Drive and Interface Technologies for System x Servers

Last Update: December 2013

Describes the drive interfaces available in System x servers

Covers SAS, SATA, HDD, SSD, SED, Flash Storage Adapters

Describes technolgies including RAID, encryption, tiering, caching

Provides drive selection guidelines

Ilya Krutov

LENOVO PRESS

Abstract

The choice of drives for server internal storage is broad. Therefore, this requires you to consider several factors, including connectivity interfaces, such as Serial Advanced Technology Attachment (SATA) or serial-attached SCSI (SAS); the speed of these interfaces; types of drives: hard disks, solid-state storage, or hybrid drives; rotational speeds; choice of desktop versus nearline versus enterprise drives for hard disk drives (HDDs); enterprise versus enterprise value for solid-state drives (SSDs); form-factors; and so on. So how do you choose the most appropriate drive type?

This paper describes currently available internal and external direct-attach storage (DAS) interfaces and drive types that are available for IBM System x® servers and provides suggestions on how to choose the best options for business needs and application requirements.

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

Ensuring that business-critical data is available when needed is an ever-growing need in IT. Your systems must store massive amounts of data quickly and retrieve it efficiently. At the same time, choosing the best storage technology for application data can be a complex task, because you must ensure that business requirements are met, yet contain costs under tight budget limits. Storage performance capabilities need to match the processing capabilities of the server to ensure the most efficient use of system resources. There is no “one size fits all” approach possible because different applications have different storage data access patterns.

The choice of drives for server internal storage is broad. Therefore, this requires you to consider several factors, including connectivity interfaces, such as Serial Advanced Technology Attachment (SATA) or serial-attached SCSI (SAS); the speed of these interfaces; types of drives: hard disks, solid-state storage, or hybrid drives; rotational speeds; choice of desktop versus nearline versus enterprise drives for hard disk drives (HDDs); enterprise versus enterprise value for solid-state drives (SSDs); form-factors; and so on. So how do you choose the most appropriate drive type?

This paper describes currently available internal and external direct-attach storage (DAS) interfaces and drive types that are available for IBM System x servers and provides suggestions on how to choose the best options for business needs and application requirements.

Serial Advanced Technology Attachment (SATA)

The Serial Advanced Technology Attachment (SATA or Serial ATA) is a successor of the widely used Parallel Advanced Technology Attachment (PATA) or Enhanced Integrated Drive Electronics (EIDE) interface that is used to attach separate drive options, including hard disk drives (HDDs). SATA specifications are developed and maintained by Technical Committee T13, AT Attachment, which is part of International Committee for Information Technology Standards (INCITS).

The Serial ATA International Organization (SATA-IO) was established in 2004 to promote and advance SATA connectivity technology worldwide. The most recent SATA specification adopted by the industry is Revision 3, which features 6 Gbps SATA connectivity along with the proven and the widely used SATA Revision 1.0a with Serial ATA II extensions, which feature a 3 Gbps connection speed.

For more information, see the Serial ATA International Organization home page:

http://www.sata-io.org

SATA interface

The intended use of the SATA interface in IBM System x servers is to provide internal, low-cost, entry-level host connectivity for hard disk, optical, and tape drives. Typically, internal SATA connectivity topology is based on two types of devices: SATA initiators (SATA disk controllers) and SATA targets (drives). SATA initiators and targets use the ATA command set to communicate with each other.

Entry-level System x servers (such as single-socket x3100 M4, x3250 M4, and x3250 M5 and dual-socket x3530 M4 and x3630 M4) offer a SATA HDD interface to provide low-cost internal storage.

Figure 1 shows a typical internal SATA HDD connectivity topology.

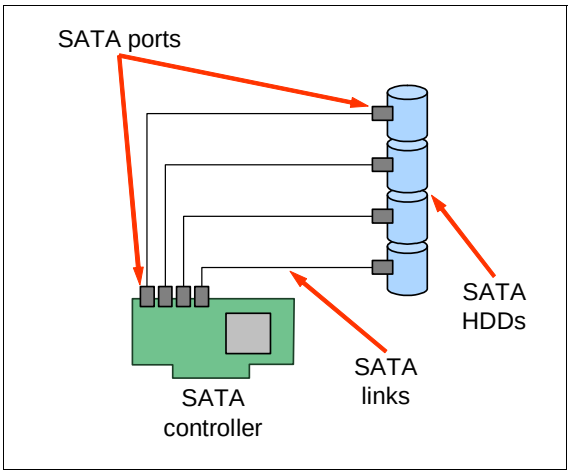

Figure 1 Entry-level IBM System x internal SATA HDD connectivity topology

SATA storage interfaces have the following characteristics:

- Serial point-to-point connection architecture
- Interface speeds of 1.5 Gbps, 3 Gbps, and 6 Gbps (150 MBps, 300 MBps, and 600 MBps of maximum theoretical throughput)
- Single-port device connections (the port for which consists of one pair of links for transmit and receive, with each link consisting of two physical wires that use low voltage differential signaling (LVDS)
- Support for narrow ports only (no wide port or port link aggregation, unlike SAS, as described in “Serial-attached SCSI (SAS)” on page 6
- Half-duplex port operation
- One target device per initiator port
- Support for port multipliers with command-based or Frame Information Structure (FIS) based switching to connect more than one target to the same initiator’s port
- Support for external SATA (eSATA) connectivity
- No multi-initiator support (no shared storage clustering possible)
- Cyclic redundancy check (CRC) for data integrity
- Support for Native Command Queuing (NCQ), which allows you to reorder disk data access command sequences. That optimizes seek time by minimizing physical movement of magnetic heads over the disk plate
- Support for SATA devices only (SAS devices are not supported)
- Support for hot-swap devices
- Support for SATA compatibility with an earlier version
- Seven-pin, L-shaped physical connector

SATA controllers

The IBM ServeRAID™ C100 and C105 are SATA controllers with software Redundant Array of Independent Disks (RAID) assistance capabilities that are built into the chipset on the system board. They are a cost-effective way to provide reliability, performance, and

fault-tolerant disk subsystem management to help safeguard your valuable data and enhance availability.

The ServeRAID C100 and C105 controllers have the following features:

- Auto-resume on array rebuild or array reconstruction after loss of system power

  Auto-resume uses nonvolatile RAM (NVRAM) to save rebuild progress during a host reboot or power failure to automatically resume from the last checkpoint. Auto-resume ensures that data integrity is maintained throughout the process.

- Fast initialization for quick array setup

  Fast initialization quickly writes zeros to the first and last sectors of the virtual drive. This function allows you to start writing data to the virtual drive while the initialization is running in the background.

- Consistency check for background data integrity

  This check verifies that all stripes in a virtual disk with a redundant RAID level are consistent. The consistency check mirrors data when an inconsistent stripe is detected for a RAID 1.

- Extensive online configuration options and advanced monitoring and event notification

  Management tools provide convenience for the configuration of logical volumes and alert you when errors occurred or are about to occur.

- Global hot spare support

  A hot spare rebuilds data from all virtual disks within the disk group in which it is configured. ServeRAID can define a physical disk as a hot spare to replace a failed drive. A global hot spare allows any physical drive to be designated as a hot spare for all drive groups that are defined on the controller.

- Human Interface Infrastructure (HII) Configuration Utility for pre-boot array configuration and management

  You can use the HII Configuration Utility to configure drive groups and logical drives before you install or boot the operating system.

- MegaRAID Storage Manager management software

  MegaRAID Storage Manager is an easy-to-use advanced RAID management application that is used across the entire family of ServeRAID controllers. It allows you to configure, monitor, and maintain drive groups, virtual drives, and advanced features with an intuitive GUI, which reduces administration tasks and simplifies troubleshooting.

Table 1 compares specifications of the ServeRAID C100 and C105 controllers for the SATA internal storage.

Table 1 ServeRAID C100 and C105 specification comparison.

Specification	**C100 (M4)**	**C100 (M5)**	**C105**
Interface type	SATA	SATA	SATA
Number of ports	Up to 6[a]	Up to 6[a]	Up to 8[b]
Port speed	3 Gbps	6 Gbps	3 Gbps
Number of physical drives supported[c]	Up to 4	Up to 4	Up to 8[b]
Number of virtual drives supported	8	8	8
Virtual drive size support	> 2 TB	> 2 TB	> 2 TB

Specification	C100 (M4)	C100 (M5)	C105
Stripe unit size	64 KB fixed	64 KB fixed	64 KB fixed
RAID levels	0, 1, 10	0, 1, 10, 5[d]	0, 1, 10
SATA HDD support	Yes	Yes	Yes
SAS HDD support	No	No	No
SSD support	No	No	No
Simple-swap HDD support	Yes	Yes	Yes
Hot-swap HDD support	No	No	Yes
Optical drive support	Yes	Yes	No
Tape drive support	Yes	Yes	No
Internal connector type	Up to six 7-pin L-shape SATA	Up to six 7-pin L-shape SATA	2x mini-SAS (SFF-8087) x4
Supported servers	x3100 M4, x3250 M4	x3250 M5	x3530 M4, x3630 M4

a. Up to four ports are used to connect hard disk drives, and up to two ports are used to connect optical or tape drives (server-dependent).
b. Eight HDD support requires the optional 8-pack enabler Features on Demand (FoD) upgrade, 90Y4349.
c. The maximum number of physical drives that are supported depends on the server model.
d. RAID 5 support requires the optional C100 Series RAID 5 FoD upgrade, 81Y4406.

For more information, see the Product Guide for ServeRAID C100 and C105 for System x:

http://lenovopress.com/tips0855

Serial-attached SCSI (SAS)

The serial-attached SCSI (SAS) connectivity technology is an evolution of the parallel SCSI Interface. It overcomes performance and scalability limitations of bus topology yet still provides enterprise-class reliability and software stack compatibility.

SAS specifications are developed and maintained by Technical Committee T10, SCSI Storage Interfaces, which are part of International Committee on Information Technology Standards (INCITS). The SCSI Trade Association (STA) was established in 1996 to promote and advance SCSI connectivity technology worldwide.

For more information about SCSI technology, see the website for the SCSI Trade Association:

http://www.scsita.org

SAS interface

SAS was introduced to the market in 2004 as a 3 Gbps connectivity technology (SAS-1.1). That evolved into Revision 2 (SAS-2) with support for 6 Gbps. The 6 Gbps SAS technology is the mainstream SAS technology that is used in most current servers. The most recent SAS specification adopted by the industry is Revision 3 (SAS-3), which features 12 Gbps SAS connectivity.

