IBM

Redpaper

Elaine Wood
Thomas Peralto

XIV Storage System: A look into Fibre Channel and iSCSI Performance

With their inherent high performance and lossless behavior, Fibre Channel storage area networks (SAN) have emerged as the standard technology for access to block-based storage. Although iSCSI over Ethernet has been an option for lower-cost storage networking, the performance limitations and packet loss that are associated with Ethernet have limited the usefulness of iSCSI in demanding storage environments. With the advent of 10 Gbps Ethernet, and the development of lossless Ethernet technology, internet Small Computer System Interface (iSCSI) is now a viable alternative to Fibre Channel for the deployment of networked storage.

Many customers have implemented iSCSI for host attachment to storage. The adoption of iSCSI by IBM® XIV® customers is highest in the VMware space. At the IBM Edge 2013 XIV Focus Group, 43% of customers indicated that they were using 1 Gigabit Ethernet (GbE) iSCSI for host attachment in their production environments, compared to a 15% adoption rate in 2012. The success of existing deployments and the availability of 10 GbE has led many customers to examine the usage of iSCSI attachment for all of their workloads, including the most demanding high IOPS transactional applications. For 10 GbE iSCSI to be viable in these environments, it must provide performance that is comparable to 8 Gb Fibre Channel.

This IBM Redpaper™ publication presents the results of head-to-head performance testing of 10 Gb iSCSI and 8 Gb Fibre Channel accessing a Gen3 XIV and running various workloads. The goal of these tests was to determine the relative performance of these two technologies, and to understand the processor impact of iSCSI over Ethernet.

Test environment

The tests were performed at the IBM Littleton, MA XIV test lab. Two Intel based host configurations were available for testing.

1. IBM 3650-M4 (host xiv02) running RedHat Enterprise Linux (REL) 5.9 with the following adapters:
   - Fibre Channel HBA: Emulex Corporation Saturn-X: LightPulse 8 Gbps
   - Ethernet NIC: Emulex Corporation OneConnect OCe11102 10GbE
2. IBM 3650-M4 (host xiv04) running Windows 2012 with the following adapters:
   - Fibre Channel HBA: Emulex Corporation Saturn-X: LightPulse 8 Gbps
   - Ethernet NIC: Emulex Corporation OneConnect OCe11102 10GbE

The XIV storage that was used was a fully configured 15 module Gen3 model 214 system with flash drive (SSD) cache and 2 TB disk drives. The XIV storage was at code level 11.3.0.a. There are twelve 10 GbE iSCSI ports on the XIV (two on each interface module).

As shown in Table 1, the XIV Storage System is modular in design and starts with a six module configuration that includes 72 disks (55 with up to 112 TB usable capacity), 114 GB with up to 288 GB of cache, 2.4 TB of SSDs, and eight 8 Gbps Fibre Channel interfaces and four 10 GbE iSCSI interfaces. A fully configured 15 module XIV includes 180 disks (161 with up to 320 TB usable capacity), 360 GB with up to 720 GB of cache, 6 TB of SSDs, and twenty-four 8 Gbps Fibre Channel interfaces and twelve 10 GbE iSCSI interfaces.

The XIV system can support 2,000 concurrent hosts that are attached through a Fibre Channel worldwide port name (WWPN) or iSCSI Qualified Name (IQN).

An upgrade option exists that can double the amount of SSD cache available, with up to 12 TB of capacity.

Table 1 XIV configurations - 6 - 15 modules

Number of data modules	**6**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Number of disks	72	108	120	132	144	156	168	180
Usable capacity in terabytes (2 TB drives)	55	88	102	111	125	134	149	161
Usable capacity in terabytes (3 TB drives)	84	132	154	168	190	203	225	243
Usable capacity in terabytes (4 TB drives)	112	176	204	223	252	270	292	320
Fibre Channel (FC) ports (8 GB)	8	16	16	20	20	24	24	24
iSCSI ports (10 GB)	4	8	8	10	10	12	12	12
Memory (GB) at 24 GB	144	216	240	264	288	312	336	360

Number of data modules	6	9	10	11	12	13	14	15
Memory (GB) at 48 GB	288	432	480	528	576	624	672	720
SSD (TB) at 400 GB	2.4	3.6	4	4.4	4.8	5.2	5.6	6
SSD (TB) at 800 GB	4.8	7.2	480	8.8	9.6	10.4	11.2	12

Per XIV preferred practices for host attachment (see the IBM Redbooks® publication *IBM XIV Storage System: Host Attachment and Interoperability*, SG24-7904), each host is attached to all six interface modules of the XIV for a total of six paths. In the lab environment, only a single fabric was used. However, in a production environment, dual fabrics should be configured to eliminate the Ethernet or Fibre Channel switch as a single point of failure.

Figure 1 and Figure 2 on page 4 show the Fibre Channel and iSCSI test environments. Figure 1 shows the Littleton Lab iSCSI configuration.

