ThinkSystem DE Series Performance Tuning Best Practices

Last Update: April 2023

Introduces performance tuning considerations on DE Series storage systems

Explains DE Series tuning terminologies

Covers ThinkSystem System Manager tuning options

Provides the reference resources and the best practices for specific workloads

Vincent Kao

LENOVO PRESS

Abstract

Lenovo® ThinkSystem™ DE Series storage array is designed to provide high performance I/O. ThinkSystem System Manager is the browser-based management software embedded on the DE Series storage array and is a versatile and easy to use for performance tuning purposes. In this paper, we explain the options for tuning the storage array and the background behind the tuning. We also provide recommendations for quick reference. Operating system-related tuning practices are not covered, however.

This paper is intended for technical specialists, sales specialists, sales engineers, and IT architects who want to learn more about the performance tuning of the ThinkSystem DE Series storage array. It is recommended that users have basic ThinkSystem System Manager knowledge.

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

Lenovo ThinkSystem DE Series Storage Arrays are SAN storage systems that are designed to provide performance, simplicity, capacity, security, and high availability for medium to large businesses. ThinkSystem DE Series Storage Arrays deliver hybrid and all flash storage with enterprise-class storage management capabilities and a wide choice of host connectivity options, flexible drive configurations, and enhanced data management features.

Figure 1 Lenovo ThinkSystem DE4000H

The best practices in this paper apply to Lenovo ThinkSystem System Manager Version 11.50 and later. ThinkSystem DE Series systems support multiple interface protocols, including Fibre Channel, iSCSI and SAS, however the tuning considerations are generally protocol independent.

This document is intended for DE Series, which is the block storage, only. The illustration of DE Series solution in SAN fabric is shown in Figure 2.

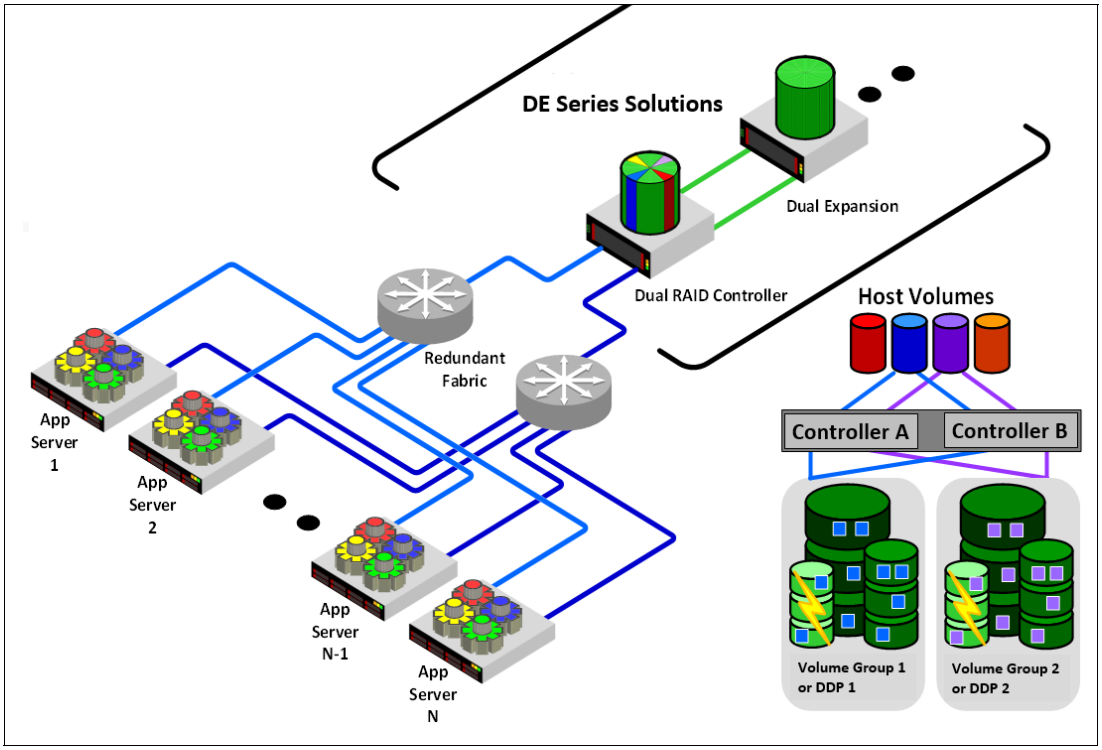

Figure 2 DE Series solution in SAN fabric

For more information about ThinkSystem DE Storage, see the Lenovo Press product guide:

https://lenovopress.com/lp0940-lenovo-thinksystem-de-series-storage-arrays

Performance metrics and measurement units

The following three performance metrics are recommended whenever measuring the performance of storage arrays:

- IOPS

  Input/output operations per second (IOPS) is the measure of how many I/O communications pass between the hosts and the storage systems in one second.

- Latency

  Latency is the time interval between a request, such as a read or a write, and the response from the storage system.

- Throughput/bandwidth

  Throughput is a measure of how much data pass between the host and the storage system in one second.

The measurement units used are as follows:

- Decimal units

  The capacity of the storage media commonly uses decimal units.

  - A kilobyte (KB) is equal to 1,000 (10^3) bytes
  - A megabyte (MB) is equal to 1,000,000 (10^6) bytes

- Binary units

  Many operating systems use binary units rather than decimal units. This document and ThinkSystem System Manager use the binary representation.

  - A kibibyte (KiB) is equal to 1,024 (210) bytes.
  - A mebibyte (MiB) is equal to 1,048,576 (220) bytes.
  - A gibibyte (GiB) is equal to 1,073,741,824 (230) bytes.

The importance of queue depth

The HBA queue depth (QD) affects the performance of the DE Series array. Queue depth is the number of I/O requests that the HBA can send/receive per LUN. A higher queue depth usually yields better performance. However, if the storage array's maximum queue depth is reached, the storage array rejects the I/O requests and the performance is degraded.

