Best Practices

Best Practices for VMware® ESX 3.5 and the LSI CTS2600 Storage System

August 16, 2010

Table of Contents

Chapter 1

Overview of the LSI CTS2600 Configurable Storage Components and VMware ESX 3.5

Many businesses and enterprises have implemented VMware® or have plans to implement VMware. VMware provides more efficient use of assets and lower costs by consolidating servers. Applications that previously had been running in under-utilized dedicated physical servers are migrated to their own virtual machine or virtual server that is part of a VMware ESX cluster or a virtual infrastructure.

As part of this consolidation, asset use can increase from a typical 10 percent to at least 85 percent. Applications that previously had dedicated internal storage now use a shared or networked storage system that serves storage to all of the virtual machines and their applications. As a result of this server and storage consolidation, many VMware customers find that the storage demand shifts from the physical server to the storage system or network. The storage system must deliver balanced performance and high performance in support of these multiple consolidated applications and servers.

LSI CTS2600 Configurable Storage Components are designed to deliver reliable performance for mixed applications including transaction and sequential workloads. These workloads include applications that are typical of a virtual infrastructure including email, database, webserver, data warehouse, and backup profiles. LSI offers a complete line of storage systems from entry-level systems to mid-range systems to enterprise-level systems that are certified to work with VMware ESX. The LSI CTS2600 Configurable Storage Components offer as an optional premium feature Snapshot, Volume Copy and Remote Volume Mirroring (across FC host ports).

The following items describe the storage systems available from LSI.

- LSI CTS2600 Configurable Storage Components is an entry-level storage system that is easy to use and easy to configure and can be direct attached or connected through a network.

Purpose of This Document

This document describes the optimum performance settings for the LSI CTS2600 Configurable Storage Components with VMware ESX 3.5 operating environment in support of virtual machines.

This document identifies parameters for optimizing a high-performance storage system. For each parameter, this document explains how to monitor, evaluate, adjust, and make sure that the adjustment was appropriate and positive. The process of keeping the parameters tuned involves the following tasks:

- Identify the relevant parameters.
- Take a baseline to determine the benchmark value for each relevant parameter.
- Continuously monitor each parameter on an ongoing basis. Only *continuous* monitoring can isolate the triggers that impact performance. Also, continue monitoring after any adjustment so that the effectiveness of the adjustment can be evaluated.
- Adjust parameters while the system remains in production.
- Watch how adjustments in one parameter are affecting other parameters.

This document provides important information about how to tune for optimum performance when using LSI CTS2600 Configurable Storage Components with the VMware ESX 3.5. You also must be familiar with your specific operating system and any additional applications running in your VMware ESX 3.5 environment to get the most performance out of your system.

Disclaimer

Because of the highly customizable nature of a VMware ESX 3.5 environment, you must take into consideration your specific environment and equipment to achieve optimal performance from an LSI CTS2600 Configurable Storage Components storage system. When weighing the recommendations in this document, start with the first principles of I/O performance tuning:

- There are no absolute answers. Each environment is unique and the correct settings depend on the unique goals, configuration, and demands for the specific environment.
- Results vary widely because conditions vary widely.

IMPORTANT Attempt the procedures within this document only if you are a trained storage specialist with intimate knowledge of the working environment.

Chapter 2

Planning the Design of the LSI CTS2600 Configurable Storage Components

Before you start any configuration of the LSI CTS2600 Configurable Storage Components, you must understand the following concepts to guide you in your planning.

- Recognizing the LSI CTS2600 storage system feature set
- Balancing drive-side performance
- Understanding the segment size of volumes
- Knowing about storage system cache improvements
- Comprehending file system alignment
- Knowing how to allocate volumes for ESX 3.5
- Recognizing server hardware architecture
- Identifying specific ESX 3.5 settings

The following sections in this chapter assist you in planning for the optimal design of your implementation.

Basing the Segment Size on File I/O Operations

Base the segment size on the type of data and on the expected I/O size of the data. Store sequentially read data on volumes with small segment sizes and with dynamic prefetch enabled to dynamically read-ahead blocks. For the procedure for setting up the appropriate disk segment size, see "Calculating Optimal Segment Size" on page 3-2.

Oracle

Very little I/O from Oracle is truly sequential in nature except for processing redo logs and archive logs. Oracle can read a full table scan all over the drive. Oracle calls this type of read a *scattered read*. Oracle's sequential data read is for accessing a single index entry or a single piece of data. Use small segment sizes for an OLTP with little or no need for a read-ahead data. Use larger segment sizes for a Decision Support System (DSS) environment where you are doing full table scans through a data warehouse.

Remember three important things when considering block size:

- Set the database block size lower than or equal to the drive segment size. If the segment size is set at 2 KB and the database block size is set at 4 KB, this procedure takes two I/O operations to fill the block, resulting in performance degradation.
- Make sure that the segment size is an even multiple of the database block size. This practice prevents partial I/O operations from filling the block.
- Set the parameter `db_file_multiblock_read_count` appropriately. Normally you want to set the `db_file_multiblock_read_count` as shown:

  segment size = `db_file_multiblock_read_count * DB_BLOCK_SIZE`

You also can set the `db_file_multiblock_read_count` so that the result of the previous calculation is smaller but in even multiples of the segment size.

For example, if you have a segment size of 64 KB and a block size of 8 KB, you can set the `db_file_multiblock_read_count` to 4, which equals a value of

32 KB—an even multiple of the 64 KB segment size.

SQL Server

For SQL Server, the page size is fixed at 8 KB. SQL Server uses an extent size of 64 KB (eight 8-KB contiguous pages). For this reason, set the segment size to 64 KB. See "Calculating Optimal Segment Size" on page 3-2.

Exchange Server

Set the segment size to 64 KB or multiples of 64. See "Calculating Optimal Segment Size" on page 3-2.

Improvements in Cache

There are two improvements for drive cache included in the LSI CTS2600 Configurable Storage Components feature set that are worth describing. These improvements are the permanent cache back-up and the cache mirroring.

The new permanent cache back-up provides a cache hold-up and de-staging feature to remove cache and processor memory to a permanent device. This feature replaces the reliance on batteries to keep the cache alive for a period of time when power is interrupted.

Drive cache has permanent data retention in a power outage. This function is accomplished through the use of USB flash drives. The batteries must power the cache only until data in the cache is written to the USB flash drives before the drive cache powers down. When the LSI CTS2600 storage system is powered back up, the contents are re-loaded to cache and flushed back to the volume.

When you turn off an LSI CTS2600 storage system, the storage system does not shut down immediately because the storage system uses USB flash drives for cache. The storage system writes the contents of cache to the USB modules. Depending on the amount of cache, the LSI CTS2600 storage system takes up to several minutes to actually power off. Cache upgrades include both DIMMs and USB modules.

The dedicated cache mirroring system is new for the LSI CTS2600 storage system and is implemented to improve the performance of the storage system when cache mirroring is enabled. When cache mirroring is enabled, there is no impact to performance.

Enabling Cache Settings

Always enable read cache. Enabling read cache allows the controllers to service reads from cache for any additional read requests to the data stored within the cache.

Enable write cache to let the controllers acknowledge writes as soon as the data reaches the cache instead of waiting for the data to be written to the physical media. For other storage systems, a trade-off exists between data integrity and speed. LSI CTS2600 Configurable Storage Components were designed to store data on both controller caches before being acknowledged. To protect data integrity, cache mirroring must be enabled to permit for dual controller cache writes.

The LSI CTS2600 storage system has a cache battery backup, which can alleviate some of the need for write cache mirroring because the data in cache is protected from loss of power. Disabling write cache mirroring can provide an increase in performance.

Whether you need to prefetch cache depends on the type of data that is stored on the volumes and how that data is accessed. If the data is accessed randomly (by way of table spaces and indexes), disable prefetch. Disabling prefetch prevents the controllers from reading ahead segments of data that most likely will not be used—unless your volume segment size is smaller than the data read size requested.

Aligning File System Partitions

Align partitions to stripe width. Calculate stripe width by the following formula:

```
segment_size / blocks * num_disks
```

In this formula, 4+1 RAID5 with 512 KB segment equals 512KB / 512 * 4 = 4096.

Here is the command line syntax used for ORION for the test comparisons:

```
orion -run advanced -testname rd5_v1 -num_disks 10 -size_small 8 -
size_large 1024 -type rand -write 10 -duration 60 -simulate concat -
matrix basic -cache_size 0
```

