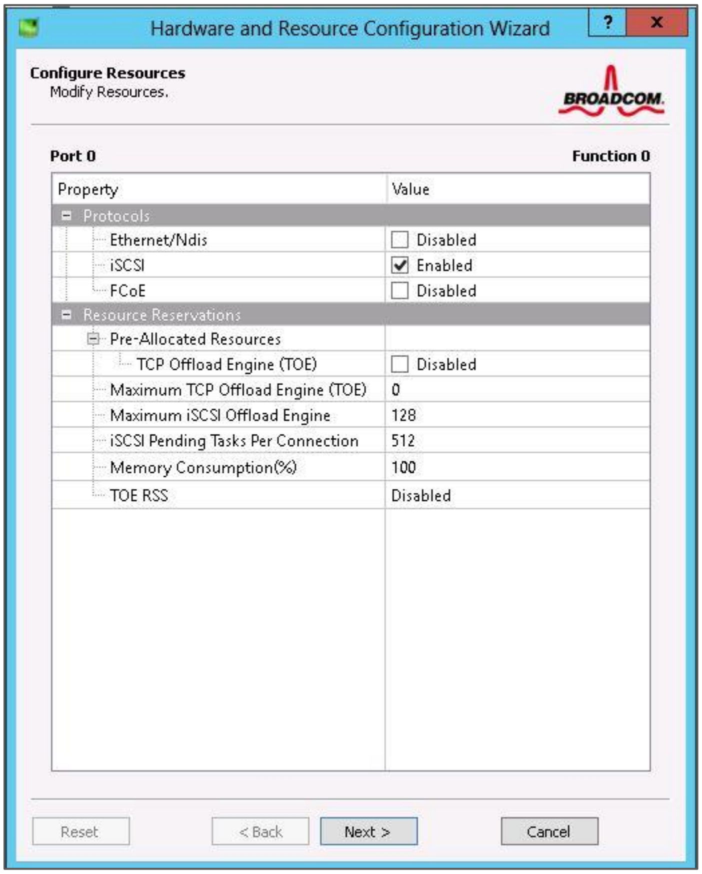
Setting	Default value
Flow control	Auto
MTU*	1500

* Equivalent to the Jumbo packet option listed in Table 10.

Once iSOE is enabled, only the Flow control and MTU options are available for configuration.

A.4 Configuring Broadcom BCM57810 adapter properties in iSOE mode

Adapter properties for the Broadcom BCM57810 iSOE adapter must be set in BACS. After enabling iSOE mode with BACS, Jumbo frames and flow control settings can be established.