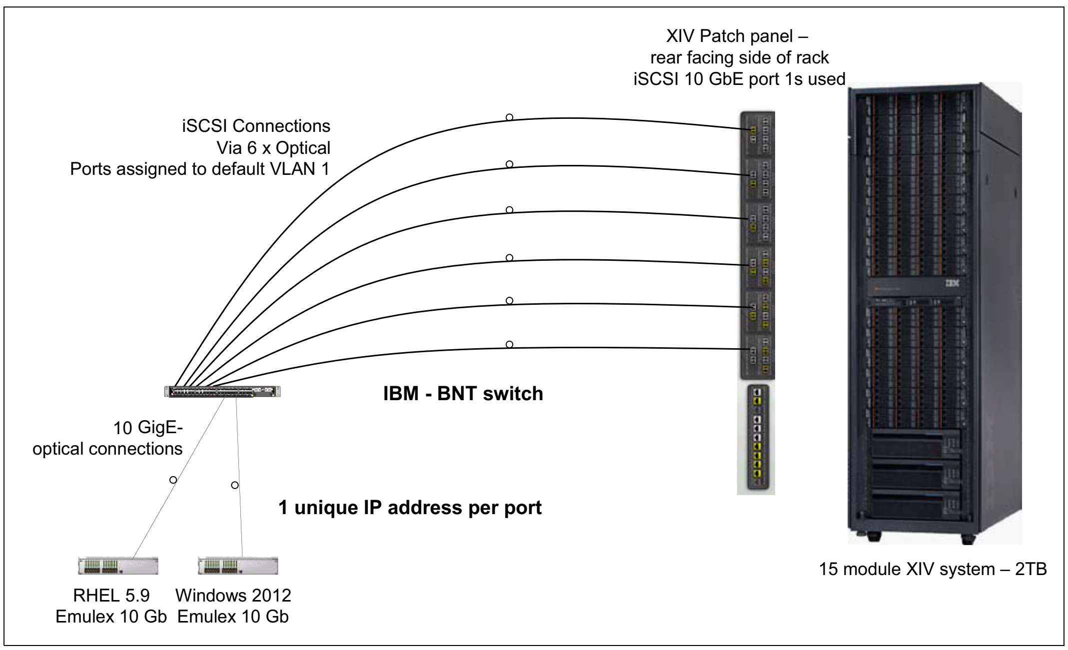

Figure 1 Littleton Lab iSCSI configuration

Figure 2 shows the Littleton Lab Fibre Channel configuration.

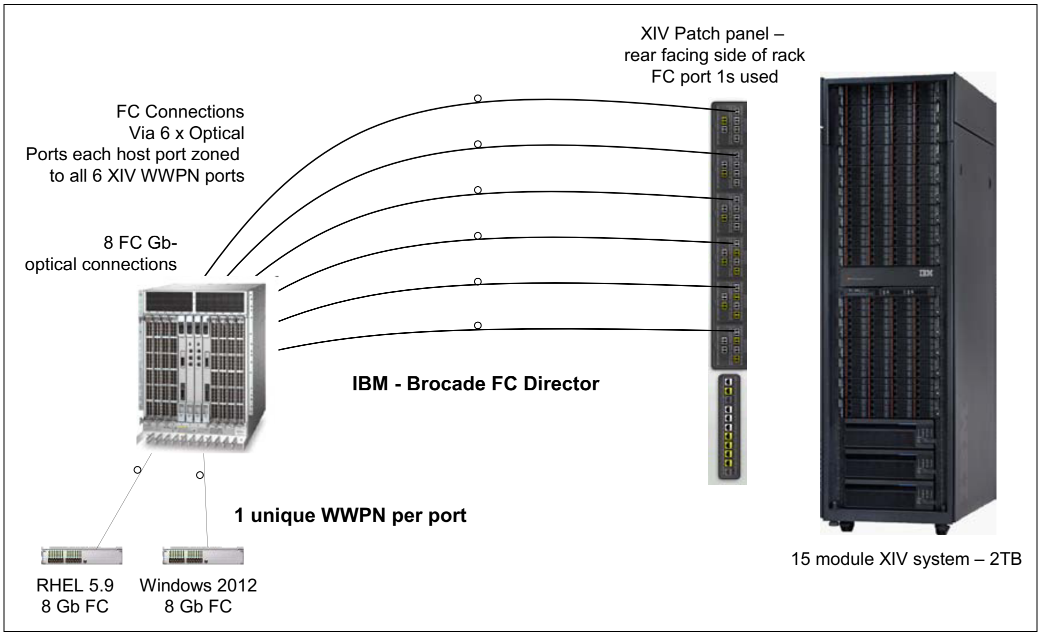

Figure 2 Littleton Lab Fibre Channel configuration

Test methodology

Performance tests were run on both Windows and Linux operating systems. Load generators were used to generate I/O workloads: IOmeter for Windows and IOrate for Linux.

Various workloads were generated representing high-IOPS and high-bandwidth workloads. The load generation tools were set up to generate the maximum IOPS possible for the workload. The workloads were purely sequential because the intent was to drive as much data as possible through the network and compare results with the two technologies. All patterns were run with 50% reads and 50% writes.

Table 2 shows the workloads that were used.

Table 2 Test workloads

Pattern number	Read Block Size (KB)	Write Block Size (KB)
1	4	4
2	4	8
3	4	16
4	8	16
5	16	32
6	32	64
7	32	128
8	64	128
9	256	512

Pattern number	Read Block Size (KB)	Write Block Size (KB)
10	512	1024
11	1024	1024

The same workloads were run with the host connected through FCP and then through iSCSI. Modifications to the connectivity were made through the XIV GUI by modifying host port attachments. Latency, throughput, and IOPS were monitored through XIV TOP and processor usage was monitored by using the **top** command on the Linux host.