Because of its high performance, reliability, and scalability features, the SAS interface is used in the System x server systems for both internal and external storage connectivity for a wide range of applications and use patterns.

In general, there are three types of devices that form SAS topology:

- SAS initiators
- SAS or SATA targets
- SAS expanders

The *initiators* are the SAS controllers, that is, the SAS RAID controllers or SAS Host Bus Adapters (HBAs).

The *targets* are the end-point devices, such as disk or tape drives. SAS targets can be directly connected to the SAS initiator ports or indirectly through SAS expanders (or even a sequence of SAS expanders).

Essentially, a SAS *expander* is a switch device to connect more target devices to the initiator than the number of ports that the initiator has. That dramatically increases SAS fabric scalability without sacrificing reliability and performance. Expanders also support wide SAS links (or aggregated links) that consist of several narrow SAS links for expander-expander or expander-initiator connections to improve performance of the fabric.

A *narrow* SAS port is an interface that has one pair of transmit-receive links. This pair of transmit-receive links is commonly referred as a *PHY*, for the physical layer.

A *wide* SAS port has more than one pair of transmit-receive links (up to eight). Each wide port represents one aggregated link with a single worldwide name (WWN) address.

Each SAS device has a WWN address that is used to uniquely identify this device in the SAS fabric. SAS expanders maintain the WWN address routing tables to forward control and data traffic between the targets and initiators, which is similar to traditional network switch operations.

A SAS expander can be implemented as an integrated device on the disk backplane (as in many current storage-rich servers, including System x3500 M4, x3630 M4, x3650 M4, and x3650 M4 HD) or into the expansion enclosure (like in the System Storage DS2500 series external storage enclosures).

SAS specifications define three protocols that are used for communications between initiators, targets, and expanders:

- Serial SCSI Protocol (SSP), which is used to communicate between initiators and SAS target devices, such as hard disk drives
- Serial Management Protocol (SMP), which supports SAS expanders
- Serial ATA Tunneled Protocol (STP), which supports SATA targets in SAS fabric

The following figures show most common SAS connectivity topologies that are used in System x servers. In these figures, each physical SAS connector incorporates four SAS PHYs. Figure 2 shows how each SAS PHY can be a separate SAS port that is connected to an end-point device, such as a disk drive.

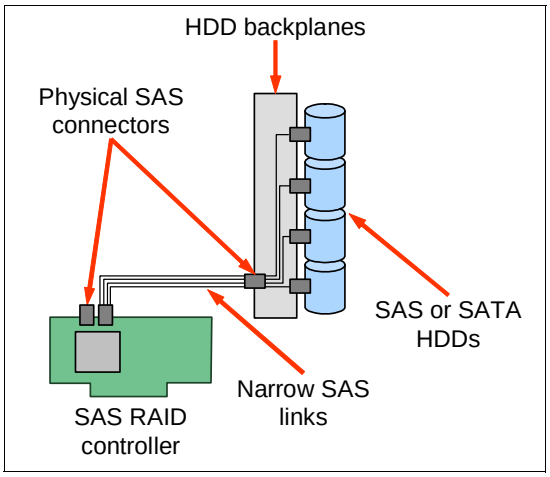

Figure 2 Typical internal SAS connectivity without expanders

Figure 3 shows how the four SAS PHYs can be combined into single x4 wide SAS port that is connected to the SAS expander.

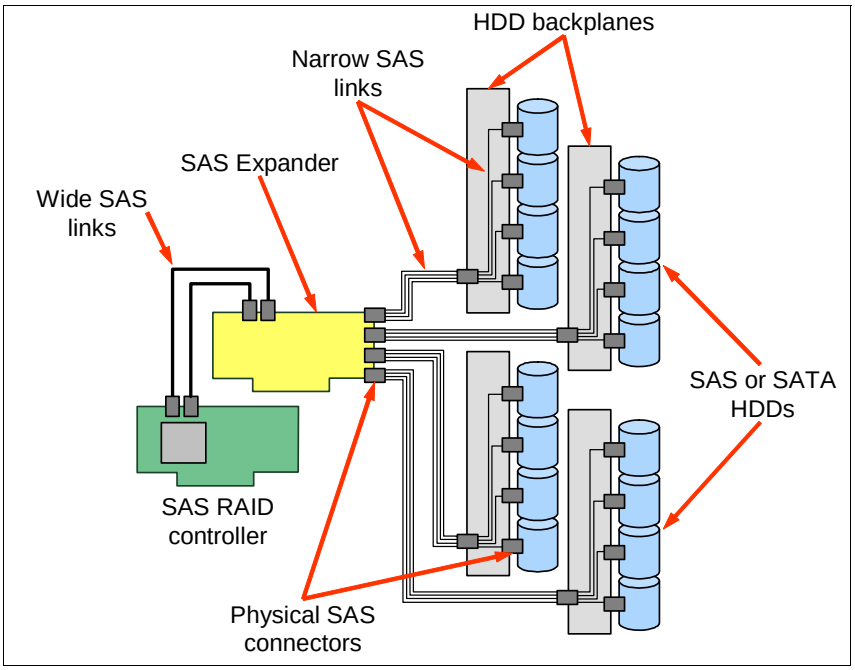

Figure 3 Typical internal SAS Connectivity with expanders

External SAS connectivity can be extremely scalable, because it allows SAS expanders to be connected to each other. For example, a single ServeRAID M5125 RAID Controller can handle up to 18 System Storage DS2512 expansion enclosures that serve up to 216 3.5-inch hard disk drives. Typical connectivity topology in that case is shown in Figure 4.

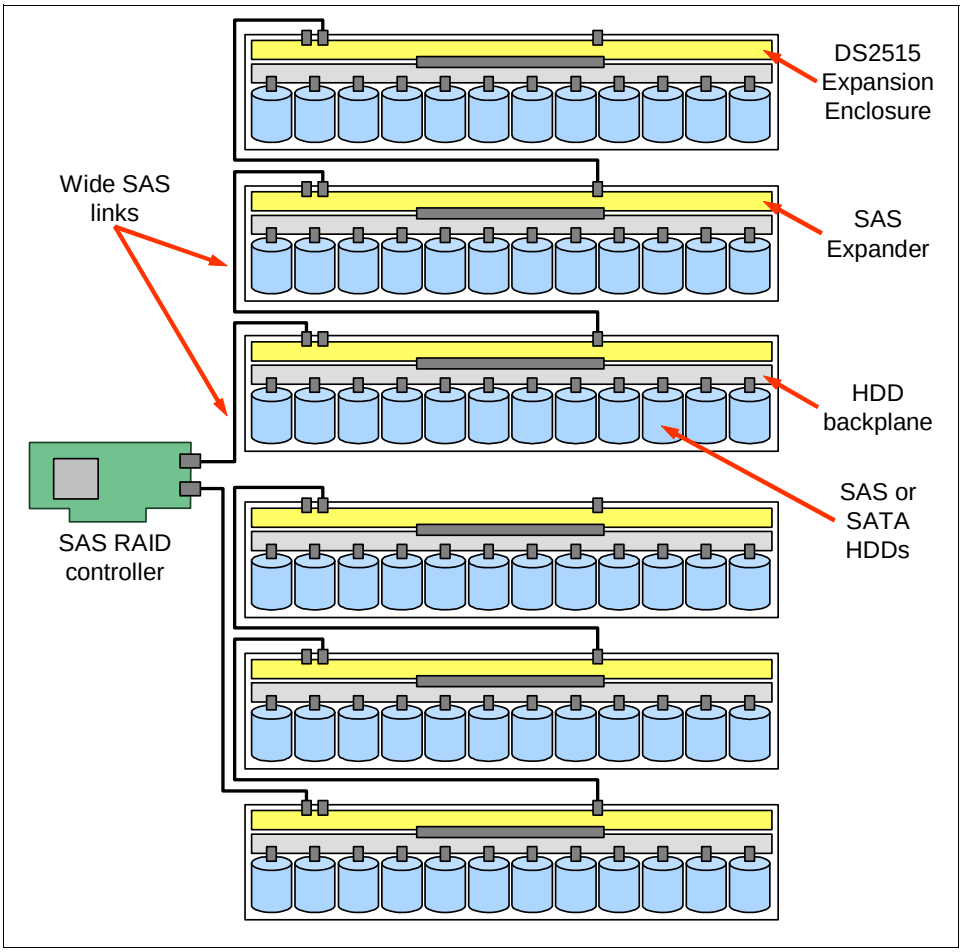

Figure 4 Typical external SAS connectivity with expanders

The SAS storage interface has the following characteristics:

- Serial point-to-point connection architecture
- Connection-oriented data transmission
- Interface speeds of 3 Gbps, 6 Gbps, and 12 Gbps (300 MBps, 600 MBps, and 1.2 GBps of maximum theoretical throughput)
- Dual-port device connections, with the port consisting of a pair of transmit-receive links and each link consisting of two physical wires that use Low Voltage Differential Signaling (LVDS)
- Support for narrow and wide (aggregated) ports, with aggregated wide port throughput up to 4800 MBps (four 12 Gbps PHYs or x4 link)
- Full-duplex port operation
- More than one target device per initiator's port with SAS expanders
- Support for internal and external connectivity
- Multi-initiator support
- Zoning support
- Cyclic redundancy check (CRC) for data integrity
- Enterprise-level error recovery
- Support for SAS and SATA devices

- Support for hot-swap devices
- Support for SAS and SATA compatibility with an earlier version
- Physical connectors: SFF-8087 (internal) and SFF-8088 (external)

SAS controllers

Lenovo offers a broad range of SAS host bus adapters (HBAs) and RAID controllers for System x servers, from entry-level to advanced configurations, supporting both internal and external storage attachments.

Table 2 shows the positioning of the currently available SAS controllers for System x.

Table 2 SAS controller positioning

	Internal connectivity			**External connectivity**	
SAS speed	Host bus adapters (non-RAID)	Basic RAID (cost-optimized)	Enterprise RAID (performance optimized)	Host bus adapters	Enterprise RAID (performance optimized)
HDD and SSD connectivity					
12 Gbps	► N2215		► M5210 ► M5210e		
6 Gbps	► N2115 ► IBM 6 Gb Performance Optimized HBA	► M1115 ► H1110 ► M1015	► M5110 ► M5110e ► M5016 ► M5015 ► M5014		► M5120
3 Gbps		► BR10il v2			► M5025
Tape connectivity					
6 Gbps	► IBM 6 Gb SAS HBA			► N2125 ► IBM 6 Gb SAS HBA	
External shared SAS storage connectivity					
6 Gbps				► N2125 ► IBM 6 Gb SAS HBA	

The ServeRAID M Series SAS/SATA controllers have the following features:

- Auto-resume on array rebuild or array reconstruction after the loss of system power

  Auto-resume uses non-volatile RAM (NVRAM) to save the rebuild progress during a host reboot or power failure to automatically resume from the last checkpoint. Auto-resume ensures that data integrity is maintained throughout the process. The card supports several features that can be implemented without rebooting the server. One benefit is that applications, such as email and web server, avoid downtime during the transition.

