Lenovo

Comparing the Effect of PCIe Host Connections on NVMe Drive Performance

Provides data and analysis of NVMe SSD scaling tests

Validates storage performance of NVMe SSDs with three different PCIe configurations

Describes suitable test parameters of storage benchmark for different workloads

Explains the importance of correctly configuring and sizing your NVMe drive subsystem

Travis Liao

LENOVO PRESS

Abstract

With the rapid improvement of flash technology and the recent implementation of the Non-Volatile Memory express (NVMe) protocol replacing the SATA/SAS protocol, it is to the customer's advantage to select NVMe SSDs instead of conventional drives to gain significant improvements to internal storage performance. However, to gain the maximum benefit of these new drives, it is important that the components and connections in the storage subsystem be properly selected.

This paper demonstrates why the use of x4 PCIe lanes is crucial to the performance of the NVMe storage subsystem and the potential performance impact when a reduced number of PCIe lanes is used. This paper is aimed at users planning to deploy servers with high performance NVMe SSDs. It is expected that readers will have an understanding of the basic concepts of storage performance benchmarking.

At Lenovo® Press, we bring together experts to produce technical publications around topics of importance to you, providing information and best practices for using Lenovo products and solutions to solve IT challenges.

See a list of our most recent publications at the Lenovo Press web site:

http://lenovopress.com

Do you have the latest version? We update our papers from time to time, so check whether you have the latest version of this document by clicking the **Check for Updates** button on the front page of the PDF. Pressing this button will take you to a web page that will tell you if you are reading the latest version of the document and give you a link to the latest if needed. While you're there, you can also sign up to get notified via email whenever we make an update.

Contents

Introduction

Non-Volatile Memory express solid-state drives (NVMe SSDs) are PCIe SSDs supporting the NVMe protocol, which designed specifically to reduce latency and utilize internal parallelism of the SSDs. It is set to replace SAS and SATA interfaces that were primarily designed for mechanical hard disk drives.

With the technology improvement on fast NVMe SSD in recent years, a single NVMe SSD can deliver more than 2,000 MB/s with a sequential workload. The NVMe drive's throughput is no longer the only bottleneck of storage performance. The connection from drive to processor need to be taken into consideration to ensure the throughput of storage subsystem reach its true potential.

NVMe SSDs are available in the industry in multiple form factors:

- A traditional drive form factor (2.5-inch or 3.5-inch)
- Standard PCIe adapter form factor (also known as an add-in card or AIC)
- M.2 form factor

This paper is focused on the implementation using the traditional drive form factor as implemented in Lenovo ThinkSystem™ servers.

The purpose of this paper is to show the impact on performance by comparing the performance of NVMe storage devices attached to a ThinkSystem server using each of three methods:

- Direct cable connection from NVMe ports on the system board of the server
- Connection through 1610-4P NVMe Switch Adapter with 16 PCIe lanes
- Connection through 810-4P NVMe Switch Adapter with eight PCIe lanes

These three connections are shown in Figure 1.

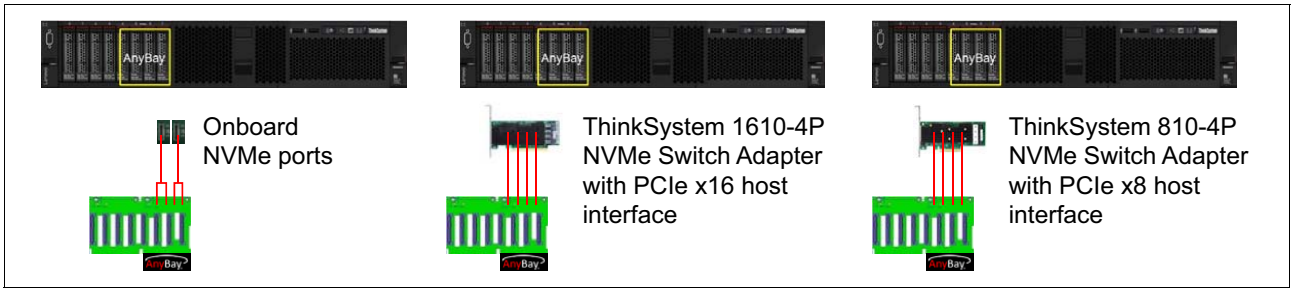


Figure 1 The NVMe connections to test

Architecturally, the PCIe connections for these three configurations are shown in Figure 2.