On Linux, iSCSI daemons were created when doing iSCSI I/O, and the usage of those daemons was recorded. All tests were run with 300 GB, 1 TB, and 2 TB LUNS. Because there was no difference in results for different LUN sizes, we present the 2 TB LUN results for comparison.

TCP and iSCSI offload

When using iSCSI storage networking, TCP/IP and iSCSI protocols introduce additional protocol processing. If it is not handled by the network adapter, this protocol processing impact is incurred by the processors on the host. TCP/IP Offload Engines (TOEs) are available in the marketplace, and the Emulex card that was used in the test is capable of offload functions. Some or all of the following operations might be offloaded to a processor on the network card, and different network cards support different combinations of the following offload functions:

- IP, TCP, and user datagram protocol (UDP) checksum offloads
- Large and giant send offload (LSO and GSO)
- Receive side scaling (RSS)
- TCP Segmentation Offload (TSO)

Vendors have modified Linux to support TOE, but the Linux foundation rejected requests to include a TOE implementation in the kernel. As such, TOE features are both Linux operating system and NIC dependent. The **ethtool** command displays (and allows modification of) supported offload functions in Linux. For more information about the **ethtool** command, see “Appendix 1: The ethtool command on RHEL 5.9” on page 10.

The Linux operating system has the OFFLOAD functions enabled by default; all the benchmark testing uses the default configuration. For the purposes of comparison, the OFFLOAD functions were disabled and compared later in the paper.

Microsoft Windows provides adapter-dependent offload support. In our test, the following parameters were available for the Emulex adapter:

- Large Send Offload
- Checksum Offload
- Receive Side Scaling
- Receive Side Coalescing
- TCP Checksum Offload
- UDP Checksum Offload

It is possible to purchase network cards that offload the iSCSI initiator function. These cards have a higher cost than network cards that do not provide this function. In our tests, the iSCSI initiator function was provided by operating system software and not by the network cards.

MTU settings and jumbo frames

For optimum performance, set the maximum transmission unit (MTU) to 9000. Instructions about setting the MTU size for Linux are provided in “Appendix 2: Configuring the NIC in Linux” on page 11, and instructions for Windows are provided in “Appendix 3: Windows 2012 MTU configuration” on page 12.

Test results

We present out test results from three perspectives:

- IOPS and bandwidth comparisons
- Impact of TCP/IP protocol offloads on iSCSI performance
- Processor impact of iSCSI

IOPS and bandwidth comparisons

Figure 3 shows the maximum IOPS for each of the workload patterns across three different tests:

- Fibre Channel on Linux
- iSCSI on Linux with offload turned on
- iSCSI on Windows 2012 with the default offload values (see “Appendix 2: Configuring the NIC in Linux” on page 11)

Figure 3 shows that at high IOPS levels, Fibre Channel has the best overall performance. However, at lower IOPS, there is much less of a difference and iSCSI gives better results for the higher block size I/O workloads.

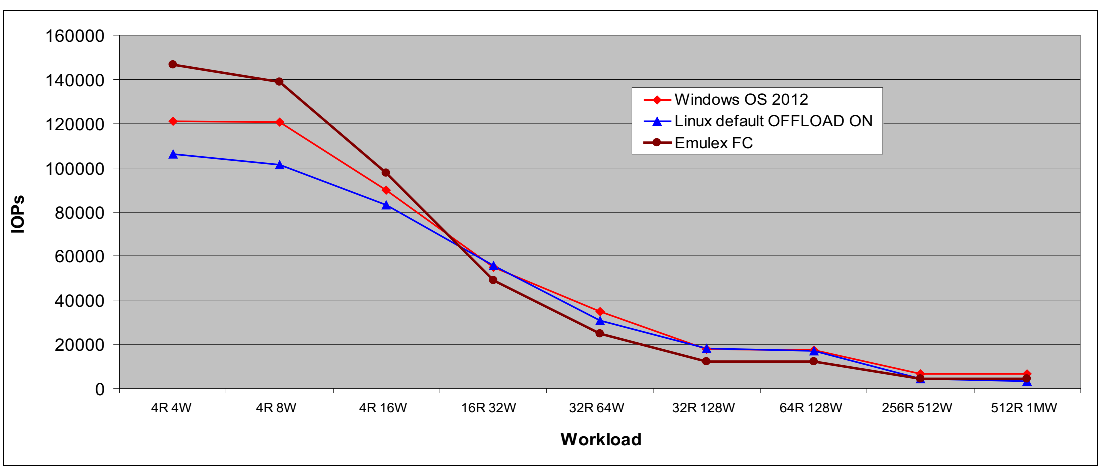

Figure 3 High IOPS levels

Figure 4 shows more detail about IOPS differences for the large-block I/O. As shown in this chart, iSCSI on Windows out-performed Fibre Channel for these workloads.