- Online Capacity Expansion

  Online Capacity Expansion (OCE) allows the capacity of a virtual disk to be expanded by adding new physical disks or by using unused space on existing disks, without requiring a reboot.

- Online RAID Level Migration

  Online RAID Level Migration, which is also known as *logical drive migration*, can migrate a virtual disk from any RAID level to any other RAID level without requiring a reboot. System availability and application functionality remain unaffected.

- Fast initialization for quick array setup

  Fast initialization quickly writes zeros to the first and last sectors of the virtual drive. This feature enables you to start writing data to the virtual drive while the initialization is running in the background.

- Consistency check for background data integrity

  This check verifies that all stripes in a virtual disk with a redundant RAID level are consistent. The consistency check mirrors data when an inconsistent stripe is detected for RAID 1 and re-creates the parity from the peer disks for RAID 5 or RAID 6. Consistency checks can be scheduled to take place periodically.

- Extensive online configuration options and advanced monitoring and event notification

  Management tools provide convenience for the configuration of logical volumes and alerts when errors occurred or are about to occur.

- Patrol read for media scanning and repairing

  Patrol read is a background sentry service that proactively discovers and corrects media defects (bad sectors) that arise normally as a disk drive ages. The service issues a series of verify commands. If a bad block is discovered, the card's firmware uses RAID algorithms to re-create the missing data and remap the sector to a good sector. The task is interruptible, based on controller activity and host operations. The firmware also provides an interface where the patrol read task can be initiated, set up for continuous operation, and terminated from a management application. Patrol read can be activated by a manual command or automatically.

- Global and dedicated hot spare with revertible hot spare support

  A *hot spare* rebuilds data from all virtual disks within the disk group in which it is configured. ServeRAID can define a physical disk as a hot spare to replace a failed drive. Hot spares can be configured as either global or dedicated. A *global* hot spare allows any physical drive to be designated as a hot spare. A *dedicated* hot spare allows the user to assign a hot spare drive to a particular array of the same drive type.

- WebBIOS configuration utility for pre-boot array configuration and management

  WebBIOS is a utility that is built into the ServeRAID controller. It allows you to configure drive groups and logical drives before you install or boot the operating system.

- MegaRAID Storage Manager management software

  MegaRAID Storage Manager is an easy-to-use advanced RAID management application that is used across the entire family of ServeRAID M controllers. It allows you to configure, monitor, and maintain drive groups, virtual drives, and advanced features with an intuitive GUI, so it reduces administrative tasks and simplifies troubleshooting.

The following features are optional and require the particular upgrade to be purchased (not all upgrades might be available for a specific controller, so be sure to check the specific product guide for details):

- MegaRAID SafeStore support for self-encrypting drive (SED) services

  MegaRAID SafeStore encryption services offer instant secure erase and local key management for self-encrypting drives. This technology represents a step forward in securing data on a disk drive from any unauthorized access or modification that results from theft, loss, or repurposing of drives.

  - Instant secure erase permanently removes data when repurposing or decommissioning SEDs.
  - SafeStore local key management provides the necessary management and protection of SEDs by using a simple pass phrase, security key identifier, and security key file that can be set and applied to all SEDs that are assigned to a ServeRAID adapter. This feature removes the complexity of managing each SED's unique encryption key, and it essentially relieves the administrator of most of the daily tasks of securing data. SafeStore is a part of any M Series RAID 5 upgrade that is available.

- MegaRAID flash cache protection

  MegaRAID flash cache protection uses NAND flash memory, which is powered by a supercapacitor, to protect data that is stored in the controller cache. This module eliminates the need for a lithium-ion battery, which is commonly used to protect DRAM cache memory on PCI RAID controllers. To avoid the possibility of data loss or corruption during a power or server failure, flash cache protection technology transfers the contents of the DRAM cache to NAND flash by using power from the offload power module. After the power is restored to the RAID controller, the content of the NAND flash is transferred back to the DRAM, which is flushed to the disk.

- MegaRAID FastPath SSD performance acceleration

  MegaRAID FastPath software provides high-performance I/O acceleration for SSD-based virtual drives by using a low latency I/O path to increase the maximum I/O per second (IOPS) capability of the controller. This feature boosts the performance of applications with a highly random data storage access pattern, such as transactional databases. The feature is activated by enabling the M Series Performance Accelerator.

- MegaRAID CacheCade SSD caching for traditional hard disk drives

  MegaRAID CacheCade read/write software accelerates the performance of hard disk drive (HDD) arrays with only an incremental investment in solid-state drive (SSD) technology. The software enables SSDs to be configured as a dedicated pool of controller cache to help maximize the I/O performance for transaction-intensive applications, such as databases and web services. CacheCade software tracks data storage access patterns and identifies the most frequently accessed data. The hot data is then automatically stored on the solid-state storage devices that are assigned as a dedicated cache pool on a ServeRAID controller that has the M Series SSD Caching feature enabled.

Table 3 on page 13 compares key specifications of the ServeRAID M Series controllers for the SAS/SATA internal storage.

Table 3 ServeRAID M Series specification comparison

Specification	M1015	M1115	M5014 or M5015	M5016	M5110 M5110e	M5210 M5210e
Interface type	SAS	SAS	SAS	SAS	SAS	SAS
Port speed	6 Gbps	6 Gbps	6 Gbps	6 Gbps	6 Gbps	12 Gbps
Number of ports	8	8	8	8	8	8
Host interface	PCIe 2.0 x8	PCIe 2.0 x8	PCIe 2.0 x8	PCIe 2.0 x8	PCIe 3.0 x8	PCIe 3.0 x8
SAS controller	LSI SAS2008	LSI SAS2008	LSI SAS2108	LSI SAS2208	LSI SAS2208	LSI SAS3108
Maximum cache, MB	No cache	No cache	256/512	1024	2048	2048
Battery backup	No	No	Optional/Yes	No	Optional	No
Flash backup	No	No	No	Yes	Optional	Optional
Maximum stripe unit size	64 KB fixed	64 KB fixed	Up to 1 MB	Up to 1 MB	Up to 1 MB	Up to 1 MB
RAID (standard)	0, 1, 10	0, 1, 10	0, 1, 10	0, 1, 10, 5, 50, 6, 60	0, 1, 10	0, 1, 10
RAID (optional)	5, 50	5, 50	5, 50, 6, 60	5, 50, 6, 60	5, 50, 6, 60	5, 50, 6, 60
HDD support	Yes	Yes	Yes	Yes	Yes	Yes
SED support	Optional	Optional	Optional	Yes	Optional	Optional
SSD support	Yes	Yes	Yes	Yes	Yes	Yes
FastPath	No	No	Optional	No	Optional	Optional
CacheCade	No	No	Optional	No	Optional	Optional
Tape drive support	No	No	No	No	No	No
Internal connector type	2x mini-SAS x4 SFF-8087	2x mini-SAS x4 SFF-8087	2x mini-SAS x4 SFF-8087	2x mini-SAS x4 SFF-8087	2x mini-SAS x4 SFF-8087	2x mini-SAS HD x4 SFF-8643

For more information, see the list of Product Guides in the RAID adapters category:

https://lenovopress.com/servers/options/raid

Hard disk drives

A typical disk subsystem consists of the physical hard disk and the disk controller. A disk is made up of multiple *platters* that are coated with a magnetic material to store data. The entire *platter assembly*, which is mounted on a *spindle*, revolves around the central axis. A *magnetic head assembly* that is mounted on an *arm* moves to and from (linear motion) the spindle to read the data that is stored on the magnetic coating of the platter.

The platters are divided into *tracks* and *sectors*, and the sector size represents a minimum amount of data that can be transferred to or from the disk drive during one I/O operation. During a disk I/O operation, the magnetic head is positioned over a track and waits until the addressed sector appears under it.

The linear movement of the head is referred to as the *seek*. The time that it takes to move to the exact track where the data is stored is called *seek time*. The rotational movement of the platter to the correct sector to present the data under the head is called *latency*. As a rule, higher drive rotation per minute (RPM) means lower access latency.

This section covers the following topics:

- "Advanced HDD format"
- "SATA HDDs" on page 17
- "SAS HDDs" on page 17
- "Nearline HDDs" on page 18

Advanced HDD format

Historically, an HDD was physically formatted by using the sector size of 512 bytes. However, the need for higher-capacity storage and better data integrity led to the changes in how the data is structured and stored on the disk platter. These changes resulted in the adoption of the *Advanced Format* by the industry.

Advanced Format introduces the sector size of 4,096 bytes (4 KB) and longer error-correcting code (ECC), which increases efficiency and integrity of storing the data, as shown in Figure 5.

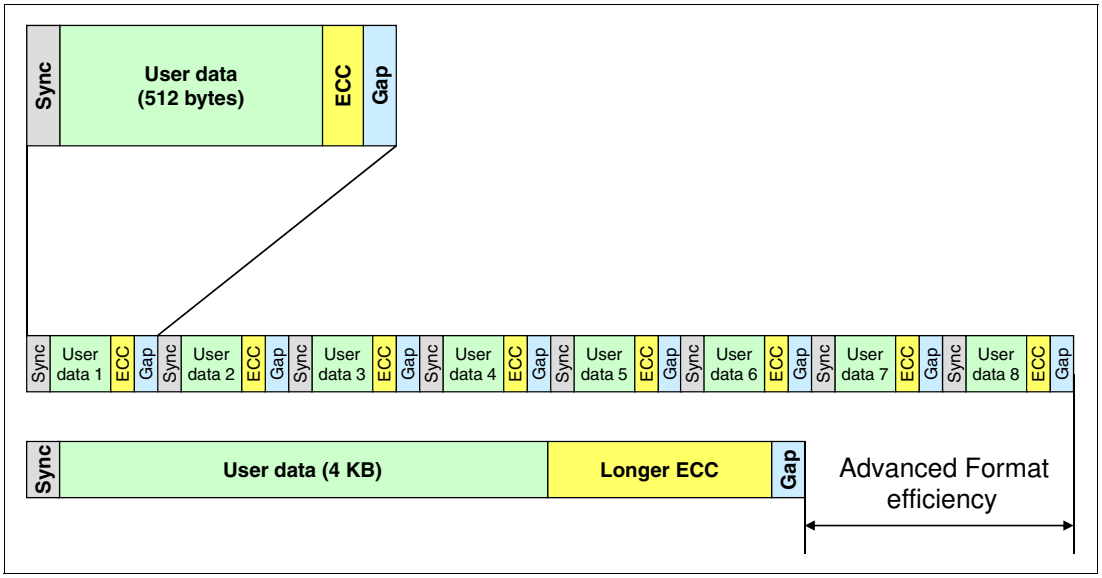

Figure 5 Advanced HDD format

From the data storage point of view, eight existing data sectors are now stored in a single 4 KB sector. This helps to eliminate multiple control fields, such as *Sync* and *Gap*, so it improves storage efficiency. In addition, larger and more powerful ECC algorithms can be used to improve data integrity.