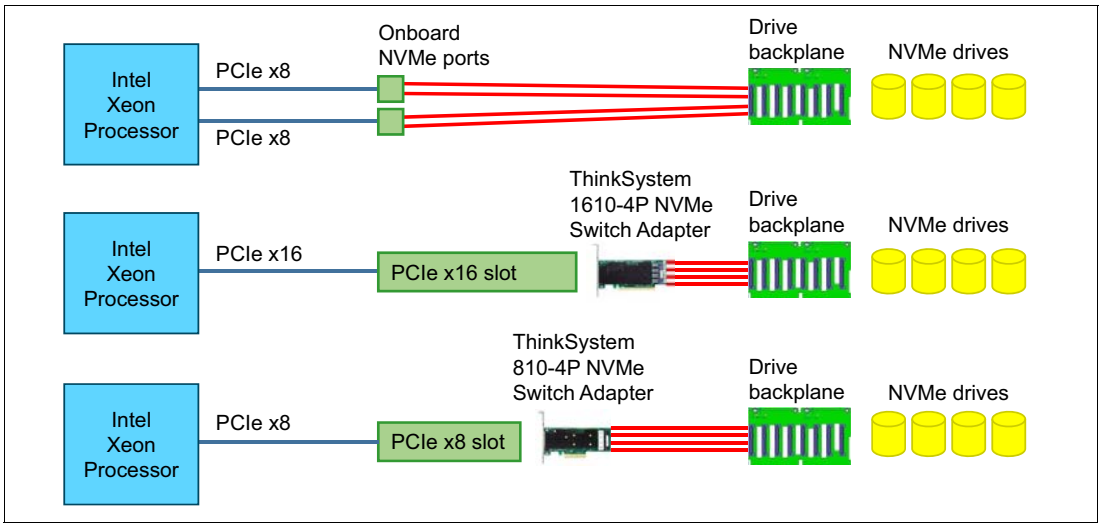


Figure 2 Block diagrams showing the PCIe connections and cables of the configurations

For direct cable and x16 adapter configurations, each NVMe SSD in the system has bandwidth of four PCIe lanes. For x8 adapter configuration, each NVMe SSD only gets bandwidth of two PCIe lanes.

Lab configuration

The system we tested with was a four-socket Lenovo ThinkSystem SR850 server with the following configuration:

- 4x Intel Xeon Platinum 8176 processors
- 24x 16GB memory DIMMs operating at 1600 MHz
- 1x ThinkSystem 1610-4P NVMe Switch Adapter
- 1x ThinkSystem 810-4P NVMe Switch Adapter*
- 4x ThinkSystem PX04PMB 960GB Mainstream NVMe SSDs
- 1x ThinkSystem M.2 Enablement Kit
- Red Hat Enterprise Linux 7.3

*** Note:** ThinkSystem 810-4P NVMe Switch Adapter is currently not officially supported for use in the ThinkSystem SR850 server. The use of this adapter for this test is for illustrative purposes only.

RHEL was installed on an M.2 drive rather than on the NVMe SSDs to avoid interference on target disks. The Operating mode in UEFI was set to maximum performance to prevent CPU frequency fluctuation during test period.

FIO test parameters

FIO (Flexible I/O Tester) is a synthetic benchmark tool that can simulate behaviors of storage workloads with specific access patterns, such as sequential/random, read/write, block size, thread numbers, queue depth. We use FIO to test the performance of the different configurations using different workloads.

Table 1 shows the test parameters used on each workload.

Table 1 Listing of FIO parameters

Patterns	Block Size	Workers	Queue Depth
Random Write	4KB	8	256
Random Read	4KB	8	256
OLTP	4KB	8	256
Sequential Write	256KB	1	64
Sequential Read	256KB	1	64

Test results

In each test, we scaled from one NVMe SSD to four NVMe SSDs to observe the performance as loading to system increase.