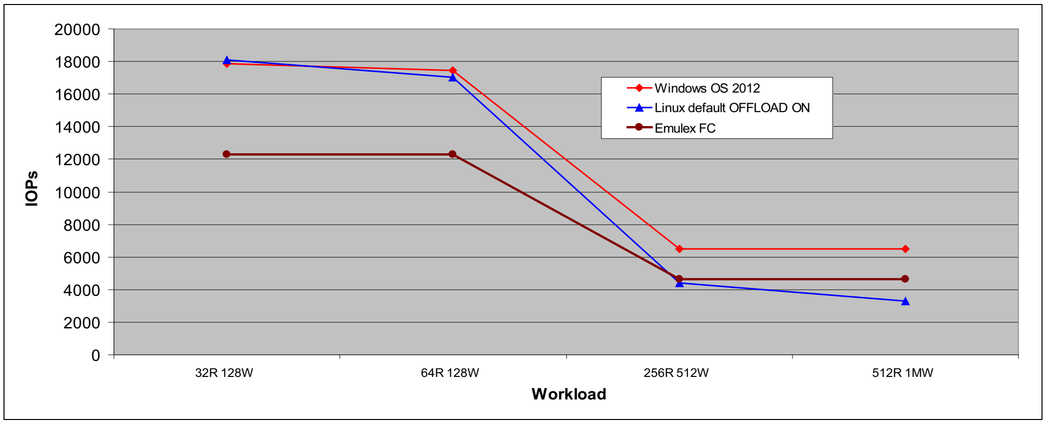

Figure 4 Detailed information about IOPS differences for large-block I/O

Bandwidth results are similar to IOPS. For the workloads with smaller block sizes, Fibre Channel has a slight performance edge. However, iSCSI outperforms Fibre Channel for the higher-block-size workloads, as shown in Figure 5.

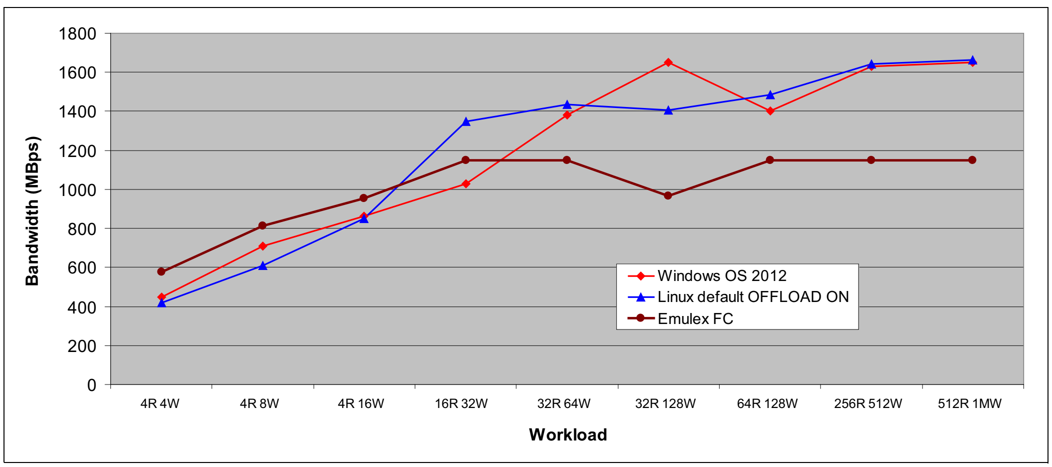

Figure 5 Bandwidth results

Impact of TCP/IP protocol offloads on iSCSI performance

By using the **ethtool** command on Linux, various offloads were toggled and the performance results were compared.

TCP offload is enabled by default in Linux; for the purposes of comparison, the functions were disabled. The preferred practice, however, is to keep the default configuration with the OFFLOAD functions enabled.

Figure 6 shows iSCSI on Linux IOPS results with and without TCP/IP offload. It was expected that there would not be much difference in performance results because the main host processors can handle protocol processing.

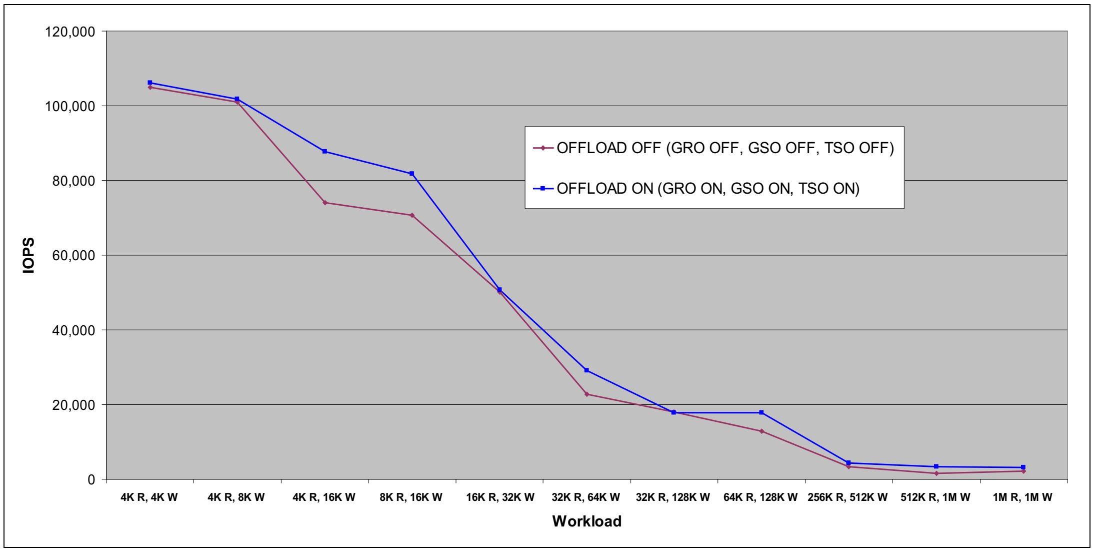

Figure 6 iSCSI results on Linux with and without TCP/IP offload

As shown in Figure 6, performance with and without offloads is almost identical.