There are two modes of operation for the advanced format HDDs:

- 4 KB native mode (4 Kn)

  In the 4 Kn mode, an HDD directly maps 4 KB logical blocks from the operating system to the 4 KB physical sectors.

- 512-byte emulation mode (512e)

  In 512e mode, an HDD transparently translates 512-byte logical block I/O requests into 4 KB physical sector operations, and each physical sector contains eight logical blocks.

Many modern operating systems already implement support for the advanced HDD format by allocating file system space in blocks (also known as *clusters*) of 4 KB. However, many existing hardware and software components are still designed around 512-byte sectors and expect the data be addressed, sent, and received by using the 512 byte I/O blocks. Advanced format HDDs support 512e mode to maintain compatibility with existing applications.

In 512e mode, the drive transparently maps logical 512-byte blocks to the 4 KB physical sectors, where each physical sector contains eight logical blocks. This approach is shown in Figure 6.

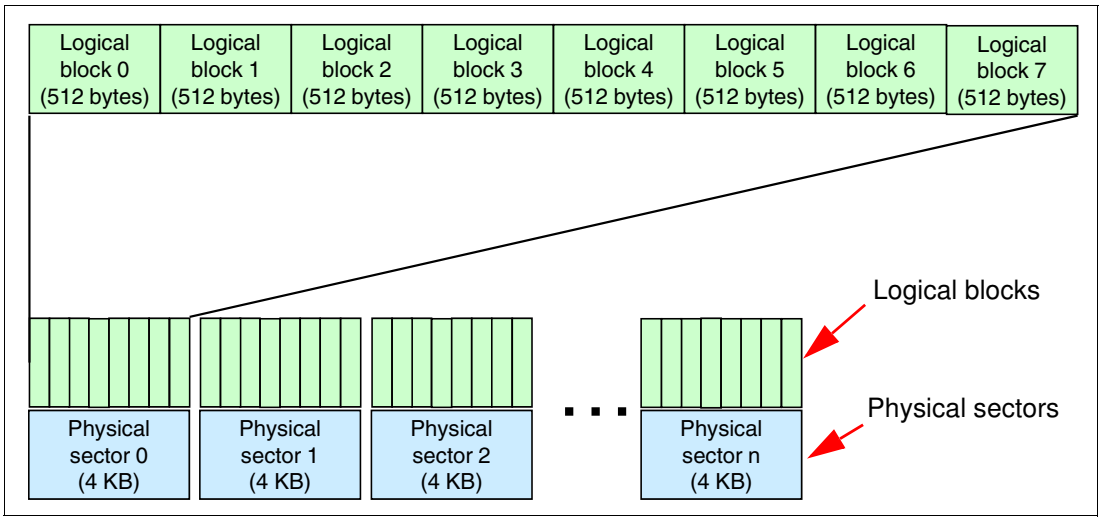

Figure 6 512-byte logical block mapping

With the Advanced Format HDDs, when the application issues a 512-byte READ operation, the HDD reads the entire 4 KB physical sector that contains this requested logical block, and then passes this 512-byte block to the application. When a 512-byte write operation is requested, a *read-modify-write* sequence is initiated. It includes three steps:

1. The entire 4 KB physical sector that contains the addressed 512-byte logical block is read from the HDD platter to the HDD buffer.
2. The HDD locates the 512-byte logical block that needs to be overwritten within the 4 KB block and overwrites it.
3. The entire 4 KB sector is written back to the HDD platter.

Because of this behavior of the 512e mode, certain performance considerations must be taken into account:

- Ideally, the operating system and applications interact with a disk drive by using the 4 KB blocks.
- It is optional but desirable for each 4 KB I/O operation to be *aligned* with the 4 KB physical sector. Therefore, it is better for each 4 KB read or write request to involve reading from or writing to only one physical sector.

Misalignment, where 4 KB operations are split across two 4 KB sectors, can have a significantly negative performance impact because each 4 KB I/O request forces the drive to manipulate with data from two physical sectors. Proper alignment is important for write operations, where each operation performs two I/O transfers (read and write).

The alignment concept is illustrated in Figure 7.

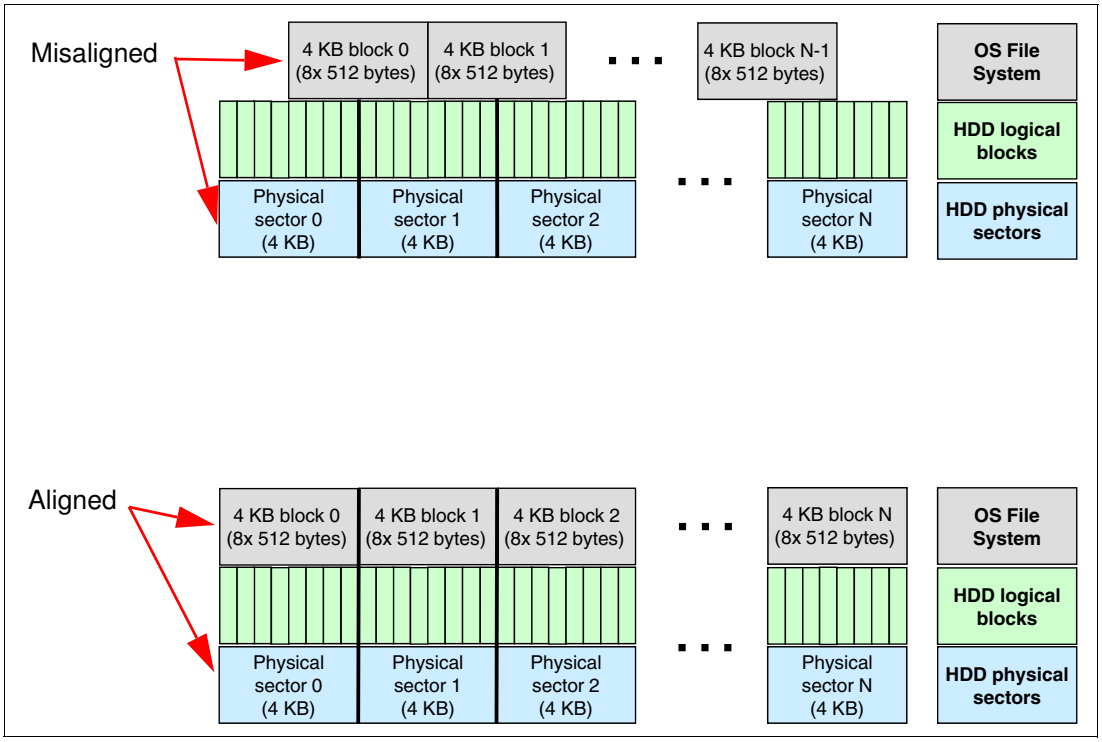

Figure 7 Physical sector and I/O block alignment

Logical partitions that are created by the operating system must start on the 4 KB physical sector boundary to achieve proper alignment. Many modern operating systems recognize the advanced format, and they automatically align their partitions to start on the physical sector boundary during installation. Table 4 on page 16 summarizes the information about whether the operating system supports 4 KB file system logical blocks and automatic partition alignment or not.

Table 4 Advanced format aware and unaware server operating systems

Server operating system	4 KB file system block	Automatic partition alignment
Microsoft Windows Server 2012	Yes	Yes
Microsoft Windows Server 2008	Yes	Yes
Microsoft Windows Server 2003	Yes	No
Red Hat Enterprise Linux 6	Yes	Yes
Red Hat Enterprise Linux 5	Yes	No[a]
SUSE Linux Enterprise Server 11	Yes	No[a]
SUSE Linux Enterprise Server 10	Yes	No[a]

a. You can use Linux Partitioning Utility to create partitions.

If the operating system does not support automatic partition alignment during installation, third-party disk partitioning and alignment tools can be used to align partitions.

SATA HDDs

Lenovo SATA HDDs have the following characteristics:

- Interface speeds of 3 Gbps and 6 Gbps (300 MBps and 600 MBps of maximum theoretical throughput)
- Rotational speed of 7,200 RPM
- Single hard disk drive capacities of 250 GB, 500 GB, 1 TB, 2 TB, 3 TB, or 4 TB
- Support for Native Command Queuing (NCQ)
- Support for Self-Monitoring, Analysis, and Reporting Technology (S.M.A.R.T.)
- Lower power consumption than SAS drives
- 2.5-inch and 3.5-inch form factor
- Simple-swap and hot-swap HDDs

Traditional or desktop SATA drives are intended for use in 8x5 hours of operation (8 hours per day, 5 days per week) and low I/O single user desktop workload environments, so they do not fit well into server environments. Because of this, System x servers always use Enterprise SATA drives (also known as *nearline* or *NL SATA* drives), which offer better reliability, better support of multi-drive RAID array deployments, and capacity for 24x7 hours of operation (24 hours per day, 7 days per week) in multi-user workload environments without significant costs, as compared to desktop SATA drives. See “Nearline HDDs” on page 18 for more information.

SATA HDDs provide the most reliable, energy-efficient, and cost-effective storage per Gb for lightly loaded departmental applications that store user data, such as file servers and email servers, or for infrequent bandwidth-intensive sequential workloads, such as archives, imaging, multimedia libraries, and disk backups.

SAS HDDs

SAS hard disk drives have the following characteristics:

- Interface speeds of 3 Gbps and 6 Gbps (300 MBps and 600 MBps of maximum theoretical throughput)
- Rotational speed of 10,000 RPM or 15,000 RPM
- Single hard disk drive capacities of 73 GB, 146 GB, 300 GB, 600 GB, 900 GB, or 1.2 TB
- Support for Tagged Command Queuing (TCQ), which you can use to reorder disk data access command sequences and to optimize seek time by minimizing physical movement of magnetics heads over disk plate
- Support for Predictive Failure Analysis (PFA)
- 2.5-inch and 3.5-inch form-factor
- Simple-swap and hot-swap HDDs

SAS HDDs are designed for 24x7 hours of operations (24 hours per day, 7 days per week) for continuous multi-user I/O-intensive workloads such as OLTP databases, data warehouses, and heavily loaded file servers and email servers. SAS connectivity provides high performance, high availability, and scalability for mission-critical enterprise applications and establishes a foundation for building multi-tiered storage architectures.