If a large number of HBAs are accessing a storage array, the QD of each HBA should be set carefully to avoid flooding the array. All HBAs in the same environment should have similar QDs.

For details on modifying the queue depth, please refer to the individual HBA vendor document.

ThinkSystem System Manager tuning

ThinkSystem System Manager provides several options to configure the whole array and the individual volume to improve performance for specific workloads. For information on the use of ThinkSystem System Manager, see the User's Guide, available from the Information Center:

https://thinksystem.lenovofiles.com/help/topic/thinksystem_storage_de_series/overview.html

General guidelines on Volume Groups (RAID)

RAID technology provides various levels of redundancy and performance. RAID 0 improves the volume performance, but no redundancy. RAID 10 gets the benefit of RAID 0 with data mirroring for redundancy.

RAID 5 uses striped volumes with parity distributed evenly across all drives. The data is protected when the single drive fails. The entire volume would be lost with the second drive failure during the rebuild of the first failed drive.

To minimize the risk during lengthy rebuild time, choosing RAID 6 over RAID 5 is safer if the individual drive capacity of the volume groups is larger than 2 TB. RAID 6 provides double parity protection instead of single, and the volume would survive with two drive failures.

For performance-sensitive scenarios, the recommended minimum and maximum numbers of drives are listed in Table 1.

Table 1 Recommended minimum and maximum drives for RAID and DDP

RAID Level or DDP	Recommended minimum drives[a]	Recommended maximum drives[a]
RAID 0	24	48
RAID 1 or 10	24	48
RAID 5	9	17
RAID 6	10	18
DDP	24	48

a. Thes recommended drive quantities are recommended in terms of performance tuning. They are not the system specification nor the limitation.

Dynamic Disk Pools (DDP)

Dynamic Disk Pools randomly spread data across all drives in the pool. Unlike traditional RAID volume groups, which suffer long rebuild time and degraded performance after drive failure, DDP dynamically rebalances the data across all drives. Each data stripe resides on 10 drives regardless of the pool size, and a different set of 10 drives is used for each stripe.

Figure 3 shows the segments of each stripe by the same color. Multiple stripes can be written concurrently.

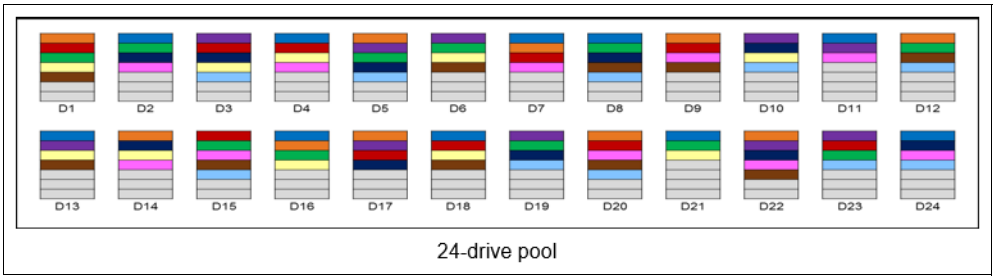

Figure 3 DDP with 24 drives

In the event of the drive failure, the segments from the failed drive would be rebalanced to other drives, Figure 4. The performance impact during the rebalance operation is less than during the drive rebuild in traditional RAID.

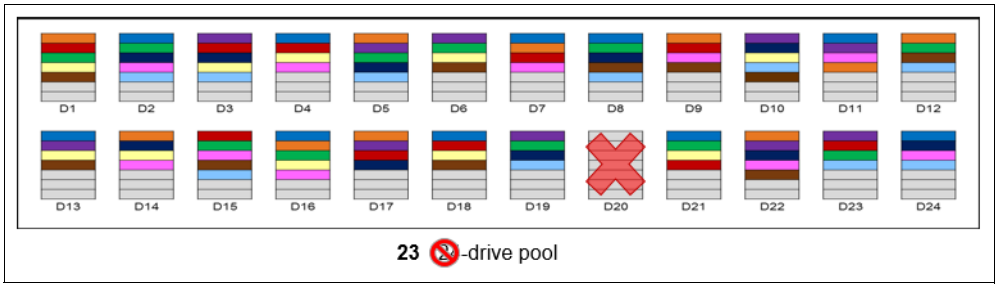

Figure 4 DDP with 23 drives and 1 failed drive

For performance-sensitive scenarios, the recommended minimum and maximum numbers of drives for DDP are listed in Table 1 on page 5.

Volume Groups and DDP comparison

Table 2 shows the summary of the differences between Volume Groups and DDP.

Table 2 Comparison between Volume Groups and DDP

Feature	Volume Groups	Dynamic Disk Pools
RAID selection	RAID 0, 1, 10, 5 or 6	RAID 6 only
Volume Segment Size selection	16, 32, 64, 128, 256 or 512 KiB	128 KiB fixed
Thin provisioning	No	Yes
Rebuild speed	Slow	Fast
Suited I/O type	Random or sequential	Random

Thick provisioning and thin provisioning

A thick volume is fully-provisioned, which means all of the capacity is allocated when the volume is created. The capacity of a thin volume is allocated as the data is being written to the volume. You can create thick volumes from a DDP or a volume group. You can create thin volumes only from a DDP.

Thin provisioning helps to avoid wasted allocated capacity and save on storage costs. Data archiving applications may benefit from deploying smaller physical drive capacity to begin with and adding more as the demand increases. Thin provisioning is not recommended for performance demanding situations. It could impact the I/O performance up to 75% comparing to thick provisioning due to the processing overhead.

DDP has the following restrictions that make thin volumes not tunable for performance:

- Volume Segment size change not available
- Dynamic Read Cache Prefetch disabled
- Write Caching with Mirroring is enabled and not changeable
- Pools are RAID 6 only

Cache block size

Cache block size is the maximum size of each cache block, which is an organizational unit for cache management. The cache block size is by default 8 KiB, but can be set to 4, 8, 16 or 32 KiB. Ideally the cache block size should be set to the predominant I/O size of your applications. All volumes share the same pool of cache on the controller, therefore the size is constant for all volumes.