The effect on sequential workloads

First, we looked at the impact of the PCIe connection on sequential workloads.

Figure 3 shows bandwidth in MB/second for a 100% sequential write workload using 256 KB blocks.

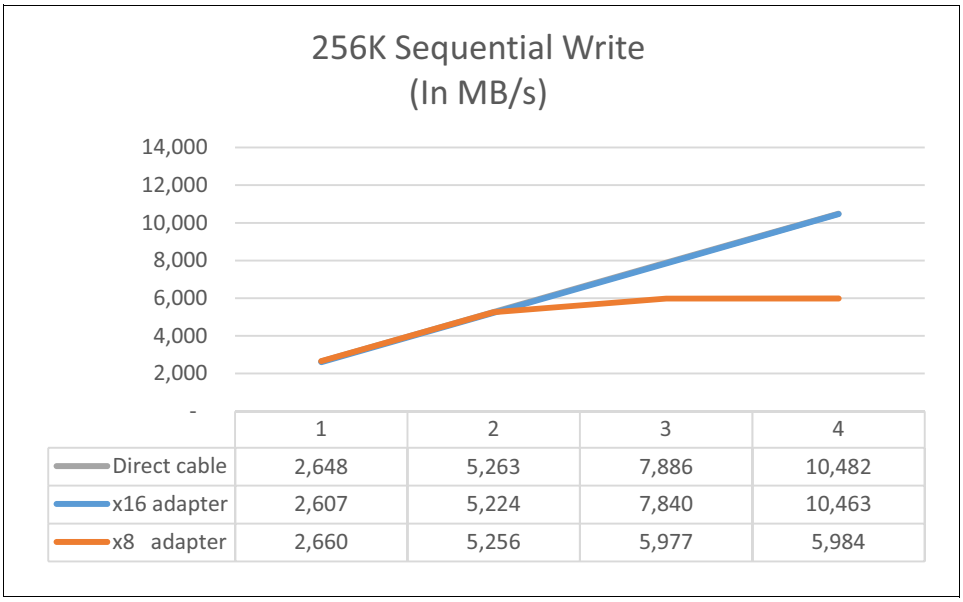


Figure 3 Results of 256K sequential write

Figure 4 shows bandwidth in MB/second for a 100% sequential read workload using 256 KB blocks.

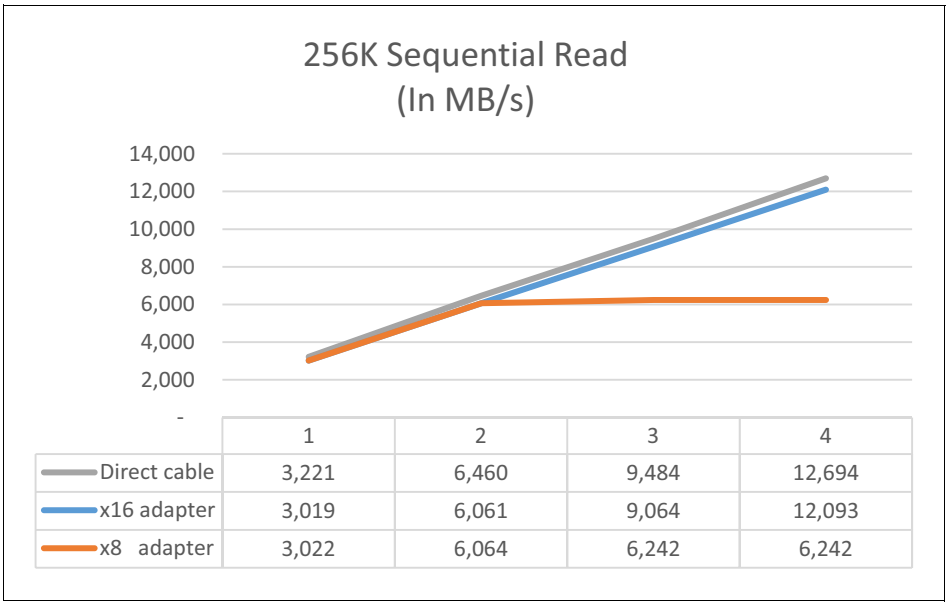


Figure 4 Results of 256K sequential read

As the charts show, the scaling effect of direct cable and 1610-4P x16 adapter configurations work out as expected on sequential workloads. As each additional NVMe drive is added, the performance increased linearly.

