Lenovo

Measuring the Impact of Memory on RDMA, TCP and UDP Network Performance

Describes the performance analysis conducted by the Lenovo Performance Lab

Explains the network performance analysis on various memory populations

Illustrates the different memory demands for RDMA, TCP and UDP workload

Provides recommendations regarding memory population

Jocelyn Wang

LENOVO PRESS

Abstract

From a network performance perspective, the recommended memory population for Lenovo® ThinkSystem™ servers is usually one DIMM per memory channel (1 DPC) at a minimum. However, customers may want a smaller quantity of DIMMs because they do not require the memory capacity.

The purpose of this paper is to measure the precise effect on network performance with various memory populations and then to provide a guide for customers to follow to select a suitable memory configuration according to their workload.

RDMA, TCP and UDP network performance data using 64K packet size is measured on two Lenovo ThinkSystem SD530 servers with a direct point-to-point connection. Tests were performed using a Mellanox ConnectX-4 EDR VPI adapter and each test was performed 30 times. The results show that RDMA performance is independent of the memory population, but a balanced memory configuration across the populated memory channels is a key factor for TCP/UDP performance.

This paper is for customers who run applications with high-performance networking needs and need to know how the memory configuration may affect performance. Readers should have a basic understanding of memory and networking technologies.

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

TCP/IP and UDP/IP over Ethernet are traditional and widely used network protocols. Both use socket-based communication. To transmit data, an application copies the data from an application buffer to a socket buffer. The data is processed through the network protocol stacks under kernel space. Lastly, the NIC copies the data from the main memory to its buffer to transmit onto the network. Multiple memory copies and kernel involvement make the traditional transmission susceptible to performance issues such as high CPU utilization and high network latency.

To mitigate these performance issues, high-performance network technologies such as RDMA (Remote Direct Memory Access) have been developed. RDMA is a technology that allows an application to read and write to virtual memory in a remote host.

Compared to TCP/IP, RDMA has three advantages:

- Zero copy
- Kernel bypass
- No CPU involvement

The data channel is created between the host application memory and the NIC with the RDMA engine. Data is transferred through the channel without multiple memory copies in the network protocol stacks. The OS kernel is involved only for the control path such as resource allocation and connection establishment rather than the data path. When doing RDMA reads and writes, the remote host CPU is not involved.

There are 4 types of RDMA techniques:

- InfiniBand (IB)
- iWARP (Internet Wide Area RDMA Protocol)
- RoCEv1 (RDMA over Converged Ethernet)
- RoCEv2

These can be divided into two groups according to the link layer. IB uses the InfiniBand link layer. iWARP, RoCEv1, and RoCEv2 use the Ethernet link layer.

For the lab experiments, the Mellanox ConnectX-4 1x100GbE/EDR IB QSFP28 VPI Adapter (part number 00KH924) was selected for the RDMA, TCP and UDP performance tests in this paper. It is one of the highest speed network adapters supported on the Lenovo ThinkSystem SD530. The Virtual Protocol Interconnect (VPI) enables EDR to support both 100 Gb/s InfiniBand and 100 Gb/s Ethernet. Two modes can be switched to EDR: ib mode (link layer: InfiniBand) and eth mode (link layer: Ethernet).

Memory population for optimal high-speed network adapter performance in various tuning guides suggests customers have 1 DPC or 2 DPC balanced memory populations. However, customers may choose a DIMM population lower than 1 DPC due to budget concerns. The motivation of this paper is to measure the network performance including traditional and high-speed network protocols and to observe the performance impacts resulting from different memory populations.

Methodology

In our lab, two Lenovo ThinkSystem SD530 servers with the same configuration were connected back to back. One system was used as a server, and the other one was as a client. IB mode or eth mode was switched by using Mellanox Software Tools (MST) v4.13.0, which is a service that runs on Linux. The tests commands were run on the server first, and then on the client.

There are 16 memory DIMM slots in SD530 (see Figure 1): 12 slots for DRAM and 4 slots for Intel Optane DC Persistent Memory (DCPMM). In this paper, only the DRAM slots were used.

- For a one processor socket configuration, DRAM is installed in the following order: slots 6, 3, 7, 2, 8, 1.
- For a 2 processor socket configuration, DRAM is installed in slots 6, 14, 3, 11, 7, 15, 2, 10, 8, 16, 1, 9.