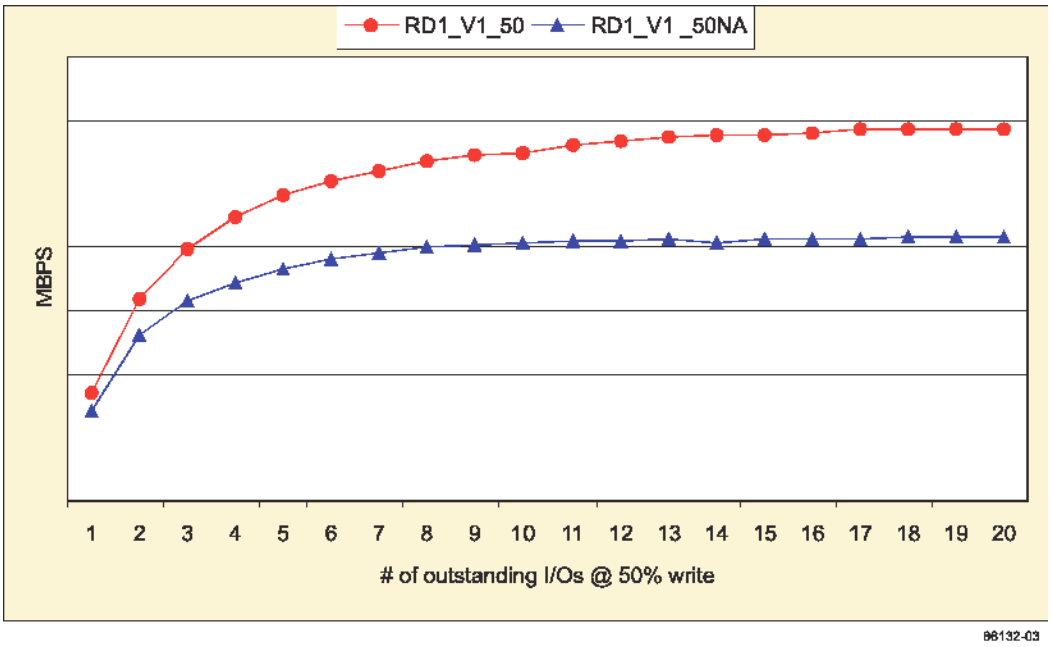

Figure 1 Comparing the Performance of Aligned RAID1 to Non-Aligned RAID1

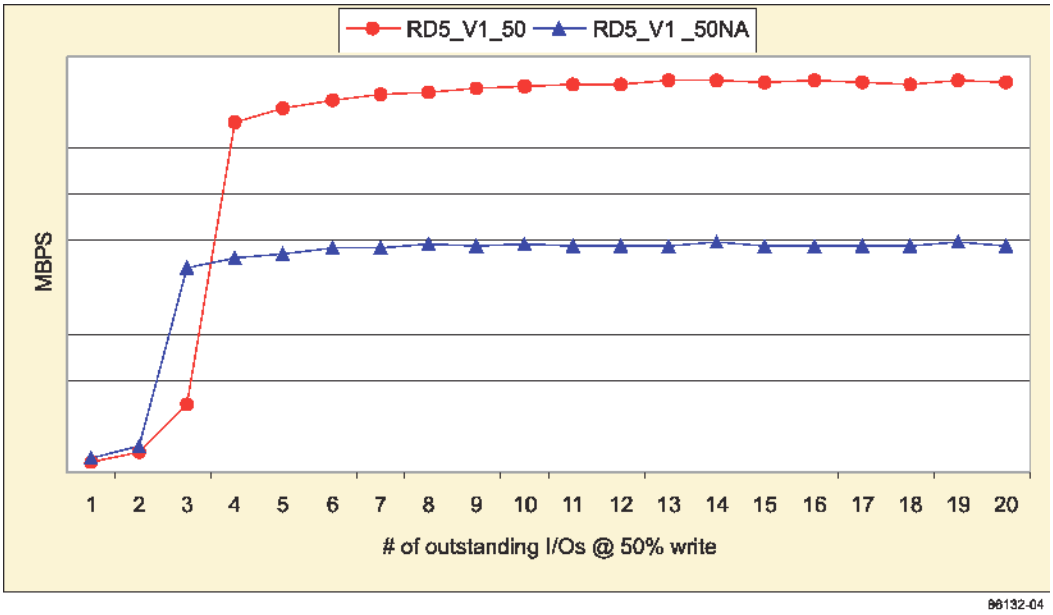

Figure 2 Comparing the Performance of Aligned RAID5 to Non-Aligned RAID5

Laying Out Volumes and Drives

When tuning for optimal performance, consider the specific applications being used. First, identify certain important characteristics of the applications. Then, based upon those characteristics, choose an appropriate RAID level, choose the number of drives to put in a volume group, and set the cache parameters.

Deciding on an VMware Design Strategy

The reasons for implementing a virtual environment can vary, depending on the requirements for optimal performance. Some virtualization efforts focus on consolidation of large quantities of significantly under-utilized servers. IT staff with

consolidation as a primary goal are interested in reducing power, cooling, space, hardware, and support footprints in the data center.

Other environments pursue advanced objectives, such as migrating critical business applications into the virtual environment. For IT staff moving critical applications into a virtual environment, the IT staff wants to simplify management and increase availability while maintaining high levels of storage performance in support of key business applications.

VMware recommends separating application data from the guest operating system boot disk. You can select from two types of placement of application data:

- A VMware Virtual Disk file (VMDK) residing on a Virtual Machine File System (VMFS)
- A raw disk device called Raw Device Mapping (RDM)

The key difference between the two choices is the elimination of the VMFS layer. Eliminating the additional layer might lead to improved I/O performance, but eliminating that layer permits the use of SAN replication features. The question to ask is when can you use a VMDK file and when can you use a raw device?

A presentation to VMware User Group provided a general recommendation for when to use VMDK files and when you can use an RDM device.[1] For VMDK files less than 250 GB, the VMFS, (where the VMDK file resides), must be between 300 GB and 500 GB in size. For application data disks greater than 250 GB, you have several options to consider. The RDM device is one of those options.

If you want to consolidate under-utilized servers, create separate VMFSes to store application data. You can group the application's virtual disk files of these servers until the VMFS is relatively full. You might anticipate supporting between five and ten virtual machines. The quantity of virtual machines actually supported by the VMFS depends on the size of the volume and the storage performance requirements of the individual virtual machines.

As you add in critical business or higher performance applications, you must think about storage performance. You can create dedicated VMFS file systems for an application’s data disk or you can use a dedicated raw device manager (RDM) for the application’s data.

Adding Premium Features

Premium features, like Snapshot and Volume Copy, are available for both the virtual disk and for the RDM device. For virtual disks, VMware has tools for providing these functions. For RDM devices, the LSI CTS2600 storage system provides the following premium features:

- Volume Copy

[1] RAID 1 on an LSI CTS2600 storage system is implemented as a RAID10 mirrored stripe configuration. Although SANtricity Storage Manager calls this configuration RAID1, this configuration is functionally equivalent to a RAID10.

- Snapshot
- Remote Volume Mirroring across FC host ports
- Storage Partitioning

Considering Individual Virtual Machines

Before you can effectively design your volume group and volumes, you must determine the primary goals of the configuration—performance, reliability, growth, manageability, or cost. Each goal has positives, negatives, and trade-offs. After you have determined what goals are best for your environment, follow the guidelines to implement those goals. To get the best performance from the LSI CTS2600 storage system, you must know the I/O characteristics of the files to be placed on the storage system. After you know the I/O characteristics of the files, you can set up a correct volume group and a correct volume to service these files.

Web Servers

Web server storage workloads typically contain random small writes. RAID 5 provides good performance and has the advantages of protecting the system from drive loss and having a lower cost by using fewer drives.

Backup and File Read Applications

The LSI CTS2600 storage system performs very well for a mixed workload. There are ample resources as described by IOPS and throughput to support backups of virtual machines while not impacting other applications in the virtual environment. Addressing performance concerns for individual applications takes precedence over backup performance.

However, there are applications that read large files sequentially. If performance is important, consider using RAID 10. If cost is also a concern, RAID 5 protects from drive loss with the least amount of drives.

Databases

Frequently updated databases: If your database is frequently updated and if performance is a major concern, your best choice is RAID 10, even though RAID 10 is the most expensive because of the number of drives and drive trays. RAID 10 provides the least drive overhead and provides the highest performance from the LSI CTS2600 storage system.

Low-to-medium updated databases: If your database is updated infrequently or if you must maximize your storage investment, choose RAID 5 for the database files. RAID 5 lets you create large storage volumes with minimal redundancy of drives.

Remotely replicated environments: If you plan to remotely replicate your environment, carefully segment the database. Segment the data on smaller volumes and selectively replicate these volumes. Segmenting limits WAN traffic to only what is absolutely needed for database replication. However, if you use large volumes in replication, initial establish times are larger and the amount of traffic through the WAN might increase, leading to slower than necessary database performance. The LSI CTS2600 storage systems premium features, Remote Volume Mirroring (across FC host ports), Volume Copy, and Snapshot are extremely useful with replicating remote environments.

Determining the Best RAID Level for Volumes and Volume Groups

In general, RAID 5 works best for sequential large I/Os (> 256 KB), while RAID 5 or RAID 1 works best for small I/Os (< 32 KB). For I/O sizes in between, the RAID level can be dictated by other application characteristics. Table 2-1 shows the I/O size and optimal RAID level.

Table 2-1 I/O Size and Optimal Raid Level

I/O Size	RAID Level
Sequential, large (>256 KB)	RAID 5
Small (<32 KB)	RAID 5 or RAID 1
Between 32 KB and 256 KB	RAID level does not depend on I/O size

RAID 5 and RAID 1 have similar characteristics for read environments. For sequential writes, RAID 5 typically has an advantage over RAID 1 because of the RAID 1 requirement to duplicate the host write request for parity. This duplication of data typically puts a strain on the drive-side channels of the RAID hardware. RAID 5 is challenged most by random writes, which can generate multiple drive I/Os for each host write. Different RAID levels can be tested by using the SANtricity Dynamic RAID Migration feature, which allows the RAID level of a volume group to be changed while maintaining continuous access to data.

Table 2-2 shows the RAID levels that are most appropriate for specific file types.

Table 2-2 Best RAID Level for File Type

File Type	RAID Level	Comments
Oracle Redo logs	RAID 10	Multiplex with Oracle
Oracle Control files	RAID 10	Multiplex with Oracle
Oracle Temp datafiles	RAID 10, RAID 5	Performance first / drop recreate on drive failure
Oracle Archive logs	RAID 10, RAID 5	Determined by performance and cost requirements
Oracle Undo/ Rollback	RAID 10, RAID 5	Determined by performance and cost requirements
Oracle Datafiles	RAID 10, RAID 5	Determined by performance and cost requirements
Oracle executables	RAID 5	
Oracle Export files	RAID 10, RAID 5	Determined by performance and cost requirements
Oracle Backup staging	RAID 10, RAID 5	Determined by performance and cost requirements
Exchange database	RAID 10, RAID 5	Determined by performance and cost requirements

Exchange log	RAID 10, RAID 5	Determined by performance and cost requirements
SQL Server log file	RAID 10, RAID 5	Determined by performance and cost requirements
SQL Server data files	RAID 10, RAID 5	Determined by performance and cost requirements
SQL Server Tempdb file	RAID 10, RAID 5	Determined by performance and cost requirements

Use RAID 0 volume groups only for high-traffic data that does not need any redundancy protection for device failures. RAID 0 is the least used RAID format but provides for high-speed I/O without the additional redundant drives for protection.