All I/O of the system must pass through the cache, and the block size affects how many blocks are required for each I/O. If a server issues a 12 KiB I/O and the cache block size is set at 8 KiB, two blocks are required, and the second block has 4 KiB of waste space. File systems or database applications generally use smaller sizes, while a larger size is good for video surveillance or multimedia applications requiring large data transfer or sequential I/O.

To change the setting in ThinkSystem System Manager, go to **Settings** → **System** → **Additional Settings**, and click on **Change Cache Settings** to show Figure 5.

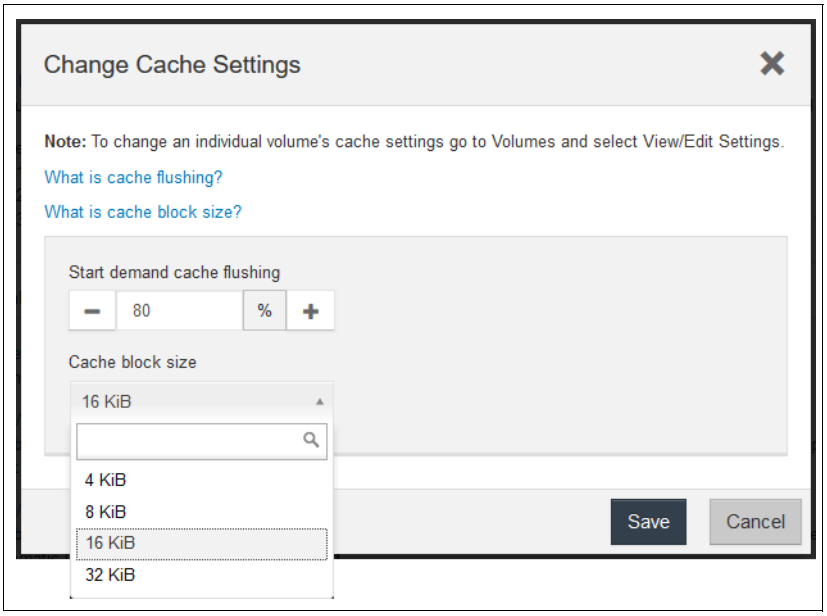

Figure 5 Changing the cache block size

Volume segment size

A segment is the amount of data in kibibytes (KiB) that is stored on a drive before the storage array moves to the next drive in the stripe (RAID group). The default segment size is 128 KiB,

and can be set to 16, 32, 64, 128, 256 or 512 KiB. Segment size applies only to volume groups, not Dynamic Disk Pools (DDP).

If an application typically uses small, random reads and writes (IOPS), a smaller segment size typically works better. If the application has large, sequential reads and writes (throughput), a large segment size is generally better.

In the traditional RAID volume, performance is maximized when a single data transfer request is served by a single data stripe, i.e., multiple drives are used for the same request, but each drive is accessed only once. The size of a data stripe is the segment size multiplied by the number of drives in the volume group used for data transfer, e.g. in Figure 6.

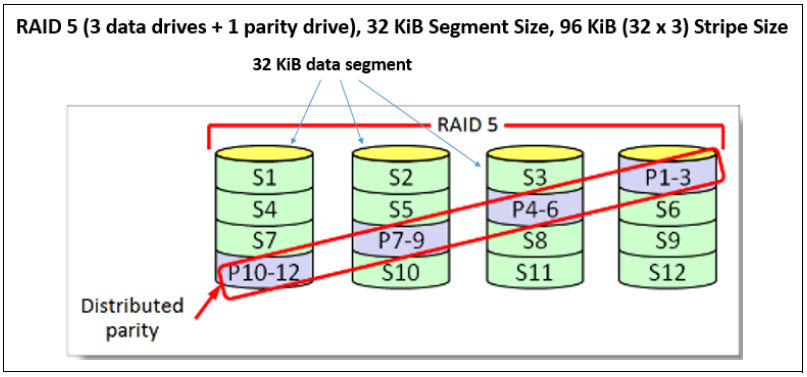

Figure 6 Segment size and data stripe size in RAID 5

Stripe alignment is required to avoid partial segment updates, which incur read modify write (RMW) penalties. This is especially important for SSD write operations in RAID 5 or RAID 6.

In other words, divide the application I/O size by the number of data drives to determine the optimal segment size. For common block sizes, aligning the I/O size with the segment size to consistently produce full stripe writes is possible when the number of data drives is a power of 2 (2, 4, 8, 16, etc.).

See “Dynamic Disk Pools (DDP)” on page 5 for the DDP stripe, which is different from traditional RAID stripe.

The segment size should be specified when creating the volumes. To change the setting dynamically, go to **Storage** → **Volumes**, select the volume to be changed then click **View/Edit Settings** button, **Advanced** tab, and select the value as shown in Figure 7.

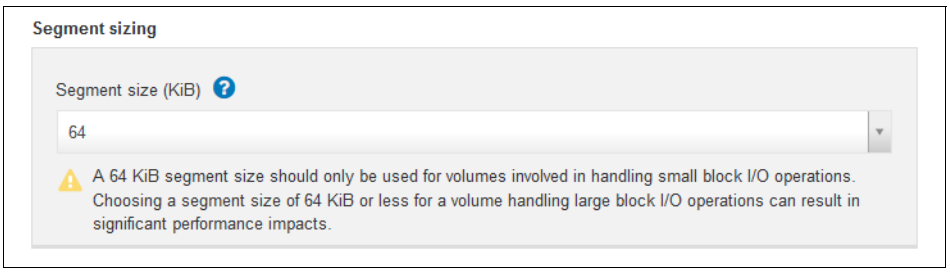

Figure 7 Changing the segment size dynamically

Dynamic read cache prefetch

Dynamic read cache prefetch allows the controller to copy additional sequential data blocks into the cache while it is reading data blocks from a drive to the cache. This caching increases

the chance that future requests for data can be filled from the cache. Dynamic read cache prefetch is important for multimedia applications that use sequential I/O. The rate and amount of data that is prefetched into cache is self-adjusting based on the rate and request size of the host reads. Random access does not cause data to be prefetched into cache. This feature does not apply when read caching is disabled.