Nearline HDDs

Data stored on a nearline disk drive is accessed infrequently but requires 24x7 availability. From the application perspective, these drives are commonly used for storing archives, document images, multimedia libraries, and backups. There are two types of nearline drives:

- NL SATA, which is also commonly referred as *Enterprise SATA*
- NL SAS

NL SATA drives use native SATA interfaces and have the same capacity and performance characteristics as traditional or desktop SATA drives. However, NL SATA drives have better reliability, tolerance to vibration, and design for 24x7 hours of operation than traditional SATA drives. In general, reliability and tolerance to vibration of NL SATA drives are twice that of traditional SATA drives. Vibration tolerance is important for deployment of multi-drive arrays, such as RAID arrays, to eliminate risk of read/write errors and retry cycles because of HDD rotational vibration interference.

NL SAS has the same performance, capacity, and reliability characteristics as NL SATA drives. The only difference is that NL SAS drives provide native SAS interface capabilities, including dual-port connectivity, full-duplex data transfer, data integrity, and SCSI command support.

Flash internal storage

Solid-state drives (SSDs) use nonvolatile flash memory rather than spinning magnetic media to store data. They provide outstanding performance, endurance, reliability, and energy efficiency for random, storage I/O-intensive enterprise workloads, such as databases, data warehouses, corporate email and collaboration, actively connected users, caching and tiering applications, and others.

This section includes the following topics:

For additional information about Flash internal storage offerings, see the list of Product Guides in the PCIe SSD Adapters category:

https://lenovopress.com/servers/options/ssdadapter

SSD technology

SSDs differ from traditional HDDs in many ways, but there is one key difference: no moving parts. Where HDDs contain spinning disks and movable heads that read and write data on the disks, SSDs use solid-state (chip-based) memory to store data. This difference provides SSDs with the following advantages over HDDs:

- High-performance input/output operations per second (IOPS): Significant increases in performance of storage I/O subsystem
- Durability: Less susceptible to physical shock and vibration
- Longer lifespan: SSDs are not susceptible to mechanical wear
- Lower power consumption: As little as a few watts of power per drive
- Quieter and cooler running capabilities: Less floor space and lower energy costs
- Lower access times and latency rates: Up to 100 times faster than the spinning HDDs

SSDs use NAND-based nonvolatile flash memory, which is the same technology that is used by USB storage devices, memory cards, mobile phones, and other portable electronic devices that require data storage. However, the type of NAND flash memory that an SSD employs for data storage and retrieval is a key factor for determining the appropriate environment for which the device is used. Where one methodology might be adequate for the type of use and environment that the device is intended for (such as a notebook model that is designed for the consumer market), it might not be feasible for enterprise-class markets where high-performance standards and reliability are key factors for data storage.

Two methods currently exist for facilitating NAND flash memory:

- Single-level cell (SLC)
- Multi-level cell (MLC)

SLC flash memory stores data in arrays of floating-gate transistors, or cells, with 1 bit of data to each cell. This single-bit-per-cell methodology results in faster transfer speeds, higher reliability, and lower power consumption than HDDs. SLC SSDs are two to three times more expensive to manufacture than MLC devices.

The basic difference between SLC flash memory and MLC flash memory technologies is storage density. Unlike SLC flash memory, which allows only two states (one bit of data, either 0 or 1) to be stored in each cell, MLC flash memory can store up to four states (two bits of data: 00, 01, 10, or 11) per cell, so MLC stores twice as much data.

MLC flash memory can be further delineated into two categories:

- Commercial-grade MLC (cMLC or simply MLC): Used in consumer (single user) devices, such as notebooks, USB storage devices, memory cards, mobile phones, and so on
- Enterprise-grade MLC (eMLC): Designed specifically for use in commercial (multiple-user) enterprise environments

Both MLC and eMLC flash memory have the advantage of higher data density and the resultant lower cost-per-bit ratio than SLC flash memory. At the same time, the high-density storage model that is employed by both technologies results in lower write endurance ratios and higher rates of cell degradation than SLC flash memory, which greatly reduces the lifetime of the device.

For the traditional cMLC devices, this does not pose problems, because the lifetime expectancies are considered adequate for consumer-grade devices. This makes cMLC flash

memory ideal for lower-cost, consumer-targeted devices, such as memory cards and mobile devices, where cost and market factors outweigh performance and durability.

However, for enterprise environments where performance and durability are requirements, eMLC or MLC, combined with the advanced NAND flash management techniques, provides longer endurance through trimming of components and optimizing certain parameters in the firmware. In addition, MLC SSDs employ over-provisioning of data storage capacity and wear-leveling algorithms that evenly distribute data when the drives are not being heavily used. This results in a significant increase in write cycles and reduces concerns about cell degradation.

Currently, available SSDs are based on SLC, eMLC, or MLC with the advanced NAND flash management techniques, because they provide cost-effective IOPS performance and required endurance levels.

SSD-based offerings

The SSD-based offerings include High IOPS SSD PCIe Adapters (also known as PCIe SSD adapters), SSDs in a conventional 2.5-inch server drive bay, and 1.8-inch SSDs in the eXFlash™ unit.

High IOPS MLC adapters use the latest enterprise-level solid-state storage technologies in a standard PCIe form factor. They include sophisticated advanced features to optimize flash memory and deliver consistently high levels of performance, endurance, and reliability.

Enterprise SSDs for System x are flexible in delivering outstanding performance, reliability, and endurance at an affordable cost across a wide variety of enterprise workloads.

eXFlash technology is a server-based, high-performance internal storage solution that is based on SSDs and performance-optimized disk controllers (both Redundant Array of Independent Disks, or RAID, and non-RAID).

A single eXFlash unit accommodates up to eight hot-swap SSDs, and can be connected to up to two performance-optimized controllers. eXFlash is supported on System x3690 X5™, x3850 X5, x3950 X5, x3750 M4, and x3650 M4 servers.

Table 5 on page 20 compares features of these offerings.

Maximums: The maximums that are listed in Table 5 are independent from each other. That is, if the maximum capacity row shows up to 3.2 TB and the maximum read IOPS row shows up to 285,000 IOPS, that does not necessarily mean that the device that offers 3.2 TB supports 285,000 IOPS, or vice versa. Check the respective product guide for details.

Table 5 flash storage offerings

Feature	2.5-inch SSDs	PCIe SSD adapters	eXFlash SSDs
Form factor	2.5-inch drive	PCIe adapter	1.8-inch drive
Interface	6 Gbps SAS or SATA	PCIe 2.0 x8	6 Gbps SATA
Capacity per unit	Up to 1.6 TB	Up to 3.2 TB	Up to 400 GB
Maximum random read IOPS	Up to 100,000	Up to 285,000	Up to 75,000
Maximum sequential read rate	Up to 520 MBps	Up to 3.0 GBps	Up to 500 MBps

Feature	2.5-inch SSDs	PCIe SSD adapters	eXFlash SSDs
Write latency	Less than 100 µs	15 µs	65 µs
Hot-swap capabilities	Yes	No	Yes
RAID support	Yes	Chip-level redundancy	Yes

As a general rule, 2.5-inch SSDs can be a cost-effective solution that is optimized for commodity servers with a conventional HDD drive tray, with moderate storage IOPS and bandwidth performance requirements.

Both PCIe SSD adapters and eXFlash SSDs are optimized for the storage I/O intensive enterprise workloads, although PCIe adapters offer significantly lower write latency and the eXFlash SSDs offer better IOPS density and use the convenience of traditional hot-swap drives with hardware RAID protection. However, PCIe SSD adapters do not support hot-swap capabilities, and they can use the operating system's software RAID capabilities to offer data protection if required.

Solid-state drives

Solid-state drives are flash memory devices that are attached to the system through the traditional storage interface, such as SAS or SATA. The SSDs are typically installed in a conventional 2.5-inch or 3.5-inch drive bay of a supported server. There are 1.8-inch SSDs installed in specialized bays of selected systems or in eXFlash units (see "eXFlash" on page 23).

There are two types of SSDs:

- Enterprise SSDs
- Enterprise Value SSDs

Although both Enterprise SSDs and Enterprise Value SSDs typically have similar read IOPS characteristics, the key difference between them is their write IOPS performance and their endurance (or lifetime expectancy).

An SSD has a huge but finite number of program/erase (P/E) cycles. That affects how long it can perform write operations and, therefore, its life expectancy. An enterprise SSD has significantly better endurance but a higher cost/IOPS ratio than an Enterprise Value SSD. SSD write endurance is typically measured by the number of program/erase cycles that the drive can incur over its lifetime, which is listed as TBW (total bytes written) in the device specification.

Because of this behavior of Enterprise Value SSDs, careful planning is required to use them only in read-intensive environments to ensure that the TBW of the drive is not exceeded before the end of the expected lifespan. In other words, for an optimal cost:IOPS ratio, use Enterprise SSDs for write-intensive workloads and Enterprise Value SSDs for read-intensive workloads, as shown in Table 6.

Table 6 Application workload by SSD type

SSD type → ↓ Application type	Enterprise	Enterprise Value
OLTP database	Yes	
Data warehouse		Yes
Email server	Yes	
Medical imaging		Yes
Video on demand		Yes
Web, Internet		Yes
Web 2.0	Yes	

Table 7 summarizes key specifications of the available SSDs for Systems products.

Table 7 SSD specification comparison

Specification	IBM SATA Enterprise SSDs	IBM SAS EnterpriseSSDs	IBM SATA Enterprise Value SSDs	S3700 SATA Enterprise SSDs	S3500 SATA Enterprise Value SSDs
SSD class	Enterprise	Enterprise	Enterprise Value	Enterprise	Enterprise Value
NAND technology	eMLC	MLC with FlashGuard	MLC	MLC with HET[a]	MLC
Interface	6 Gbps SATA	6 Gbps SAS	6 Gbps SATA	6 Gbps SATA	6 Gbps SATA
Form factor	1.8";2.5"	2.5";3.5"	1.8";2.5";3.5"	1.8";2.5"	1.8";2.5"
Maximum capacity	400 GB	1.6 GB	512 GB	800 GB	800 GB
Maximum endurance	6 PB TBW	29.2 PB TBW	350 TB TBW	14.6 PB TBW	450 TB TBW
Maximum IOPS reads[b]	60,000	100,000	50,000	75,000	75,000
Maximum IOPS writes[b]	40,000	50,000	7,500	36,000	11,500
Maximum read rate[c]	520 MBps	500 MBps	350 MBps	500 MBps	500 MBps
Maximum write rate[c]	500 MBps	500 MBps	140 MBps	460 MBps	450 MBps
Typical power	2.8 - 4.6 W	7 W	2.5 - 3.5 W	6 W	5 W

a. HET = High Endurance Technology
b. 4 KB blocks random access.
c. 128 KB blocks sequential transfer.

