Cisco UCS C-Series I/O Characterization

White Paper

May 2013

Contents

Executive Summary

This white paper outlines the performance characteristics of the Cisco Unified Computing System™ (Cisco UCS®) C-Series Rack Servers using different RAID (Redundant Array of Independent Disks) controllers, such as LSI MegaRAID 9266CV-8i (UCS-RAID-9266CV), LSI MegaRAID 9271CV-8i (UCS-RAID9271CV-8i), and the embedded RAID controller. It presents the performance data of the various hard disk drives (HDDs), solid state drives (SSDs), and LSI MegaRAID controllers (with their impact on cache settings and RAID configuration). The benchmark activity detailed in this white paper also records the I/O operations per second (IOPS) characteristics of Cisco UCS C-Series Rack Servers. The performance characteristics are evaluated on RAID 0, 5, and 10 volumes. For maximum performance benefits, it is crucial to find the right RAID configuration for a storage system before hosting any application or data. This white paper aims to assist customers in making an informed decision when choosing the right RAID configuration for any given I/O workload.

Objectives

This document has the following objectives:

- Describe the I/O characterization study of various types of HDDs (15,000 and 10,000 rpm SAS disks) and SSDs (SATA) offered on Cisco UCS rack-mount servers as local storage
- Describe the I/O scaling and sizing study between Cisco UCS C220 M3 and Cisco UCS C240 M3 servers
- Describe the evaluation of the embedded RAID option to provide a comparative analysis of embedded and hardware RAID performance

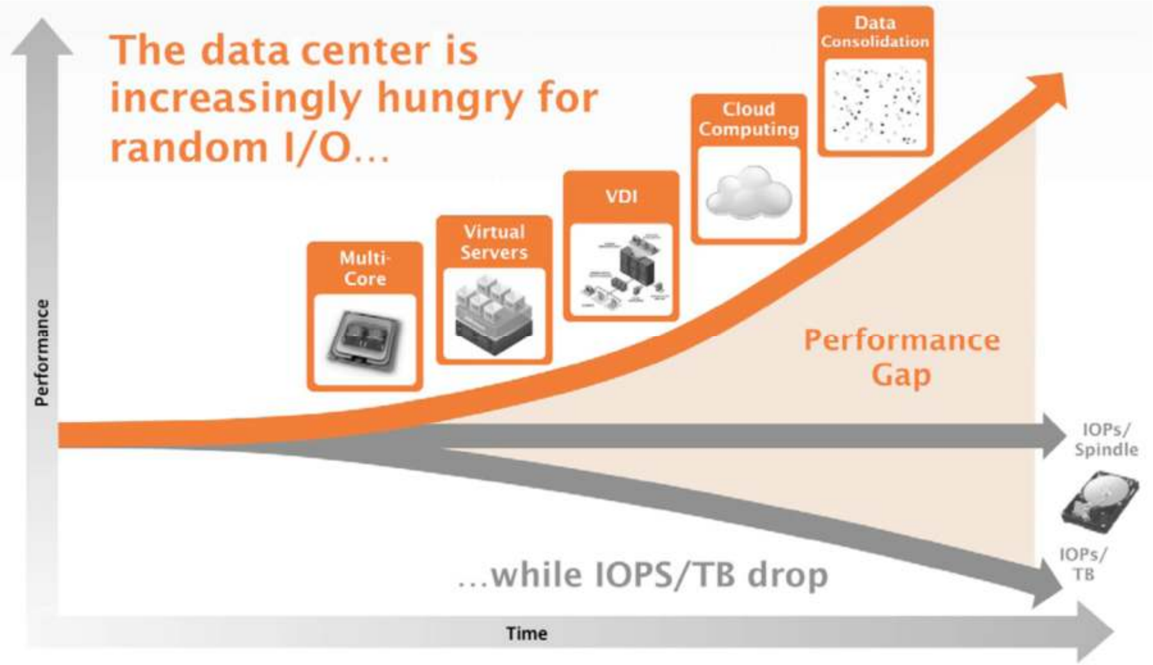
Introduction

The widespread adoption of virtualization and data center consolidation technologies has had a profound impact on the efficiency of the data center. Virtualization brings with it added challenges for the storage technology, requiring the multiplexing of distinct I/O workloads across a single I/O “pipe.” From a storage perspective, this results in a sharp increase in random IOPS. For storage disks, random I/O is the toughest to handle, requiring costly seeks and rotations between microsecond transfers. The hard disks not only add a security factor but also are the critical performance components in the server environment. Therefore, it is important to bundle the performance of these components through intelligent technology so that they do not cause a system bottleneck and also so they will compensate for any failure of an individual component. RAID technology offers a solution by arranging several hard disks in an array so that any hard disk failure can be compensated. Conventional wisdom holds that data center I/O workloads are either random (many concurrent accesses to relatively small blocks of data) or streaming (a modest number of large sequential data transfers). Historically, random access has been associated with a transactional workload, which is an enterprise’s most common type of workload. Currently, data centers are dominated by random and sequential workloads, brought in by the scale-out architecture requirements in the data center.

IOPS—Why Is It Important?

IOPS, or I/O operations per second, is an attempt to standardize the comparison of disk speeds across different environments. When the computer is first turned on, everything must be read from disk, but thereafter things are retained in memory. Enterprise-class applications, especially relational databases, involve high levels of disk I/O operations, which necessitate optimum performance of the disk resources.

Figure 1. I/O Trends for the Data Center



The I/O Dilemma

The rise of technologies such as virtualization, cloud computing, and data consolidation poses new challenges for the data center and requires enhanced I/O requests (Figure 1). This leads to an increased I/O performance requirement and the need to maximize available resources in ways that support the newest requirements of the data center and reduce the performance gap observed industrywide.

The following are the major factors leading to an I/O crisis:

- Increasing CPU utilization = increasing I/O

  Multicore processors with virtualized server and desktop architectures drive up processor utilization, raising the demand for I/O per server. In a virtualized data center, it is the I/O performance that limits server consolidation ratios, not CPU or memory.

- Randomization

  Virtualization has the effect of multiplexing multiple logical workloads across a single physical I/O path. The greater the virtualization achieved, the more random the physical I/O requests.

Technology Components

Hardware Components

Cisco Unified Computing System Server Platform

The Cisco Unified Computing System (Cisco UCS) is a next-generation data center platform that unites computing, network, and storage access. The platform, optimized for virtual environments, is designed within open industry-standard technologies and aims to reduce total cost of ownership (TCO) and increase business agility.

Cisco UCS C-Series Rack Servers extend Cisco Unified Computing System innovations to a rack-mount form factor, including a standards-based unified network fabric, support for Cisco® VN-Link virtualization, and Cisco Extended Memory Technology. Designed to operate both in standalone environments and as part of the Cisco UCS, these servers enable organizations to deploy systems incrementally—using as many or as few servers as needed—on a schedule that best meets the organization's timing and budget. This flexibility provides investment protection for customers.

Cisco UCS C220 M3

The Cisco UCS C220 M3 Rack Server is a 1-rack-unit (1RU) server designed for performance and density over a wide range of business workloads, from web serving to distributed database. Building on the success of the Cisco UCS C200 M2 High-Density Rack Servers, the enterprise-class Cisco UCS C220 M3 server further extends the capabilities of the Cisco Unified Computing System portfolio. And with the addition of the Intel® Xeon® processor E5-2600 product family, it delivers significant performance and efficiency gains.

The Cisco UCS C220 M3 also offers up to 512 GB of RAM, eight drives or SSDs, and two Gigabit Ethernet LAN interfaces built into the motherboard, delivering outstanding levels of density and performance in a compact package.

Cisco UCS C240 M3

The Cisco UCS C240 M3 Rack Server is a 2RU server designed for both performance and expandability over a wide range of storage-intensive infrastructure workloads, from big data to collaboration. Building on the success of the Cisco UCS C210 M2 General-Purpose Rack Server, the enterprise-class Cisco UCS C240 M3 further extends the capabilities of the Cisco Unified Computing System portfolio. The addition of the Intel Xeon processor E5-2600 product family delivers an optimal combination of performance, flexibility, and efficiency gains.

The Cisco UCS C240 M3 offers up to 768 GB of RAM, 24 drives, and four Gigabit Ethernet LAN interfaces built into the motherboard to provide outstanding levels of internal memory and storage expandability along with exceptional performance.

Table 1 provides the comparative specifications of the Cisco UCS C220 M3 and the Cisco UCS C240 M3 servers.

Table 1. Cisco UCS C-Series Server Specifications

Model	Cisco UCS C220 M3 Rack Server	Cisco UCS C240 M3 Rack Server
Image		
Processors	Up to 2 Intel Xeon E5-2600 processors	Up to 2 Intel Xeon E5-2600 processors
Form factor	1 RU	2 RU
Maximum memory	512 GB	768 GB
Internal disk drives	Up to 8	Up to 24
Maximum disk storage	Up to 8 TB	Up to 24 TB
Built-in RAID	0, 1, 5, and 10	0, 1, 5, and 10
RAID controller	• Cisco UCS RAID SAS 2008M-8i mezzanine card (RAID 0, 1, 10, 5) • On-board (embedded) LSI RAID controller UCSC-RAID-ROM55 • LSI MegaRAID 9266CV-8i • LSI MegaRAID 9285CV-8e	• LSI MegaRAID SAS 9266-8i with FTM +LSI CacheVault Power Module (RAID 0, 1, 10, 5, 6, 50, 60) • LSI MegaRAID 9285CV-8e • LSI MegaRAID 9266CV-8i
Disks supported	4 or 8 SAS, SATA, and SSD	16 or 24 SAS, SATA, and SSD
Integrated networking	• 2 GE ports • 10-Gbps unified fabric	• 4 GE ports • 10-Gbps unified fabric
I/O using PCI Express (PCIe)	PCIe Generation 3 slots: • 1 x8 half height and half length • 1 x16 full height and three-quarter length	• 2 PCIe Generation 3 x16 slots: both full height, 1 half and 1 three-quarter length • 2 PCIe Generation 3 x8 slots: both full height and 1 half length • 1 PCIe Generation 3 x8 slots: half height and half length

Storage Hardware Requirements

Solid State Drives (SSDs)

A growing number of organizations are deploying solid state drives (SSDs) to accelerate data storage performance. The organizations can choose one of three interfaces—SATA, SAS, or PCIe—for their SSD storage solution interface. Each interface has its own advantages and limitations and offers added value differentiation depending on the applications deployed.

Serial Attached SCSI (SAS)

Serial Attached SCSI (SAS) is the traditional interface for enterprise storage. It can handle up to 256 outstanding requests and is highly scalable. SAS SSDs are compatible with contemporary RAID architecture. For applications that demand high performance, SAS SSDs are a logical choice, offering support for 6-Gbps speed. SanDisk offers workload-optimized Lightning 6-Gbps SAS enterprise SSDs in a range of capacities to meet a wide variety of data center applications.

In our performance test, the following SAS drives were used:

- SAS 15,000-rpm drives (300 GB)
- SAS 10,000-rpm drives (600 GB)
- SSD (100 GB)

Peripheral Component Interconnect Express (PCIe)

Peripheral Component Interconnect Express (PCIe) is an I/O interface between various peripheral components inside a system. Unlike SAS and SATA, PCIe is designed to be an I/O expansion interface and not a storage interface. In addition, cards require a driver to function, so they are not widely used as primary storage.

Of the three interface types, PCIe connectors are the fastest and sit closest to the CPU, making them ideal for I/O-intensive application acceleration or as a caching solution. Today's second-generation PCIe interfaces offer speeds up to 5 GTps (gigatransfers per second), with a roadmap to 8 GTps and beyond. As PCIe becomes more popular for performance acceleration in servers and workstations, manufacturers are working to improve the PCIe interface to meet the storage and serviceability requirements of enterprises.

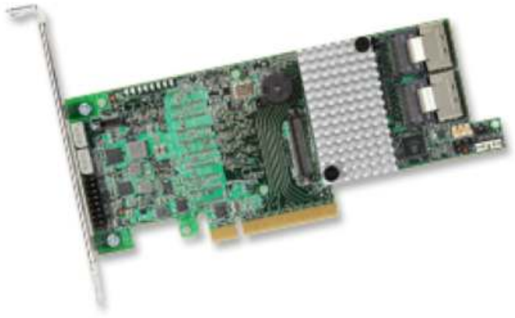
In our performance testing, the PCIe slots on Cisco UCS servers were used to install the LSI MegaRAID cards.

LSI MegaRAID 9266CV-8i

LSI MegaRAID 9266CV-8i is a PCIe Generation 2.0 MegaRAID SATA plus SAS RAID controller built on 6-Gbps SAS technology, offering high levels of performance. The low-profile MD2 controller features the dual-core LSISAS2208 SAS RAID-on-a-chip (ROC) interface card with two 800-MHz PowerPC processor cores and 1-GB DDR3 cache memory. The controller cards support up to 128 SATA and SAS HDDs or SSDs and hardware RAID levels 0, 1, 5, 6, 10, 50, and 60.

Figure 2 depicts the LSI MegaRAID 9266CV-8i controller card.

Figure 2. LSI MegaRAID 9266CV-8i Controller



The second-generation 6-Gbps SAS MegaRAID controller is well equipped to address high-performance storage requirements for both internal and external connections to the server. It includes the following:

- 8-port internal RAID controller with top connectors (MegaRAID SAS 9265CV-8/SAS 9270CV-8i)
- 8-port internal RAID controller with side connectors (MegaRAID SAS 9266CV-8i/SAS 9271CV-8i)
- 8-port external RAID controller with external SAS connectors (MegaRAID SAS 9285CV-8e/SAS 9286CV-8i)

Note: MegaRAID controllers support CacheVault module and remote Supercap for backup.