Use RAID 1 for the best performance while providing data protection by mirroring each

physical drive. Create RAID 1 volume groups with the most drives possible (30 maximum) to achieve the highest performance.

Use RAID 5 to create volume groups with either 4+1 drives or 8+1 drives to provide the best performance while reducing RAID overhead. RAID 5 offers good read performance at a reduced cost of physical drives compared to a RAID 1 volume group.

Use RAID 10 (RAID 1+0) to combine the best features of data mirroring of RAID 1 plus the data striping of RAID 0. RAID 10 provides fault tolerance and better performance compared to other RAID options. A RAID 10 volume group can sustain multiple drive failures and losses as long as no two drives form a single pair of one mirror.

Considering the Server Platform

The server platform contains the server hardware and the system software. When considering the hardware and operating system on which you want to run the Oracle database, there are many issues to consider:

High availability – Is Oracle Real Application Clusters (Oracle RAC) needed to provide HA capabilities? Are other clustering solutions, like Microsoft Clustering Services, required for virtual machines? Is DRS or VMotion needed to support high availability?

Scalability – If the database is expected to grow and requires more hardware resources to provide future performance that the customer needs, Oracle can provide a scalable approach to accommodate growth potential in Oracle Databases. VMware HA cluster, DRS, and VMotion can accommodate scalability for virtual machines.

Number of concurrent sessions – Determine the number of concurrent sessions and the complexity of these transactions before deciding what virtual hardware and operating system to use for the database.

Amount of disk I/Os per second (IOPS) – If the database is performing a large amount of IOPS, consider ESX server hardware that supports multiple HBAs. Also consider the number of drive spindles that you must provide the necessary IOPS that are forecasted by the application.

Size – If you have a small database or a small number of users, a small-to-medium size hardware platform could be justified.

Cost – If cost is a factor for purchasing hardware, the x86 platform could be a less expensive platform. The x86 typically provides outstanding performance for the money.

Considering the Server Hardware Architecture

Available bandwidth depends on the server hardware. The number of buses adds to the aggregate bandwidth, but the number of HBAs sharing a single bus can throttle the bandwidth.

Calculating Aggregate Bandwidth

An important limiting factor in I/O performance is the I/O capability of the server that hosts the application. The aggregate bandwidth of the server to the storage system is measured in MB/s and contains the total capability of the buses to which the storage system is connected. For example, a 64-bit PCI bus clocked at 133MHz has a maximum bandwidth calculated by the following formula:

PCI Bus Throughput (MB/s) = PCI Bus Width / 8 * Bus Speed 64-bit /8

* 133 MHz = 1062 MB/s ~ = 1GB/s

See Table 2-3 for FCI-X bus throughput.

Table 2-3 PCI-X Bus Throughput

MHz	PCI Bus	Throughput
66	64	528
100	64	800
133	64	1064
266	64	2128
533	64	4264

Sharing Bandwidth with Multiple HBAs

Multiple HBAs on a bus share this single source of I/O bandwidth, and each HBA might have multiple FC ports, which typically operate at 1Gb/s, 2 Gb/s, 4 Gb/s, or 8 Gb/s. As a result, the ability to drive a storage system can be throttled[2] by either the server bus or by the HBAs. Therefore, whenever you configure a server or whenever you analyze I/O performance, you must know how much server bandwidth is available and which devices are sharing that bandwidth.

[2] *Throttle* - To slow down I/O processing during low memory conditions, typically processing one sequence at a time in the order that the request was received.

ESX 3.5 Server Path Failover and Load Distribution

ESX 3.5 has a built-in failover driver to manage multiple paths. At startup, or during a rescan that might be issued from the Virtual Center 2.5 Console, all LUNs or volumes are detected. When multiple paths to a volume are found, the failover driver is configured and uses the default Most Recently Used (MRU) policy.

The LSI CTS2600 storage system is an Active/Passive storage system where volume ownership is distributed between the two controllers. The individual volumes are presented to the ESX server by both controllers. The ESX 3.5 server configures both controllers as possible owners of a LUN, even though only one controller owns the LUN. ESX 3.5 is able to distinguish between the active controller, the controller which owns a volume, and the passive controller. The active controller is the preferred controller.

NOTE Additional multi-path drivers, such as RDAC, are not supported by ESX 3.5.

The ESX 3.5 failover driver provides three policies:

Fixed – The fixed policy is intended for Active/Active devices and is not recommended for the LSI CTS2600 Configurable storage system. If the fixed policy is selected for volumes presented by the LSI 2600 storage system, thrashing could result.

MRU – The MRU policy is intended for Active/Passive devices. The MRU failover policy is recommended for the LSI CTS2600 Configurable Storage Components.

Round Robin – Round Robin is an experimental policy. This failover policy can be

selected, but does not offer any advantage. The experimental Round Robin policy sends a measure of throughput, (for example, quantity of 100 I/Os), over one path before using the next available path. With a single LUN, the effect is the same as using just one path. Where there are multiple LUNs or volumes, the likely behavior is that all I/ O to all volumes tends to traverse the same path rather than distributing the load.

Concerns and Recommendations

A single path to both controllers can lead to either unbalanced volume ownership or thrashing under certain conditions. The ownership of all volumes can be forced to one of the controllers. Depending on which path that the ESX 3.5 server finds first, the single active controller on that path could be forced to assume ownership of all LUNs, even those for which that controller is not the preferred owner. This process limits the storage performance for the ESX 3.5 server.

In configurations involving multiple ESX 3.5 servers attached to the LSI CTS2600 storage system (with FC host ports in this example), the above behavior is exacerbated. When one ESX 3.5 performs LUN discovery, volume ownership could lead to *thrashing*, or bouncing ownership between the controllers.

To avoid these problems, VMware advises that you set up four paths between the server and the storage system as shown in the following figure. At least two ESX 3.5 server HBA ports must be used and both HBA ports should see both controllers. Assuming that the SAN switches are interconnected, the configured zone would include all four connections. Each ESX 3.5 server HBA port sees both controllers in the LSI CTS2600 storage system.

A loss of one of the paths can lead to less than optimal performance, because volumes owned by the controller on the lost path are transferred to the other controller with the surviving path.

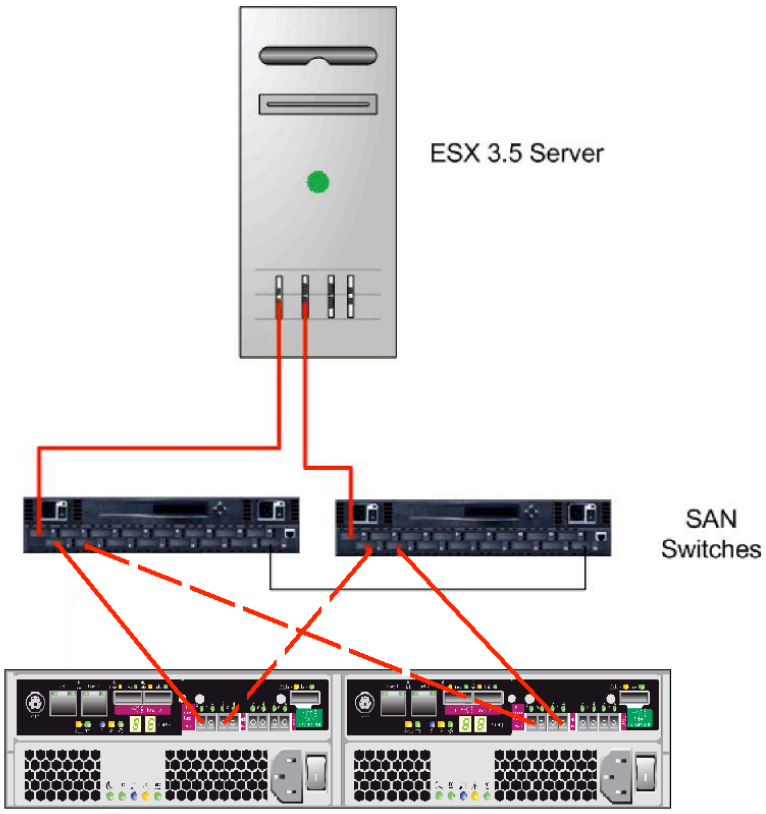

Figure 3 Additional Connections Added to the Switch

To preserve volume ownership, each controller is cross-connected to the other switch. The disadvantage of this type of switching is that the additional storage system host ports are consumed for the zone and cannot be used to address other performance concerns. If you are seeking to prevent volume ownership transfer, consider using the additional controller to switch connections in multiple zones.

The previous recommendations prevent thrashing but do not sufficiently address performance concerns. Only one of the paths can be active, because the first HBA port configured by ESX 3.5 server is used to communicate with both controllers. To maximize performance, you must spread the load between more paths.

Example of Server Path Failover and Load Distribution

An ESX 3.5 server has eight paths consisting of eight server FC HBA ports (four dual port FC HBA), eight storage system host ports, and a pair of switches. In a simple configuration depending only on ESX 3.5, the MRU failover policy implements all individual paths. However, the additional ESX 3.5 server's HBA ports do not add benefit because only two of the eight paths are used.

To increase the I/O performance, spread the load across more ESX 3.5 server HBA ports and more storage system host ports. You can implement this process by creating multiple groups of four-path configurations.

There are several elements needed to perform this task:

1. Combine pairs of ESX 3.5 HBA ports with pairs of LSI CTS2600 storage system host ports through the use of zoning on the SAN switches.

2. Logically divide the ESX 3.5 server's pairs of HBA ports into separate storage partitions on the storage system.

3. Assign specific volumes, which are balanced between controllers, to the storage partition.

Zoning the switches defines a specific path to the storage system. This path is refined with the storage partitioning and the creation of the logical host definition. After specific LUNs are presented to the logical host, the path definition is complete.