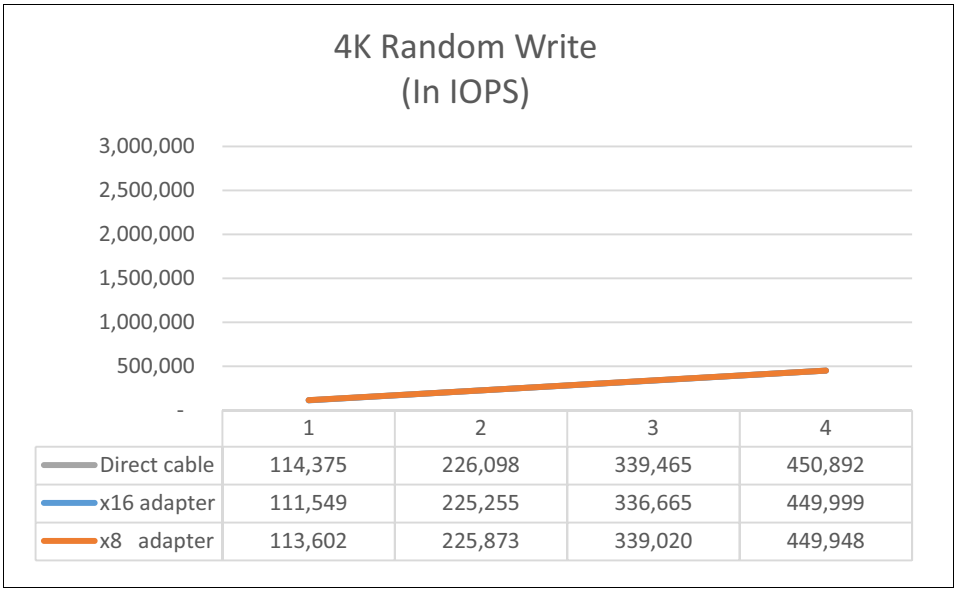
There is a slightly difference between direct cable and x16 adapter configuration on sequential read workload. As the relatively low stress level of sequential workload, the performance difference exists but is insignificant.

However, with the 810-4P x8 adapter configuration, due to the x8 PCIe lane constraint, bandwidth is capped to 6,200 MB/s once more than two disks are attached.

The effect on random workloads

We then looked at the impact of the PCIe connection on random workloads.

Figure 5 shows throughput in I/O operations per second (IOPS) for a 100% random write workload using 4KB blocks.



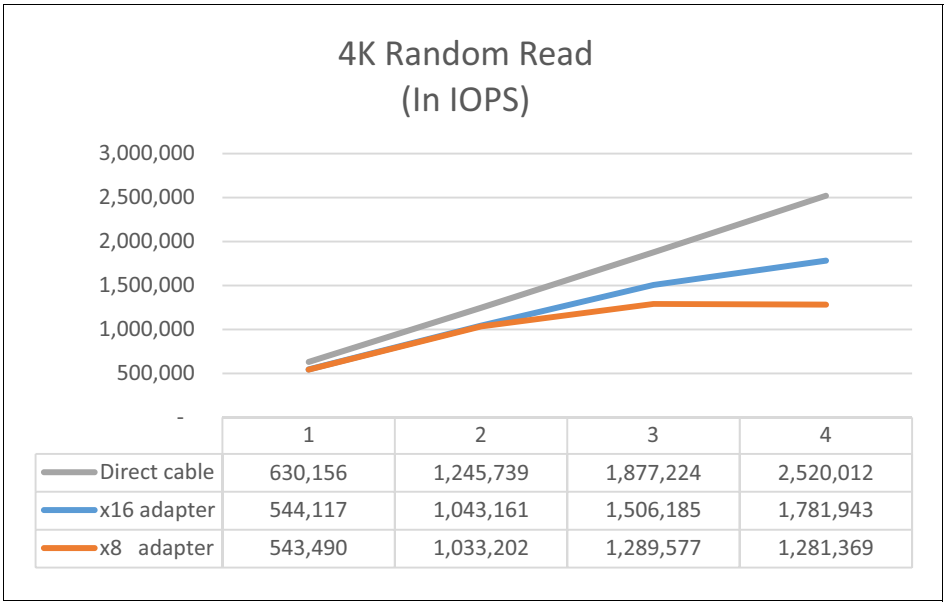
	1	2	3	4
Direct cable	114,375	226,098	339,465	450,892
x16 adapter	111,549	225,255	336,665	449,999
x8 adapter	113,602	225,873	339,020	449,948