SAS Expanders

SAS expanders enable you to maximize the storage capability of the SAS controller card. Many SAS controllers support up to 128 hard drives, which can be done only with a SAS expander solution.

In this performance test, two X4 wide SAS cables were connected to SAS expander ports/connectors located on the midplane. If the SAS cable is connected to a single port, only 16 drives will be active. When both SAS 0 and 1 ports are connected, 24 drives will be active.

Note: This configuration is applicable only to the Cisco UCS C240 M3 server.

Software Components

LSI MegaRAID Storage Manager

The LSI MegaRAID Storage Manager (MSM) is an application that enables you to configure, monitor, and maintain storage configurations on LSI SAS controllers. The MSM GUI enables the user to create and manage storage configurations.

MSM offers the following features:

- N-to-1 management. In an environment that enables communication through TCP/IP, more than one system can be managed remotely.
- Ability to create or delete the following arrays (packs) through the GUI:
  - RAID 0 (data striping with more than one hard disk drive)
  - RAID 1 (data mirroring with two hard disk drives)
  - RAID 5 (data and parity striping with more than two hard disk drives)
  - RAID 1 spanning (same as RAID 10. Data mirroring and striping with more than three hard disk drives)
  - RAID 5 spanning (same as RAID 50. Data striping, including parity and striping with more than five hard disk drives)

Iometer

Iometer is an I/O subsystem measurement and characterization tool for single and clustered systems. It is used as a benchmark and troubleshooting tool and is easily configured to replicate the behavior of many popular applications. One commonly quoted measurement provided by the tool is IOPS. Iometer is based on a client-server model and performs asynchronous I/O operations, such as accessing files or blocking devices (allowing one to bypass the file system buffers).

Iometer allows the configuration of disk parameters such as the maximum disk size, starting disk sector, and the number of outstanding I/O requests. This allows the user to configure a test file upon which the access specifications configure the I/O types of the file. Configurable items within the access specifications are as follows:

- Transfer request size
- Percentage random/sequential distribution
- Percentage read/write distribution
- Aligned I/O
- Reply size
- TCP/IP status

- Burstiness

In addition to defining the access specifications, Iometer allows the specifications to be cycled with incremental outstanding I/O requests, either exponentially or linearly.

RAID Levels

In this performance test characterization, we used RAID 0, RAID 5, and RAID 10 options with the LSI MegaRAID 9266CV-8i. Table 2 describes the various RAID levels and their characteristics.

Table 2. RAID Levels and Characteristics

RAID Level	Characteristics	Parity	Redundancy
RAID 0	Striping of two or more disks to achieve optimal performance	No	No
RAID 1	Mirroring data on two disks for redundancy with slight performance improvement	No	Yes
RAID 5	Striping data with distributed parity for improved fault tolerance	Yes	Yes
RAID 6	Striping data with dual parity and dual fault tolerance	Yes	Yes
RAID 10	Mirroring and striping the data for redundancy and performance improvement	No	Yes
RAID 5+0	Block striping with distributed parity for high fault tolerance	Yes	Yes
RAID 6+0	Block striping with dual parity for performance improvement	Yes	Yes

Controllers Supported in Cisco UCS C-Series Servers

Cisco UCS C-Series Rack Servers offer various caching and RAID configurations with various RAID controllers. Table 3 describes the LSI MegaRAID controllers and their corresponding features.

Table 3. Controllers Supported in the Cisco UCS C-Series Servers

RAID Card Controller	Number of Drives Supported Internal	Number of Drives Supported External	Types of Drives Supported	RAID Supported	Operating Speed	DDR3 Cache	Data Cache Backup
LSI MegaRAID 9266CV-8i controller	16 and 24	0	SAS, SATA, SSD	0, 1, 5, 6, 10, 50, and 60	6 Gbps	1 GB	Yes
LSI MegaRAID 9285CV-8e controller	8	240	SAS, SATA, SSD	0, 1, 5, 10, 50, and 60	6 Gbps	1 GB	Yes
Cisco UCSC RAID SAS 2008M-8i controller	8 and 16	0	SAS, SATA, SSD	0, 1, 5, 10, and 50	6 Gbps	• 4 MB flash part for IR code • 32 K of NVS RAM for write journaling	No
Embedded RAID (on motherboard)	8	0	SAS, SATA, SSD	0, 1, 5, and 10	3 Gbps	N/A	No

Note: In the Cisco UCS C240 M3 server with the 24-drive backplane or 16-drive backplane, the PCIe RAID controllers are installed by default in slot 3 for a server that hosts one CPU and in slot 4 for a server that hosts two CPUs.

RAID Summary for Cisco UCS C240 M3 Servers

The Cisco UCS C240 M3 Small Form-Factor (SFF) server can be ordered with a 16-drive backplane or a 24-drive backplane.

- ROM 5 and ROM 55 embedded RAID upgrade options support up to 8 drives with the 16-drive backplane.
- Mezzanine cards (UCSC-RAID-11-C240 and Cisco UCSC-RAID-MZ-240) support up to 8 drives for the 16-drive backplane and up to 16 drives for the 24-drive backplane.
- SAS 9266-8i and SAS 9266CV-8i PCIe cards support up to 8 drives for the 16-drive backplane and up to 24 drives for the 24-drive backplane.
- LSI MegaRAID SAS 9285CV-8e supports up to 8 external SAS ports (240 external drives).

Table 4 and Table 5 detail the RAID configuration options supported by the Cisco UCS C240 M3 and Cisco UCS C220 M3 server, respectively.

Table 4. Summary of the Supported RAID Configurations on Cisco UCS C240 M3

Server	Number of CPUs	Embedded RAID (slot 2)	Mezzanine RAID (slot 3)	Internal PCIe RAID #2 (slot 4)	Internal PCIe RAID (slot 5)	External PCIe RAID	Number of Drives Supported	PCIe Slots				
								1	2	3	4	5
Cisco UCS C240 M3 SFF 24 HDD	1	**Not allowed**	**Not allowed**	***Installed in slot 3 (default)***	**Not allowed**	*Card* absent	24 internal	A	A	***O***	U	U
Cisco UCS C240 M3 SFF 24 HDD	1	**Not allowed**	**Not allowed**	Card *absent*	**Not allowed**	***Installed in slots 1, 2, or 3***	8 internal	A	A	A	U	U
Cisco UCS C240 M3 SFF 24 HDD	1	**Not allowed**	**Not allowed**	***Installed in slot 3 (default)***	**Not allowed**	***Installed in slots 1 and 2***	24 internal, 240 external	A	A	***O***	U	U
Cisco UCS C240 M3 SFF 24 HDD	2	**Not allowed**	***Installed***	**Not allowed**	**Not allowed**	*Card absent*	16 internal	A	A	A	A	A
Cisco UCS C240 M3 SFF 24 HDD	2	**Not allowed**	**Not allowed**	***Installed in slot 4 (default)***	**Not allowed**	*Card absent*	24 internal	A	A	A	***O***	A
Cisco UCS C240 M3 SFF 24 HDD	2	**Not allowed**	*Card absent*	*Card absent*	*Card absent*	***Installed in any slot***	0 internal, 240 external	A	A	A	A	A
Cisco UCS C240 M3 SFF 24 HDD	2	**Not allowed**	***Installed***	**Not allowed**	**Not allowed**	***Installed in any slot***	16 internal, 240 external	A	A	A	A	A
Cisco UCS C240 M3 SFF 24 HDD	2	**Not allowed**	**Not allowed**	***Installed in slot 4 (default)***	**Not allowed**	***Installed in any slot (expect slot 4)***	24 internal, 240 external	A	A	A	***O***	A

Table 5. Summary of the Supported RAID Configurations on Cisco UCS C220 M3

Server	Number of CPUs	Embedded RAID	Mezzanine RAID	Internal PCIe RAID #1	Internal PCIe RAID #2	External PCIe RAID #	Number of Drives Supported	PCIe Slots	
								1	2
Cisco UCS C220 M3 SFF	1	Enabled	**Not allowed**	**Not allowed**	**Not allowed**	**Not allowed**	8 internal	A	U
Cisco UCS C220 M3 SFF	1	**Not allowed**	**Not allowed**	***Installed in slot 1 (default)***	**Not allowed**	**Not allowed**	8 internal	***O***	U
Cisco UCS C22 M3 SFF	1	**Not allowed**	**Not allowed**	**Not allowed**	**Not allowed**	***Installed in slot 1***	240 external	***O***	U
Cisco UCS C220 M3 SFF	2	Enabled	**Not allowed**	**Not allowed**	**Not allowed**	**Not allowed**	8 internal	A	A
Cisco UCS C220 M3 SFF	2	**Not allowed**	***Installed***	**Not allowed**	**Not allowed**	*Card absent*	8 internal	A	A
Cisco UCS C220 M3 SFF	2	**Not allowed**	**Not allowed**	***Installed in slot 2 (default)***	**Not allowed**	**Not allowed**	8 internal	A	***O***
Cisco UCS C220 M3 SFF	2	**Not allowed**	*Card absent*	**Not allowed**	**Not allowed**	***Installed in slot 1 or slot 2***	240 external	A	A
Cisco UCS C220 M3 SFF	2	**Not allowed**	***Installed***	**Not allowed**	**Not allowed**	***Installed in slot 1 or slot 2***	8 internal, 240 external	A	A

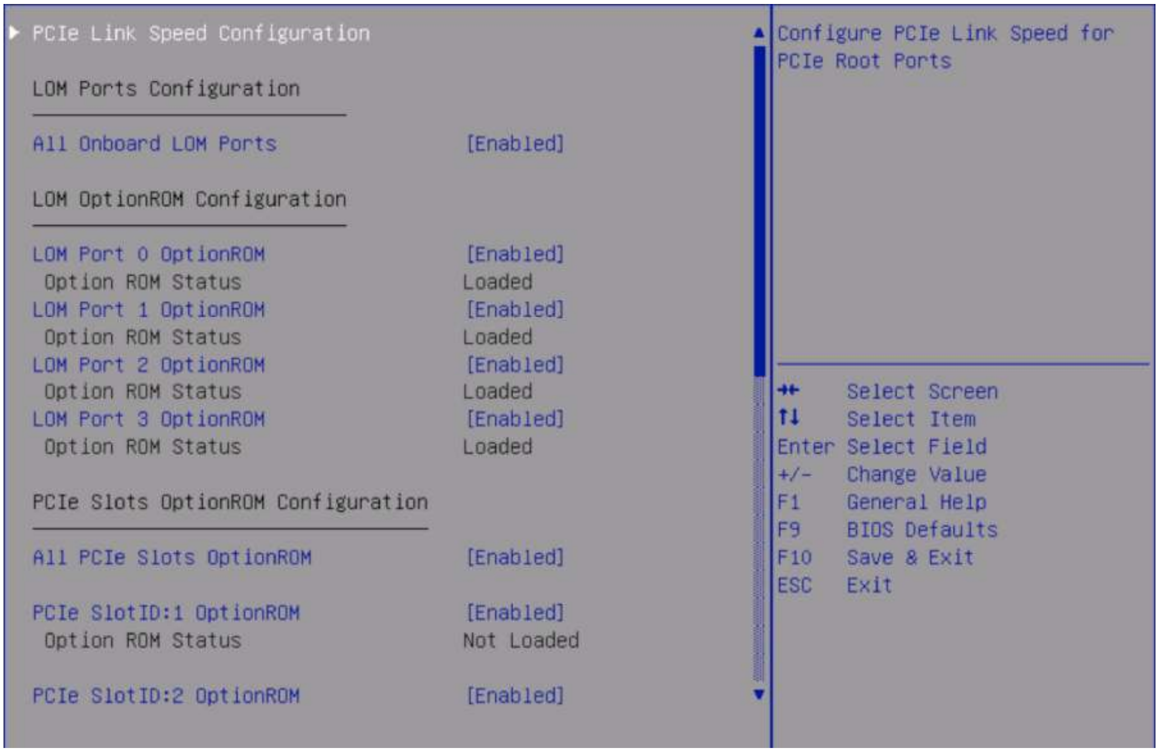
Note: A = Available slot, O = Occupied slot, U = Unsupported slot (slots 4 and 5 are not supported in 1-CPU systems).

The following points provide additional information for installing the LSI MegaRAID controller in Cisco UCS C-Series M3 servers:

- The RAID types cannot be mixed. Only one type—embedded RAID, mezzanine RAID, or PCIe RAID—can be used at a time.
- Embedded RAID is compatible with the 16-HDD backplane, and it cannot be used with the 24-HDD backplane.
- Do not disable the Option ROM (OPROM) for a mezzanine slot if the mezzanine card is present. If OPROM is disabled, the system will not boot. If you remove the mezzanine card and disable the OPROM, you can boot from another bootable device (from a RAID card, from embedded RAID, or from the SAN via HBA or CNA card). When booting from a device, make sure that the OPROM is enabled, that there is a proper boot sequence, and that the BIOS is configured for a bootable device.
- To boot from a device other than the 9266-8i or 9266CV-8i, leave the cards installed and disable the OPROM.
- The external RAID card used is the 9285CV-8e. The 9285CV-e can be installed simultaneously with either one mezzanine RAID controller card or one internal RAID controller card (9266-8i or 9266CV-8i).
- The mezzanine card is not supported in a 1-CPU configuration.
- The OPROM is enabled for the default PCIe RAID controller slots. If you want to enable a different slot, log in to the BIOS and enable the OPROM option for the desired slot, and disable the OPROM for the default PCIe slot.