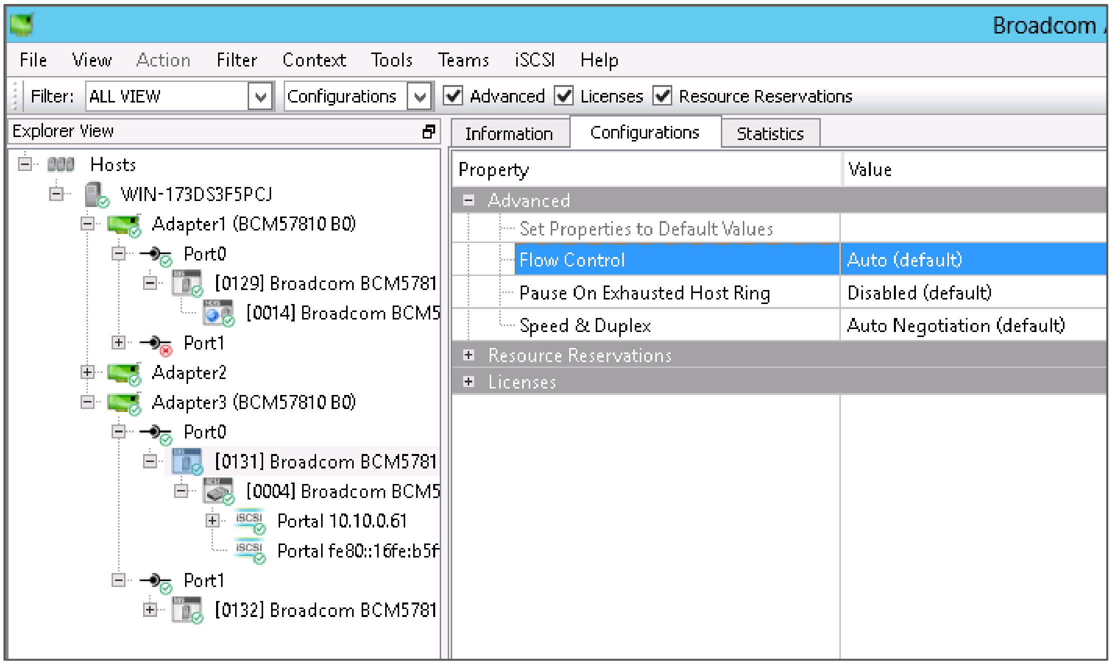


Configure Resources
Modify Resources.

Port 0 **Function 0**

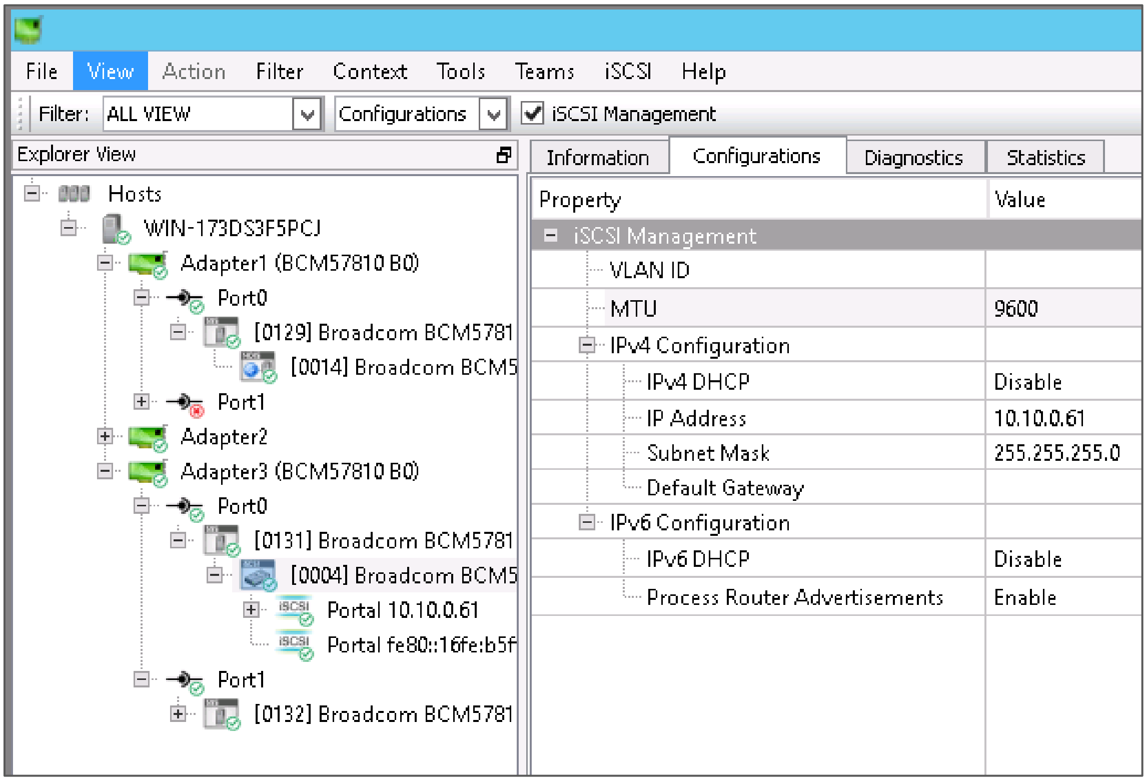
Property	Value
⊟ Protocols	
Ethernet/Ndis	□ Disabled
iSCSI	☑ Enabled
FCoE	□ Disabled
⊟ Resource Reservations	
⊟ Pre-Allocated Resources	
TCP Offload Engine (TOE)	□ Disabled
Maximum TCP Offload Engine (TOE)	0
Maximum iSCSI Offload Engine	128
iSCSI Pending Tasks Per Connection	512
Memory Consumption(%)	100
TOE RSS	Disabled

Figure 16 Enabling iSOE mode in the Broadcom Advanced Control Suite



Property	Value
⊟ Advanced	
Set Properties to Default Values	
Flow Control	Auto (default)
Pause On Exhausted Host Ring	Disabled (default)
Speed & Duplex	Auto Negotiation (default)
⊞ Resource Reservations	
⊞ Licenses	