Figure 5 Results of 4K random write

Note: Figure 5 only appears to show the orange line for the x8 adapter. That is because the grey and blue lines are nearly identical and are hidden underneath the orange line.

Figure 5 shows with the slower write than read behavior on NAND flesh technology and small access block size, the disks do not hit any limitation on 4KB random write. All three configurations perform equally with good scaling outputs.

Figure 6 shows throughput in IOPS for a 100% random read workload using 4KB blocks.



	1	2	3	4
Direct cable	630,156	1,245,739	1,877,224	2,520,012
x16 adapter	544,117	1,043,161	1,506,185	1,781,943
x8 adapter	543,490	1,033,202	1,289,577	1,281,369

Figure 6 Results of 4K random read

In Figure 6, the direct cable configuration shows no performance degradation and outperforms the 1610-4P x16 adapter configuration with considerable performance gap. Even

with the same x16 PCIe bandwidth to CPU, the processing capacity of the adapter itself limits the IOPS from matching the direct connections.

Scaling from one to four disks, the random read IOPS of 810-4P x8 adapter is capped below 1.3 million IOPS. Due to both bandwidth and processing power limitation, the performance reduced by half compared to direct cable with four NVMe SSDs attached.

Conclusion

When the high performance NVMe SSDs used as the internal storage, high throughput on both bandwidth and throughput are expected. It is crucial to check every potential bottleneck to ensure that performance is not restricted by any factor to maximize the return of investment on the storage subsystem.

This paper demonstrates that the number of PCIe lanes used to connect NVMe SSDs to the processors can affect performance in many use cases with the exception of random write workloads. Performance drops of up to 50% could happen four NVMe drives are connected to the 810-4P adapter with only eight PCIe lanes to the processor.

The analysis also showed that the performance of NVMe drives connected via an NVMe switch adapter may also be impacted compared to NVMe drives directly connected to the processor via onboard NVMe ports. A performance impact of up to 30% may be incurred by the processing overhead of the switch adapter.

In conclusion, the use of NVMe ports on the system board of the server for connecting NVMe drives is a preferable approach to avoid any impact on performance for high performance NVMe SSDs. This NVMe direct connection provides a PCIe x4 link for each drive plus there is no switch in the path to limit performance. Using a configuration where only a PCIe x2 link is available for each drive should only be used when other devices limit the total number PCIe lanes.

Authors

Travis Liao is a Performance Engineer in the Lenovo Data Center Group Performance Laboratory based in Taipei. His focus is modelling and validating performance of server storage subsystem including RAID controllers, SSDs and software RAID. Travis holds a Master's Degree in Electronic Engineering from National Taiwan University in Taiwan.

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

- David Watts, Lenovo Press

Notices

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

Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
1009 Think Place - Building One
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO 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 jurisdictions 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. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

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

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

Any performance data contained herein was determined in a controlled environment. Therefore, the result 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.

This document was created or updated on April 13, 2018.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/lp0865

Trademarks

Lenovo, the Lenovo logo, and For Those Who Do are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. These and other Lenovo trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by Lenovo at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of Lenovo trademarks is available on the Web at http://www.lenovo.com/legal/copytrade.html.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Lenovo® Lenovo(logo)® ThinkSystem™

The following terms are trademarks of other companies:

Intel, Xeon, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.