Processor impact of iSCSI

Although offloads do not impact performance, they do remove processing burden from the host processors. iSCSI requires additional processing in two areas:

1. TCP/IP protocol processing
2. iSCSI initiator function

Our Emulex adapter, along with RHEL 5.9, was able to provide offload of some of the TCP/IP processing from the host processor to the card, but the iSCSI initiator function on Linux was handled by the host processor. Figure 7 shows the results of this test.

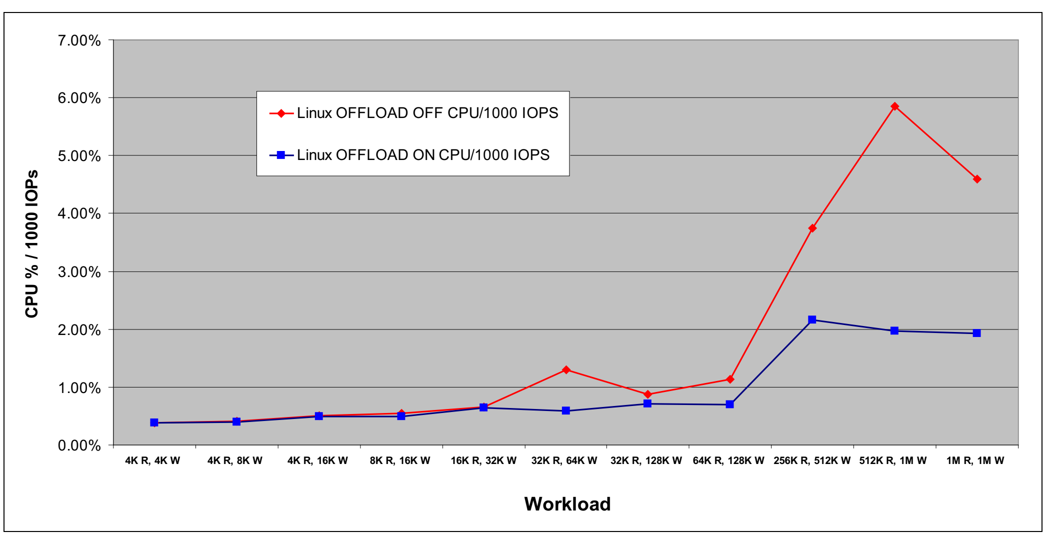

Figure 7 Processor usage with and without offload

This chart shows the percent of a single core that was used by the iSCSI daemons when processing 1000 IOPS for each workload. The processor impact per 1000 IOPS for smaller block workloads is almost negligible, and topped out just over 2% per 1000 IOPS when running the higher block sizes (blue line). This means that 50,000 IOPS running iSCSI uses the equivalent processor capacity of a single core.

The red line in Figure 7 on page 8 shows that there is more significant TCP/IP protocol processing impact at the high block sizes; the red line represents the impact when offloads are turned off. The impact for large blocks reached as high as 6% of a single core per 1000 IOPS. This translates to needing an additional three cores (6% * 50 = 300%) when processing 50,000 IOPS.

Conclusions

Based on the tests we performed, we conclude that 10 Gb iSCSI can deliver very high levels of performance, and is comparable to Fibre Channel across all types of workloads. For large block workloads, iSCSI outperforms Fibre Channel.

The amount of processor impact must be considered, especially for large-block reads and writes. However, the processor impact is balanced by the cost savings that are possible with iSCSI implementations. The performance of iSCSI, and the resulting processing impact, varies by environment and by workload. Based on the results of these tests, 10 Gbps iSCSI is a viable alternative to Fibre Channel for all workloads.

Customer case study: 10 GbE iSCSI with XIV

A MySQL transactional cloud-based application was converted from using direct-attached storage to using a XIV Gen3 system attached through 10 GbE. Although high levels of performance were achieved with direct-attached storage, the implementation was unable to scale to support a larger number of concurrent users without an unacceptable increase in response time (latency).

With the XIV and 10 GbE connectivity, the application showed massive performance improvements, as shown in Figure 8. The customer was able to scale the users by 900% while keeping latency under 2 ms.