Figure 17 Configuring Flow Control in the Broadcom Advanced Control Suite



Property	Value
⊟ iSCSI Management	
VLAN ID	
MTU	9600
⊟ IPv4 Configuration	
IPv4 DHCP	Disable
IP Address	10.10.0.61
Subnet Mask	255.255.255.0
Default Gateway	
⊟ IPv6 Configuration	
IPv6 DHCP	Disable
Process Router Advertisements	Enable

Figure 18 Configuring Jumbo frames (MTU) in the Broadcom Advanced Control Suite

A.5 Intel X520 adapter options

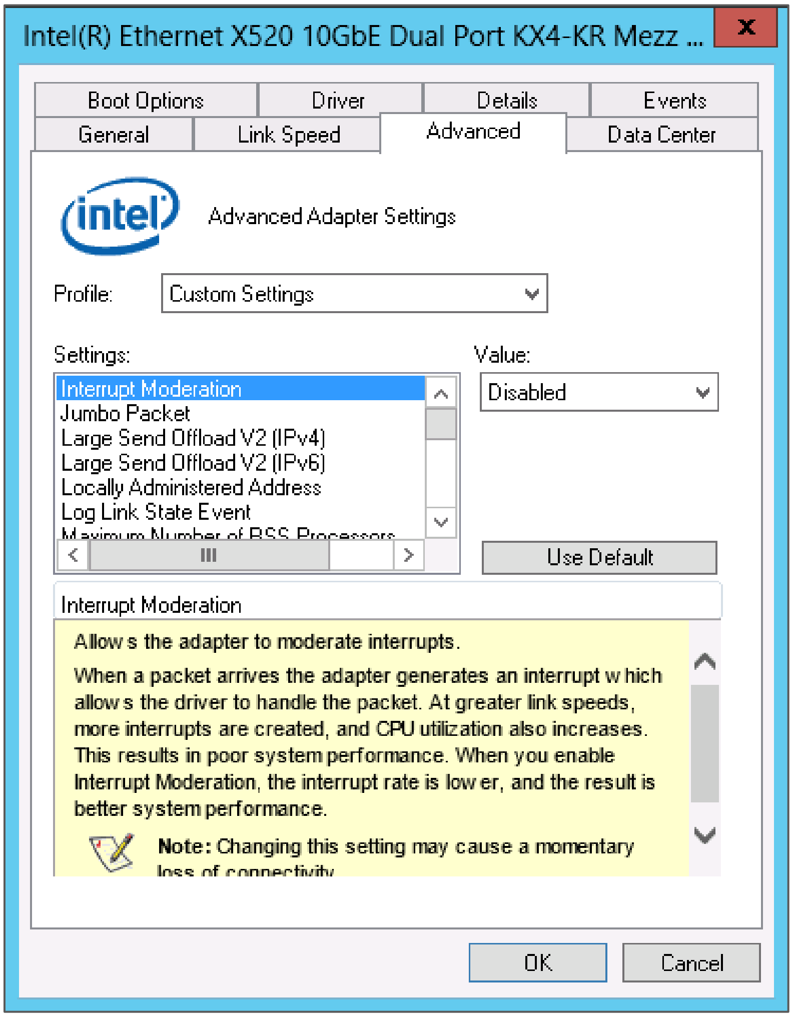
Table 12 lists the tested adapter options for the Intel X520 10 GigE NIC along with the default value.

Table 12 Intel X520 adapter options

Setting	Default value
Interrupt Moderation	Enabled
Jumbo packet	Disabled
Large Send Offload V2	Enabled
Maximum Number of RSS Queues	8
Flow Control	Rx and Tx Enabled
Receive Buffers	512
Transmit Buffers	512
Receive Side Scaling	Enabled
Recv Segment Coalescing	Enabled
RSS load balancing profile	Closest Processor
IPv4 Checksum Offload	Enabled
TCP Checksum Offload	Enabled

A.6 Configuring Intel X520 adapter properties

Adapter properties for the Intel X520 NDIS adapter can be set in the traditional Windows Server adapter properties window in the **Advanced** tab.