Disabling the dynamic cache read prefetch option is recommended for highly random read workloads.

Read and write caching

Read cache is a buffer that stores data that has been read from the drives. The data for a read operation might already be in the cache from a previous operation, which eliminates the need to access the drives. The data stays in the read cache until it is flushed.

The Read Caching caches the current host requested data, while Dynamic Read Cache Prefetch caches the sequential data in advance without hosts' interventions.

Write cache is a buffer that stores data from the host that has not yet been written to the drives. The data stays in the write cache until it is written to the drives. Write caching can increase I/O performance.

Write cache mirroring

Write caching with mirroring occurs when the data written to the cache memory of one controller is also written to the cache memory of the other controller. Therefore, if one controller fails, the other can complete all outstanding write operations. Write cache mirroring is available only if write caching is enabled and two controllers are present. Write caching with mirroring is the default setting at volume creation.

Disabling write cache mirroring is not recommended. If one controller fails, the unflushed data would be lost as the other controller does not have the cached date. This would cause the corruption of the user data.

To change the above caching settings, go to **Storage** → **Volumes**, click on the **More** button, click **Change cache settings** and check/uncheck desired boxes on the **Basic** and **Advanced** tabs as shown in the following figures.

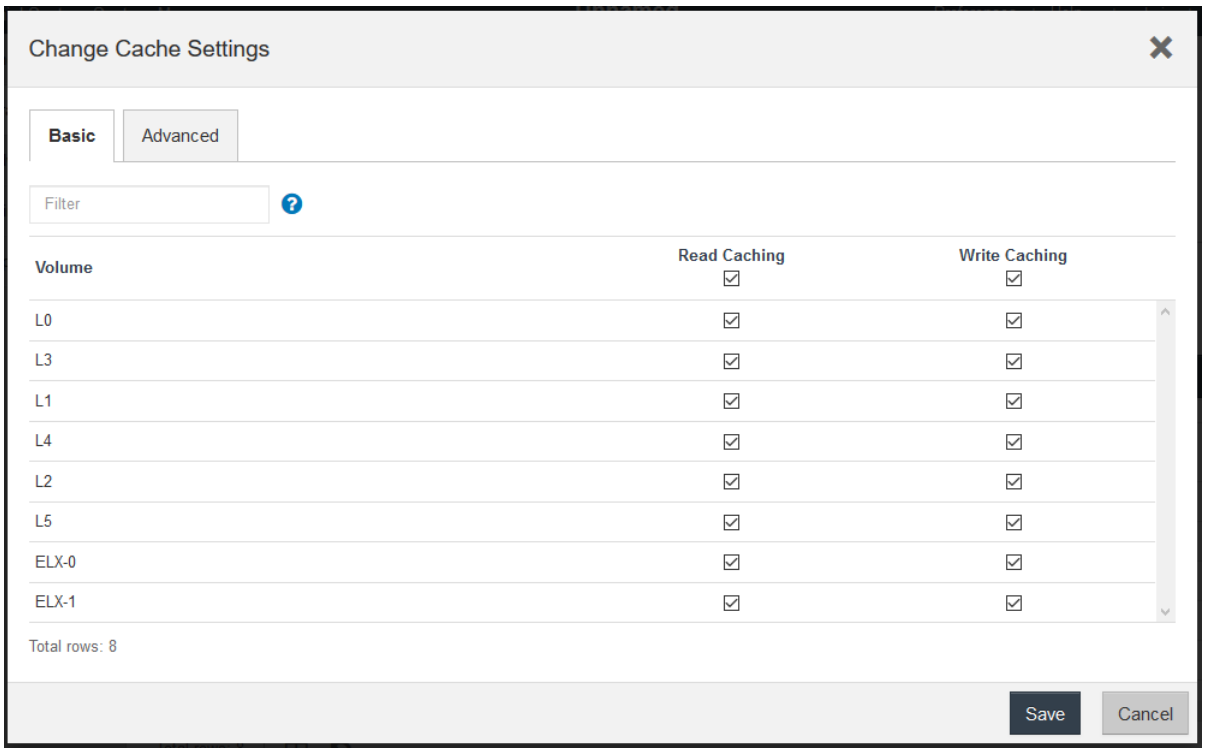

Figure 8 Change Cache Settings - Basic tab

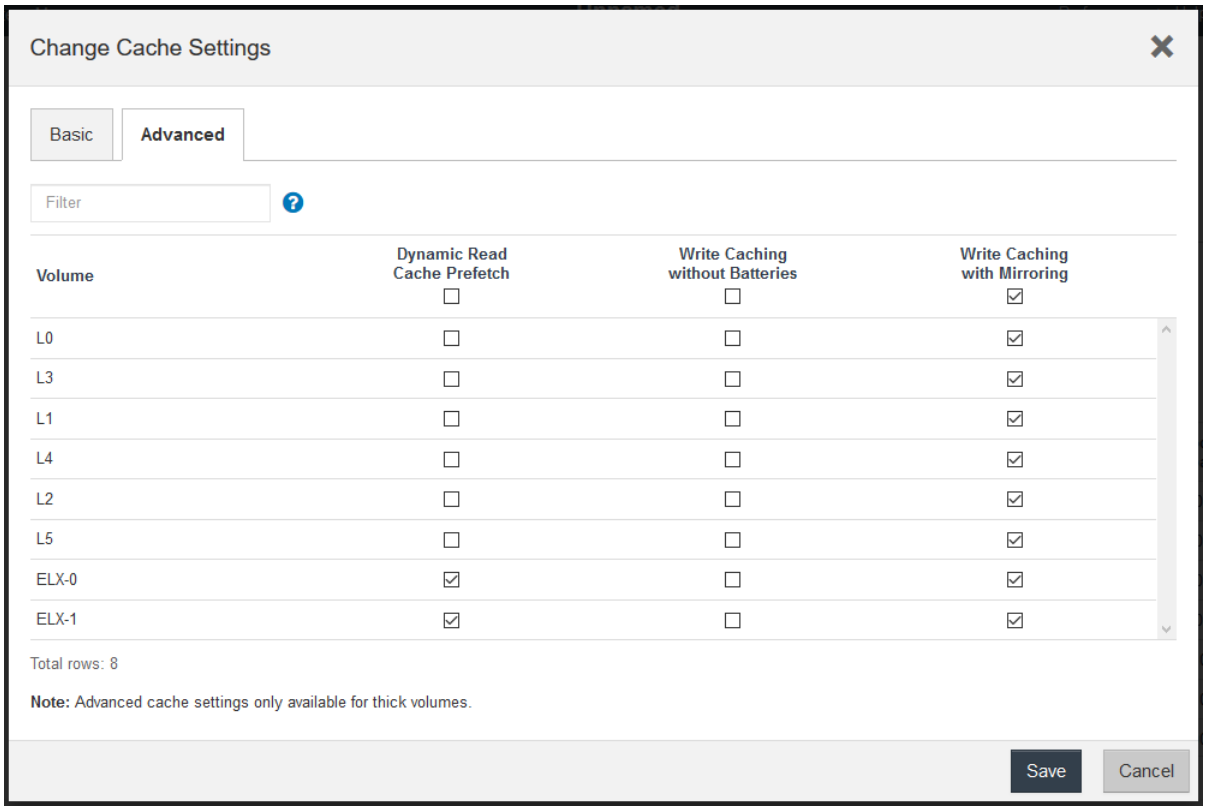

Figure 9 Change Cache Settings - Advanced tab

SSD read cache

The SSD read cache feature uses SSD storage to hold frequently accessed data from user volumes. It is intended to improve the performance of the situations where read performance is limited because of the use of HDDs.