The server contains a finite amount of OPROM for PCIe slots, which enables it to boot devices. In the BIOS, disable OPROM on the PCIe slots not used for booting. This releases resources to the slots that are used for booting devices. Figure 3 shows the OPROM BIOS screen.

Figure 3. OPROM BIOS Screen



```
▶ PCIe Link Speed Configuration                              ▲ Configure PCIe Link Speed for
                                                               PCIe Root Ports
  LOM Ports Configuration

  All Onboard LOM Ports                  [Enabled]

  LOM OptionROM Configuration

  LOM Port 0 OptionROM                   [Enabled]
   Option ROM Status                     Loaded
  LOM Port 1 OptionROM                   [Enabled]
   Option ROM Status                     Loaded
  LOM Port 2 OptionROM                   [Enabled]
   Option ROM Status                     Loaded                →←    Select Screen
  LOM Port 3 OptionROM                   [Enabled]             ↑↓    Select Item
   Option ROM Status                     Loaded                Enter Select Field
                                                               +/-   Change Value
  PCIe Slots OptionROM Configuration                           F1    General Help
                                                               F9    BIOS Defaults
  All PCIe Slots OptionROM               [Enabled]             F10   Save & Exit
                                                               ESC   Exit
  PCIe SlotID:1 OptionROM                [Enabled]
   Option ROM Status                     Not Loaded

  PCIe SlotID:2 OptionROM                [Enabled]           ▼
```

LSI MegaRAID 9266CV-8i/9271CV-8i Controller Cache

The LSI MegaRAID 9266CV-8i controller comes with read/write caching software to accelerate the I/O performance of the underlying HDD arrays using SSDs as a high-performance cache.

Caching Mechanisms

The LSI MegaRAID SAS controller 9266CV-8i provides a controller cache in read/write caching versions that also offers extra protection against power failure through a battery-powered backup unit (BBU). The controller cache is used to accelerate write and read performance, which is influenced through the following read/write caching mechanisms:

- Read policies
- Controller cache
- Write policies
- Disk cache

Read Policies

The read policies used in this performance test are Always Read Ahead and No Read Ahead.

- **Always Read Ahead**

This caching policy causes the controller to always read ahead if the two most recent I/O transactions are sequential in nature. If all I/O requests are ahead of the current request, the subsequent I/O request is kept in the controller cache (the incoming I/O is always kept in the cache). The controller reads ahead for all data until the last stripe of the disk.

Note: If all read I/O requests are random in nature, the caching algorithm reverts to No Read Ahead (Normal Read).

- **No Read Ahead (Normal Read)**

In the No Read Ahead caching policy mode, only the requested data is read and the controller does not read ahead any data.

Controller Cache

- **Direct I/O**

When the Direct I/O caching policy is turned on, all read data is dumped directly into the host memory, bypassing the RAID controller cache. If Always Read Ahead is set as the caching policy, all reads are cached even if the controller mode is set to Direct I/O. All write data is transferred directly to the disk if Write Through mode is turned on.

- **Cached I/O**

In the Cached I/O policy mode, all read and write data passes through the controller cache memory to the host memory. When Cached I/O mode is turned on, all writes are cached, even when the write cache mode is set to Write Through.

Write Policies

- **Write Through**

In Write Through mode, data is written directly to the disk before an acknowledgment from the host is received. Write Through mode is always recommended for RAID 0 and RAID 10 to improve performance in the sequential workloads (since the data is moved directly from host to disks).

- **Write Back**

In the Write Back mode, data is first written to the controller cache, and when it receives acknowledgment from the host, data is flushed to the disks. Data is written to the disks when commit happens at the controller cache. Write Back mode is more efficient when there are bursty write activities. The BBU can be used for additional data protection in case of power failure. Write Back mode is highly recommended for transactional workloads on RAID 5 and RAID 6.

Disk Cache

Disk Cache mode improves the read operation performance, and the write operations also show considerable improvement.

Recommended MegaRAID Cache Settings

This section defines the MegaRAID cache settings recommended for optimum performance on Cisco UCS C-Series servers. For any streaming workload in a RAID 0 and RAID 10 configuration, Cisco recommends turning on Write Through mode, because the host places the data directly on the disk and stores it in the controller cache. So every write access requires another write to the disk to activate that particular sector of the disk. For sequential applications, Write Back mode does not provide much benefit, because data is not reused from the cache. Enabling Disk Cache mode improves the drive performance and minimizes instances of drive-limited performance.

Table 6 lists the recommended settings for RAID 0, 1, and 10.

Table 6. Recommended Settings for RAID 0, 1, and 10

RAID 0, 1, and 10	
Stripe size	Greater than 64 K
Read policy	Always Read Ahead
Write policy	Write Through (for streaming sequential performance)
	Write Back (for random workloads)
I/O policy	Direct I/O (for sequential applications)
	Cached I/O (for random application)
Disk cache policy	Enabled

In RAID 5 and 6 configurations, it is always recommended to have Write Back mode on, because caching improves I/O performance, since it can reorder writes and can write multiple hard drive sectors simultaneously. Table 7 lists the recommended settings for RAID 5 and 6 configurations.

Table 7. Recommended Settings for RAID 5 and 6

RAID 5 and 6	
Stripe size	64 K or lower
Read policy	No Read Ahead
Write policy	Write Back (for random workloads)
I/O policy	Direct I/O (for sequential applications)
Disk cache policy	Enabled
Initialization state	Full Initialization (or write data across the entire volume)

Always do a full initialization of a RAID 5 volume before running any host I/O. In this mode the controller is fully employed in performing the initialization process and blocks any host I/O. For a RAID 5 volume the fast initialization process is not recommended because the background verification does not initialize the lower block address, resulting in a slower response time. Full initialization also ensures that the parity calculation is not affected and that the logical block address (LBA) and the logical row are read before the physical drives are read.

Table 8 and Table 9 list the recommended settings for HDD and SSD performance testing.

Table 8. Recommended Settings for HDD Performance Testing

HDD Disk Environment		Recommended Settings		
RAID Type	**I/O Benchmarking**	**Controller Write Cache**	**Controller Read Cache mode**	**Stripe Size**
0	Transactional	Enabled	No Read Ahead	64K to 256K
1/10	Transactional	Enabled	No Read Ahead	64K to 256K
5/50	Transactional	Enabled	No Read Ahead	64K
6/60	Transactional	Enabled	No Read Ahead	64K to 256K
0	Sequential	Disabled	Always Read Ahead	256K or higher
1/10	Sequential	Disabled	Always Read Ahead	256K or higher
5/50	Sequential	Enabled	Always Read Ahead	256K or higher
6/60	Sequential	Enabled	Always Read Ahead	256K or higher

Table 9. Recommended Settings for SSD Performance Testing

SSD Disk Environment		Recommended Settings		
RAID Type	**I/O Benchmarking**	**Controller Write Cache**	**Controller Read Cache mode**	**Stripe Size**
0	Transactional	Disabled	No Read Ahead	64K
1/10	Transactional	Disabled	No Read Ahead	64K to 256K
5/50	Transactional	Disabled	No Read Ahead	Lowest stripe size
6/60	Transactional	Disabled	No Read Ahead	Lowest stripe size
0	Sequential	Enabled	Always Read Ahead	64K
1/10	Sequential	Enabled	Always Read Ahead	64K
5/50	Sequential	Enabled	Always Read Ahead	64K
6/60	Sequential	Enabled	Always Read Ahead	64K

Note: For any SSD benchmark, it is recommended that you use the smallest stripe size on the controller.

When the write cache is enabled, data that is written to the drive is cached in its RAM before it is stored permanently. There is a slight risk that a power outage will wipe out the data stored temporarily in RAM, and you can lose the entire file system in case of a crash. There is also a risk of metadata getting written out of order with data, which can destroy entire directories and large parts of the file system, even destroying files that have not been touched or updated for months.

PHASE I

This section elaborates on the test plan and the raw performance results achieved on the various hard drives and RAID controllers on Cisco UCS C-Series M3 Rack Servers. The performance tests described in this document were carried out in two phases, in which I/O characterization of SAS and SSD disks was validated on Cisco UCS C-Series servers.

Test Plan

In Phase I, the SAS and SSD (for mainstream performance) was characterized using Iometer. The SAS and SSD’s performance in various RAID combinations was benchmarked against a series of I/O block sizes until the disk was saturated. In every test scenario, the IOPS, the throughput, and the latency were captured. The memory configuration was not taken into account while evaluating the efficiency of the I/O subsystem.

In Phase I the scaling test is performed by applying various block sizes and RAID combinations to the disk. The analysis of the test results shows the optimum specifications that can be configured to maximize throughput, after which diminishing returns were observed. The various data points collected in Phase I helped to identify the sweet spot for various combinations (see Table 10). The data points, such as the workload, the access pattern, the RAID type, the disk type, and the I/O block size, help the customer identify the limitations posed by the system for a specific workload and thereby make the best selections.

Table 10 describes the combination of workloads and access patterns that were chosen to run on the system using the Iometer.

Table 10. Workload and Access Pattern

I/O Mode	I/O Mix ratio (Read: Write)			
Sequential	100:0		0:100	
	RAID 0	RAID 5	RAID 0	RAID 10
Random	70:30		50:50	
	RAID 5		RAID 5	

Table 11 lists the I/O block size that was used to generate the workload.

Table 11. I/O Block Size

I/O Block Size										
4K	8K	16K	32K	64K	128K	256K	512K	1M	2M	4M

Table 12 lists the metrics collected on each selected server platform with different RAID types. These metrics were captured while ensuring that the response time threshold for read was within the range of 10 to 12 milliseconds and the threshold for write was within the range of 8 to 10 milliseconds. Bandwidth is relevant for sequential workloads, and IOPS is relevant for random workloads.

Table 12. I/O Metrics Against Server Platform and RAID Types

I/O Metrics	Server Platform	RAID Configuration	Disk Type
Bandwidth **IOPS** **Response time**	C220 M3 C240 M3	RAID 0, 5, 10	SAS (15,000 rpm) SAS (10,000 rpm) SATA SSD

Software vs. Hardware RAID Performance

In the above combination (see Table 10), RAID functionality is provided by an external controller (hardware RAID). However, there is an option to use the embedded RAID option in these platforms. To understand the performance difference between the embedded and the hardware RAID, we tested the Online Transaction Processing (OLTP) workload (70R:30W, 8K block size) and the sequential workload with the embedded RAID option.

Benchmark Configuration

In this test, 24 internal SAS (15,000 and 10,000 rpm) and SSD drives were used for I/O validation on the Cisco UCS C240 M3 with the LSI MegaRAID 9266-8i controller. Additionally, 8 internal 15,000-rpm SAS disks were validated on the Cisco UCS C220 M3 with the embedded RAID option.

On the Cisco UCS C240 M3 server, we set up a SAN boot and installed Windows 2008 R2 to ensure that the boot drive was independent of the LSI MegaRAID controller. Also, we ensured that all 24 SAS drives were connected to the LSI MegaRAID controller to fully saturate it.

Controller Settings

The controller settings (LSI MegaRAID 9266-8i) in this test used RAID 0/5/10 as a single volume carved out of 24 drives with 64K (Kilobytes) stripe size, with specific settings for the RAID volume. Table 13 lists the various RAID types and controller settings used for performance testing.

Table 13. RAID Types and Controller Settings

RAID	Controller Settings
0	Write Through Direct I/O No Read Ahead Disk cache disabled
5	
10	

The controller settings defined in Table 13 were used in the first phase of the testing to ensure that there was no caching effect on the workload iterations run. The controller setting was kept constant for all the IOPS testing in Phase I. These settings ensured that the disks and the LSI MegaRAID controller were stressed while running the various I/O workloads. The results provide the raw IOPS (of the disk) and throughput (controller bandwidth) values.

Iometer Settings

The following Iometer settings were used in this performance test:

- 70 percent of the volume was used as a dataset (to avoid caching effect at the OS host)
- Outstanding I/O requests were tuned for each specification per drive set and RAID level
- All worker threads

Performance Results

This section presents the test results, performance analysis, and throughput of the disks using the LSI MegaRAID controller in Phase I. The primary metrics are the throughput rate and the I/O rate (IOPS) measured by Iometer. Table 14 describes the component details of the measuring environment in this performance test.