You can benefit from this strategy by the number of supported LUNs. ESX 3.5 supports a maximum of 256 LUNs or paths to LUNs. Relying on just the failover driver's MRU policy severely limits the actual number of LUNs found. In practice, only sixteen actual LUNs could be supported in an eight-server port configuration.

In a configuration with 44 physical LUNs, a given path shows 88 LUNs, including active LUNs and standby LUNs. If there are eight FC HBA ports, 88 LUNs are available on each port. The resulting 704 LUNs greatly exceeds ESX 3.5 server capabilities. By following the recommended practice, you can increase the quantity of supported LUNs to 128.

The multiple zone and storage partitioning configuration better distributes the load by using four of eight available paths to the storage system. You can scale this strategy by adding additional pairs of ESX 3.5 server HBA ports, zones, storage system host ports and storage partitions.

ESX 3.5 Server Configuration

To configure the ESX 3.5 server, set the ESX 3.5 Server Advanced options.

1. Disable the `Disk.UseDeviceReset` option for the LSI CTS2600 storage system volumes by typing the following command.

   ```
   Disk.UseDeviceReset=0
   ```

2. Enable `Disk.UseLunReset` for the LSI CTS2600 storage system volumes.

   ```
   Disk.UseLunReset=1
   ```

3. Disable `Disk.ResetOnFailover` for the LSI CTS2600 storage system if the volumes are not being used for either RDM or Microsoft Cluster.

   ```
   Disk.ResetOnFailover=0
   ```

4. If Using Raw Device Mapping, RDM, or Microsoft Cluster nodes across multiple ESX 3.5 servers, enable `Disk.ReseOnFailover`. Change the virtual machine configuration file to use the SCSI target address, `vmhba`*`X`*`.`*`X`*`.`*`X`*, instead of the VMFS volume label.

```
Disk.ResetOnFailover=1

Enable Disk.RetryUnitAttention
Disk.RetryUnitAttention=1
```

5 Enable logging on ESX host.

```
Scsi.LogMultiPath = 1
Scsi.PrintCmdErrors = 1
```

6 If working with Snapshot or RVM volumes, enable LVM.EnableResignature.

```
LVM.EnableResignature = 1
```

Chapter 3

Operating System Considerations

This chapter describes items to consider when using a particular operating system and how that operating system affects partition alignments.

SANtricity Storage Manager in a Guest Operating System

Although not required for performance, you can install SANtricity Storage Manager, the storage management software for the LSI CTS2600 storage system, inside a guest operating system. You can manage a storage system out-of-band as long as the Virtual Machine has a configured network adapter.

Using SANtricity Storage Manager with a guest operating system lets you manage the storage system in-band using the SAN network. To manage the storage system in-band, you must map the access volume to the virtual machine as a RDM device. The access LUN is a small volume, approximately 20 MB in size, which is used to transfer data to the LSI CTS2600 storage system through the SAN.

The RDM mapping shows the access volume as another guest operating system SCSI device. After the access volume is mapped and visible inside the guest operating system, you can use the `SMdevices.bat` command to alert SMagent of the available LUN. Although the volume appears as a SCSI device, SMagent does recognize the access volume and lets SANtricity Storage Manager use the volume for in-band management.

Buffering the I/O

The type of I/O—buffered or unbuffered—provided by the operating system to the application is an important factor in analyzing storage performance issues. Unbuffered I/O (also known as *raw* I/O or *direct* I/O) moves data directly between the application and the drive devices. Buffered I/O is a service provided by the operating system or by the file system. Buffering improves application performance by caching write data in a file system buffer, which the operating system or the file system periodically moves to permanent storage.

Buffered I/O is generally preferred for shorter and more frequent transfers. File system buffering might change the I/O patterns generated by the application. Writes might coalesce so that the pattern seen by the storage system is more sequential and more write-intensive than the application I/O itself. Direct I/O is preferred for larger, less frequent transfers and for applications that provide their own extensive buffering (for example, Oracle).

Regardless of I/O type, I/O performance generally improves when the storage system is kept busy with a steady supply of I/O requests from the host application. Become familiar with the

parameters that the operating system provides for controlling I/O (for example, maximum transfer size).

Clustering

So-called *shared*, *clustered*, or *SAN* file systems such as CXFS, StorNext, and GPFS provide file sharing for multiple hosts in a SAN. All such multi-node systems introduce additional I/O performance issues that require a complete understanding of the data flow, I/O alignment, and I/O sizes of the specific file system. For information about setting the segment size, see "Basing the Segment Size on File I/O Operations".

Oracle uses the term *block size* instead of the more common term *page size*. A block is the smallest unit of work.

When you conduct a performance tuning session on the database, test the performance with backups that are running synchronously with the daily jobs of rebuilding database objects or defragmenting database objects.

Calculating Optimal Segment Size

The LSI term *segment size* refers to the amount of data written to one drive in a volume group before writing to the next drive in the volume group. For example, in a RAID 5, 4+1 volume group with a segment size of 128 KB, the first 128KB of the LUN storage capacity is written to the first drive and the next 128 KB to the second drive. For a

RAID 1, 2+2 volume group, 128 KB of an I/O would be written to each of the two data drives and to the mirrors. If the I/O size is larger than the number of drives times 128 KB, this pattern repeats until the entire I/O is completed.

For very large I/O requests, the optimal segment size for a RAID volume group is one that distributes a single host I/O across all data drives. The formula for optimal segment size is as follows:

LUN segment size = LUN stripe width ÷ number of data drives

For RAID 5, the number of data drives is equal to the number of drives in the volume group minus 1. For example:

RAID 5, 4+1 with a 64KB segment size => (5-1) * 64 KB = 256 KB stripe width

For RAID 1, the number of data drives is equal to the number of drives divided by 2. For Example:

RAID 10, 2+2 with a 64 KB segment size => (2) * 64 KB = 128 KB stripe width

For small I/O requests, the segment size must be large enough to minimize the number of segments (drives in the LUN) that must be accessed to satisfy the I/O request, that is, to minimize segment boundary crossings. For IOPS environments, set the segment size to 64 KB or 128 KB or larger, so that the stripe width is at least as large as the median I/O size.

When using a volume manager to collect multiple storage system LUNs into a Logical Volume Manager (LVM) volume group (VG), the I/O stripe width is allocated across all of the segments of all of the data drives in all of the LUNs. The adjusted formula becomes as follows:

LUN segment size = LVM I/O stripe width / (# of data drives/LUN * # of LUNs/VG)

To learn the terminology so that you can understand how data in each I/O is allocated to each LUN in a logical volume group, see the vendor documentation for the specific Logical Volume Manager.

Aligning Host I/O with RAID Striping

For all file systems and operating system types, you must avoid performance degrading segment crossings. You must not let I/O span a segment boundary. Matching I/O size (commonly, by a power-of-two) to volume group layout helps maintain aligned I/O across the entire drive. However, this statement is true only if the starting sector is correctly aligned to a segment boundary. Segment crossing is often seen in the Windows operating system, where partitions created by Windows 2000 or Windows 2003 start at the 64th sector. Starting at the 64th sector causes misalignment with the underlying RAID striping and allows the possibility for a single I/O operation to span multiple segments.

Aligning Partitions for ESX 3.5 Server

Ideally, VMFSes should be aligned when they are created. VMware recommends that you create the VMFS using the VI client interface that automatically aligns the partition.

NOTE Guest operating system boot partitions are generally not as active as data partitions. For this reason, do not align guest operating system boot partitions. All data disks must be aligned because these partitions are active and require alignment. Make sure that highly active application data is stored on its own disk. This recommendation to align is true for both RDM devices and for the *.vmdk disk files used for application data.

Use this procedure to start the partition at 64 to align the partition with the CTS2600 storage system volume.

NOTE These are not complete steps for file system alignment. This section only provides an overview of the process.

---

1 To see if your file system is aligned, type the following command. `disk –lu /dev/sd*`

```
Device boot Start End BlocksId System
   /dev/sdj1    64     167766794 83883333+ fb Unknown
```

NOTE The stripe size equals the segment size divided by 512 and multiplied by the number of disks. For example:

512 KB (524,288) / 512 * 4 = 4096

If the Start value is 63 (the default), the partition is not aligned.

NOTE Perform manual partition alignment on the service console only if you are an expert familiar with the `fdisk` tool. Use the `t` option to specify `fb` to identify the partition as a VMware partition. Use the `b` option to adjust the starting block number, which is 64 for LSI CTS2600 Configurable Storage Components .

2

After you have aligned your VMware VMFS partitions, align the data file system partitions within your virtual machines.

Aligning Partitions on a Microsoft Windows Operating System

Microsoft provides the `diskpar.exe` utility as part of the Windows 2000 Resource Kit (`diskpart.exe` in the Windows 2003 Service Pack 1). Using `Diskpar.exe`, you can set the starting sector in the master boot record to a value that makes sure of sector alignment for all I/Os. Use a multiple of 64, such as 64 or 128. Sector alignment is especially important for Exchange. For Microsoft's usage details on `diskpar`, go to:

"How to Align Exchange I/O with Storage Track Boundaries"
http://technet.microsoft.com/en-us/library/0e24eb22-fbd5-4536-9cb4-2bd8e98806e7.aspx

Aligning Partitions on a Linux Operating System

IMPORTANT Adjust a Linux operating system for correct alignment only if you are an expert in the Linux operating system. Only a Linux expert can use the extra-functionality `x` mode with the `[disk` command.

In `x` mode, experts can use the `b` option to set the starting block of a partition as an absolute address. For example, an application has a 2MB block size. If the first stripe group occupies blocks 0-4095, setting the starting block to 4096, that is, one block + 1, gives the correct alignment.

To move the starting offset of data in a partition, follow these steps.

WARNING Do not attempt to partition a live volume that has data on the volume. The data will be lost.