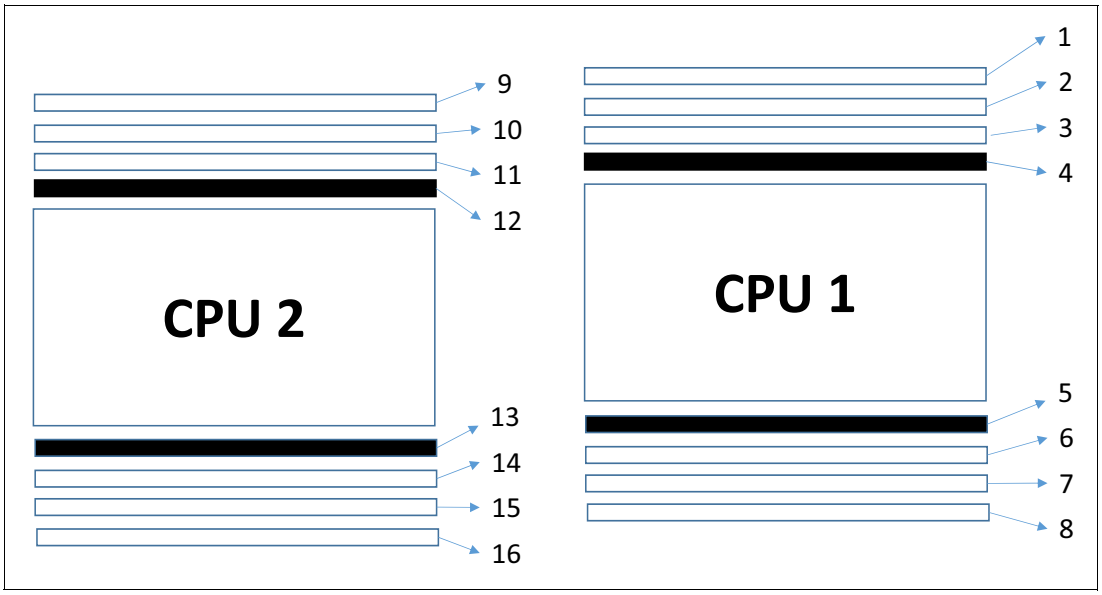

Figure 1 Memory DIMM slots in the ThinkSystem SD530

RDMA, TCP and UDP network performance was measured with each memory configuration, using 64K packet sizes. In order to avoid bias during the test, 30 test iterations were executed with each memory configuration.

One-processor configuration:

- CPU: 1x Intel Xeon Platinum 8280 processor, 2.60 GHz
- Memory: 32 GB DDR4 2933 MHz RDIMM,
- DIMM quantity: 1-6 in slots: 6, 3, 7, 2, 8, 1
- NIC: Mellanox ConnectX-4 1x100GbE/EDR IB QSFP28 VPI Adapter
- Cable: 100Gb/s QSFP28 Direct Attach Copper Cable
- OS: RHEL7.6
- NIC Driver: 4.5-1.0.1
- NIC firmware: 12.20.1010
- PCIe slot: 1 (device numa_node: 0)
- BIOS setting: Maximum Performance Mode

Two-processor configuration:

- CPU: 2x Intel Xeon Platinum 8280 processors, 2.60 GHz
- Memory: 32 GB DDR4 2933 MHz RDIMM,
- DIMM quantity: 1-12 in slots: 6, 14, 3, 11, 7, 15, 2, 10, 8, 16, 1, 9
- NIC: Mellanox ConnectX-4 1x100GbE/EDR IB QSFP28 VPI Adapter
- Cable: 100Gb/s QSFP28 Direct Attach Copper Cable,
- OS: RHEL7.6
- NIC Driver: 4.5-1.0.1
- NIC firmware: 12.20.1010
- PCIe slot: 1 (device numa_node: 0)
- BIOS setting: Maximum Performance Mode

RDMA test tools, parameters, and commands:

- Tool for RDMA over IB: ib_write_bw (perftest rpm), under ib mode.
- Tool for RDMA over Ethernet: ib_write_bw (perftest rpm), under eth mode, type: RoCEv1
- Packet size: 64K byte
- Test duration: 30 seconds
- Iteration: 30 runs for each test
- Command on RDMA server:

  ```
  #numactl --cpunodebind=0 ib_write_bw -s packet_size -D 30 -d mlx5_0 -i 1
  --report_gbits -F
  ```

- Command on RDMA client:

  ```
  #numactl --cpunodebind=0 ib_write_bw -s packet_size -D 30 -d mlx5_0 -i 1
  --report_gbits –F server_ip
  ```

TCP/UDP test tools, parameters, and commands:

- Tool for TCP/UDP: iperf-2.0.9, under eth mode.
- Packet size: 64K byte
- Test duration: 30 seconds
- Iteration: 30 runs for each test
- Command on TCP server:

  ```
  #numactl --cpunodebind=0 --membind=0 iperf -s
  ```

- Command on TCP client:

  ```
  #numactl --cpunodebind=0 --membind=0 iperf -c server_ip -l packet_size -P 4 -t
  30
  ```

- Command on UDP server:

  ```
  #numactl --cpunodebind=0 --membind=0 iperf -s -u -w 416K
  ```

- Command on UDP client:

  ```
  #numactl --cpunodebind=0 --membind=0 iperf -c server_ip -u -l packet_size -b 1G
  -P 120 -t 30 -w 416K
  ```

Plot tool:

- R version 3.6.1, Boxplot (ggplot package)

Results

This section analyzes the results of the one-processor and two-processor tests using a range of quantities of installed memory DIMMs.

Single-processor tests

The one socket IB/RoCEv1/TCP/UDP performance test result is shown in Figure 2 using boxplot.

Six different memory configurations from 1 DIMM to 6 DIMMs installed were evaluated. Each box in the boxplot presents 30 data points. The middle line in the box indicates the median of the dataset. The top edge of the box shows the Q3 (75th percentiles) data, and the bottom edge shows the Q1 (25th percentiles) data. The circles out of the box refer to outliers which are defined by “data > Q3+1.5*(Q3-Q1)” or “data < Q1-1.5*(Q3-Q1)”.

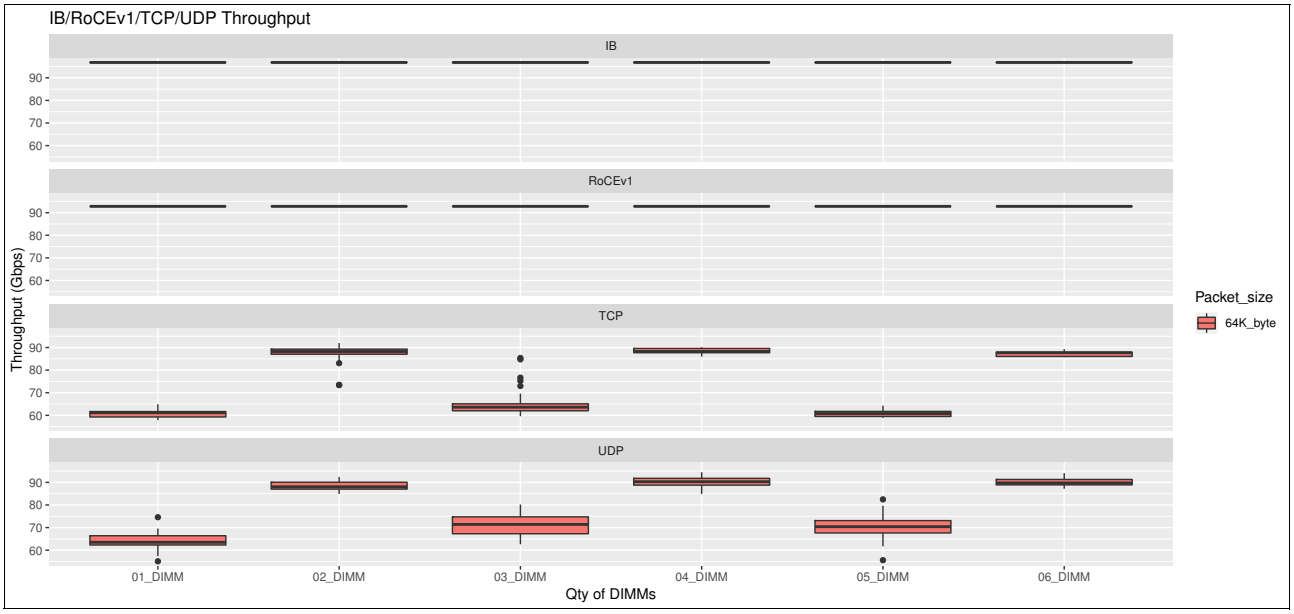

Figure 2 IB/RoCEv1/TCP/UDP throughput on 1-6 memory DIMM configurations with one processor socket installed

As Figure 2 shows, IB and RoCEv1 performed consistently independent of the memory population. In general, the median of the IB dataset is 96.8 Gbps, and the median of RoCEv1 dataset is 92.79 Gbps.