Table 14. Measuring Environment

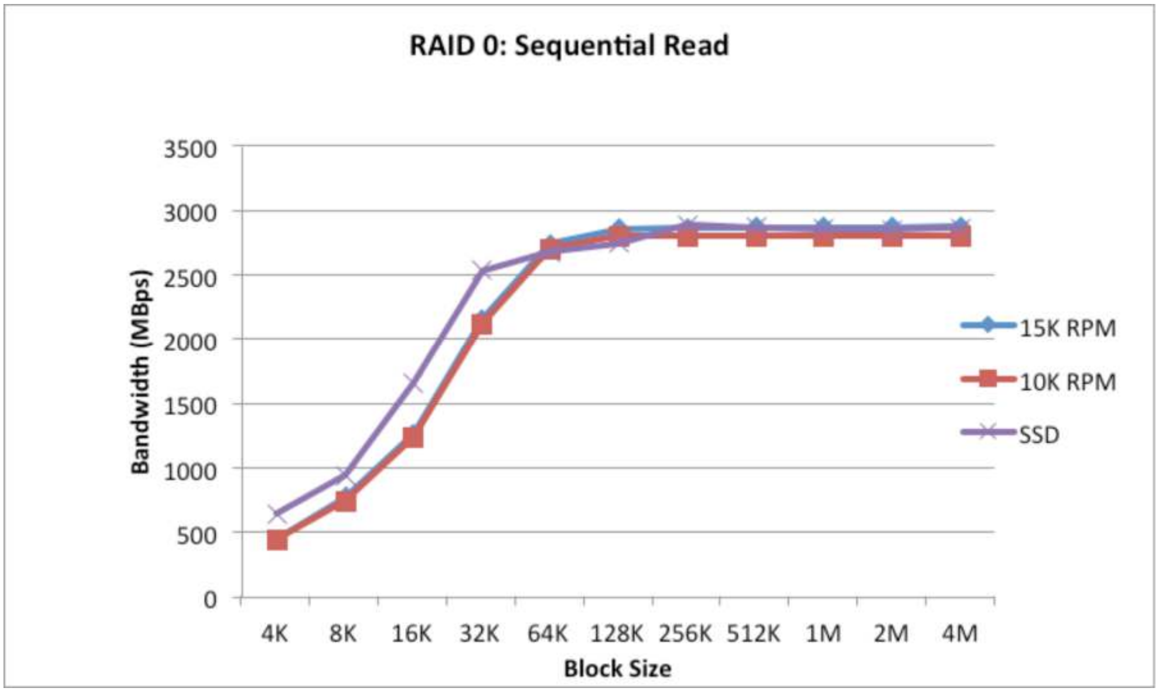
Component	Details
Server	Cisco UCS C240 M3 and Cisco UCS C220 M3
LSI MegaRAID 9266-8i controller, embedded RAID (Cisco UCS C220 M3)	Firmware version: 3.220.75-2196
Operating system	Microsoft Windows 2008 R2—Data Center version
Hard disk SAS, 2½ in., 15,000 and 10,000 rpm and SSD	300 GB (15,000 rpm), 600 GB (10,000 rpm), and 100 GB (Intel R710 SSD)
Device type	Raw device
Test SATA	24-disk volume (C240 M3) and 8-disk volume (C220 M3)
I/O block size	512 K to 4 MB
I/O mix	• Sequential: 100% read, 100% write • Random: 70% read, 30% write

Component	Details
	50% read, 50% write

Sequential Read/Write Bandwidth with RAID 0, 5, and 10

This section presents the performance results and analysis of the sequential read/write bandwidth with RAID 0, 5, and 10.

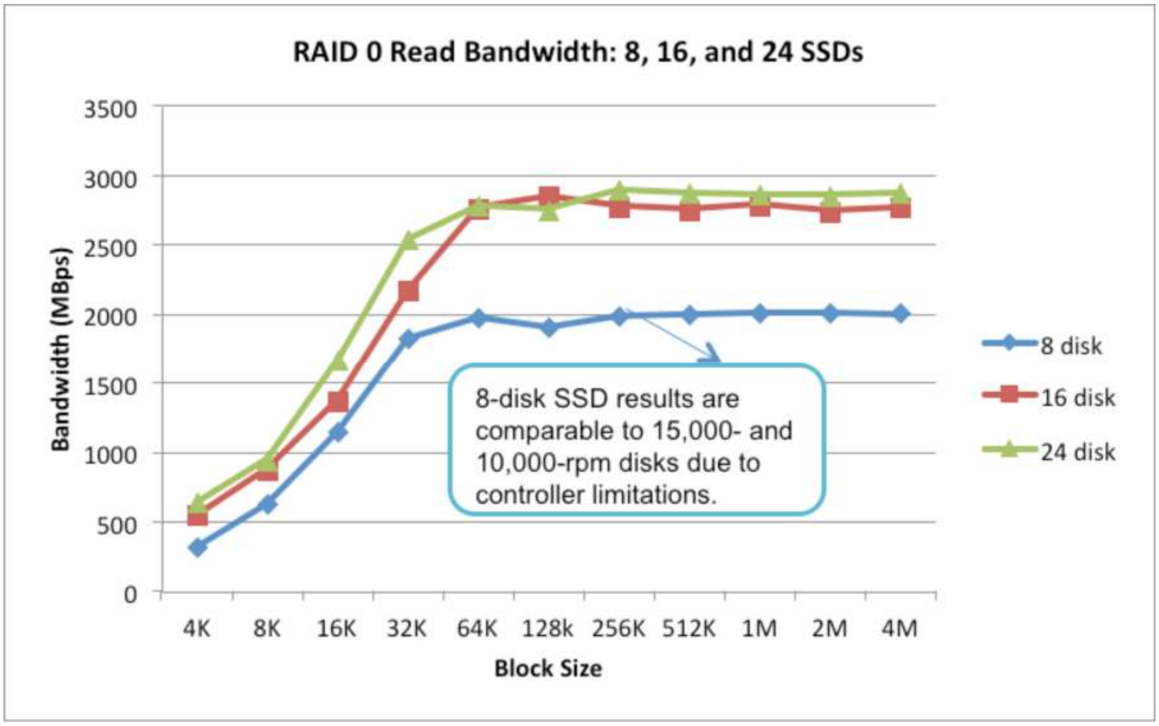
Figure 4. Sequential Read Bandwidth on RAID 0 Configuration



As Figure 4 shows, the LSI MegaRAID controller achieved a sequential read bandwidth of 2.8 GBps with 15,000-rpm and 10,000-rpm disks and SSD drives that match the controller specification. The block size of 64K was observed as the sweet spot at which the controller peaked at a maximum read bandwidth of 2800 MBps. The controller limitation also dictated the SSD’s read bandwidth, which peaked at 2800 Mbps with the 18-disk configuration, and in a 8 disk SSD configuration, single SSD achieved approximately 250 MBps of read bandwidth, which matches the disk specification (see Figure 5). The same controller limitation behavior is observed in the sequential write bandwidth with SSD disks.

Iometer was set to send a stream of read requests (using various block sizes), maintaining the queue depth at 24 outstanding I/O requests. The results of the sequential access pattern indicate that the data was being read from all 24 disks, leading to high performance that also saturated the controller. With the controller set at Always Read Ahead and Direct I/O, we observed a higher transfer rate on RAID 0.

Figure 5. Sequential Read Bandwidth Comparison



8-disk SSD results are comparable to 15,000- and 10,000-rpm disks due to controller limitations.

The controller registered saturation after enabling 18 disks. Beyond that, the sequential read performance stayed constant with the increasing number of disks. The sequential read bandwidth shows faster response time when Always Read Ahead and Direct I/O mode are selected.

Figure 6. Sequential Write Bandwidth on RAID 0

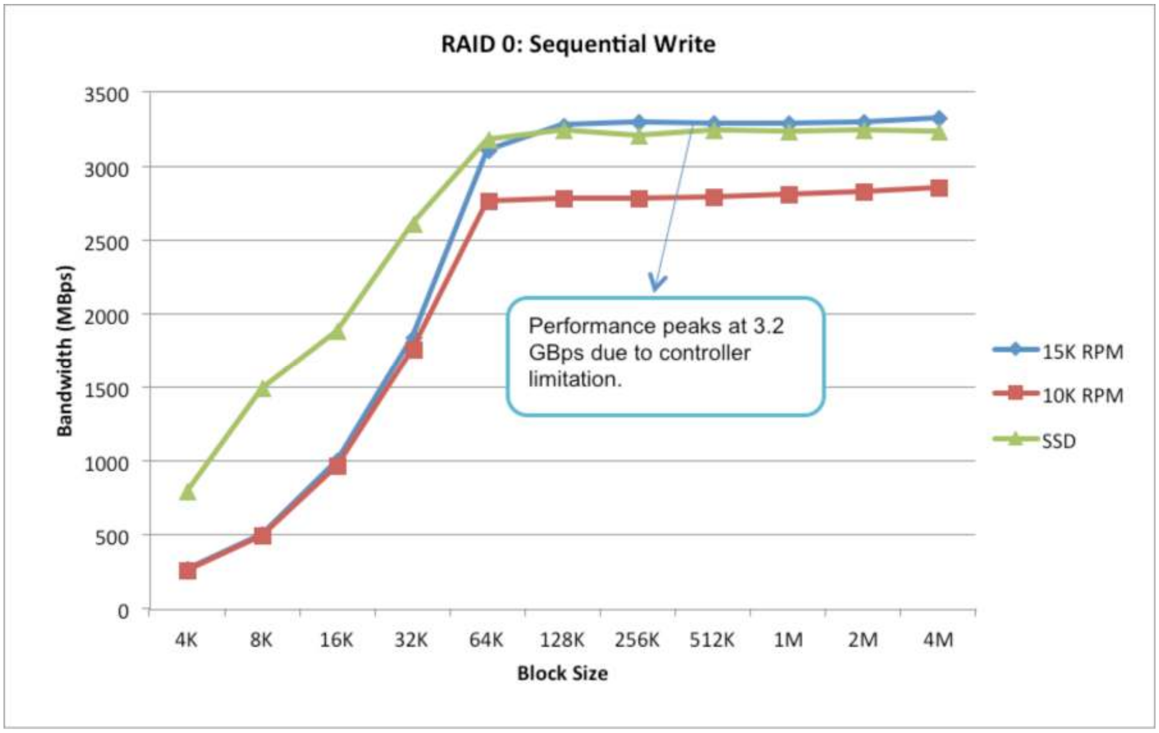


Figure 6 illustrates the maximum sequential write bandwidth achieved by the LSI MegaRAID controller at 3200 MBps, which matches the controller specification. In this test the 15,000-rpm SAS and SSD drives saturated the controller bandwidth. The write throughput remained constant beyond the 64K block size, owing to the saturation of the LSI controller bandwidth.

Figure 7. Sequential Write Bandwidth on RAID 10

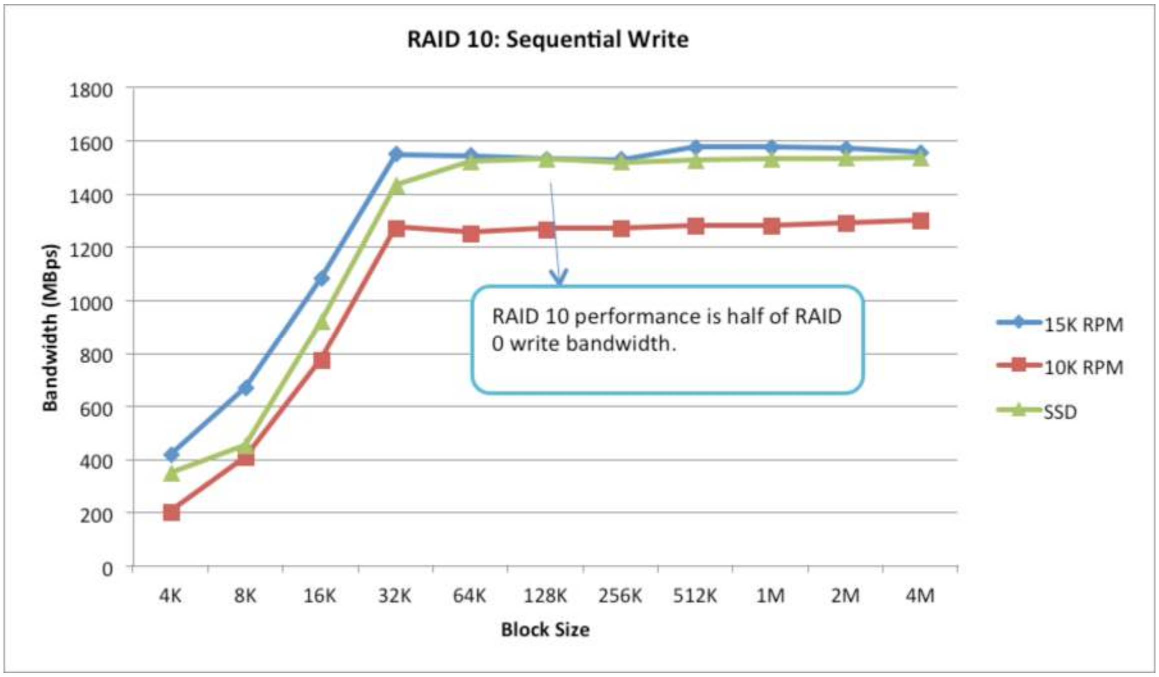


Figure 7 illustrates the RAID 10 sequential write bandwidth performance observed on the controller. RAID 10 bandwidth peaked at 1500 MBps, which is half the bandwidth of RAID 0. This is expected behavior and matches the controller specification.

Data written sequentially floods the cache and makes caching less effective during a sequential write operation. It becomes expensive, as it writes to the cache first and then to the disk. However, with the Direct I/O policy, there is no overhead of caching, and the data directly hits the disk.