Advanced Adapter Settings

Profile: Custom Settings

Settings:

- Interrupt Moderation
- Jumbo Packet
- Large Send Offload V2 (IPv4)
- Large Send Offload V2 (IPv6)
- Locally Administered Address
- Log Link State Event
- Maximum Number of RSS Processors

Value: Disabled

Interrupt Moderation

Allow s the adapter to moderate interrupts.

When a packet arrives the adapter generates an interrupt w hich allow s the driver to handle the packet. At greater link speeds, more interrupts are created, and CPU utilization also increases. This results in poor system performance. When you enable Interrupt Moderation, the interrupt rate is low er, and the result is better system performance.

Note: Changing this setting may cause a momentary loss of connectivity

Figure 19 Windows 2012 adapter properties window for the Intel NDIS adapter

A.7 Windows Server 2012 TCP stack options

Table 13 lists the tested TCP stack options for Windows Server 2012 along with the default value.

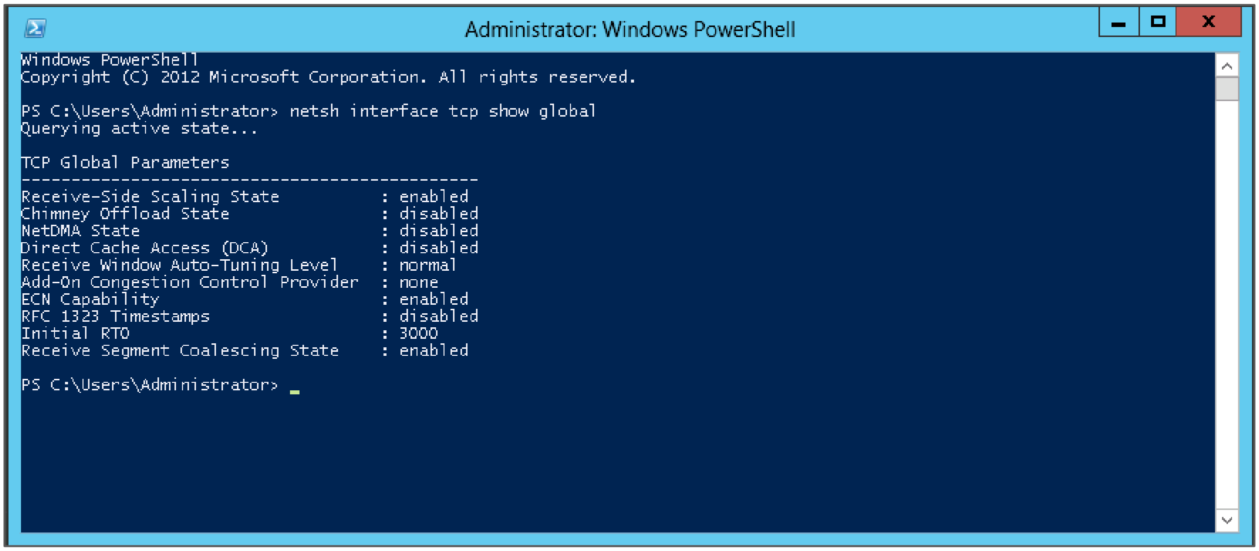
Table 13 Windows Server 2012 TCP stack options

Setting	Default value
Receive-Side Scaling State*	Enabled
Chimney Offload State*	Disabled
Receive Window Auto-Tuning Level	Normal
Receive Segment Coalescing State*	Enabled
Delayed ACK algorithm	Enabled
Nagle's algorithm	Enabled

*There are network adapter options which correspond to these OS options. During test cases, the option’s test case value was achieved by changing the option setting at both the network adapter and the OS.

A.8 Configuring the Windows Server 2012 TCP stack

Windows Server 2012 TCP stack options can be configuring using the `netsh` command in PowerShell.



```
Windows PowerShell
Copyright (C) 2012 Microsoft Corporation. All rights reserved.

PS C:\Users\Administrator> netsh interface tcp show global
Querying active state...

TCP Global Parameters
----------------------------------------------
Receive-Side Scaling State          : enabled
Chimney Offload State               : disabled
NetDMA State                        : disabled
Direct Cache Access (DCA)           : disabled
Receive Window Auto-Tuning Level    : normal
Add-On Congestion Control Provider  : none
ECN Capability                      : enabled
RFC 1323 Timestamps                 : disabled
Initial RTO                         : 3000
Receive Segment Coalescing State    : enabled

PS C:\Users\Administrator> _
```

Figure 20 Using the `netsh` command in Windows Server 2012 PowerShell

Note: For Windows Server 2012 R2 see appendix B for instructions.