---

1 Create a new partition.

2 Using `[disk`, select the **x** option.

3 Select option **b** to move the beginning of the data in a partition.

4 Select the partition number.

5 Select a new beginning for the data.

You usually can put in one stripe width in the **`vdShow [LUN_Number]`** >> **Stripe Size** >> **Sectors**.

6 Select option **w** to write the partition table.

NOTE The stripe size equals the segment size divided by 512 multiplied by the number of LUNs. For example:

512KB (524,288) / 512 * 4 = 4096

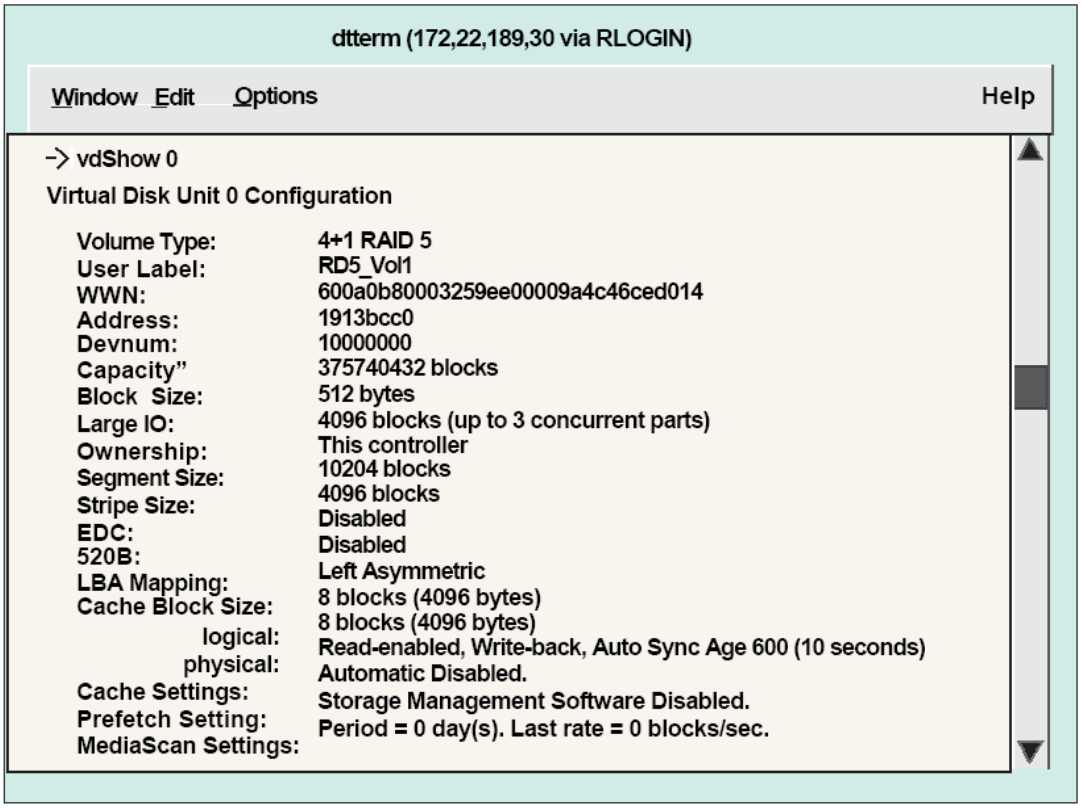

Figure 4 Diagnostic Display of RD5.v1

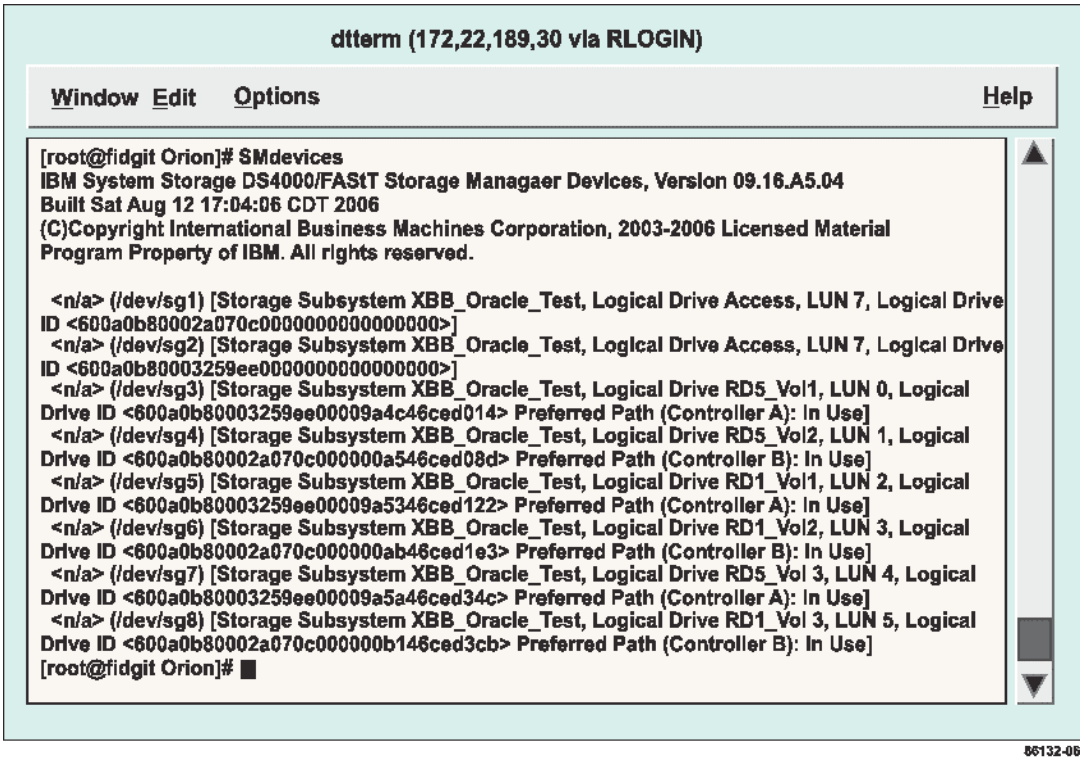

Figure 5 List of mapped volumes in SMdevices

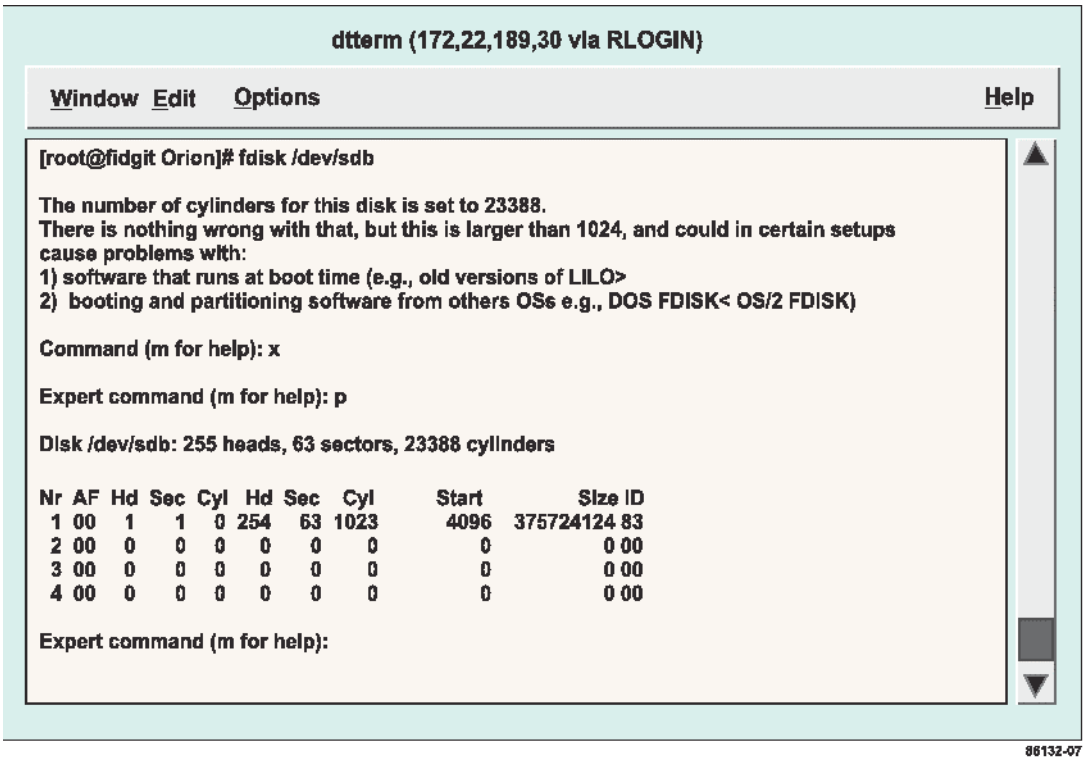

Figure 1 fdisk display RD5.V1 volumes

Aligning Partitions on Other Operating Systems

I/O alignment is equally important for other operating systems and their associated file systems (for example, UFS, VxFS, QFS, ZFS, CXFS, and SNFS). For details about how to provide alignment in each situation, refer to the respective vendor's documentation.

Locating Recommendations for Host Bus Adapter Settings

Use the default HBA settings of the HBA vendor. Use the same model of HBA in the ESX 3.5 server. Mixing HBAs from different vendors in the same ESX 3.5 server is not supported.

Recommendations for Fibre Channel Switch Settings

Table 3-1 shows the recommended Fibre Channel switch settings.

Table 3-1 Fibre Channel Switch Settings

Fibre Channel Switch Setting	Description
Enable In-Order Delivery	Recommended settings are available from the supplier of the storage system. For example, on Brocade switches, verify that the `In-Order Delivery` parameter is enabled.

Inter-switch Links	In a multi-switch SAN fabric, where I/O traverses inter-switch links, make sure to configure sufficient inter-switch link bandwidth.
Disable Trunking on the Fibre Channel switch	When using a Cisco Fibre Channel switch, the CTS2600 storage system host ports and the Fibre Channel HBA ports on the server cannot be configured on the switch with the trunking enabled. The use of the trunking feature can cause thrashing of volume ownership on the storage system. Trunking is set to automatic by default. You can change trunking to non-trunk under the Trunk Config tab.