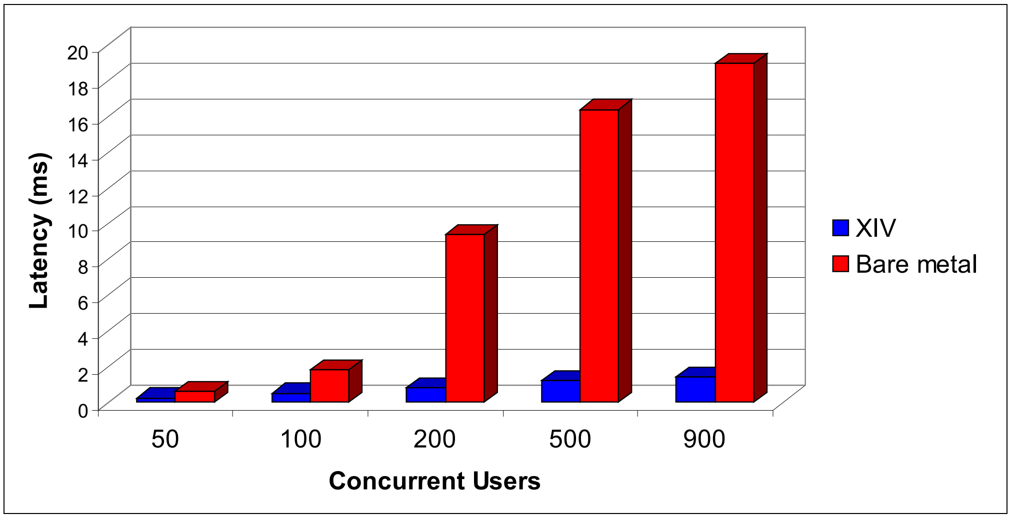

Figure 8   10 GbE iSCSI with XIV

Without the combination of a high-performance iSCSI implementation on the XIV, and the inherent ability of the XIV to support many transactions, the scale-up could not have been possible. This customer is standardized on 10 GbE iSCSI for storage connectivity, and is realizing excellent performance without the high costs that are associated with Fibre Channel implementations.

Appendix 1: The ethtool command on RHEL 5.9

This appendix describes the use of the **ethtool** command.

The command format is:

```
ethtool -k ethx
```

x is the device number for your network interface card (NIC).

Figure 9 shows the default **ethtool** output for the Linux hosts that were used in our test.

```
Offload parameters for eth2:
Cannot get device udp large send offload settings: Operation not supported
rx-checksumming: on
tx-checksumming: on
scatter-gather: on
tcp segmentation offload: on
udp fragmentation offload: off
generic segmentation offload: off
generic-receive-offload: on
```

Figure 9   Host xiv2, Emulex adapter, RHEL 5.9

The **ethtool** command can also be used to turn these offload parameters on and off:

ethtool -K eth*x* gro on gso on rx on tx on lso off …

Hardware-based iSCSI initiator functionality can be provided by specialized network adapters. iSCSI offload was not provided by the Emulex card that was used in the test. As such, there was host impact that was incurred when testing iSCSI. The top command output showed CPU processing impact for iSCSI daemons.

Appendix 2: Configuring the NIC in Linux

During initial installation, the configuration file for the iSCSI NIC must be edited. In addition to the network settings, the MTU size should be set to 9000 to enable jumbo frames.

Go to the following directory:

```
/etc/sysconfig/network-scripts
```

Display the card configuration, as shown in Figure 10.

```
[root@xiv2 network-scripts]# ifconfig
eth0      Link encap:Ethernet  HWaddr 00:90:FA:2B:B6:02
          inet6 addr: fe80::290:faff:fe2b:b602/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:43217 errors:0 dropped:0 overruns:0 frame:0
          TX packets:34 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:10554754 (10.0 MiB)  TX bytes:6701 (6.5 KiB)

eth1      Link encap:Ethernet  HWaddr 00:90:FA:2B:B6:06
          inet addr:10.1.1.110  Bcast:10.1.1.255  Mask:255.255.255.0
          UP BROADCAST MULTICAST  MTU:9000  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:0 (0.0 b)  TX bytes:0 (0.0 b)

eth2      Link encap:Ethernet  HWaddr 00:90:FA:2B:B5:E2
          inet addr:10.1.1.112  Bcast:10.1.1.255  Mask:255.255.255.0
          inet6 addr: fe80::290:faff:fe2b:b5e2/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:9000  Metric:1
          RX packets:43064 errors:0 dropped:0 overruns:0 frame:0
          TX packets:52 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:10540979 (10.0 MiB)  TX bytes:6892 (6.7 KiB)

eth4      Link encap:Ethernet  HWaddr 6C:AE:8B:3C:26:FA
          inet addr:9.32.248.13  Bcast:9.32.248.255  Mask:255.255.255.0
          inet6 addr: fe80::6eae:8bff:fe3c:26fa/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:12073844 errors:0 dropped:0 overruns:0 frame:0
          TX packets:732869 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:870723575 (830.3 MiB)  TX bytes:42655076 (40.6 MiB)
          Memory:c4580000-c45a0000
```

Figure 10   NIC configuration

Ensure that you edit the correct card. Figure 11 shows how to configure eth2.

```
[root@xiv2 network-scripts]# vi icfg-eth2
```

Figure 11   Configuration command

Figure 12 shows the edited configuration file.

```
# Emulex Corporation OneConnect 10Gb NIC (be3)
DEVICE=eth2
BOOTPROTO=none
HWADDR=00:90:fa:2b:b5:e2
ONBOOT=yes
HOTPLUG=no
DHCP_HOSTNAME=xiv2
TYPE=Ethernet
IPADDR=10.1.1.112
NETMASK=255.255.255.0
MTU=9000
~
~
~
~
```

Figure 12 Configuration file

The configured parameters were:

```
MTU=9000
ONBOOT=yes
IPADDR=10.1.1.112
NETMASK=255.255.255.0
```

Changes that are made to the configuration file persist across system boots.