Figure 8. Sequential Read Bandwidth on RAID 5 Configuration

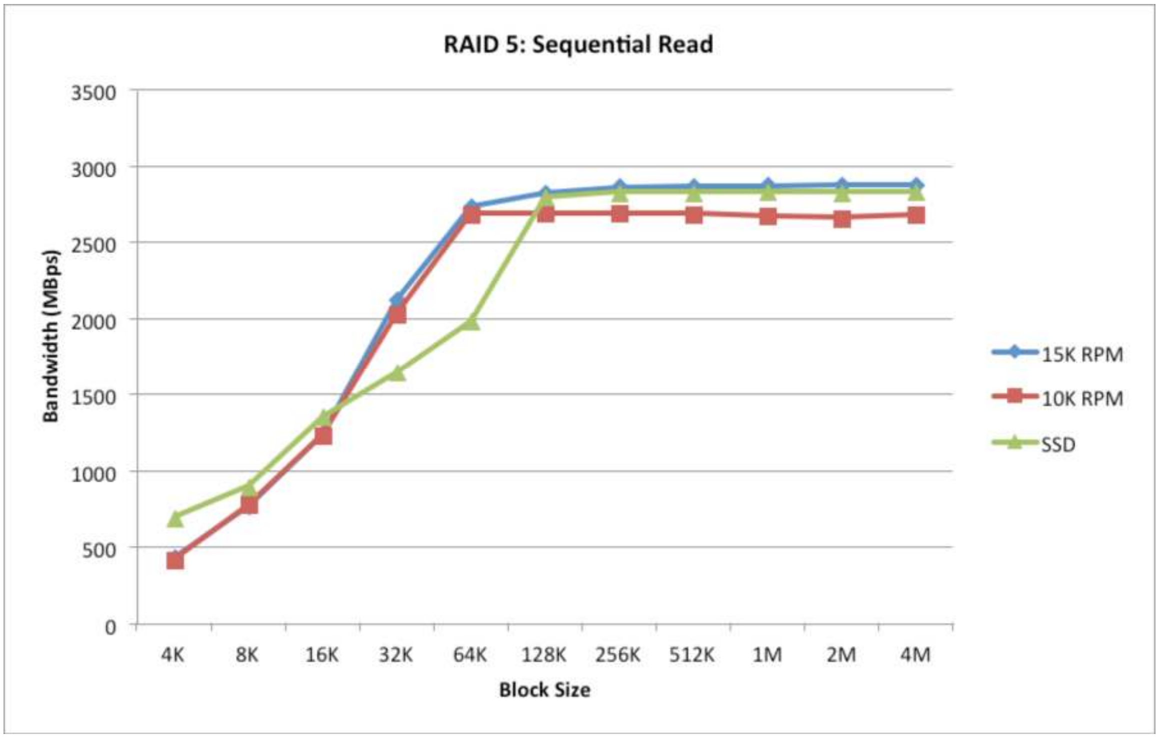


Figure 8 illustrates the RAID 5 sequential read bandwidth performance. A performance peak was observed at 2.8 GBps, which matches the controller specification.

Random IOPS Performance on RAID 0 and RAID 5

This section presents the performance test results and analysis of the random IOPS on RAID 0 and 5.

Figure 9. Random IOPS Performance on RAID 0 and RAID 5

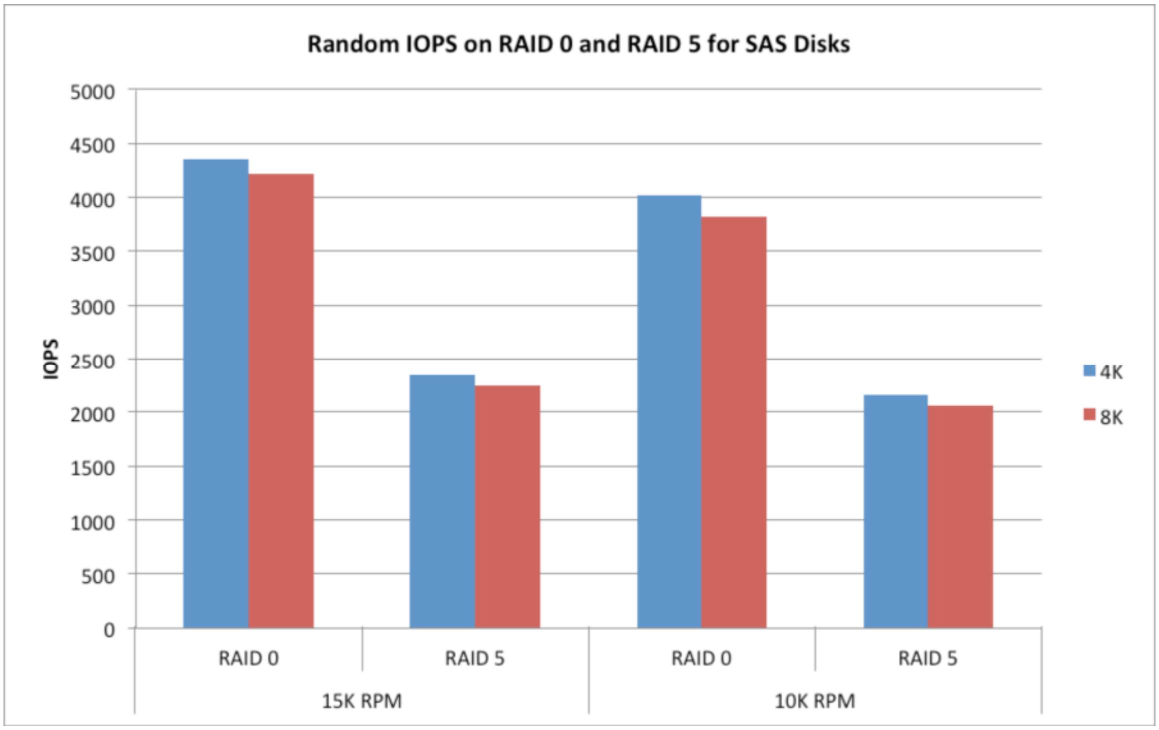


Figure 9 illustrates the maximum IOPS achieved by the 15,000-rpm and 10,000-rpm drives.

Since RAID 0 does not have any I/O overhead (compared to RAID 5), the disks achieved IOPS that matched the specifications under a random I/O workload. The sweet spot was observed to be 48 outstanding I/O requests, achieving optimal IOPS for 15,000-rpm and 10,000-rpm drives. Beyond 48 outstanding I/O requests, the IOPS for the drives was constant, but the response time gradually increased, with increasing queue depth.

The performance testing illustrated that the RAID 5 IOPS values were less than the RAID 0 IOPS values. This behavior is expected, because RAID 5 has four I/O penalties, and writes are expensive in a RAID 5 configuration. To check the raw performance of the disks, we did not use any cache controller settings in these tests.

In RAID 0, all the disks participating in a RAID volume are used in parallel for I/O processing, which results in maximum performance. This also diminishes the chance that any disk in the array will be idle due to outstanding I/O requests. In RAID 5, caching of write I/O is more important than caching of read I/O. When Write Back, which is a write cache optimization mechanism, is enabled, writes are stored in the controller cache while parity is calculated before the data with parity is written back to the disk. A write cache avoids the latency caused by parity calculation, ensuring a faster response time for writes.

Figure 10. Random IOPS Performance in a RAID 0 and RAID 5 Configuration for SSD

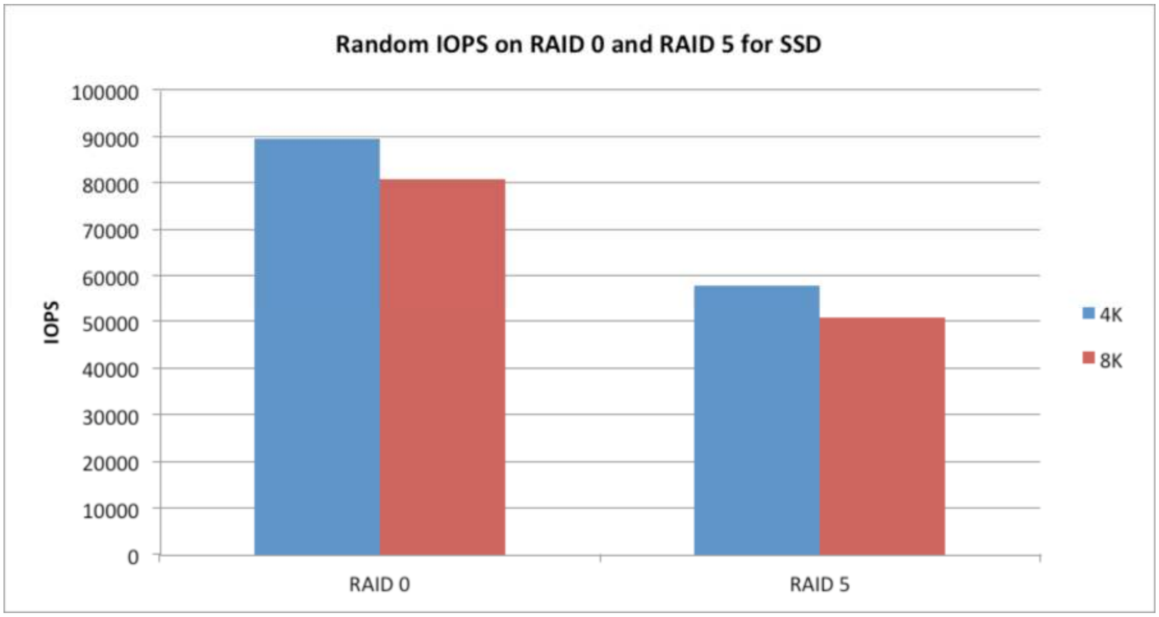


Figure 10 illustrates the SSD OLTP performance on 24-drive RAID 0 and RAID 5 configurations. According to the SSD specification, a 4K random read workload (with 20 percent overprovisioning) should sustain 4000 IOPS per disk. The test results showed IOPS on a random access workload that matched the SSD specification (70R:30W, 4K and 8K block size. Similar results were obtained during the performance testing of the 15,000-rpm and 10,000-rpm SAS drives.

Figure 11. Random IOPS on RAID 5 with BBU Mode on 15,000-rpm Drives

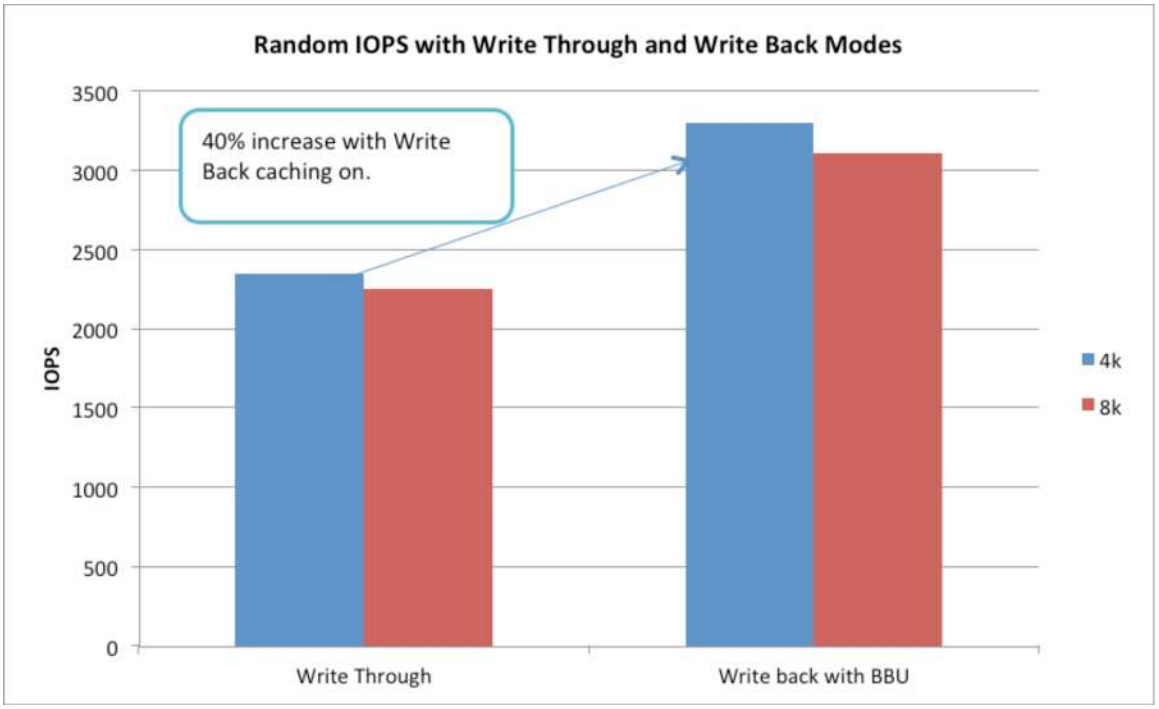


Figure 11 illustrates the difference between Write Through and Write Back with BBU mode set for the controller cache.

The applications must be tuned properly for the random I/O applications storage. Owing to the overhead with RAID 5, writes take a long time, since parity calculation occurs on every drive used in the volume (distributed parity). In a RAID 5 configuration the time taken to read a particular stripe that was not written on each drive is also calculated, which yields a higher response time.

Turning on Write Back mode for RAID 5 in all the OLTP workloads produced an increase in I/O operations that are buffered by the cache. RAID 5 is not recommended for small writes, because it has to read each block on each drive before writing to the disk. The LSI MegaRAID controller has a battery backup cache unit that avoids wait time on disks required to seek data from a particular sector. Enabling Write Back with BBU helps small writes, such as the OLTP transactions.