Using Command Tag Queuing

Command Tag Queuing (CTQ) refers to the controller's ability to line up multiple SCSI commands for a single LUN and run the commands in an optimized order that minimizes rotational and seek latencies. Although CTQ might not help in some cases, such as single-threaded I/O, CTQ never hurts performance and therefore is generally recommended. The LSI models vary in CTQ capability, generally up to 2048 per controller. Adjust the CTQ size to service multiple hosts. CTQ is enabled by default on LSI CTS2600 storage system, but you also must enable CTQ in the host operating system and on the HBA. Refer to the documentation from the HBA vendor.

The capability of a single host varies by the type of operating system, but you can generally calculate CTQ as follows:

CTQ Depth Setting = Maximum OS queue depth (< 255) /Total # of LUNs

NOTE If the HBA has a lower CTQ capacity than the result of the previously mentioned calculation, the HBA's CTQ capacity limits the actual setting.

VMware provides guidance for configuring the Emulex and Qlogic HBAs in the *Fibre Channel SAN Configuration Guide, ESX Server 3.5, ESX Server 3i version 3.5, Virtual Center 2.5*.

Setting the maximum outstanding disk requests per virtual machine is described in the following link:

"*Setting the Maximum Outstanding Disk Requests per Virtual Machine*"
http://kb.vmware.com/selfservice/search.do?cmd=displayKC&docType=kc&externalId=1268&sliceId=1&docTypeID=DT_KB_1_1&dialogID=936508&stateId=0%200%20938651

Analyzing I/O Characteristics

Analyze the application to determine the best RAID level and the appropriate number of drives to put in each volume group:

- Is the I/O primarily sequential or random?
- Is the size of a typical I/O large (> 256 KB), small (< 64 KB), or in-between?

  If this number is unknown, calculate a rough approximation of I/O size from the statistics reported by the SANtricity Performance Monitor using the following formula:

  Current KB/second ÷ Current I/O/second = KB/I/O

- What is the I/O mix, that is, the proportion of reads to writes? Most environments are primarily Read.
- What read percent statistic does SANtricity Performance Monitor report?
- What type of I/O does the application use—buffered or unbuffered?
- Are concurrent I/Os or multiple I/O threads used?

In general, creating more sustained I/O produces the best overall results, up to the point of controller saturation. Write-intensive workloads are an exception to this general rule.

Using VMS for Spanning Across Multiple LUNs

Although ESX 3.5 supports using several smaller LUNs for a single VMFS, spanning LUNs is not recommended. You can improve performance by using a single, correctly-sized LUN for the VMFS. Fewer larger LUNs are easier to manage.

Chapter 4

Setting Up the Storage System

After you have considered the requirements for the operating system and the application, you can now set up the storage system for optimal performance.

Factors Influencing Storage Performance

The LSI CTS2600 storage system, with its intelligent controllers, delivers the highest possible performance from volume groups. LSI CTS2600 storage system was designed for open system I/O. LSI CTS2600 storage system technology provides extremely high performance in open systems, and are designed to provide the best possible drive-based performance—a requirement for today's transaction-intensive applications. Drive-based performance is accomplished with a combination of attentive controller design, custom integrated circuits to accelerate RAID XOR, and efficient cache management.

Drive I/O capacity is a critical part of storage system performance. For LSI CTS2600 storage system, the number of drives in a configuration usually establishes the upper bound for storage system performance. Many various interacting factors determine how much of the raw performance of a group of drives that a specific application can use. These factors include the following:

- The size of the cache
- The algorithms that manage the cache
- The number and type of host and drive-side channels
- The performance of RAID parity calculations
- The way that SCSI commands are queued for optimized execution by the controllers
- The way that the controllers choose data paths

Estimating Capacity Limits

When setting up a storage system, first estimate its capacity limits. To establish a framework for tuning, estimate the upper limit for performance for the LSI CTS2600 storage system, based upon the specifications for the particular model.

For IOPS environments, the number of drives in the volume group largely determines performance. The maximum IOPS (from drive) for a storage system is typically specified with a full complement of drives. Many factors determine IOPS, including:

- Drive type (SAS mechanical drives of SSDs)
- Drive RPM
- Data layout
- Varying I/O sizes
- Volume group layout
- Controller architecture and workload

Performance in bandwidth environments is not quite so directly dependent on drive count, and the full bandwidth rating of the storage system often can be realized with less than a full configuration, for example, with as few as four full drive trays.

Automatic Volume Transfer

Auto-Volume Transfer (AVT) is an LSI failover method that does not require specialized host software to actively watch or manage host paths to the storage. Path failover occurs simply by sending I/O to the controller that does not own the volume. LSI chose to use the AVT failover method to work with fast host failover implementations.

LSI chose to limit the amount of dirty cache allowed for a LUN, so that cache flush required for a LUN failover does not take too long. The dirty cache limit is 16 MB / LUN. When the amount of dirty data in cache exceeds 16 MB for a single LUN, all of the data for that LUN is marked with an age of zero, making the dirty data in cache available for immediate flush. Any new data coming into that LUN is also marked with an age of zero, rather than the usual default age of ten seconds. This cache flush strategy can cause read delays while longer than normal writes to drive complete, and cache flushing might occur at inopportune times.

If you have AVT enabled in *any* host region, the AVT cache flush rules apply to all host regions. If you want to totally disable AVT cache flush logic, you must disable AVT in all host regions, and then restart the controllers. The decision on flush strategy is made at boot time.

If you are using RDAC or other multi-path software, like that found in ESX 3.5, disable AVT on the LSI CTS2600 storage system. This feature, when enabled, automatically moves volumes back to the preferred controller after a problem with the controller or paths to the controller have been fixed. If this situation occurs, the ESX 3.5 server starts its own path failover and the volume ownership would bounce between the two controllers, causing thrashing. VMware recommends that multiple paths for each server be available. At least two ESX 3.5 server HBA ports must be used and both HBA ports should see both controllers. This practice prevents several thrashing conditions.

AVT is disabled by default on all volumes configured as the VMware host type. To disable AVT on all host regions, copy the AVT disable script to a `*.SCT` file and run the `*.SCT` file on the LSI CTS2600 storage system. For the AVT disable script, see "Appendix A: AVT Disable Script" on page A-1.

Choosing the Number of Drives to Put in a Volume Group

For high-bandwidth applications, use enough drives to enable a full stripe write for the typical application I/O size, while still allowing for a segment size of 64 KB or larger. Host I/O sizes of a power of two are typical, such as 512 KB, 1MB, and 2 MB. A RAID 5 volume group of 4+1 or 8+1 is a good match for those host I/O sizes. Therefore, for a typical host I/O size of 1MB, use a RAID 4+1 with a 256 KB segment size, or a RAID 8+1 with a 128 KB segment size.

For IOPs or transaction-oriented applications, the number of drives becomes more significant because drive random I/O rates are relatively low. Select a number of drives that matches the per volume group I/O rate needed to support the application. Make sure to account for the I/Os required to implement the data protection of the selected RAID level. Make the segment size at least as large as the typical application I/O size. The reason is to avoid segment crossings, which place additional I/O demand on the drives. A segment size of 128 KB is a reasonable starting point for most applications. The higher the spin speed of the drive, the better. The spindle count of an existing volume group can be increased using the Dynamic Capacity Expansion feature of SANtricity Storage Manager.

To get the best performance from the storage system, use as many drives as possible within the RAID 1 volume groups. Figure 4-1 on page 4 demonstrates the performance difference between the following volume groups at ten percent writes.

RAID Level	Drives	Volume Name
RAID 1	4+4 = 8 total	RD1_V1_10
RAID 1	8+8 = 16 total	RD1_V2_10
RAID 5	4+1 = 5 total	RD5_V1_10
RAID 5	8+1 = 9 total	RD5_V2_10

Figure 7 shows how the number of drives affects performance of RAID 1 and RAID 5, measured in MB/s. In Figure 7, see that the performance level of the RAID 5, 8 + 1 volume group is almost the same (-5 percent max) as a RAID 1, 4 + 4 volume group, with twice the volume space with only nine drives for the RAID 5 compared to nine drives for the RAID 1 volume group. For light write access data, RAID 5 is a good economical choice for large amounts of data.

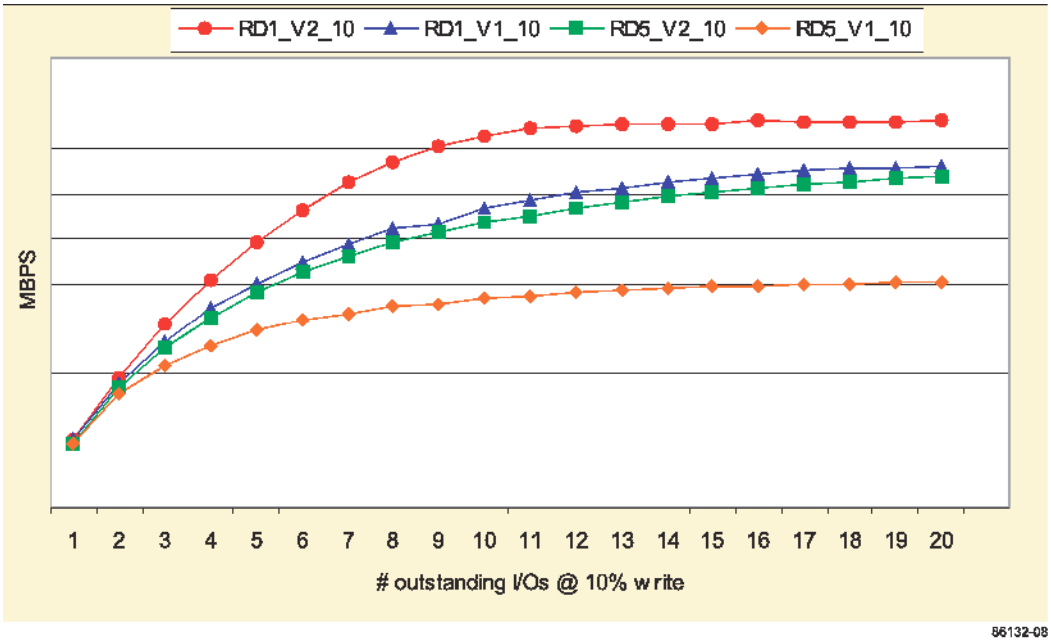

Figure 7 Impact in MB/s of the Number of Drives on the Performance of RAID 1 and RAID 5

Figure 8 shows how the number of drives affects the performance of RAID1 and RAID5, measured in IOPS.