Note: SSD read cache is only available on DE6000H, DE4000H and DE2000H.

SSD cache is a secondary cache that is used with the primary cache in the controller's memory (DRAM). The data is copied from user volumes and stored on two internal RAID 0 volumes (one per controller) that are automatically created when an SSD cache is created. The SSD cache is not the tiering implementation. Data in the SSD cache is not permanent and may be aged. The SSDs are used for accessing speed not for its non-volatility. The main data copy always comes from HDDs. As a result, these volumes are not accessible nor displayed in the user interface.

Following a host read or write, the SSD cache feature copies data from an HDD volume to the SSD volume. A subsequent host read of the same logical block addresses (LBAs) can be read directly from the SSD volume with lower response time than being reread from the HDD volume.

The operations of controller read cache and SSD cache are illustrated in Figure 10.

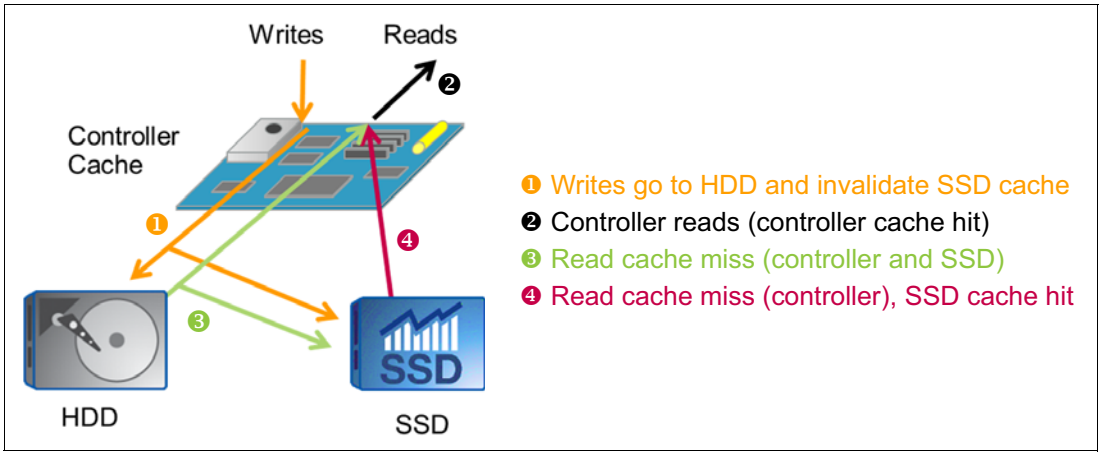

Figure 10 SSD read cache operations

We recommend using the SSD read cache if:

- Random read workloads >80% (performance could reduce with >15% writes or sequential workloads)
- Large number of reads that are repeat reads to the same or adjacent areas of the drives
- Application working size set is smaller than the SSD cache capacity
- The cost-effective approach that uses a mix of HDDs and an SSD cache is desired

Volume controller ownership

A volume will have a preferred controller that owns the volume. Plan the volume ownership distribution to balance I/O loading between controllers. Enable Automatic Load Balancing if the workloads tend to shift between controllers frequently. It is recommended to distribute the ownership of the volumes evenly between the two controllers.

To change the controller ownership, go to **Storage** → **Volumes**, click on the **More** button, click **Change ownership** and choose a Preferred Owner as shown in Figure 11.