Figure 12. Random Access Workload with Equal Mix of Read and Write on RAID 5 Configuration

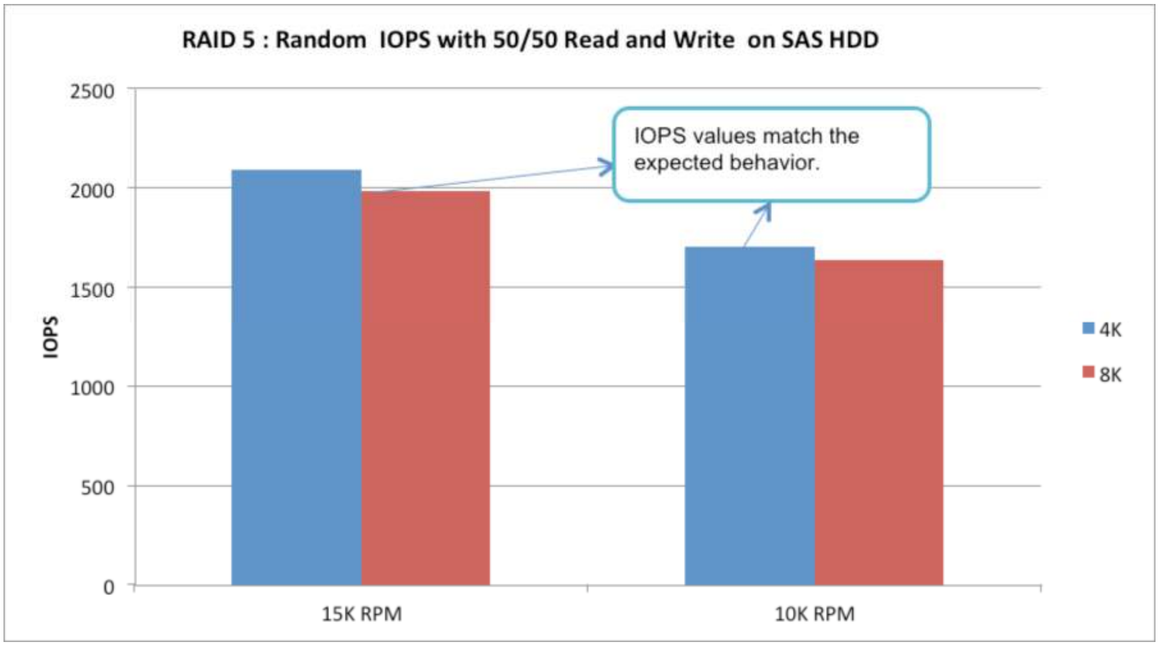
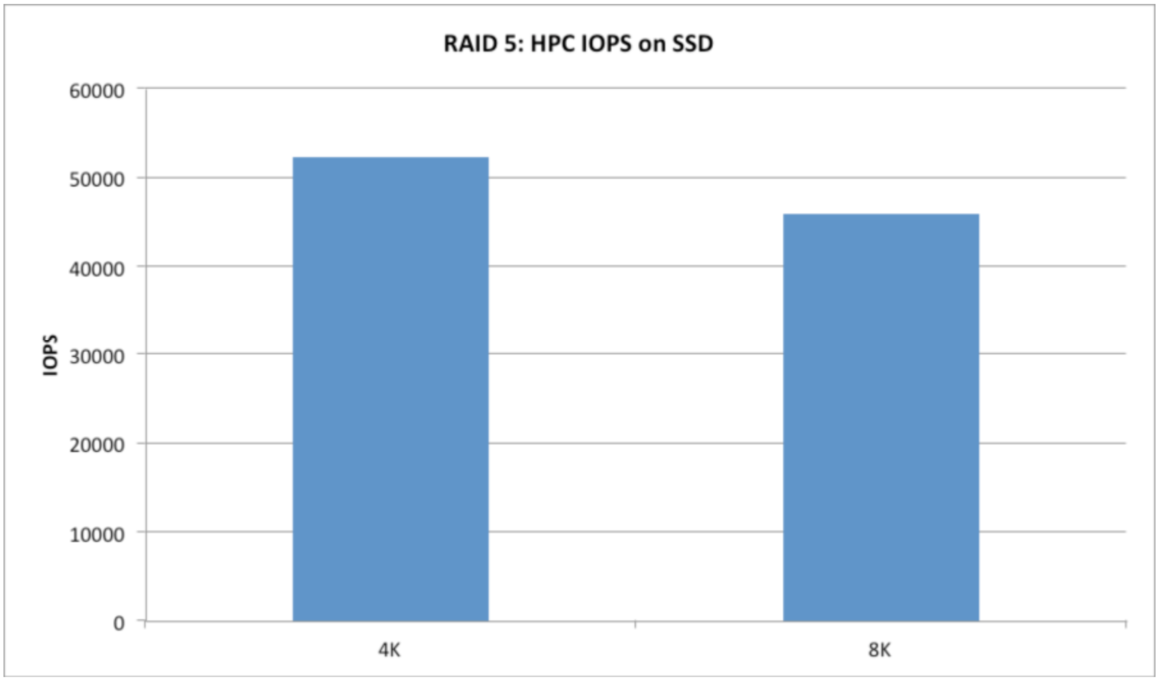


Figure 12 illustrates the RAID 5 performance for a workload similar to a high-performance computing (HPC) workload on the SAS HDD for the 15,000-rpm and 10,000-rpm drives.

Figure 13 illustrates the RAID 5 performance for an HPC workload for 4K and 8K blocks. The IOPS achieved on the SSD with a high-throughput workload is expected from SSD drives when compared to OLTP on RAID 5.

Figure 13. HPC-like Workload for RAID 5 Configuration



LSI MegaRAID (SAS9271CV-8i) Generation 3 Performance

This section presents the performance test results and analysis of the LSI MegaRAID PCIe Generation 3 controllers. It also includes the performance difference between PCIe Generation 2 and Generation 3 controllers.

Figure 14. Sequential Read Performance Using 15,000-rpm SAS HDD

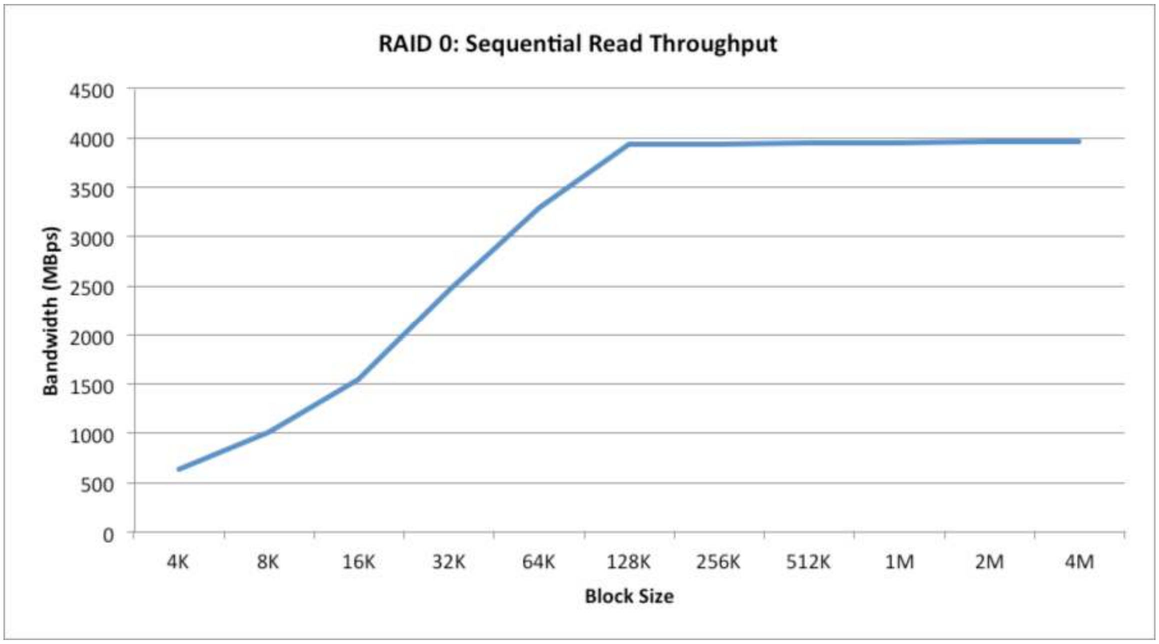


Figure 14 illustrates the sequential read performance of a RAID 0 configuration with 24 15,000-rpm SAS HDDs. The performance peaked at 4000 MBps bandwidth.

As Figure 15 shows, the sequential read bandwidth peaked at 4 GBps, which matches the controller specification. A 42 percent gain in throughput was registered on the Generation 3 LSI MegaRAID controller compared to the Generation 2 MegaRAID controllers. No gain on write throughput in Generation 3 cards was observed during the test.

Figure 15. Sequential Read Bandwidth Comparison

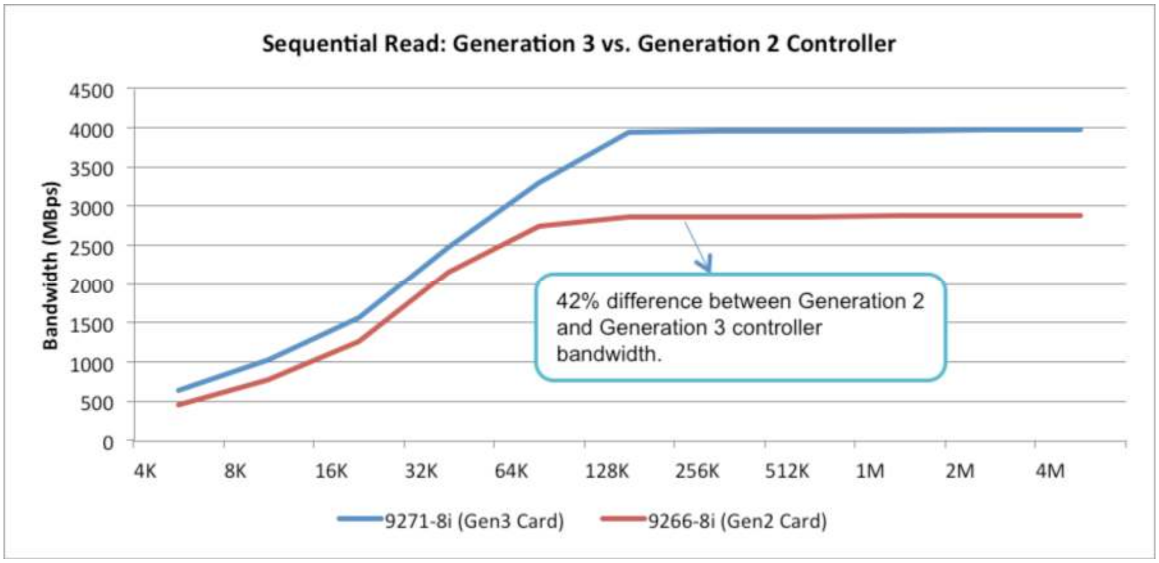


Figure 16. RAID 5 OLTP Performances on Generation 2 and Generation 3 Cards

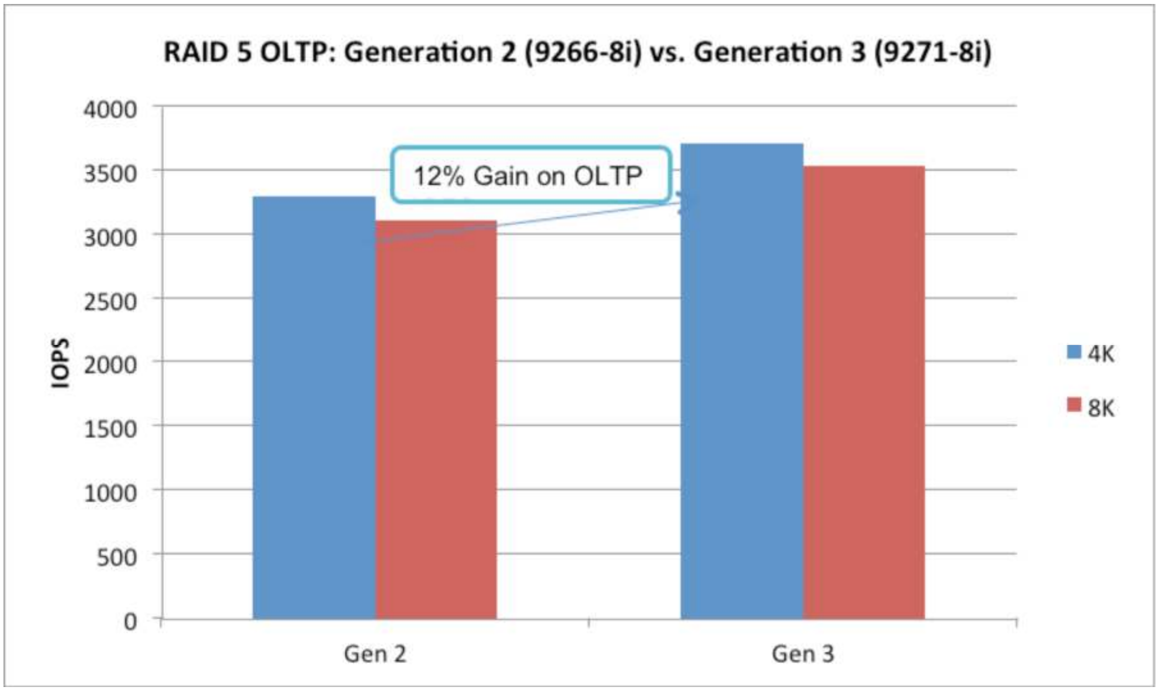


Figure 16 illustrates the RAID 5 OLTP performance on the Generation 2 and Generation 3 cards. With a 24-drive (15,000-rpm SAS) RAID 5 volume, a 12 percent gain was registered on the Generation 3 MegaRAID controller compared to the Generation 2 controller running an OLTP workload. We used Normal Read, Write Back with BBU, Cached I/O, and Disk Cache Enabled controller settings to run the OLTP workload on the RAID configurations.

Since writes are expensive on a RAID 5 configuration, we ensured that Write Back mode was turned on to enhance the OLTP performance.

Embedded RAID vs. Hardware RAID Performance

This section presents the comparative performance results and analysis of the embedded (software) RAID and the controller-based (hardware) RAID.