The Q3 and Q1 are very close in each IB and RoCEv1 datasets -- in Figure 2 the IB and RoCEv1 boxes are more like thick lines than boxes -- however the TCP and UDP data shows a wider variation. The TCP/UDP results can be divided into two groups: high performance and low performance. The TCP/UDP performance is correlated with the memory population.

To study the relationship between TCP/UDP and memory population further, Table 1 shows whether the memory population was balanced or unbalanced and the median of TCP/UDP throughput for one processor socket installed.

Table 1 Median of TCP/UDP throughput on 1-6 DIMM configurations with one processor installed

Quantity of DIMMs	Memory channel (CPU1)	Memory channel (CPU2)	Median of TCP / UDP (Gbps)
1	Unbalanced	None	61.15 / 63.5
2	**Balanced**	None	88.25 / 88.05
3	Unbalanced	None	63.6 / 71.4
4	**Balanced**	None	88.25 / 90.3
5	Unbalanced	None	60.8 / 70.4
6	**Balanced**	None	87.7 / 89.75

Intel Xeon Scalable processors have two integrated independent memory controllers and each memory controller supports 3 memory channels. The memory balanced/unbalanced status indicates that the memory population is balanced when both memory controllers in one socket have the same DIMM population.

In Table 1, the TCP/UDP performance data measured on populations with 2, 4, and 6 DIMMs (median: 88~90 Gbps) are significantly higher than the data measured on populations with 1, 3, and 5 DIMMs (median: 60~71 Gbps). The 2, 4, 6 DIMM populations represent balanced memory configurations, whereas the 1, 3, 5 DIMM populations represent unbalanced memory populations.

From the results, we observe that a balanced memory population is required for high TCP/UDP performance.

Two-processor tests

Figure 3 on page 8 shows the two-processor IB/RoCEv1/TCP/UDP performance test results. Tests were performed using different quantities of memory DIMMs, from 1 to 12.

IB and RoCEv1 show the same trend as was measured on the one processor socket results. The median of the IB dataset is 96.8 Gbps and the median of RoCEv1 dataset is 92.79 Gbps. The results show that the IB/RoCEv1 network performance is independent of the memory population, i.e. independent of whether the population was balanced or unbalanced.

Regarding TCP/UDP, we can observe that the measurement on 3, 4, 8, 11, and 12 DIMM populations have higher performance scores compared with 1, 2, 5, 6, 7, 9, 10 DIMM population.