PCIe SSD adapters

The PCIe SSD adapters use flash memory as their storage medium, and they are built as block devices on a PCIe bus. They feature advanced wear-leveling, advanced ECC, and chip-level redundancy, providing exceptional reliability and efficiency.

Table 8 summarizes key specifications of the PCIe SSD adapters for System x.

Table 8 PCIe SSD adapter specification comparison

Specification	IBM High IOPS MLC adapters	IBM High IOPS Modular Adapters		IBM Flash Adapters Enterprise Value
NAND technology	MLC	eMLC	SLC	MLC
Interface	PCIe 2.0 x8(x4 or x8-wired)	PCIe 2.0 x8 (x8-wired)	PCIe 2.0 x8 (x8-wired)	PCIe 2.0 x4 or x8 (x4-wired)
Maximum capacity	2.4 TB	800 GB	300 GB	3.2 TB
Maximum endurance	34 PB TBW	30.7 PB TBW	76.8 PB TBW	20 PB TBW
Maximum IOPS reads[a]	285,000 (512 bytes)	218,000 (4 KB)	179,000 (4 KB)	115,000 (512 bytes)
Maximum IOPS writes[b]	725,000 (512 bytes)	75,000 (4 KB)	100,000 (4 KB)	535,000 (512 bytes)
Maximum read rate[b]	3.0 GBps (1 MB)	2.0 GBps (256 KB)	1.5 GBps (256 KB)	1.5 GBps (1 MB)
Maximum write rate[c]	2.5 GBps (1 MB)	1.0 GBps (256 KB)	850 MBps (256 KB)	1.3 GBps (1 MB)

a. Random access, block size is in brackets.
b. Sequential transfer, block size is in brackets.

eXFlash

eXFlash technology is a server-based, high-performance internal storage solution. It is based on SSDs and performance-optimized disk controllers (both RAID and non-RAID).

A single eXFlash unit accommodates up to eight hot-swap SSDs and can be connected to up to two performance-optimized controllers. eXFlash is supported on System x3690 X5, x3850 X5, x3950 X5, x3750 M4, and x3650 M4 servers. Figure 8 shows an eXFlash unit with the status lights assembly on the left side.

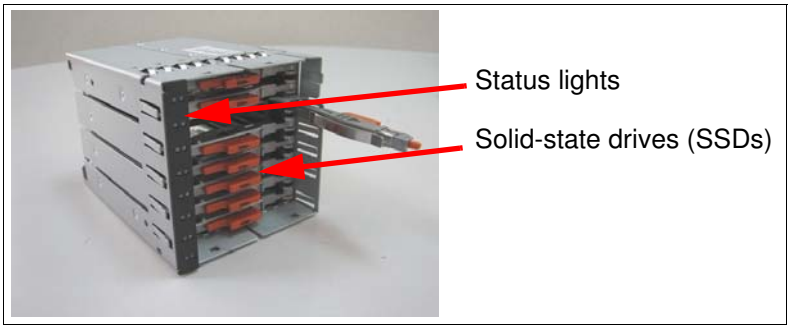

Figure 8 eXFlash unit

Each eXFlash unit can accommodate up to eight 1.8-inch hot-swap, front-accessible SSDs, and it occupies four 2.5-inch SAS hard disk drive bays. You can install the following number of eXFlash units:

- The x3850 X5 can have up to sixteen 1.8-inch SSDs with up to two eXFlash units (up to eight SSDs per eXFlash unit).
- The x3690 X5 can have up to twenty-four 1.8-inch SSDs with up to three eXFlash units (up to eight SSDs per eXFlash unit).
- The x3750 M4 and x3650 M4 can have up to thirty-two 1.8-inch SSDs with up to four eXFlash units (up to eight SSDs per eXFlash unit).

Theoretically, the I/O performance of the single eXFlash unit might be equivalent to the storage system that consists of more than 1000 traditional spinning HDDs. Besides HDDs, building such a massive I/O-intensive, high-performance storage system requires external deployment with many more infrastructure components, including host bus adapters (HBAs), switches, storage controllers, disk expansion enclosures, and cables. This leads to more capital expenses, floor space, electrical power requirements, and operation and support costs. Because eXFlash is based on internal server storage, it does not require all of the components describe previously, and it helps to eliminate extra expenses and environmental requirements.

In summary, the eXFlash solution can provide the following benefits:

- Significantly lower implementation costs of high performance I/O-intensive storage systems with the best cost per IOPS ratio
- Significantly higher performance and better response time of storage-intensive applications with up to 10 times lower latency
- Significant savings in power and cooling with high performance per watt ratio
- Significant savings in floor space with extreme performance per unit of the rack space ratio
- Simplified management and maintenance with internal server-based configurations (no external power and information infrastructure needed)

In environments where RAID protection is required, eXFlash can be used as master data storage if you combine it with a RAID controller with the Performance Accelerator key enabled to ensure that the peak IOPS can be reached.

The main feature of M5014, M5015, M5016, M5110, and M5210 controllers that can be used in the eXFlash solutions with Performance Key for SSDs is enabling *Cut Through I/O* (CTIO). CTIO optimizes highly random read/write I/O operations for small data blocks to support the high IOPS capabilities of SSD drives and to ensure the fastest response to the application. For example, by enabling CTIO on a RAID controller with SSDs, you achieve up to two times more IOPS compared to the controller with the CTIO feature disabled.

Controller requirements: A single eXFlash unit requires a dedicated controller (or two controllers). When used with eXFlash, these controllers cannot be connected to the HDD backplanes.

In a non-RAID environment where eXFlash can be used as a high-speed read cache, use the 6 Gb performance-optimized HBA to ensure that maximum random I/O read performance is achieved. Only one 6 Gb SSD HBA is supported per single SSD backplane.

It is possible to mix RAID and non-RAID environments; however, the maximum number of disk controllers that can be used with all SSD backplanes in a single system is four.

eXFlash requires the following components:

- eXFlash hot-swap SAS SSD backplane
- 1.8-inch SSDs
- SAS/SATA disk controller

Self-encrypting drives

Data security is a growing requirement for businesses of all sizes. Although many companies invested heavily in methods to thwart network-based attacks and other virtual threats, few effective safeguards are available to protect against potentially costly exposures of proprietary data that results from a hard drive being stolen, misplaced, retired, or redeployed.

Self-encrypting drives (SEDs) can satisfy this need by providing the ultimate in security for data-at-rest and can help reduce IT drive retirement costs in the data center. When combined with the compatible RAID controllers, the 6 Gbps SAS SEDs in System x servers deliver superb performance per watt with a cost-effective, secure solution for businesses of all sizes. Self-encrypting drives are also an excellent choice if you need to comply with government or industry rules for data privacy and encryption.

SAS SEDs have the following characteristics and capabilities:

- Interface speeds of 6 Gbps (600 MBps of maximum theoretical throughput)
- Rotational speeds of 10,000 RPM and 15,000 RPM
- Single hard disk drive capacities of 146 GB, 300 GB, 600 GB, 900 GB, or 1.2 TB
- Support for Native Command Queuing (NCQ)
- Support for Self-Monitoring, Analysis, and Reporting Technology (S.M.A.R.T.)
- 2.5-inch form-factor
- Hot-swap HDDs
- Encrypt data dynamically at the drive level with no performance impact
- Provide instant secure erasure (cryptographic erasure, so data is no longer readable)
- Enable auto-locking to secure data if a drive is misplaced or stolen while in use

When the self-encrypting drive is in normal use, its owner does not need to maintain authentication keys (otherwise known as *credentials* or *passwords*) to access the data on the drive. The self-encrypting drive encrypts data that is being written to the drive and decrypts data that is being read from it, all without requiring an authentication key from the owner.

Self-encrypting drives eliminate the need to overwrite, destroy, or store retired drives. When it is time to retire or repurpose the drive, the owner sends a command to the drive to perform a cryptographic erasure. The process is nearly instantaneous, regardless of the capacity of the drive. Cryptographic erasure simply replaces the encryption key inside the encrypted drive, making it impossible to ever use the deleted key to decrypt the encrypted data.

Self-encrypting drives reduce IT operating expenses by reducing asset control challenges and disposal costs. Data security with self-encrypting drives helps ensure compliance with privacy regulations without hindering IT efficiency.

Using a self-encrypting drive when auto-lock mode is enabled requires securing the drive with an authentication key. When secured in this manner, the drive's data encryption key is locked whenever the drive is powered down. In other words, the moment the self-encrypting drive is switched off or unplugged, it automatically locks the drive's data. When the self-encrypting drive is powered on again, it requires authentication before it can unlock the encryption key and read any data on the drive. This protects against misplacement and theft.

The hardware encryption engine on the drives matches the SAS port's maximum speed and encrypts all data with no performance degradation. This performance scales linearly and automatically with each drive added to the system. No processor cycles from the host are necessary, and I/O operations occur without interruption.

ServeRAID M Series controllers offer SED support with any RAID 5 upgrade (with or without cache memory). See "SAS controllers" on page 10.

For more information, see the Product Guide for SEDs:

http://lenovopress.com/tips0761

Storage tiering

The proper planning of information infrastructure includes choosing the most cost-effective way to fulfill the application requirements for storage access regarding speed, capacity, and availability. To describe these requirements, and to establish the framework for the deployment of the storage infrastructure, the storage tiering approach was established.

Each *storage tier* defines a set of characteristics to meet the application requirements. There are four tiers, each with performance, availability, capacity, and access pattern characteristics for the data on that tier. Knowing your data access requirements helps you place data on the appropriate storage tier, which ensures that your storage infrastructure can run your workloads efficiently.

The storage tiers, their corresponding characteristics, suitable storage media types, and relative cost per gigabyte are listed in Table 9. It helps you decide what type of storage that you need, based on application requirements.