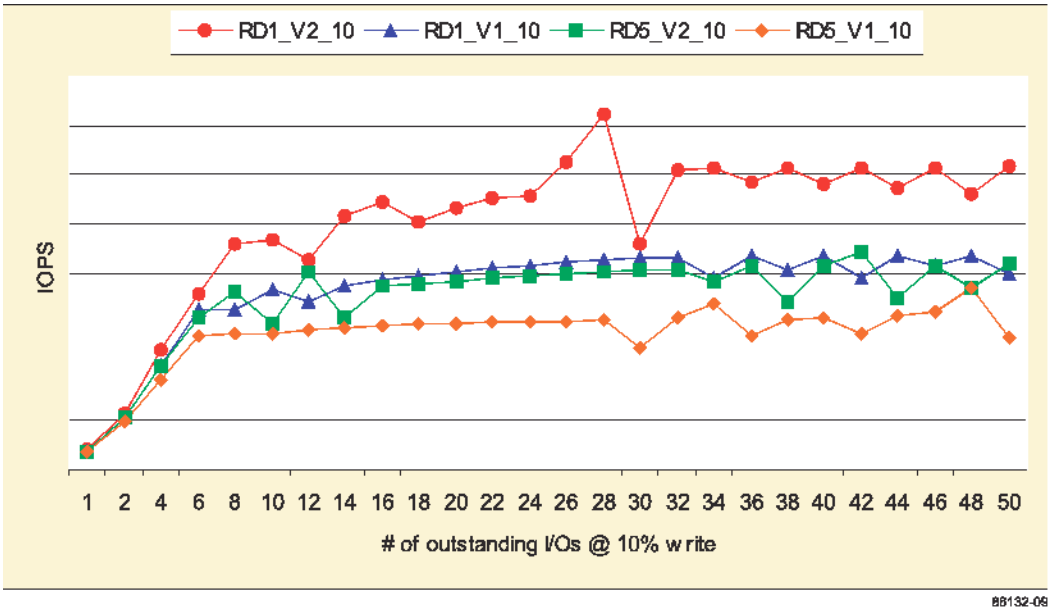

Figure 8 Impact in IOPS of the Number of Drives on the Performance of RAID 1 and RAID 5

However, Figure 9 shows that at heavy write levels, RAID 1 clearly outperforms RAID 5 at both volume sizes. Figure 9 demonstrates the need to place high write-intensive files such as redo, archives, and backup on RAID 1 volumes instead of on RAID 5 volumes.

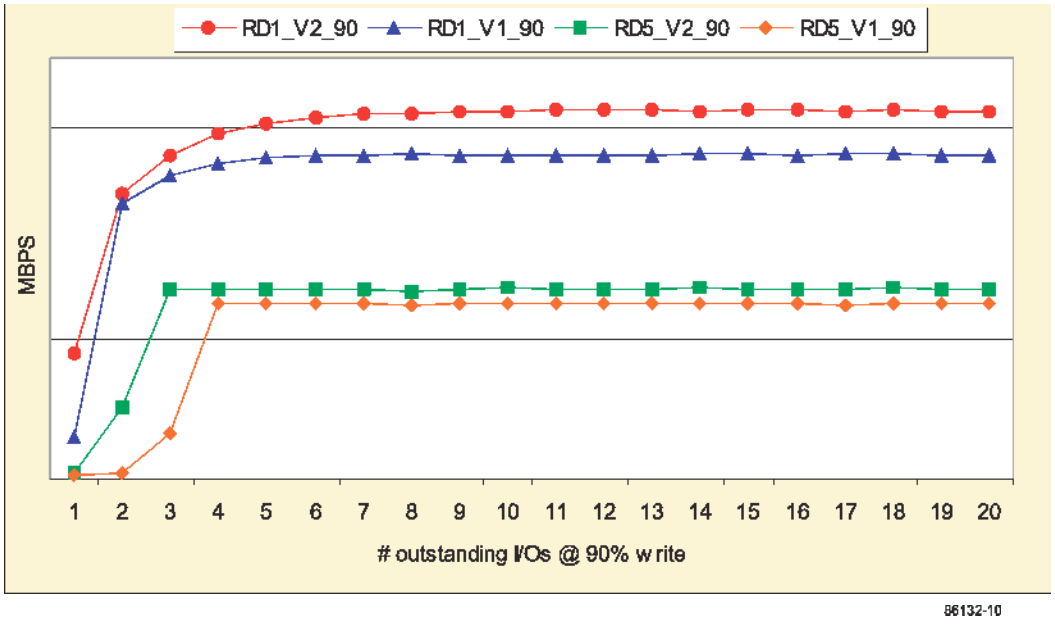

Figure 9 Impact of the Number of Drives on the Performance of RAID 1 and RAID 5 in a Write-Intensive Environment

Storage System Design Best Practices

Here are some of the most important best practices that you must follow to obtain optimal performance from an LSI CTS2600 storage system.

- Use all available host-side channels. Balance I/O across the dual controllers of the storage system (for example, with a volume manager) and strive to keep both controllers busy.
- Attach cables to the drive trays according to the latest *Hardware Cabling Guide*.
- Choose faster drives. A 15-KB RPM drive has one-third less rotational latency than a 10-KB RPM drive.
- Add more drives to the configuration for a linear increase in performance, up to the point of controller saturation. More drives provide more spindles to service I/O.
- Create volume groups across drive trays to distribute I/O across back-end loops. This distribution varies by controller tray. Try to balance the number of drives on each back-end loop. For details, see "Locating Volume Groups" on page 4-11.
- Configure the entire capacity of a volume group into a single volume. Cost considerations might require multiple drives on one volume group, but this choice typically increases seek time penalties.
- Separate random workloads and sequential workloads onto different physical drives.

- Choose an optimal segment size based on the I/O characteristics of the application.
- Volumes configured for ESX 3.5 are configured by the `vmware` host definition. By default, AVT is disabled for ESX 3.5. If AVT is not required for failover by any of the hosts that use the storage system, disable AVT in all host regions. Contact a Customer and Technical Support representative about the procedure, which must be repeated if the NVSRAM file is changed or reloaded.

Connecting the Host

LSI CTS2600 storage system was designed with high availability and performance in mind. LSI CTS2600 storage system features redundant power supplies, redundant controllers, redundant cache, and redundant internal architecture. To maximize this redundant technology and to prevent a single point of failure, use the following cabling guidelines in any high-available environment. Redundant switching provides the most fault-tolerant configuration available for host connectivity.

To improve performance, IOPS or throughput, add more ESX 3.5 server HBA ports and more storage system host ports to the configuration. Configure zones with a pair of ESX 3.5 server HBA ports and a pair of storage system host ports (one per controller). Use Storage Partitioning to create logical hosts and assign volumes to this host. Monitor the configuration during peak operations to make sure that I/O is balanced in terms of both IOPS and in terms of throughput. Make sure that the load on the two controllers is approximately equal and that the load on each ESX 3.5 server HBA is almost equal.

NOTE To benefit fully from this configuration, LUNs must be balanced between the two controllers.

In clustered environments, ESX servers must have shared access to LUNs. Apply this approach to all ESX servers in the cluster. When additional ESX servers are added, add the new server's HBA ports to each of the zone configurations and storage partition configurations.

Tuning an External LSI CTS2600 storage system

This section describes how to tune an external LSI CTS2600 storage system so that you get the best performance.

Basics of Performance Tuning

These basics of performance tuning steps provide information about how to understand the factors included in performance tuning and the components that influence performance.

Understanding the Context for Performance Tuning

The challenge of storage performance tuning is to understand and control these interacting factors while accurately measuring application performance. Because the performance of the

storage system accounts for only a portion of overall application performance, tuning must be completed in context. The full context includes the I/O characteristics of the application and all of the components in the data path:

- HBA
- Switches (where applicable)
- Volume manager
- File system
- Operating system
- Server

With multiple parameters to consider, the task of performance tuning even one application can seem formidable. Tuning all of the different applications that share a single storage system seems even more formidable. To reduce the complexity of tuning, LSI CTS2600 storage system features performance monitoring and flexible tuning controls in SANtricity Storage Manager.

Three Components that Influence Performance

This document provides an overall approach to tuning I/O performance and also provides specific guidelines for using the storage system tuning controls. These recommendations start with an overall analysis of the three elements that determine I/ O performance:

- Application software
- Server platform (hardware, operating system, volume managers, device drivers)
- Storage system

An Iterative Approach to Performance Tuning

Performance tuning requires you to loop repeatedly through the following steps:

1 Run benchmarking tests.

2 Measure the performance results.

3 Adjust the settings as required, changing *only one* parameter at a time.

The dynamic features in all LSI CTS2600 storage system is ideally suited for this iterative process. The first step in tuning is to establish a baseline of existing performance with a convenient and trusted metric. Compare the baseline to the estimated capability of the configuration. This document provides recommendations for this important first step.

Setting the Global Parameters

Setting global parameters lets the storage system perform these types of tasks efficiently, instead of you having to manually perform these tasks.