To disable delayed ACK and Nagle's algorithm, create the following entries for each SAN interface subkey in the Windows Server 2012 registry:

Subkey location

```
HKEY_LOCAL_MACHINE \ SYSTEM \ CurrentControlSet \ Services \ Tcpip \ Parameters
\ Interfaces \ <SAN interface GUID>
```

Entries

```
TcpAckFrequency
TcpNoDelay
```

Value type

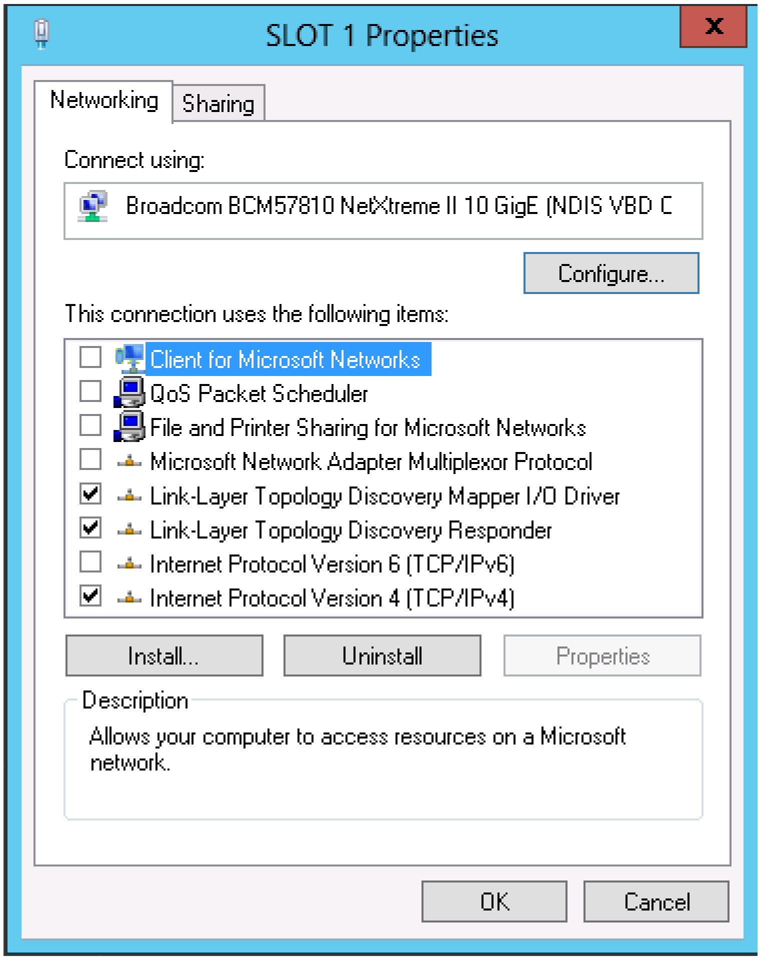
```
REG_DWORD, number
```

Value to disable

```
1
```

A.9 Disabling unused network adapter protocols

Unused protocols can be disabled in the Windows Server 2012 network adapter properties menu.



Connect using:

Broadcom BCM57810 NetXtreme II 10 GigE (NDIS VBD C

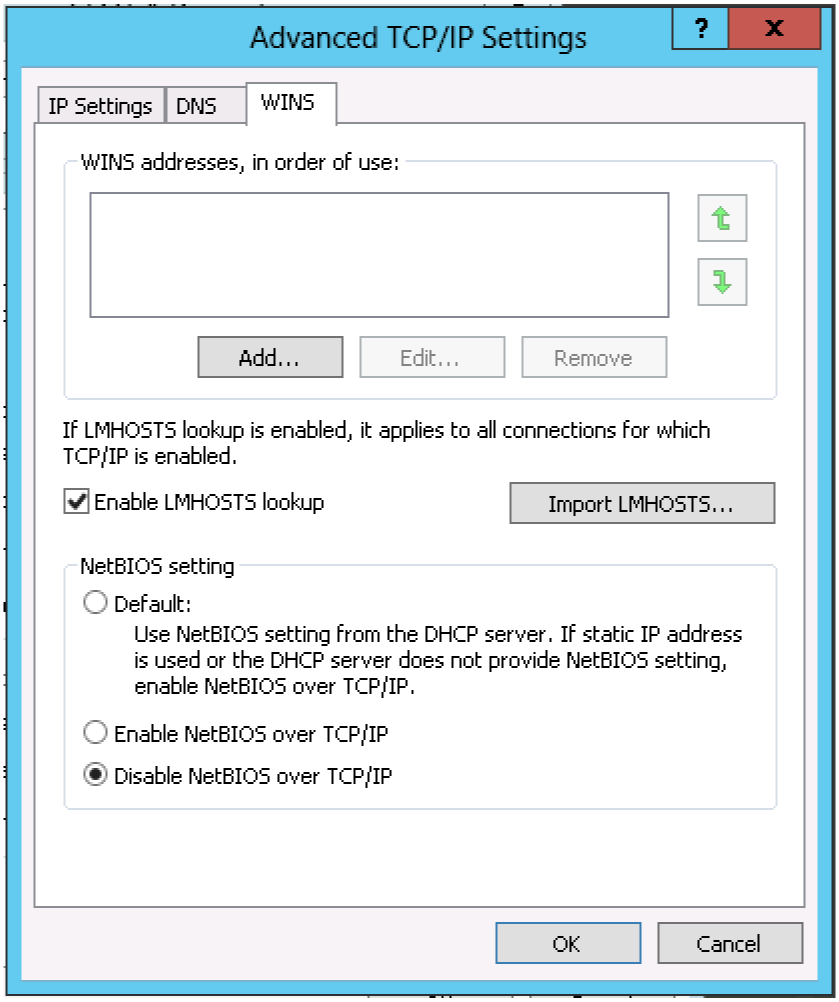
This connection uses the following items:

- ☐ Client for Microsoft Networks
- ☐ QoS Packet Scheduler
- ☐ File and Printer Sharing for Microsoft Networks
- ☐ Microsoft Network Adapter Multiplexor Protocol
- ☑ Link-Layer Topology Discovery Mapper I/O Driver
- ☑ Link-Layer Topology Discovery Responder
- ☐ Internet Protocol Version 6 (TCP/IPv6)
- ☑ Internet Protocol Version 4 (TCP/IPv4)

Description

Allows your computer to access resources on a Microsoft network.

Figure 21 Disabling unused protocols in the Windows Server 2012 network adapter properties menu



WINS addresses, in order of use:

If LMHOSTS lookup is enabled, it applies to all connections for which TCP/IP is enabled.

☑ Enable LMHOSTS lookup

NetBIOS setting

○ Default:

Use NetBIOS setting from the DHCP server. If static IP address is used or the DHCP server does not provide NetBIOS setting, enable NetBIOS over TCP/IP.

○ Enable NetBIOS over TCP/IP

● Disable NetBIOS over TCP/IP

Figure 22 Disabling NetBIOS in the Windows Server 2012 Advanced TCP/IP Settings for the network adapter

B Configuring Windows Server 2012 R2 and Windows Server 2012 R2 Hyper-V Delayed Acknowledgements

Microsoft implements network transport layer profiles and filters that allow you to customize and change TCP settings within Windows Server 2012 R2 and Windows Server 2012 R2 Hyper-V. All modifications of the profiles and filters are done using cmdlets specific for TCP/IP in Windows PowerShell. See the Microsoft TechNet article, Net TCP/IP Cmdlets in Windows PowerShell.

While a Microsoft Windows Server 2012 R2 profile is listed as available (Automatic, Datacenter, Internet, DatacenterCustom, InternetCustom, and Compat), only DatacenterCustom and InternetCustom can be customized. The profiles are described as follows:

- Internet: Optimized for networks with higher latency and lower throughput
- Datacenter: Optimized for networks with lower latency and higher throughput
- Compat: Optimized for compatibility with legacy network equipment
- Custom: Custom settings
- Automatic: The computer uses latency to select either Internet or Datacenter

These are the two key command applets used to check and modify the DatacenterCustom TCP profile parameters: Get-NetTCPSetting, and Set-NetTCPSetting.

These are the two key command applets used to check, create, and associate the network transport filters with the DatacenterCustom TCP profile: Get-NetTransportFilter, and New-NetTransportFilter.