Table 9 Storage tiers and characteristics

Storage tier	Characteristic	Storage media type	Cost per gigabyte
Tier 0 (SSD)	Random access I/O-intensive Extreme performance Extreme availability Frequent access	SSDs, eXFlash	Very high
Tier 1 (Online)	Random access I/O-intensive High performance High availability Frequent access	SAS HDDs	High
Tier 2 (Nearline)	Sequential access Throughput-intensive Capacity High availability Infrequent access	NL SAS, NL SATA HDDs	Moderate
Tier 3 (Offline)	Sequential access Throughput-intensive Large capacity Long-term retention No direct access	Tape	Low

Tiers 0, 1, and 2 are considered *primary* storage, meaning that data on these tiers can be directly accessed by the application. Tier 3 is *secondary* storage, with data that cannot be accessed directly unless it is moved to primary storage. Tier 0 is added for enterprise solid-state drives.

Data storage that is closer to the main memory (that is, closer to the application processes in the main memory) costs more to implement than storage that is farther away. In other words, the price per GB of data storage increases from Tier 3 to Tier 0. To keep costs optimized, it is best to place the most-demanded data (also referred to as *hot* data) from the working data set

closest to the main memory. Less-demanded data can be placed on a higher (more distant) storage tier.

From a planning standpoint, the rules that define the policy of placing data onto different storage tiers are part of the information lifecycle management strategy for the organization. From a technology standpoint, data management and relocation policy can be implemented either manually by administrators or automatically by management software that supports policy-based data relocation, for example, IBM General Parallel File System (GPFS).

SSD caching

The gap between CPU processing power and I/O rates increased rapidly over the last decade. This gap is an issue for fast processors that must wait for data to be moved off disks and into memory, and the gap continues to widen. To increase HDD storage access speed, different *caching* technologies can be implemented.

Despite the size of the stored data set, only a portion of its data is actively used during normal operations at certain time intervals. The data caching algorithms ensure that the most-demanded data (most frequently used) is always kept as close to the application as possible to provide the fastest response time. This principle is a foundation of caching operations.

With *caching*, the most frequently used data is kept in two copies, one on the primary data storage volume and the other in the fast cache memory. The data is duplicated in the cache memory that serves as a temporary buffer to provide fast access to the stored data. Caching begins to work immediately, with the first storage I/O requests, and it quickly and dynamically adapts to the changes in workload patterns.

Server-side caching exists at different levels. Server internal storage controllers use fast dynamic random access memory (DRAM) cache to keep the most frequently used data from disks. However, the cache size in normally limited to several GBs. Operating systems and certain applications keep their own disk cache in the fast system memory, but the cost per GB of RAM storage is high.

Solid-state storage, such as solid-state drives (SSDs) and PCIe SSD adapters, dramatically increase the performance of the disk-based storage to match the capabilities of other server subsystems. They keep costs optimized because the SSDs have a lower cost per GB ratio than DRAM memory and lower latency than traditional hard disk drives. Figure 9 on page 28 illustrates this scenario.

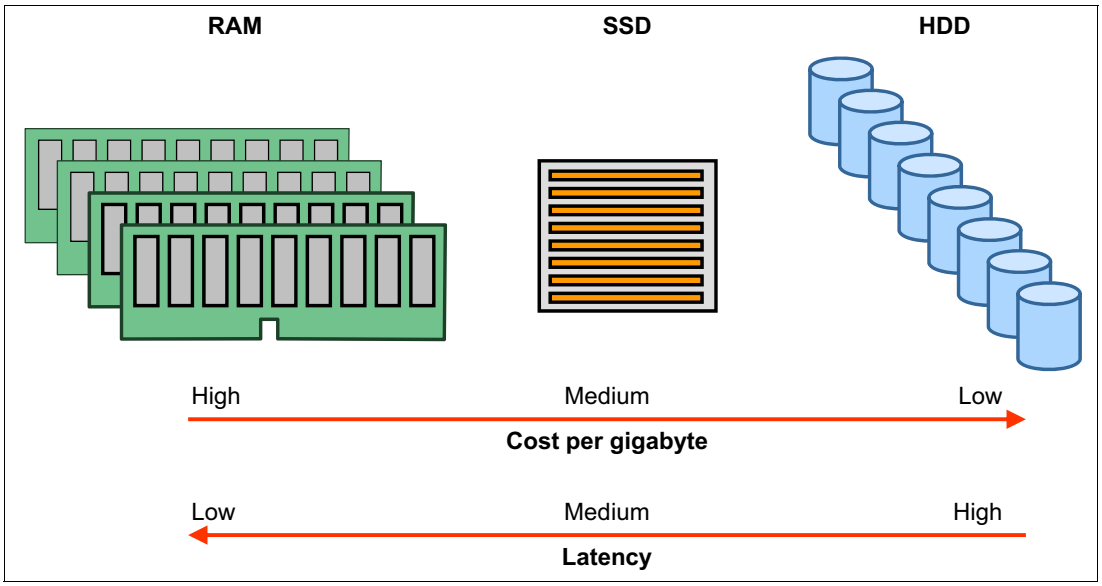

Figure 9 Cost per gigabyte and latency for RAM, SSD, and HDD

Solid-state storage can help fill the gap between processing power and storage I/O. It can help ensure that critical data is moved to the processor or memory much faster. SSDs reduce I/O wait time by initiating and completing data operations faster than spinning hard disks. For the highest level of performance, the goal is to keep the processor busy by reducing wait time and spending more time running operations.

SSD caching is as an optional feature of the ServeRAID M5000, M5100, and M5200 series controllers.

SSD Caching Enabler

MegaRAID CacheCade read/write software accelerates the performance of hard disk drive (HDD) arrays with only an incremental investment in solid-state drive (SSD) technology. The software enables SSDs to be configured as a dedicated pool of controller cache to help maximize the I/O performance for transaction-intensive applications, such as databases and web services. CacheCade software tracks data storage access patterns and identifies the most frequently accessed data. The hot data is then automatically stored on the solid-state storage devices that are assigned as a dedicated cache pool on a ServeRAID controller with the SSD Caching Enabler feature.

SSD Caching Enabler works only with server internal storage or external expansion enclosures. Connect SSDs that are used for caching and arrays that store data to the same internal RAID controller.

Caching is performed on a RAID controller level, and it is transparent to the operating system and applications.

Storage performance considerations

In general, there are two key types of storage applications that are based on workloads that they generate:

- *IOPS-intensive* applications require the storage system to process as many hosts' read and write requests (or I/O requests) per second as possible, given the average I/O request size used by this application, which is typically 4 - 8 KB. This behavior is most common for OLTP databases.
- *Bandwidth-intensive* applications require the storage system to transfer as many gigabytes of information per second as possible to or from hosts. They typically use an I/O request size of 128 - 512 KB, or more. These characteristics are commonly inherent to file servers, media streaming, and backup.

Therefore, there are two key performance metrics to evaluate storage system performance: *input/output requests per second (IOPS)* and *throughput (measured in MBps or GBps)*, depending on application workload.

Another important factor to take into account is the *response time* (or *latency*), or how much time the application spends waiting for the response from the storage system after submitting a particular I/O request. In other words, response time is the amount of time that is required by the storage system to complete an I/O request. Response time has a direct impact on the productivity of users who work with the application, such as how long it takes to get the requested information, and on the application itself. For example, a slow response to the database write requests might cause multiple record locks and further performance degradation of the application.

Key factors that affect the response time of the storage system include how quickly required data can be on the media (seek time) and how quickly the data can be read from or written to the media. Therefore response time also depends on the size of the I/O request, because reading or writing more data normally takes more time.

In addition, most applications generate many storage I/O requests at the same time, and these requests might spend some time in the queue if they cannot be immediately handled by the storage system. The number of I/O requests that can be concurrently sent to the storage system for the execution is called *queue depth*. This represents the service queue, which is the queue of requests that is being processed by the storage subsystem. If the number of outgoing I/O requests outreaches the parallel processing capabilities of the storage system (I/O queue depth), the requests are put into the wait queue and then moved to the service queue when a spot becomes available. This also affects the overall response time.

From the traditional spinning HDD perspective, improvement of latency is limited by mechanical design. Despite the increase in rotational speed of the disk plate and density of stored data, the response time of the HDD is still several milliseconds, which effectively limits its maximum IOPS (for example, a single 2.5-inch, 15,000 RPM SAS HDD is capable of ~300 IOPS, using 4 KB blocks).

With the SSD-based storage, the latency is typically measured in dozens of microseconds (almost 100 times lower than for the HDDs), which leads to significantly higher IOPS per SSD device (typically, ~50,000 IOPS or more). Higher IOPS capabilities also mean higher queue depth and, therefore, better response times for almost all types of storage I/O-intensive applications.

The knowledge of how the application accesses data, such as read-intensive or write-intensive, and random data access or sequential data access, helps to implement the most cost-effective storage that meets the required service level agreement (SLA) parameters. Table 10 summarizes typical application workload patterns in multiuser environments, depending on application type.

Table 10 Typical application workload patterns

Workload type → ↓ Application type	**Read intensive**	**Write intensive**	**IOPS intensive**	**Throughput intensive**	**Random access**	**Sequential access**
OLTP database	Yes	Yes	Yes		Yes	
Data warehouse	Yes			Yes	Yes	
File server	Yes			Yes	Yes	
E-mail server	Yes	Yes	Yes		Yes	
Medical imaging	Yes			Yes	Yes	
Video on demand	Yes			Yes	Yes	
Web, Internet	Yes		Yes		Yes	
Web 2.0	Yes	Yes	Yes		Yes	
Archives, backup		Yes		Yes		Yes

Generally, you can deploy the most efficient storage solution that satisfies application performance requirements, given the required storage capacity, by using these guidelines:

- Consider using SSDs or a higher number of HDDs for IOPS-intensive workloads with more drives of smaller capacities, because adding drives provides an almost linear increase in IOPS.
- Consider using a higher bandwidth between the host controller and storage arrays for throughput-intensive workloads, more host ports on a controller, and higher port speeds (for example, 6 Gbps rather than 3 Gbps) with enough drives in the array to put the workload on these links.

Both the number of drives and the way that the drives are connected to the controller determine throughput as a performance metric:

- Drives connected directly to the controller (typically 1 - 8 drives)
- Drives connected to the controller by using a SAS expander (more than 8 drives)

To understand the difference, assume that you have an array of eight HDDs, each drive is capable of 200 MBps of sustained throughput (which is, in fact, a best-case scenario that uses today's HDD technology), and the host is reading 8 GB of data from this array that is evenly distributed across all drives. You'll recall that *throughput* means how quickly the large sequential amount of data can be transferred from drive to host or back.