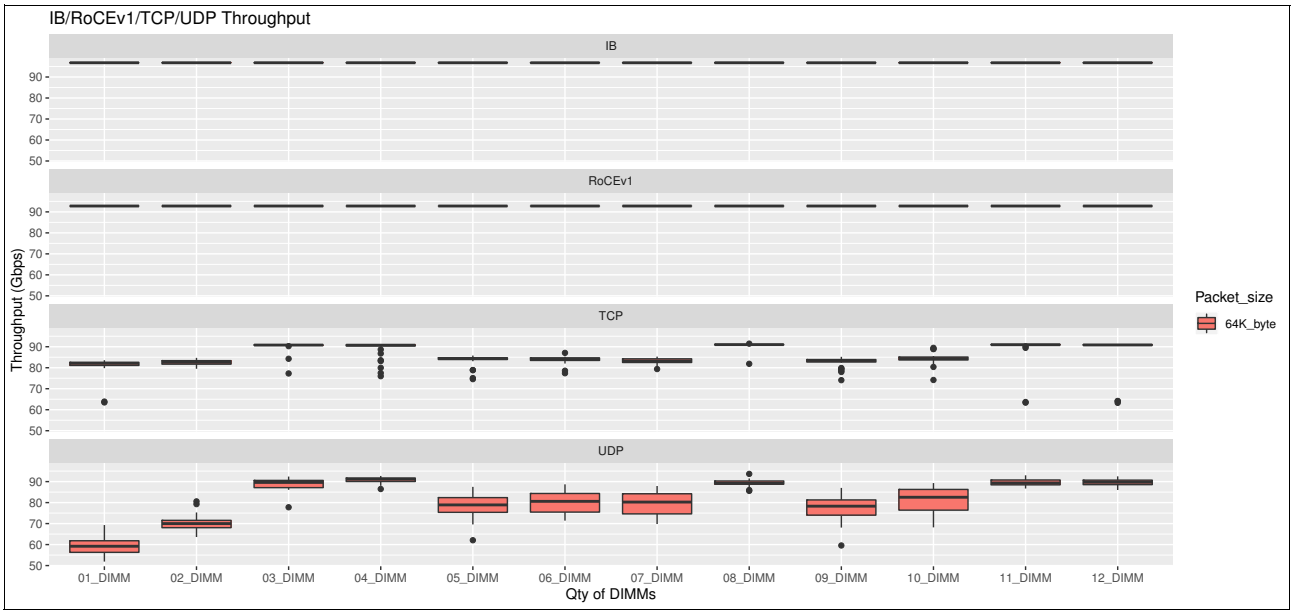

Figure 3 IB/RoCEv1/TCP/UDP throughput on 1-12 memory DIMM configurations with 2 processor sockets installed

The memory channel balance vs. unbalanced status and median of TCP/UDP performance is listed in Table 2. For DIMM quantities of 3, 4, 8, 11, and 12, the median the TCP throughput is > 90 Gbps and the median of the UDP throughput > 89 Gbps. However, for DIMM quantities of 1, 2, 5, 6, 7, 9, 10, the median of the TCP throughput ranges from 82.1 to 84.35 Gbps and the median of the UDP throughput ranges from 59.2 to 82.55 Gbps.

Table 2 Median of TCP/UDP throughput on 1-12 DIMM configurations with 2 processors

Quantity of DIMMs	Memory channel (CPU1)	Memory channel (CPU2)	Median of TCP / UDP (Gbps)
1	Unbalanced	None	82.1 / 59.2
2	Unbalanced	Unbalanced	82.6 / 70
3	**Balanced**	Unbalanced	90.8 / 89.6
4	**Balanced**	**Balanced**	90.75 / 91.15
5	Unbalanced	**Balanced**	84.35 / 78.9
6	Unbalanced	Unbalanced	84.15 / 80.6
7	**Balanced**	Unbalanced	83.2 / 80.25
8	**Balanced**	**Balanced**	91 / 89.2
9	Unbalanced	**Balanced**	83.45 / 78.3
10	Unbalanced	Unbalanced	84.35 / 82.55
11	**Balanced**	Unbalanced	90.95 / 89.35
12	**Balanced**	**Balanced**	90.85 / 89.7

According to the memory balance vs. unbalanced status in Table 2, memory configurations with higher TCP/UDP performance have balanced configurations on CPU 1. However, TCP/UDP median performance with the 7 DIMM configuration only reaches 83.2/80.25 Gbps, even if memory is balanced on CPU1.

Therefore, we recommend that all processor sockets be populated with a balanced memory configuration (configuration 4, 8 and 12) to reach high TCP/UDP performance.

Conclusion

In this paper, different memory population configurations on the Lenovo ThinkSystem SD530 server were used to measure RDMA/TCP/UDP network performance. We can clearly observe the different effects with balanced and unbalanced memory configurations between RDMA and TCP/UDP due to the characteristics of these protocols.

RDMA (IB and RoCEv1) performance is optimized regardless of the number of DIMMs installed and regardless of whether or not the memory configuration is balanced. Any memory configuration will yield the best performance for RDMA networks.

For TCP/UDP network performance on the other hand, having a balanced memory configuration plays an important role in optimizing. To optimize TCP or UDP network performance, it is recommended that all processor sockets should have a balanced memory configuration.

Our lab tests were performed using the Mellanox ConnectX-4 1x100GbE/EDR IB QSFP28 VPI Adapter, however the conclusions also apply to other network adapters. The important criteria is to determine what protocols your selected adapter supports. For Ethernet adapters, the analysis of TCP, UDP and RoCEv1 (if supported) is relevant. For other InfiniBand adapters or OPA adapters, the RDMA comparisons are relevant.

Author

Jocelyn Wang is a Hardware Performance Engineer in the Lenovo Data Center Group Performance Laboratory based in Taipei, Taiwan. She is responsible for performance validation of the PCIe bus and network subsystem, including the RDMA/TCP/UDP protocols on Lenovo ThinkSystem servers. Jocelyn holds a Master's Degree in Biomedical Electronics and Bioinformatics from the National Taiwan University in Taiwan.

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

- Joseph Jakubowski, Lenovo Principal Engineer for Performance
- Arunkumar Krishnamoorthy, Lenovo Server Performance Development Engineer
- Deeily Bonieck La Rosa, Server Performance Intern
- 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 October 30, 2019.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/lp1241

Trademarks

Lenovo and the Lenovo logo 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 from https://www.lenovo.com/us/en/legal/copytrade/.

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

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