Note: Commands in the following examples originate at the command prompt, `PS C:>\`

B.1 Customize the DatacenterCustom profile to set the DelayedAckFrequency value to 1

To configure the DatacenterCustom profile parameter DelayedAckFrequency, use this command:

```
PS C:\> Set-NetTCPSetting -SettingName DatacenterCustom -DelayedAckFrequency 1
```

Then, reconfirm the setting using the Get-NetTCPSetting command:

```
PS C:\> Get-NetTCPSetting -SettingName DatacenterCustom


SettingName                     : DatacenterCustom
MinRto(ms)                      : 20
InitialCongestionWindow(MSS)    : 4
CongestionProvider              : DCTCP
CwndRestart                     : True
DelayedAckTimeout(ms)           : 10
DelayedAckFrequency             : 1
MemoryPressureProtection        : Enabled
AutoTuningLevelLocal            : Normal
AutoTuningLevelGroupPolicy      : NotConfigured
AutoTuningLevelEffective        : Local
EcnCapability                   : Enabled
Timestamps                      : Disabled
InitialRto(ms)                  : 3000
ScalingHeuristics               : Disabled
DynamicPortRangeStartPort       : 49152
DynamicPortRangeNumberOfPorts   : 16384
AutomaticUseCustom              : Enabled
NonSackRttResiliency            : Disabled
ForceWS                         : Disabled
MaxSynRetransmissions           : 2
```

B.2 Define network transport filters to match the iSCSI traffic and associate them with the DatacenterCustom profile

Perform the following steps:

1. Check for a current filter and associated parameters in use. You may find one or more filters defined in your installation.

   ```
   PS C:\> Get-NetTransportFilter


   SettingName       : Automatic
   Protocol          : TCP
   LocalPortStart    : 0
   LocalPortEnd      : 65535
   RemotePortStart   : 0
   RemotePortEnd     : 65535
   DestinationPrefix : *
   ```

2. If the default filter is associated to the **Automatic** profile, define a new NetTransportFilter with a DatacenterCustom name. The filter will be associated with the DatacenterCustom profile which will be updated later with the desired TCP parameters.

```
PS C:\> New-NetTransportFilter


cmdlet New-NetTransportFilter at command pipeline position 1
Supply values for the following parameters:
SettingName: DatacenterCustom


SettingName       : DatacenterCustom
Protocol          : TCP
LocalPortStart    : 0
LocalPortEnd      : 65535
RemotePortStart   : 0
RemotePortEnd     : 65535
DestinationPrefix : *
```

3. Define a transport filter to target the iSCSI destination port and associate it with the DatacenterCustom profile.

```
PS C:\> New-NetTransportFilter -SettingName DatacenterCustom -
LocalPortStart 0 -LocalPortEnd 65535 -RemotePortStart 3260 -RemotePortEnd
3260

SettingName       : DatacenterCustom
Protocol          : TCP
LocalPortStart    : 0
LocalPortEnd      : 65535
RemotePortStart   : 3260
RemotePortEnd     : 3260
DestinationPrefix : *
```

4. Optionally, define a transport filter with an IP range corresponding to your iSCSI range:

Note: The **RemotePort** and **DestinationPrefix** settings are mutually exclusive, however you can define / associate multiple filters with your custom DatacenterCustom profile. In this example, we are configuring iSCSI traffic on the network 172.28.45.0 with the CIDR mask notation for subnet 255.255.255.0

```
PS C:\> New-NetTransportFilter -SettingName DatacenterCustom -
DestinationPrefix 172.28.45.0/24


SettingName       : DatacenterCustom
Protocol          : TCP
LocalPortStart    : 0
LocalPortEnd      : 65535
RemotePortStart   : 0
RemotePortEnd     : 65535
DestinationPrefix : 172.28.45.0/24
```

5. You now have two more NetTransportFilter based on the commands previously executed. Use the Get-NetTransportFilter command to verify the parameters in the DatacenterCustom profile.

```
PS C:\> Get-NetTransportFilter


[…]

SettingName       : DatacenterCustom
Protocol          : TCP
LocalPortStart    : 0
LocalPortEnd      : 65535
RemotePortStart   : 0
RemotePortEnd     : 65535
DestinationPrefix : 172.28.45.0/24

SettingName       : DatacenterCustom
Protocol          : TCP
LocalPortStart    : 0
LocalPortEnd      : 65535
RemotePortStart   : 3260
RemotePortEnd     : 3260
DestinationPrefix : *
```

These are sufficient to tie in the TCP settings from DatacenterCustom with the TCP filters set previously.