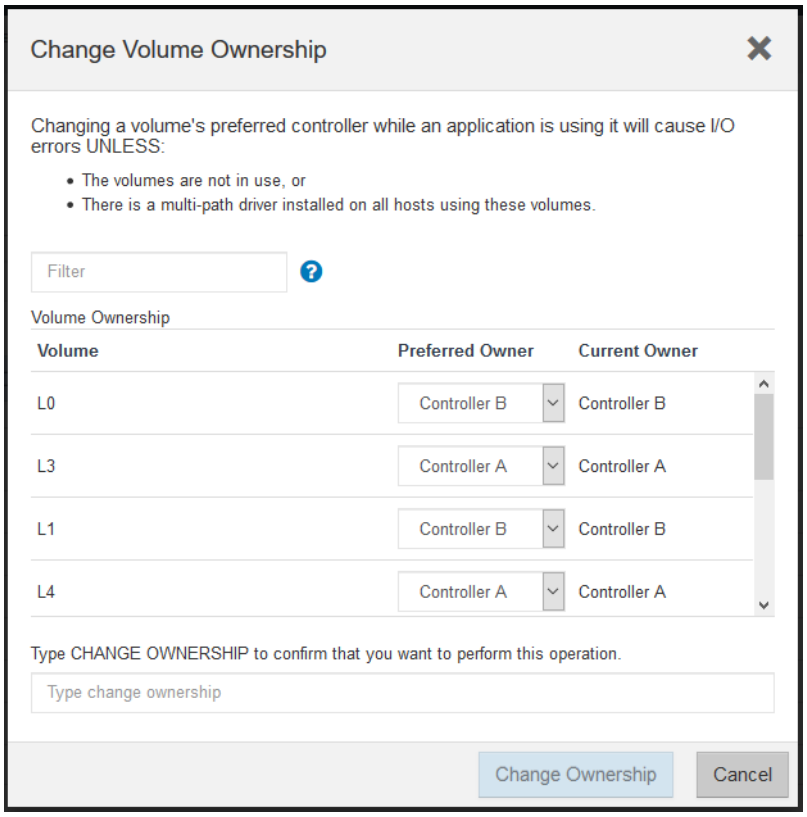

Figure 11 Change volume ownership

Automatic load balancing

This feature provides automated I/O balancing across controllers by reacting dynamically to load changes over time and automatically adjusting volume ownership to correct imbalance issue. If more than 75% of the volume traffic is shipped from the non-owning controller, the current ownership of that volume would be changed to the peer controller with the delayed decision window of 5 minutes.

To change the setting, go to **Settings** → **System** → **Additional Settings**, and click **Enable/Disable Automatic Load Balancing**. You will then be prompted as shown in Figure 12 on page 13.

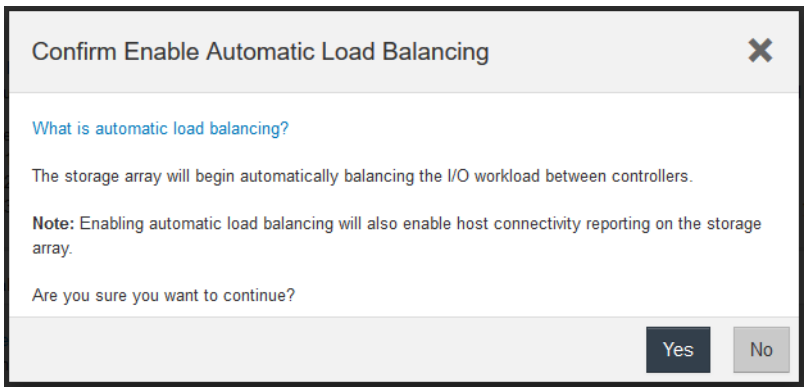

Figure 12 Confirm Enable Automatic Load Balancing

Multipath configuration in the OS

The multipath software provides the redundant path to the storage array in case one of the physical paths is disrupted. It presents a single logical or virtual device that the operating system can access by the physical paths to the storage. The multipath software also manages the failover process that up-dates the virtual device.

To configure multipath, see the multipath sections in the *ThinkSystem DE Series Hardware Installation and Maintenance Guide*, available from:

`https://download.lenovo.com/storage/thinksystem_storage_de_himg_en.pdf`

The next section will provide some guidelines of ThinkSystem System Manager settings for various applications.

Recommendations for specific workloads

You can specify the workload type during volume creation. For some applications, ThinkSystem System Manager configures the optimized volumes based on the type of application workloads. During volume creation, you specify the requirements and the optimized volumes will be created. You are able to edit the volumes as necessary.

You can create the recommended volumes with the following workload types:

- Microsoft SQL Server
- Microsoft Exchange Server
- Video Surveillance applications
- VMware ESXi

SPC-1 benchmark and general workloads

SPC-1 benchmark settings are listed here for reference. In general, the random I/O operations look for IOPS performance and the sequential I/O operations demand throughput. The random I/O workloads tend to have small I/O size, and a 16 KiB or smaller cache block size with 128 KiB volume segment size is common. On the other hand, consider a 32 KiB cache block size with 512 KiB volume segment size for the sequential I/O workloads.

Table 3 Recommended settings for SPC-1 and general workloads

Workload → Parameter	SPC-1	Highly random read	Database or File system	Multimedia application	Video Surveillance
Workload I/O Size (KiB)	8 ~ 128	4 ~ 16	16 ~ 32	64 ~ 128	>=256
Drive count	24[a]	>=24	12	>=24	30
Cache Block Size (KiB)	16	<=8	16	32	32
Volume Segment Size (KiB)	64	<=32	<=128	>=256	128
Dynamic Read Cache Prefetch	Disable	Disable	Enable	Enable	Enable
Read Caching	Enable	Enable	Enable	Enable	Enable
Write Caching	Enable	Enable	Enable	Enable	Enable
Write Cache Mirroring	Enable	Enable	Enable	Enable	Disable[b]
RAID level	RAID 10	RAID 10	RAID 10	RAID 6	RAID 6

a. Lenovo used 24 SSDs for SPC-1 submission.

b. Cache mirroring decreases the available cache by 50%, because I/O is mirrored on both controllers. It also incurs a performance penalty for the mirror operation. Because the failure of a controller incurs a loss of video recording for seconds regardless of the setting of cache mirroring, it is recommended to disable cache mirroring for video surveillance applications.

Oracle Database

The default block size setting for Oracle I/O is 8 KiB. When the data is accessed in stripes, one stripe would range from multiple 8 KiB blocks up to 64 MiB. Oracle demands the random read performance the most, followed by the random and sequential write performance.