Appendix 3: Windows 2012 MTU configuration

This appendix explains how to configure the maximum transmission unit (MTU) in the Windows 2012 operating system.

Figure 13 shows the settings in Windows Device Manager.

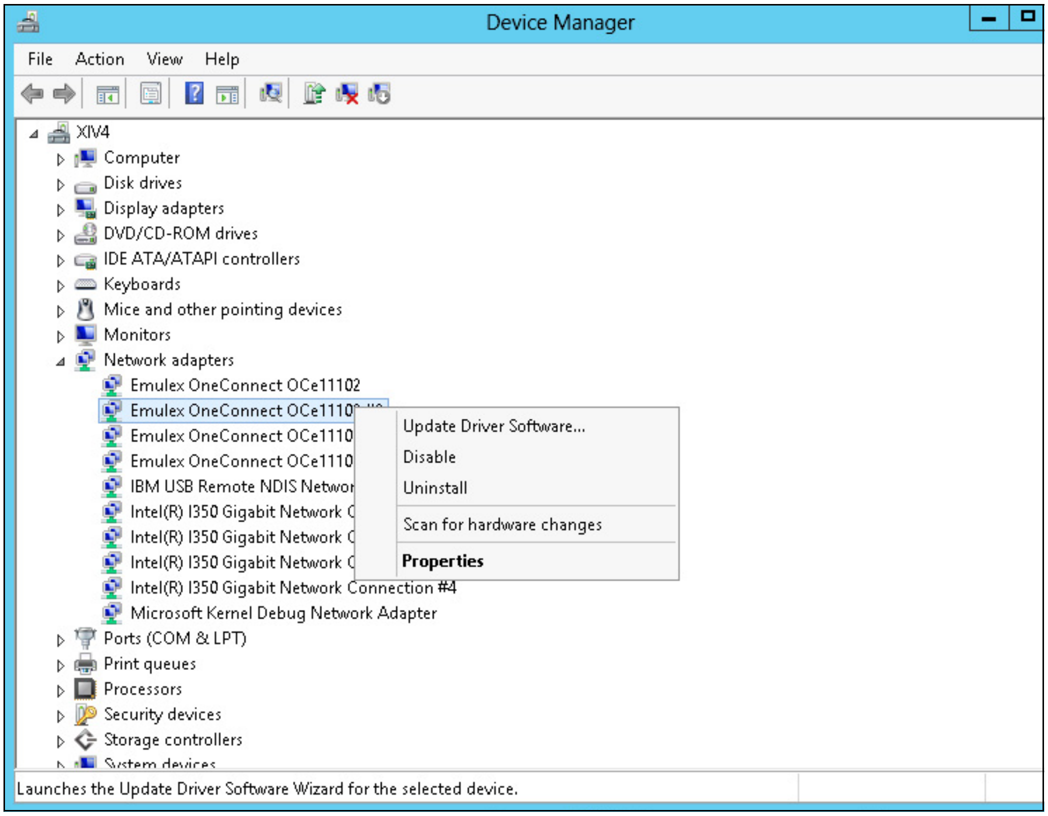

Figure 13 Network adapters

Here is the process that was used to configure the test equipment:

1. Select the appropriate iSCSI card and right-click to select **PROPERTIES**.
2. Click **Advanced** and then click **Packet Size**. The drop-down menu allows variable values. If possible, specify the maximum transmission unit (MTU) size of 9014.

Figure 14 and Figure 15 show the settings for the Emulex adapter.

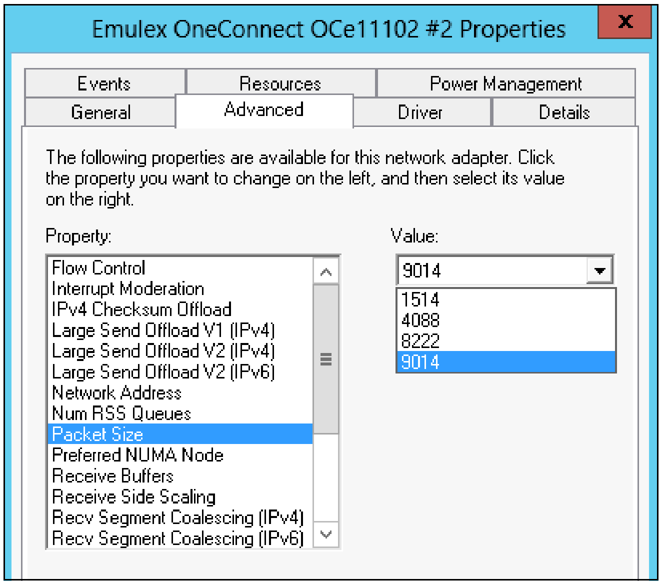

Figure 14 Network adapter properties (part 1)