Figure 17. RAID 5 OLTP IOPS Performance for Embedded RAID vs. Hardware RAID

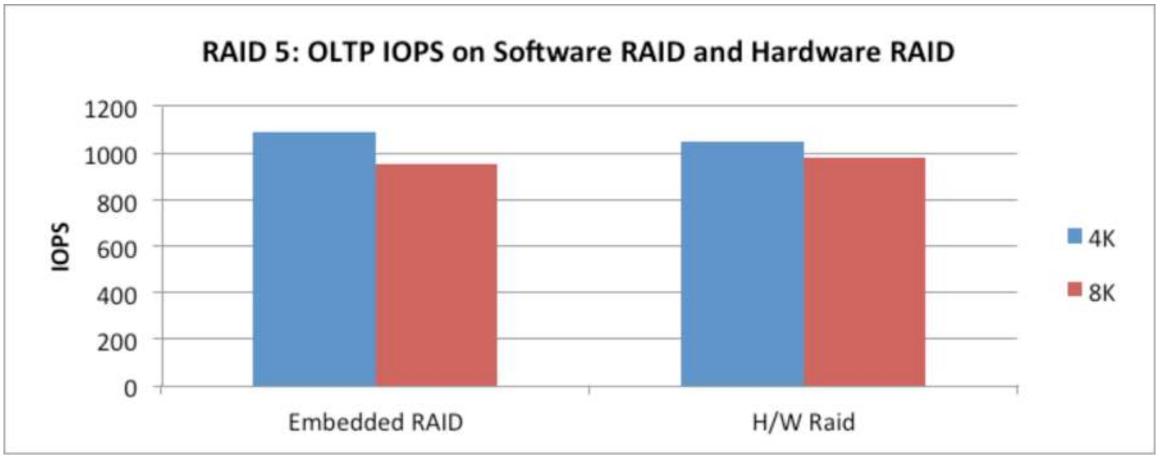


Figure 17 illustrates the OLTP workload performance of the RAID 5 configuration with embedded RAID and controller-driven RAID. Their performance (in terms of IOPS) was similar. This shows that the embedded RAID option can be a viable and cost-effective solution for a smaller number of disks without compromising on performance. Note that caching was disabled on the controller-based RAID for a comparable setup.

Figure 18. RAID 5 OLTP IOPS Performance Comparison for Hardware and Software

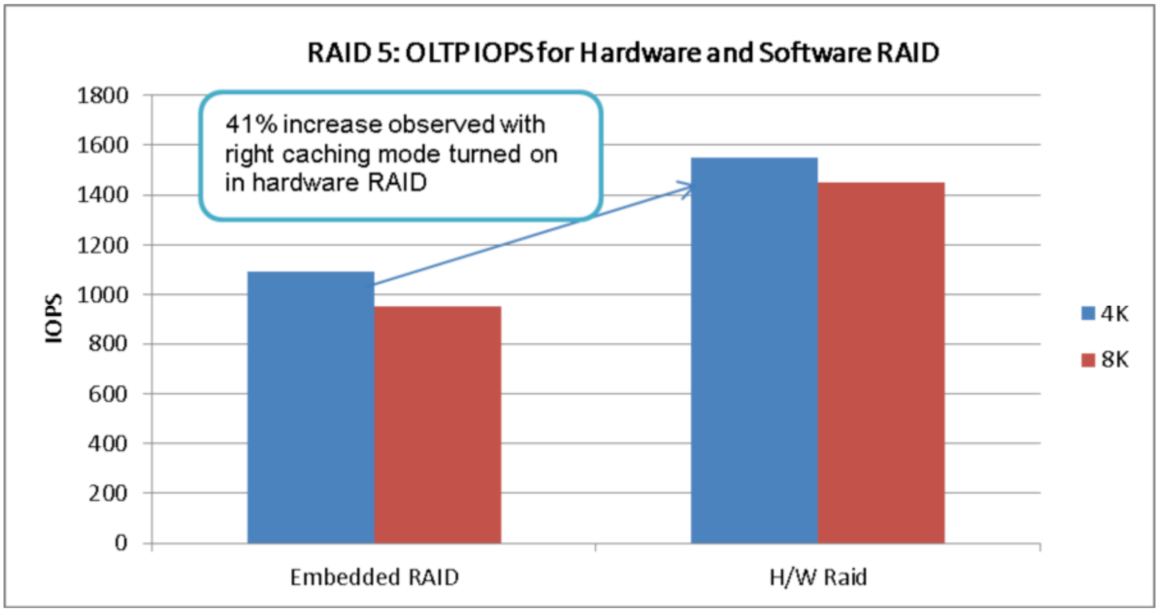


Figure 18 illustrates the OLTP workload performance for RAID 5 between embedded RAID and hardware RAID. Hardware RAID has a significant impact when appropriate caching parameters are used on the controller. A 41 percent boost in IOPS with hardware RAID was recorded when compared to the embedded RAID.

Figure 19. RAID 0 Sequential Read Bandwidth from Eight 15,000-rpm SAS HDDs

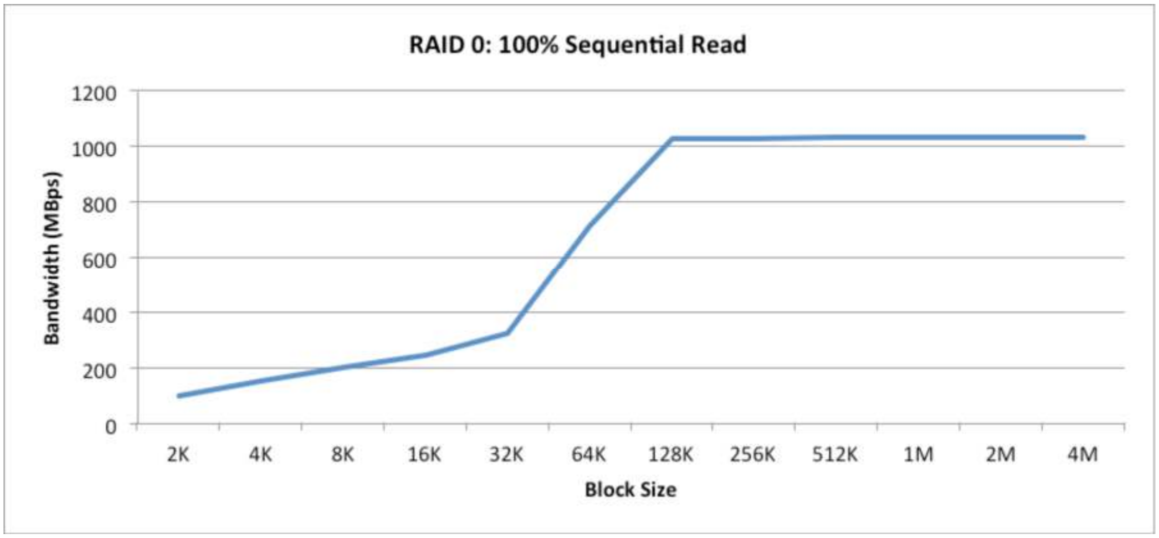


Figure 19 illustrates the maximum read throughput achieved with embedded RAID on the Cisco UCS C-Series Rack Servers. The bandwidth attained per disk (about 132 MBps) meets the disk specification. Read bandwidth on embedded RAID peaked at 1 GBps, which matches the controller specification.

PHASE II

Test Plan

The main objective of Phase II was to benchmark the SAS and SSD performance on the Cisco UCS C240 M3 server with patterns that simulate different application workloads. The I/O validation on the Cisco UCS C240 M3 server shows how small and medium-sized businesses can take optimum advantage of the internal storage for hosting applications. In this test, the IOPS (random workloads), bandwidth (sequential workloads), and response time were captured using Iometer.

The IOmeter benchmark was used to gauge each storage solution that handles a variety of storage application workloads. Table 15 shows the access pattern for each associated application. This benchmark activity was performed only on the Cisco UCS C240 M3 servers, since this server supports the maximum number of disks possible in the C-Series system.

Table 15. Benchmark Profiles Resembling Application Workloads Created in Iometer

Application Profile	RAID Type	Mode	Read: Write	Block Size	Stripe Size	Metric
Online Transaction Processing (OLTP)	5	Random	70R, 30W	8 K	64 K	IOPS and response time
Decision support system, business intelligence, video on demand	5	Sequential	100R, 0W	512 K	1 MB	Transfer rate
Database logging	10	Sequential	0R, 100W	64 K	64 K	Transfer rate
High-performance computing (HPC)	5	Random	50R 50W	64 K	1 MB	IOPS and response time
	5	Sequential	50R, 50W	1 MB	1 MB	Transfer rate
Digital video surveillance	10	Sequential	10R, 90W	512 K	1 MB	Transfer rate
Hadoop	0	Random	60R, 40W	64 K	64 K	IOPS and response time
Virtual desktop infrastructure (VDI) (boot process)	5	Random	80R, 20W	64 K	64 K	IOPS and response time
Virtual desktop infrastructure (VDI) (steady state)	5	Random	20R, 80W	64 K	1 MB	IOPS and response time

Benchmark Configuration

In this benchmark activity, 24 15,000-rpm SAS drives were used on a Cisco UCS C240 M3 server connected to an LSI MegaRAID 9266-8i controller. Microsoft Windows 2008 R2 Server was installed on the Cisco UCS C240 M3 server after performing a SAN boot. This configuration ensured that all 24 drives were connected through the LSI MegaRAID controller to the drive's internal storage.

Controller Configuration

This section describes the LSI MegaRAID controller configuration settings used for various application workloads in this performance test.

Table 16 describes the LSI MegaRAID controller settings for the various application workloads.

Table 16. LSI MegaRAID Controller Settings for Application Workloads

RAID Type	Controller Settings
10	Read policy: Always Read Ahead
	Write Through: For streaming sequential performance
	I/O policy: Direct I/O
	Disk Cache policy: Enabled
5	Write policy: Write Back with BBU
	Read policy: Normal Read
	I/O policy: Cached
	Disk Cache policy: Enabled
	Full initialization

Iometer Settings

The following Iometer settings were used in this performance testing:

- 70 percent of the volume was used as a dataset.
- Outstanding I/O requests were tuned for each benchmarking profile to attain correct numbers as per the disk and controller specifications.
- Run time is 30 minutes.
- Ramp-up time is 120 seconds.

Performance Results

The performance results listed in this section illustrate the test and inference based on a 24-drive RAID configuration with various application data sizes and LSI MegaRAID caching combinations. The best test results with correct caching combinations on different RAID subsystems are plotted and illustrated in subsequent tables and graphs. Table 17 lists the component details of the measuring environment in this performance test.

Table 17. Measuring Environment

Component	Details
Server	Cisco UCS C240 M3
LSI MegaRAID 9261-8i controller	Firmware version: 3.220.75-2196
Operating system	Microsoft Windows 2008 R2—Data Center version
Hard disk SAS, 2½ in., 15,000 rpm	300 GB (15,000 rpm)
Device type	Raw device
Test data	24-disk volume

Figure 20. IOPS from Random Workloads on RAID 5 Configurations

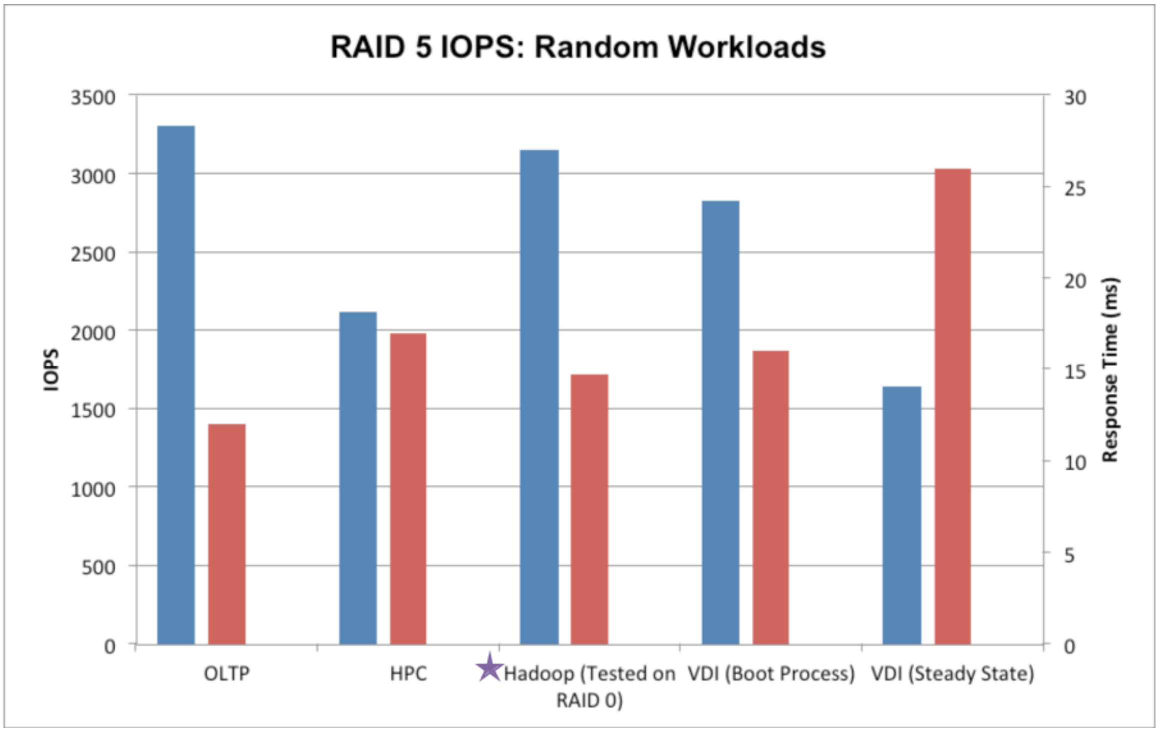


Figure 20 illustrates the application workload performance on a 24-drive (15,000-rpm SAS HDD) RAID 5 volume. See Table 15 for the block size, stripe size, and RAID levels used in the application workloads tested.