Table 4 Recommended settings for Oracle Database

Workload	Oracle Database
Workload I/O Size (KiB)	8 ~ 64 MiB
Cache Block Size (KiB)	32
Volume Segment Size (KiB)	64 or 128
Dynamic Read Cache Prefetch	Disable
Read Caching	Enable
Write Caching	Enable
Write Cache Mirroring	Enable

Microsoft SQL Server

It is recommended to select Microsoft SQL Server workload for volume creation. ThinkSystem System Manager will create volumes that are optimized for this application-specific workload. You also have the freedom to edit, add or delete the system recommended volumes.

Table 5 Recommended settings for Microsoft SQL Server

Workload	Microsoft SQL Server
Workload I/O Size (KiB)	8 ~ 64
Cache Block Size (KiB)	8
Volume Segment Size (KiB)	128
Dynamic Read Cache Prefetch	Disable
Read Caching	Enable
Write Caching	Enable
Write Cache Mirroring	Enable
RAID level	RAID 10/DDP

VMware environment setup

If you experience slower than expected VMware performance, Lenovo recommends checking the settings of various components, including HBA, switch, ESXi, server BIOS and DE storage. See “Appendix: VMware environment setup” for more detail.

Introduction to NVMe protocol

Nonvolatile Memory Express (NVMe) technology is an industry standard for PCIe SSDs that achieves reduced latency through its software stack. Mainstream networking protocols Fibre Channel Protocol (FCP) and Ethernet both use the SCSI command set for the storage protocol, however the NVMe software stack is simplified and optimized comparing to SCSI. The NVMe command set takes fewer clock cycles per I/O.

In addition, SCSI puts I/O requests into a single queue, containing a maximum queue depth (QD) of 256 commands. When the I/O requests arrive, they must wait in line while other requests are completed. Solid-state drives would benefit from parallel command queues. NVMe uses PCIe bus, which supports 64K queues and each with a queue depth (QD) of 64K commands.

Although all applications benefit from low latency, the NVMe performance boost is especially valuable in enterprise database applications that are sensitive to latency, such as Microsoft SQL Server, SAP HANA and Oracle. NVMe accelerates many modern workloads, including artificial intelligence (AI), machine learning (ML)/deep learning (DL) and internet of things (IoT).

NVMe over Fabrics (NVMe-oF)

NVMe defines access protocols and architectures for connecting local non-volatile storage to servers. NVMe over Fabrics (NVMe-oF) increases the scalability of the NVMe interface and defines how NVMe uses existing transport technologies such as FC, RoCE (RDMA over Converged Ethernet), InfiniBand and iWARP. NVMe-oF transports the NVMe protocol over greater distances and enable the use of networking switches and routers. Lenovo storage systems such as the ThinkSystem DE6000F and DE6000H support NVMe/FC and NVMe/RoCE.

Transport Layers: Fibre Channel and RoCE

NVMe-oF brings the NVMe protocol to the SAN marketplace. Many enterprise SANs use FCP for the speed and robustness of Fibre Channel. NVMe/FC can use the same data network component that FC uses for SCSI access to storage. It can coexist on the same host and fabrics that are using FCP, enabling a seamless transition to the new technology.

ThinkSystem DE6000F and DE6000H support 25/40/100 Gb NVMe/RoCE host connectivity. NVMe/RoCE host interface card (HIC) is different from iSCSI HIC. However, both iSCSI and NVMe/RoCE can coexist on the same fabric on the host side.

Differences between NVMe/FC and FCP

NVMe/FC looks very much like FCP, which encapsulates SCSI commands inside FC frames. The reason both look similar is that NVMe/FC swaps out the SCSI commands for the streamlined NVMe command set. This simple replacement improves the throughput and latency.

NVMe adds some new names for some common structures. Table 6 maps some common structures that have different names than those used in FCP.

Table 6 FCP and NVMe/FC terms

Fibre Channel Protocol (FCP)	**NVMe/FC**
World-wide Port Name (WWPN)	NVMe Qualified Name (NQN)
LUN	Namespace
Igroup, LUN mapping, and LUN masking	Subsystem
Asymmetric Logical Unit Access (ALUA)	Asymmetric Namespace Access (ANA)

The NVMe/FC terms have the following meaning:

- An NVMe Qualified Name (NQN) identifies an endpoint and is similar to the World-wide Port Name (WWPN) in both format (domain registration date, domain registered, and a serial number).
- A namespace ID is an identifier used by a controller to provide access to a namespace. This is nearly equivalent to a logical unit number (LUN) in SCSI. The accessibility of a volume by a host is configured from the management interfaces, along with setting the namespace ID for that host or host group. As with SCSI, a logical volume can be mapped to only a single host group at a time, and a given host group cannot have any duplicate namespace IDs.
- A subsystem is analogous to an initiator group (igroup), and it is used to mask an initiator so that it can see and mount a LUN or namespace.
- Asymmetric Namespace Access (ANA) is a new protocol feature for monitoring and communicating path states to the host operating system's Multipath I/O (MPIO) or multipath stack, which uses information communicated through ANA to select and manage multiple paths between the initiator and target.

NVMe/RoCE and RoCE

RoCE is the RDMA-based protocol used to transport the SCSI command set. NVMe/RoCE uses the RDMA to transport NVMe command set. The namespace ID is equivalent to LUN in SCSI as discussed in the previous section.

Like other RDMA-based protocols, NVMe/RoCE communication between devices relies on the concept of a register queue pair (QP). A QP is the combination of a submission queue (SQ) and a completion queue (CQ). The host (initiator) places commands into an SQ that is read by the storage array (target). The target places completion information relating to a received command into the CQ associated with the SQ on which the command was received. For NVMe/RoCE, there must be a 1:1 relationship between submission queues and completion queues; that is, every SQ must have a single, unique CQ associated with only that SQ.

Multipathing and failover

ThinkSystem DE6000F and DE6000H support Asymmetric Namespace Access (ANA) as part of the NVMe-oF target. Like Asymmetric Logical Unit Access (ALUA), ANA uses both an initiator-side and target-side implementation for it to be able to provide all the path and path state information that the host-side multipathing implementation to work with a storage high availability (HA) multipathing software used with each OS stack.