If an HDD array is formed from the drives that are directly connected to the ports on a RAID controller (see Figure 2 on page 8), there is a point-to-point link between every drive and the RAID controller, and this link has a bandwidth of 300 MBps for a 3 Gbps SAS and 600 MBps for 6 Gbps SAS interfaces. Because 200 MBps of single-drive throughput is less than either 300 MBps or 600 MBps, there is no difference between 3 Gbps or 6 Gbps interfaces. The link is not the limiting factor. In any case, eight drives can provide up to 1600 MBps of throughput, therefore transferring 8 GB in 3 seconds, using either a 3 Gbps or 6 Gbps interface. The use of a 6 Gbps interface provides little performance improvement.

In a case where HDDs are connected to the ports on a RAID controller through expanders, such Figure 3 on page 8 or Figure 4 on page 9 show, the bandwidth of the link between the RAID controller and expander and the drive interface speed become important. Eight drives that work simultaneously can provide up to 1600 MBps. A 3 Gbps x4 SAS link can handle up to 1200 MBps. Therefore, the time to transfer 8 GB of data is about 7 seconds, and the limiting factor is the link bandwidth. In a 6 Gbps x4 SAS link, which has 2400 MBps of bandwidth, transfer time is 5 seconds, because the link is not a limiting factor.

In this instance, the use of a 6 Gbps controller and drives offers a significant performance gain over 3 Gbps-equivalent devices. If 3 Gbps HDDs are used in this scenario, the effective bandwidth between controller and expander does not exceed 1200 MBps. Transfer time is also about 7 seconds because of the way that SAS manages connections and matches speeds of links between initiators and targets. Considering this limitation, it is important to match the speeds of RAID controller ports and drive interfaces.

The same approach works with the 12 Gbps SAS that was recently adopted by the industry. Even with 6 Gbps drives, 12 Gbps controller-expander connectivity can bring significant performance improvements for storage throughput-intensive applications.

The following points summarize drive performance considerations for direct connectivity and connectivity through an expander:

- When drives are directly connected to the controller, the drive performance is the limiting factor.
- When drives are connected to the controller by using a SAS expander, the link between the expander and controller (12 Gbps, 6 Gbps, or 3 Gbps) is the limiting factor, because the drives can fully saturate the link.

Redundant Array of Independent Disks (RAID)

To increase performance and reliability of a disk subsystem, Redundant Arrays of Independent Disks (RAIDs) are commonly used. This array is a group of physical disks that uses a certain common method to distribute data across the disks.

The data is distributed by *stripe units*. A stripe unit is the portion of data that is written to one disk drive immediately before the write operation continues on the next drive. When the last drive in the array is reached, the write operation continues on the first drive in the block that is adjacent to the previous stripe unit written to this drive, and so on.

The group of stripe units that is subsequently written to all drives in the array (from the first drive to the last) before the write operation continues on the first drive is called a *stripe*, and the process of distributing data is called *striping*. A stripe unit is a minimal element that can be read from or written to the RAID array, and stripe units contain data or recovery information.

The particular striping method that is used for data distribution is also known as the *RAID level*. This level has certain levels of availability, performance, and available storage capacity because achieving redundancy always lessens disk space that is reserved for storing recovery information.

There are basic RAID levels (0, 1, 5, and 6) and spanned RAID levels (00, 10, 50, and 60). *Spanned RAID arrays* combine two or more basic RAID arrays to provide higher performance, capacity, and availability by overcoming the limitation of the maximum number of drives per array that is supported by a particular RAID controller. For example, the ServeRAID M5014 and M5015 support up to 16 drives in a single (basic or spanned) array,

and the ServeRAID M5025 supports up to 32 drives per basic array and one spanned array that consists of up to eight basic arrays. This results in theoretical limit of 256 devices.

Table 11 and Table 12 on page 33 summarize the separate RAID levels and their characteristics. The following variables are used in these tables:

- K means number of drives in a single array.
- L means number of spans (the number of basic arrays that can be a part of single spanned array).
- N refers to several drives for redundancy. For example, N+1 means the array can sustain one drive failure and still perform I/O operations.

When redundancy is listed as L*(N+x) that means the spanned array can sustain L*x drive failures, providing that these failures happened in several basic arrays.

Table 11 RAID levels with basic arrays

RAID level → **↓ Characteristics**	**RAID 0**	**RAID 1**	**RAID 5**	**RAID 6**
Striping method	Striping	Mirroring	Striping with distributed parity	Striping with dual distributed parity
Minimum number of drives	1	2	3	4
Capacity (available space)	K*(*single drive size*)	(*single drive size*)	(K-1)*(*single drive size*)	(K-2)*(*single drive size*)
Redundancy	No	N+N	N+1	N+2
Read performance	Excellent	Very good	Excellent	Excellent
Write performance	Excellent	Very good	Satisfactory	Satisfactory

Table 12 Table 2. RAID levels with spanned arrays

RAID level → ↓ Characteristics	RAID 10	RAID 50	RAID 60
Striping method	Spanned mirroring	Spanned striping with distributed parity	Spanned striping with dual distributed parity
Minimum number of drives	4	6	8
Capacity (available space)	L*(single drive size)	L*(K-1)*(single drive size)	L*(K-2)*(single drive size)
Redundancy	L*(N+N)	L*(N+1)	L*(N+2)
Read performance	Very good	Excellent	Excellent
Write performance	Very good	Satisfactory	Satisfactory

Drive selection guidelines

The most common criteria that are used to choose the most appropriate storage solution for server environments are based on these factors:

- Application I/O workload pattern
- Cost
- Capacity
- Performance
- Scalability
- Reliability
- Power consumption
- Physical environment

Use NL SATA HDDs when these factors characterize your server environment:

- Cost per GB is a key decision factor
- Large single array capacity is required
- Infrequent sequential or light I/O workloads are planned

Use NL SAS HDDs when the following factors characterize your server environment:

- Cost per GB is a key decision factor
- Large single array capacity is required
- Reliability is important
- Infrequent sequential enterprise level I/O workloads are planned

Use 15,000 RPM SAS HDDs when these are characteristics of your server environment:

- Application performance and response time are extremely important
- Reliability is extremely important
- Continuous random enterprise level I/O workloads are planned

Use 10,000 RPM SAS HDDs when these characteristics apply:

- Application performance and response time are important
- Reliability is extremely important
- Continuous random enterprise level I/O workloads are planned

Use SSDs when these factors characterize your server environment:

- Cost per IOPS is a key decision factor
- Application performance and response time are extremely important
- Reliability is extremely important
- Storage energy efficiency is extremely important
- Continuous random enterprise level I/O workloads are planned

Use 2.5-inch drives when these factors characterize your server environment:

- GB per U density is important
- Storage performance density (IOPS per U or throughput per U) is important
- Storage energy efficiency is important

Because SAS fabric provides flexibility for both SAS and SATA devices, the SAS and SATA drives can be mixed in the same system or enclosure to better fit specific application workloads that are being deployed on it. However, mixing SAS and SATA requires server support, because there are thermal and vibration implications when you mix drive types.

Table 13 on page 35 summarizes the characteristics for specific storage connectivity types.

Note: The terms *Low*, *Moderate*, *High*, and *Very High* in Table 13 are relative indicators for comparison purposes and do not mean absolute values. For example, values in the Reliability row mean that NL SATA and NL SAS drives have better reliability than SATA drives, and SAS drives have better reliability than NL SAS or NL SATA drives.

Table 13 Feature comparison by connectivity technology and drive type

Drive feature	SATA HDD	NL SATA HDD	NL SAS HDD	SAS HDD	SATA SSD	SAS SSD
Interface speed	6 Gbps	6 Gbps	6 Gbps	6 Gbps	6 Gbps	6 Gbps
Interface bandwidth	600 MBps	600 MBps	600 MBps	600 MBps	600 MBps	600 MBps
RPM	7.2 K	7.2 K	7.2 K	10 K or 15 K	N/A	N/A
Drive ports	1	1	2	2	1	2
Duplex support	Half-duplex	Half-duplex	Full-duplex	Full-duplex	Half-duplex	Full-duplex
SATA controller support	Yes	Yes	No	No	No	No
SAS controller support	Yes	Yes	Yes	Yes	Yes	Yes
Hot-swap support	Yes	Yes	Yes	Yes	Yes	Yes
Multi-drive array suitability (tolerance to vibration)	Low	High	High	Very High	Very High	Very High
24x7 power on hours	No	Yes	Yes	Yes	Yes	Yes
24x7 I/O workloads	No	No	No	Yes	Yes	Yes
Reliability	Low	High	High	Very High	Very High	Very High
Sequential I/O performance	High	High	High	High	Very High	Very High
Random I/O performance	Low	Low	Low	High	Very High	Very High
Power consumption	Low	Low	Low	Moderate	Low	Low
Drive capacity	High	High	High	Moderate	Moderate	Moderate
Scalability	Low	Low	High	High	High	High
Cost	Low	Moderate	Moderate	High	Very High	Very High
Storage tier	N/A	Tier 2	Tier 2	Tier 1	Tier 0	Tier 0
Targeted application workloads	Single user desktop applications, such as working with documents and email, or web browsing	Lightly loaded departmental applications that store user data, such as file servers and email servers	Infrequent bandwidth-intensive sequential workloads that require reliable storage, such as archives, backups, and multimedia libraries	IOPS-intensive random mission critical workloads that require continuous (24x7) data access and fast response time, such as online transaction processing (OLTP) databases, data warehouses, and heavily loaded file and email servers		

Related publications

For more information, see the following documents:

- System x RAID products home page:

  http://www.ibm.com/systems/storage/product/systemx/scsi_raid.html
- ServeRAID software matrix:

  http://www.ibm.com/support/docview.wss?uid=psg1SERV-RAID
- System x Configuration and Options Guide:

  http://www.ibm.com/support/docview.wss?uid=psg1SCOD-3ZVQ5W
- ServeRAID Quick Reference

  http://lenovopress.com/tips0054

Authors

This paper was produced by the following team of specialists:

Ilya Krutov is a Project Leader at Lenovo Press. He manages and produces pre-sale and post-sale technical publications on various IT topics, including x86 rack and blade servers, server operating systems, virtualization and cloud, networking, storage, and systems management. Ilya has more than 15 years of experience in the IT industry, backed by professional certifications from Cisco Systems, IBM, and Microsoft. During his career, Ilya has held a variety of technical and leadership positions in education, consulting, services, technical sales, marketing, channel business, and programming. He has written more than 200 books, papers, and other technical documents. Ilya has a Specialist's degree with honors in Computer Engineering from the Moscow State Engineering and Physics Institute (Technical University).

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.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/redp4791

Trademarks