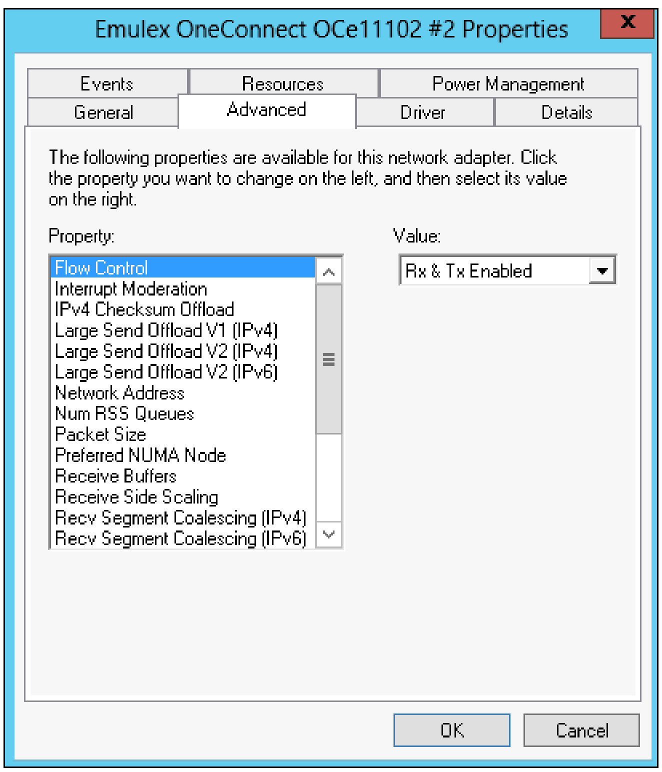

Figure 15 Network adapter properties (part 2)

Appendix 4: iSCSI replication on XIV

Beginning with XIV storage software V11.2, the iSCSI implementation has enabled excellent performance for asynchronous replication over high distances and latencies. Although prior releases saw bandwidth decline rapidly as network latency (distance) increased, the newer releases maintain high bandwidth at much higher distances.

Figure 16 shows the improvement in the iSCSI performance with XIV relative to distance and latency. The average customer running 50 ms of network latency or 2800 miles of distance sees a 250% increase in iSCSI performance with the newer XIV releases. This performance is comparable to Fibre Channel based replication over the same distances, and makes iSCSI a viable alternative to Fibre Channel for asynchronous replication.

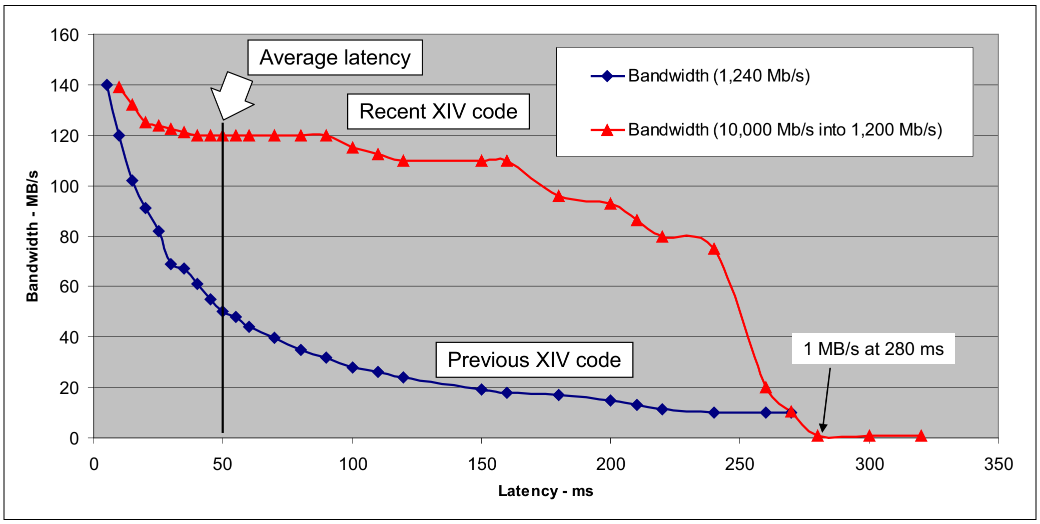

Figure 16 Bandwidth to latency comparison between version of the XIV code (storage software)

Authors

This paper was produced by the International Technical Support Organization, San Jose Center.

Elaine Wood is a Storage Client Technical Specialist and an XIV Certified Specialist. Elaine began her IT career with IBM Systems 390 Microcode Development in Poughkeepsie, NY. She has held various Technical and Sales positions across the industry and most recently was a Manager in Storage Administration with a major financial services firm. She rejoined IBM in January of 2013 to work with Global Financial Services accounts on adoption of emerging storage technologies. Elaine has a Master's Degree in Computer Science from Syracuse University.

Thomas Peralto is a principal consultant in the storage solutions engineering group. He has extensive experience in implementing large and complex transport networks and mission-critical data protection throughout the globe. He has been designing networking solutions for 20 years. He also serves as a data replication and data migration expert and speaks both at national and international levels for IBM on the preferred practices for corporate data protection.

Thanks to the following people for their contributions to this project:

Josh Blumert, Timothy Dawson, Peter Kisich, Carlos Lizzaralde, Patrick Pollard
IBM US

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

Notices

This information was developed for products and services offered in the U.S.A.

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

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

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

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

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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

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

COPYRIGHT LICENSE:

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

This document REDP-5057-00 was created or updated on February 26, 2014.

Send us your comments in one of the following ways:

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

Trademarks