ANA requires both the target and initiator to implement and support ANA to function. If either side is not available or implemented, ANA isn't able to function, and NVMe-oF will fall back to not supporting storage HA. In those circumstances, applications will have to support HA for redundancy.

NVMe-oF relies on the ANA protocol to provide multipathing and path management necessary for both path and target failover. The ANA protocol defines how the NVMe subsystem communicates path and subsystem errors back to the host so that the host can manage paths and failover from one path to another. ANA fills the same role in NVMe-oF that ALUA does for SCSI protocol.

For multipath management, ANA in NVMe-oF is similar to DM-Multipath in SCSI. Since NVMe-oF is under rapid development, the supported OS list would be updated frequently.

For the current supported OSs, see the Operating Systems section of the Lenovo Press product guide:

- ThinkSystem DE6000F All Flash Storage Array:

  https://lenovopress.com/lp0910#operating-systems
- ThinkSystem DE6000H Hybrid Storage Array:

  https://lenovopress.com/lp0883#operating-systems

ThinkSystem DE Series NVMe support

Some products on the market focus on adding NVMe-enabled drives to the back-end storage while keeping the SCSI-based host ports to the front end. ThinkSystem DE6000F and DE6000H have taken a different approach. With DE Series, the NVMe protocol is supported from the host to the front end of the storage array, while the back end is still with SCSI-based SAS drives.

The tuning considerations discussed in the rest of this document apply to the NVMe-oF environment. The front-end NVMe-oF implementation with back-end SCSI keeps the cost low and enhances performance with lower latency.

ThinkSystem DE Series Sizing Tool

The ThinkSystem DE Series Sizing Tool provides the storage capacity planning and the estimated performance under specified configurations. The tool is available to our partners and is accessed from the following location (Lenovo Partner Hub login required):

https://storagesizing.tools.lenovo.com/

Appendix: VMware environment setup

Lenovo servers use UEFI to adjust the power management policies. To improve VMware performance, processor power management has to be disabled. Update to the latest server firmware before you proceed to ensure you have the latest features.

To change power management policies on ThinkSystem and System x servers, do the following:

1. Turn on or reboot the server
2. Press F1 to enter System Setup when prompted during boot.
3. Select **System Settings** (select UEFI Setup first and then System Settings if in XClarity)
4. Select **Operating Modes**.
5. Change to **Maximum Performance** in Choose Operating Mode field
6. Change to **Custom Mode** in Choose Operating Mode field, and change the following options:
   - Memory Power Management = Disable
   - CPU P-state Control = None
   - C1 Enhanced Mode = Disable
   - Turbo Mode = Enable
   - C-States = Disable
   - Power/Performance Bias = Platform Controlled
   - Platform Controlled Type = Maximum Performance
   - MONITOR/MWAIT = Disable
7. Save the changed settings and reboot the system.

To change power management policies on ThinkServer systems:

1. Turn on the server
2. Press F1 when prompted to enter setup.
3. On the main setup screen, press the right arrow key to go to the **Advanced Settings** tab
4. Select **Advanced Power Settings**
5. Select **Performance Profile** and select **Custom**, and change the following options:
   - Enhanced Intel SpeedStep Technology = Enabled
   - Turbo Mode = Enabled
   - C1E Support = Disabled
   - Core C3 = Disabled

- Core C6 = Disabled
- CPU Performance and Energy Bias = Disabled
- Thermal Profile = Max Performance
- Memory Power Savings = Disabled

6. Press F10 to Save and Reset. Select **Yes** when prompted to confirm. The server will now reboot.

ESXi client

In some scenarios, it may be required the power management is done at VMware level as opposed to host UEFI. Apply all of the critical fixes for ESXi before you proceed.

To change power management policies, do the following:

1. Select the host from inventory, click the **Manage** tab and then **Settings** tab
2. On the left pane under Hardware, select **Power Management**
3. Click **Edit** on the top of the right pane. The Edit Power Policy Settings box appears
4. Choose **High performance** and click **OK** to save

See more details on power management policy at the following VMware web page:

https://docs.vmware.com/en/VMware-vSphere/6.7/com.vmware.vsphere.resmgmt.doc/GUID-4D1A6F4A-8C99-47C1-A8E6-EF3865603F5B.html

Host HBA

Make sure the latest HBA drivers are installed.

Network switch

Set the switch MTU to 9000 bytes to support jumbo frame payloads, if you are using iSCSI for your SAN attach method.

DE Series host interface card (HIC)

Try to use the HIC ports instead of the base ports when connecting to VMware hosts. The HIC ports generally perform better than the base ports.

Author

Vincent Kao is a Performance Engineer on the Lenovo Storage Development Team, based in Taipei. He is responsible for the performance analysis of RAID storage systems. Vincent earned a Master's Degree in Electrical Engineering from San Jose State University, CA and a Bachelor's Degree in Electrical Engineering from National Central University, Taiwan.

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

- Ted Vojnovich, CTO External Storage
- Yuwen Yang, Storage Development
- Shawn Andrews, Storage Development
- James Stewart, Storage Development
- David Watts, Lenovo Press

Change history

Changes to the April 2023 update:

- Added a note to Table 1 on page 5 that the drive quantities listed are recommended in terms of performance tuning. They are not the system specification nor the limitation.

Changes to the March 2021 update:

- Updates to “Thick provisioning and thin provisioning” on page 6

Changes to the November 2020 update:

- Added “Introduction to NVMe protocol” on page 15
- Added “NVMe over Fabrics (NVMe-oF)” on page 15

Changes to the November 2019 update:

- Added “VMware environment setup” on page 15
- Added “Appendix: VMware environment setup” on page 18

Changes in the October 2019 update:

- Added Table 1 on page 5, Recommended drive counts for RAID and DDP
- New section, “SSD read cache” on page 11
- New section, “Multipath configuration in the OS” on page 13
- Removed volume mapping section
- Other small changes throughout the document

Changes in the September 2019 update:

- Updated the link the the DE Series Sizing Tool, page 18

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 17, 2023.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/lp1220

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:

Microsoft, SQL Server, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

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