Iometer was set to send read and write requests by maintaining a queue depth of 48 requests on a 24-drive RAID 5 volume for various application workloads.

The data center workloads were either random or streaming, and random access was associated with some of the critical applications, such as database, VDI, and HPC.

OLTP systems contain the operational data to control and run some of the most important business tasks. These systems are characterized by their ability to complete various concurrent database transactions and process real-time data. They are designed to provide optimal data processing speed. OLTP systems are often decentralized to avoid single points of failure. Spreading the work over multiple servers can also maximize transaction processing volume and minimize response times.

In a typical OLTP transactional workload that is random in nature, with 70 percent read and 30 percent write (8K block size, 64K stripe size), RAID 5 is preferred. Figure 20 illustrates that the OLTP workload yielded 3300 IOPS and achieved a 10-ms response time from a 24-drive RAID 5 volume. The full initialization for a RAID 5 volume is done to ensure that no background verification occurs when a host tries to write to the RAID 5 volume. Enabling disk cache, controller cache, and write back cache ensured that the small random write was buffered in the controller cache, resulting in maximum IOPS. In RAID 5, response time is on the high side, since the longer wait queries increase response time for any I/O device. To minimize slower response time, Cisco recommends enabling Write Back and controller caching, so that all writes are buffered in the controller cache. However, caution should be exercised to enable BBU for those applications that demand data consistency and durability (such as database devices).

The **Hadoop** framework transparently provides both reliability and data motion to applications. Hadoop implements a computational paradigm named MapReduce, in which the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system that stores data on the computing nodes, providing very high aggregate bandwidth across the cluster.

Figure 20 illustrates the best IOPS obtained from a Hadoop workload on a 24-drive RAID 0 volume (64 K block size, 64 K stripe size, RAID 0). In a Hadoop environment, normally a RAID 0 configuration is given priority over RAID 5 because in Hadoop data is replicated thrice for data redundancy. Need for disk level redundancy is not necessary in Hadoop environment and having RAID 0 gives better performance compare to RAID 5. Cisco recommends using RAID 0 configurations for Hadoop workloads.

HPC refers to cluster-based computing, which uses many individual, connected nodes that work in parallel in order to reduce the amount of time required to process large datasets that would otherwise take exponentially longer to run on any one system. Figure 20 illustrates the HPC IOPS and response time on a 24-drive RAID 5 volume. The HPC workload performs a lot of in-memory calculation, so we recommend enabling the Write Back and controller cache policies. These settings are advisable because more random parallel I/O requests can be processed. I/O requests are buffered in the controller cache. Any data that is requested by the user is read from the cache and not from the disks. This results in increased HPC performance.

Virtual desktop infrastructure (VDI) is a desktop-centric service that hosts users’ desktop environments on remote servers and/or rack servers, which are accessed over a network using a remote display protocol. A connection brokering service is used to connect users to their assigned desktop sessions. For users, this means they can access their desktop from any location, without being tied to a single client device. Since the resources are centralized, users moving between work locations can still access the same desktop environment with their applications and data.

Figure 20 illustrates the VDI workload performance in terms of IOPS with a RAID 5 configuration. VDI workloads are characterized by two scenarios, one at boot storm and the other during steady state. Reads in a boot process are typically heavier (80 percent read, 20 percent write, 64K stripe size, 64K block size), because the OS image is read from a file, requiring a high IOPS during an I/O storm. Enabling the controller cache allows the boot images to be read from the cache, which enhances the VDI performance during a boot storm.

During steady state the VDI workload flips, and there are more writes than reads. To accommodate bursty writes, the Write Back with BBU caching policy should be enabled, so that the maximum number of write operations is buffered in the controller cache before being written to the disks. Ensure that the disk cache is enabled to increase throughput for write operations.

Figure 21. Transfer Rate from Sequential Operations on RAID 10 Configurations

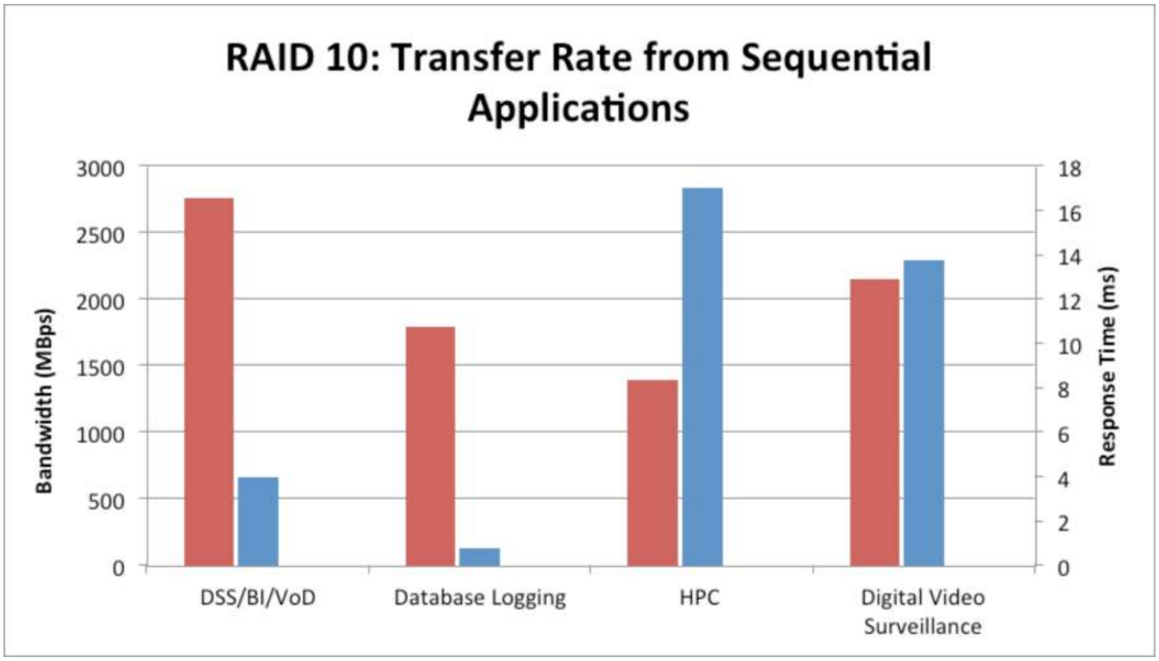


Figure 21 illustrates the maximum throughput achieved by the sequential applications in RAID 10 configurations. See Table 15 for the block size, stripe size, and RAID levels used in the application workload performance data.

Iometer was set to send sequential read and write requests by maintaining a queue depth of 24 requests on a 24-drive RAID 5 volume for various sequential application workloads. The MegaRAID settings listed in Table 18 ensured that the maximum performance was achieved.

Table 18. Recommended Controller Settings for Sequential Workloads

RAID Type	Controller Settings
5 and 10	Read policy: Always Read Ahead
	Write Through: For streaming sequential performance
	I/O policy: Direct I/O, I/O Cached (for RAID 5)
	Disk Cache policy: Enabled
	Read policy: Normal Read

Decision support systems, business intelligence, and video on demand have similar I/O characteristics, and their performance is also comparable to a sequential read workload. The RAID 5 volume was selected to run the I/O benchmarks because RAID 5 performs well with read operations.

Decision support system (DSS) applications are designed to help make decisions based on data that is picked from a wide range of sources. DSS applications are not single information resources, such as databases or programs that graphically represent sales figures, but involve integrated resources working together.

Business intelligence (BI) is a set of theories, methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information for business purposes. BI technologies provide historical, current, and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics, and prescriptive analytics.

Video on demand (VoD) consists of systems that allow users to select and watch or listen to video or audio content on demand. IPTV technology, often used to bring video on demand to televisions, is a form of video on demand. Television VoD systems either stream content through a set-top box, a computer, or other device, allowing viewing in real time, or download it to a device such as a computer, digital video recorder (also called a personal video recorder), or portable media player for viewing at any time.

Figure 21 illustrates the results from the sequential access pattern, which indicates that the data is read from all 24 drives, resulting in high performance. Performance numbers registered in this experiment illustrate that the controller peaked at 2.8 GBps, which is the maximum controller limit. The highest bandwidth during the testing was attained using the Always Read Ahead and Direct I/O options. Enabling the Write Back option for streaming applications reduced the performance, because data read and written sequentially floods the cache and makes caching less effective. When Always Read Ahead mode is enabled, the subsequent read I/O request is kept in the buffer, increasing the sequential read performance. In a VoD workload, since the request size is larger than 512 K, the largest stripe size is recommended, for better performance.

In the field of databases in computer science, a transaction log (also transaction journal, database log, binary log, or audit trail) is a history of actions executed by a database management system to guarantee atomicity, consistency, isolation, and durability (ACID) properties after crashes or hardware failures. Physically, a log is a file of updates done to the database, stored in stable storage.

Figure 21 illustrates the maximum throughput achieved by a database logging workload on a 24-drive RAID 10 configuration. The database log file, which is a write-intensive application with a large stripe size and no caching effect, gives more throughput and good performance on a RAID 10 subsystem. Cisco recommends having a larger stripe size to yield better throughput performance on RAID 10 applications. RAID 0 gives the best throughput among all the RAID options, but it is not recommended for hosting applications, since it is not a redundant subsystem. Cisco also recommends controller settings of Normal Read, Write Through, and Direct I/O for sequential write applications on RAID 10, for better performance and better throughput. In Direct I/O mode, data is moved directly from the host to the disks, avoiding copying data into the cache and resulting in improved overall performance for streaming workloads.

HPC workloads are compute-intensive and typically network IO-intensive. As such, they require top-bin CPU components and high-speed, low-latency network fabrics for their MPI (message passing interface) connections. Compute clusters consist of a head node that provides a single point from which to administer, deploy, monitor, and manage the cluster, as well as an internal workload management component known as the scheduler, which manages all incoming work items, known as jobs. HPC workloads typically require large numbers of nodes with nonblocking MPI networks in order to scale. Scalability of nodes is the single biggest factor in determining the realized usable performance of a cluster.

Figure 21 illustrates that the maximum throughput was achieved by an HPC workload in a RAID 5 environment. Since an HPC workload requires a large amount of I/O bandwidth, Cisco recommends enabling the controller cache, even though Write Back mode is turned on. This is because the read/write ratio of an HPC workload is

50:50, and enabling the controller cache allows it to accommodate enough streaming writes on RAID 5. It is also recommended that you use the largest stripe size, to achieve maximum performance in a RAID 5 configuration. Throughput achieved in this workload is expected behavior in a RAID 5 configuration.

Digital video surveillance is an appliance that enables embedded image capture capabilities so that video images or extracted information can be compressed, stored, or transmitted over communications networks or digital data links. Digital video surveillance systems are used for any type of monitoring.

Figure 21 illustrates the throughput achieved for digital video surveillance in a RAID 10 configuration (for 24 drives with 15,000-rpm SAS HDDs). Using the Normal Read and Direct I/O options achieved the highest bandwidth in this performance test. The sequential write workload on the controller peaked at 2.1 GBps in a RAID 10 configuration, which is the expected behavior when compared to RAID 0 throughputs (see Figure 4. Since this is a sequential write workload, it is recommended that you define the largest stripe size to yield maximum throughput. Enabling Direct I/O for a sequential workload on the controller ensures that there is no cache I/O overhead and that data is written directly to the disks. Enabling Disk Cache on the hard drives improves the overall throughput.

Conclusion

The results presented in this document demonstrate that the Cisco C-Series Rack Servers are capable of driving various storage I/O workloads and threaded configurations with LSI MegaRAID controllers. This study also demonstrates that the Cisco UCS C-Series and LSI MegaRAID technology meet the performance and scalability demands posed by enterprise transactional applications.

To achieve high performance, an understanding of storage controllers and how host applications generate I/O is essential. The workload could be simple, resulting from a single application, or it could be complex, generated by multiple applications with different I/O profiles. To determine the expected performance, one must understand throughput, bandwidth, and required response time for I/O. The Cisco UCS C-Series servers have internal disk drives that have CPU, memory (cache), and I/O resources that need to be managed and provisioned for optimal performance and highest availability. The built-in reliability of the Cisco Unified Computing System with redundant hardware (LSI MegaRAID controller) makes it highly redundant in terms of data protection and performance.

The performance results confirmed that achieving optimum performance and throughput on the Cisco UCS C-Series Rack Servers involves following the best practices for the various RAID levels and for different applications (random and sequential) and employing the correct caching levels on the controller.

References

The following references have been used in this white paper:

- Seagate Savvio 15K.3 data sheet: www.seagate.com/files/www-content/product-content/savvio-fam/savvio-15k/savvio-15k-3/en-us/docs/savvio-15k-3-data-sheet-ds1732-5-1201gb.pdf
- Seagate Savvio 10K.5 data sheet: www.seagate.com/files/staticfiles/docs/pdf/datasheet/disc/savvio10k5-fips-data-sheet-ds1727-4-1201-us.pdf
- Intel Solid-State Drive 710 Series Product Specification: www.intel.com/content/dam/www/public/us/en/documents/product-specifications/ssd-710-series-specification.pdf
- Transactional Database Processing with the Intel SSD 720 Series: http://www.intel.com/content/www/us/en/solid-state-drives/ssd-710-transactional-database-brief.html
- LSI MegaRAID Controller Benchmark Tips: www.lsi.com/downloads/Public/Direct Assets/LSI/Benchmark_Tips.pdf

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C11-728200-00 10/13