C I/O parameters

Vdbench SAN workloads were executed using the following parameters in the parameter file.

Common parameters:

```
hd=default
hd=one,system=localhost
```

iSCSI volumes (random IO):

```
sd=sd1,host=*,lun=\\.\PhysicalDrive1,size=102400m,threads=5
sd=sd2,host=*,lun=\\.\PhysicalDrive2,size=102400m,threads=5
sd=sd3,host=*,lun=\\.\PhysicalDrive3,size=102400m,threads=5
sd=sd4,host=*,lun=\\.\PhysicalDrive4,size=102400m,threads=5
sd=sd5,host=*,lun=\\.\PhysicalDrive5,size=102400m,threads=5
sd=sd6,host=*,lun=\\.\PhysicalDrive6,size=102400m,threads=5
sd=sd7,host=*,lun=\\.\PhysicalDrive7,size=102400m,threads=5
sd=sd8,host=*,lun=\\.\PhysicalDrive8,size=102400m,threads=5
```

iSCSI volumes (sequential IO on two arrays):

```
sd=sd1,host=*,lun=\\.\PhysicalDrive1,size=30m,threads=5
sd=sd2,host=*,lun=\\.\PhysicalDrive2,size=30m,threads=5
sd=sd3,host=*,lun=\\.\PhysicalDrive3,size=30m,threads=5
sd=sd4,host=*,lun=\\.\PhysicalDrive4,size=30m,threads=5
sd=sd5,host=*,lun=\\.\PhysicalDrive5,size=30m,threads=5
sd=sd6,host=*,lun=\\.\PhysicalDrive6,size=30m,threads=5
sd=sd7,host=*,lun=\\.\PhysicalDrive7,size=30m,threads=5
sd=sd8,host=*,lun=\\.\PhysicalDrive8,size=30m,threads=5
```

iSCSI volumes (sequential IO on four arrays):

```
sd=sd1,host=*,lun=\\.\PhysicalDrive1,size=45m,threads=5
sd=sd2,host=*,lun=\\.\PhysicalDrive2,size=45m,threads=5
sd=sd3,host=*,lun=\\.\PhysicalDrive3,size=45m,threads=5
sd=sd4,host=*,lun=\\.\PhysicalDrive4,size=45m,threads=5
sd=sd5,host=*,lun=\\.\PhysicalDrive5,size=45m,threads=5
sd=sd6,host=*,lun=\\.\PhysicalDrive6,size=45m,threads=5
sd=sd7,host=*,lun=\\.\PhysicalDrive7,size=45m,threads=5
sd=sd8,host=*,lun=\\.\PhysicalDrive8,size=45m,threads=5
```

8KB random 67% read workload:

```
wd=wd1,sd=(sd1-sd8),xfersize=8192,rdpct=100,skew=67
wd=wd2,sd=(sd1-sd8),xfersize=8192,rdpct=0,skew=33
```

256KB sequential read workload:

```
wd=wd1,sd=(sd1-sd8),xfersize=262144,rdpct=100,seekpct=sequential
```

256KB sequential write workload:

```
wd=wd1,sd=(sd1-sd8),xfersize=262144,rdpct=0,seekpct=sequential
```

Runtime options:

```
rd=rd1,wd=wd*,iorate=max,elapsed=1200,interval=5
```

D Technical support and resources

Dell.com/Support is focused on meeting customer needs with proven services and support.

Dell TechCenter is an online technical community where IT professionals have access to numerous resources for Dell software, hardware and services.

Storage Solutions Technical Documents on Dell TechCenter provide expertise that helps to ensure customer success on Dell Storage platforms.

D.1 Additional resources

See the referenced or recommended publications:

- *Dell PS Series Configuration Guide*
- *Dell Storage Compatibility Matrix*
- Dell PS Series Switch Configuration Guides
- Dell PS Series firmware updates and documentation (site requires a login)
- Dell EMC Networking documentation
- Performance Tuning Guidelines for Windows Server 2012 and Windows Server 2012 R2