Setting the Global Cache Flush

Two global parameters, Start Flushing and Stop Flushing, are provided to control the flushing of write data from the controller cache to the drives. Flushing begins when the percentage of unwritten data cache exceeds the Start Flushing level and stops when the percentage hits the Stop Flushing mark. LSI recommends setting both parameters to the same value to cause a brief flushing operation to maintain a specified level of free space. Start with the default values and experiment. If you enable the per-volume failover functionality of AVT, cache management and flushing behavior can be affected. If AVT is not required for failover for the host platforms using the storage system, disabling AVT in all host regions can improve performance for some workloads.

Setting the Force Unit Access and Synchronize Cache

Two cache-related NVSRAM parameters set by the host type are related to SCSI commands from the host. They are Force Unit Access (FUA) and Synchronize Cache. FUA is a bit that is set as part of a read or write command. If enabled, FUA instructs the storage system to bypass cache and go directly to the drive. The Synchronize Cache command instructs the storage system to flush cache to the drive. The FUA bit and the Synchronize Cache command are most often used in an Windows server environment. Because LSI CTS2600 storage system usually keeps control of these functions, these NVSRAM parameters are normally set to the "Ignore" state. However, if cache behavior is not as expected, contact the supplier of the LSI CTS2600 storage system to verify the state of these parameters.

Setting the Global Media Scan

The impact of Media Scan is minimal, but the extra reads do represent a finite workload. Therefore, consider the performance demands when setting Media Scan.

- In most cases, enable Media Scan and set the scan frequency to 15 days to enable periodic scans of the surface of all drives.
- When absolute maximum performance is the objective, do not enable Media Scan.

You also can enable or disable Media Scan for each volume by setting LUN-specific parameters.

Setting LUN-Specific Parameters

Use the Performance Monitor to guide the tuning process. Observe the cache hit percentage and the read/write mix for each LUN of interest while an application is running.

Setting the LUN-Specific Media Scan

One way to limit the workload caused by Media Scan is to enable or disable Media Scan for each volume, rather than globally.

- In most cases, enable Media Scan for each volume.
- If the goal is to maximize the performance of a LUN or to take fine measurements of performance, disable Media Scan for a specific volume.

Setting the Caching Parameters

The cache block size is a global parameter for the storage system. Set the cache block size nearest to the typical I/O size. Set the cache block size to 4 KB for transactional workloads with small I/O sizes and to 16 KB for large block and sequential I/O. You easily can change the cache block size at any time to optimize for a particular workload during a specific time period.

Setting the LUN-Specific Write Cache and Write Cache Mirroring

Enabling Write Cache on a LUN generally improves performance for applications with significant write content, unless the application features a continuous stream of writes. However, write caching does introduce some small risk of data loss, in the unlikely event of a controller failure. To eliminate any chance of data loss from a controller failure, the Write Cache Mirroring option makes sure that a LUN's write data is cached in both controllers. This option historically trades write performance for the highest possible availability, although recent firmware improvements significantly reduce this penalty for bandwidth environments.

Because the cache batteries protect the controller cache for several days, a power failure alone does not threaten data.

To see the write cache settings for the test setup, see "Appendix A: AVT Disable Script." **Setting the LUN-Specific Read Cache and Read-Ahead Multiplier"**

Chapter 5

Using Performance Tools and Getting Optimal Performance from Premium Features

Performance tuning depends on measurement. Fortunately, many software measurement tools are available. SANtricity Performance Monitor comes with SANtricity Storage Manager. Many third-party tools also are readily available.

Using SANtricity Performance Monitor

SANtricity Storage Manager provides an integrated Performance Monitor that reports the following statistics for each volume in the storage system.

Table 5-1 Performance Monitoring Statistics

Statistic	Description
Total I/Os	After the start of this monitoring session
Read Percentage	Percent of Read I/Os
Cache Hit Percentage	Percent of reads that is satisfied from the cache
Current KB/sec	After the last polling interval or requested update
Max. KB/sec	Highest value following the last start
Current I/O/sec	After the last polling interval or requested update
Max. I/O/sec	Highest value following the last start

This convenient tool adds the storage system view of performance to those provided by other host-based or fabric-based monitoring tools. For detailed usage information about Performance Monitor, see the SANtricity Storage Manager online help.

Obtaining Additional Performance Tools

Table 5-2 shows a number of widely available tools, benchmarks, and utilities. Some of these benchmarks and utilities are produced by non-profit organizations and are free.

Table 5-2 Performance Tools

Name	Description	Available From
SPC-1 SPC-2	Storage Performance Council benchmarks	http://www.storageperformance.org
IOBench	I/0 throughput and fixed workload benchmark	http://portal.acm.org/citation.cfm?id=71309
IOmeter	I/O subsystem measurement and characterization tool	http://www.iometer.org
IOzone	File system benchmark tool	http://www.iozone.org
lmdd	Drive dump utility for "raw" devices	from the `lmbench` suite
sar	Unix/Linux system activity report command with numerous options	
xdd	Tool for measuring and characterizing drive subsystem I/O	http://www.ioperformance.com
esxtop	Tool for displaying ESC Server resource utilization statistics	

Getting Optimal Performance from Premium Features

You can get optimal performance by using SANtricity Storage Manager's premium features.

Getting Optimal Performance from Snapshot

For optimal performance when using the Snapshot feature, observe the following guidelines:

- Locate repository volumes on the same volume group as the base volume to minimize the copy-on-write penalty.
- Try to schedule Read I/Os to the Snapshot volume at off-peak times when I/O activity on the source LUN is lower.

Getting Optimal Performance from Volume Copy

The Volume Copy premium feature uses optimized large blocks to complete the copy as quickly as possible. Thus Volume Copy requires little tuning other than setting the copy priority to the highest level that still allows acceptable host I/O performance. Volume Copy performance is affected by other controller activity and by the RAID level and volume parameters of the source volume and the target volume. A best practice

for using Volume Copy is to disable all snapshot volumes associated with a base volume before selecting the base volume as a volume copy target volume. For more information about Volume Copy, refer to the SANtricity Storage Manager online help or refer to the *SANtricity Storage Manager Volume Copy – Feature Guide for 10.x*.

Getting Optimal Performance from Remote Volume Mirroring (across FC)

For optimal performance when using the Remote Volume Mirroring premium feature across FC host ports, observe the following guidelines:

- Upgrade both storage systems to the latest firmware levels available.
- Locate repository volumes on RAID 1 volumes separated from production volumes to isolate writes and help optimize performance.
- In general, use at least as many drives in the target volume groups as are used in the source volume groups.
- Larger segment sizes on both the source and target LUNs generally improve the performance of the Remote Volume Mirroring premium feature.
- For the target LUN, enable Write Caching, but disable Write Cache Mirroring.
- For the source LUN, enable Read Caching. Determine the write caching parameters for the source by the operational requirements, not by Remote Volume Mirroring.
- Use the highest priority level for synchronization for optimal Remote Volume Mirroring performance, assuming that the impact on host I/O performance is acceptable.
- On Brocade switches, enable the In Order Delivery option.

Chapter 6

Getting More Information

Bibliography

Maltz, Brad "*Storage Best Practices with VMware ESX 3.5: How to Make It Perform Better and More Scalable*," slide 13

LSI, 2007. *Tuning External Storage Systems*

Loaiza, Juan [undated PPT] *Optimal Storage Configuration Made Easy*

http://www.oracle.com/technology/deploy/availability/pdf/OOW2000_same_ppt.pdf

Microsoft, 2006. *How to Align Exchange I/O with Storage Track Boundaries* [Microsoft's usage details on diskpar]10

http://technet.microsoft.com/en-us/library/0e24eb22-fbd5-4536-9cb4-2bd8e98806e7.aspx

Oracle, 2006. *Optimal Flexible Architecture*

http://download.oracle.com/docs/cd/B19306_01/install.102/b15704/app_ofa.htm VMware,

2006. *Setting the Maximum Outstanding Disk Requests per Virtual Machine*

http://kb.vmware.com/selfservice/search.do?cmd=displayKC&docType=kc&externalId=1268&sliceId=1&docTypeID=DT_KB_1_1&dialogID=936508&stateId=0%200%20938651

Related Documents

For further information about Volume Copy, refer to the following document:

- *SANtricity Storage Manager Volume Copy – Feature Guide for 10.x*

For further information about Remote Volume Mirroring, refer to the following documents:

- *Remote Volume Mirroring Service Planning and Delivery Guidebook*
- *Remote Volume Mirroring Installation and Configuration Guidebook for 9.19*

For further information about Snapshot, refer to the following LSI document:

- *Protecting Data in a Changing Environment: Managing Growth with PolyServe Matrix Server and LSI Snapshot*

For further information about VMware ESX 3.5 configuration, refer to the following VMware documents.

- Maltz, Brad "*Storage Best Practices with VMware ESX 3.5: How to Make It Perform Better and More Scalable*," slide 13
- New England VMUG 3-27-08 presentation by International Computerware, Inc
- *Fibre Channel SAN Configuration Guide, ESX Server 3.5, ESX Server 3i version 3.5, Virtual Center 2.5*
- *Recommendations for Aligning VMFS Partitions*
- *ESX Server 3 Configuration Guide*
- *SAN System Design and Deployment Guide*
- *Storage/SAN Compatibility Guide for ESX Server 3.5 and ESX Server 3i*

Appendix A

AVT Disable Script

This appendix shows the disable script for AVT.

```
/* Disable AVT in all the host regions */ show
"Disabling AVT on Controller A...";
set controller[a] HostNVSRAMByte [0x00,0x24=0x00;
set controller[a] HostNVSRAMByte [0x01,0x24=0x00;
set controller[a] HostNVSRAMByte [0x02,0x24=0x00;
set controller[a] HostNVSRAMByte [0x03,0x24=0x00;
set controller[a] HostNVSRAMByte [0x04,0x24=0x00;
set controller[a] HostNVSRAMByte [0x05,0x24=0x00;
set controller[a] HostNVSRAMByte [0x06,0x24=0x00;
set controller[a] HostNVSRAMByte [0x07,0x24=0x00;
set controller[a] HostNVSRAMByte [0x08,0x24=0x00;
set controller[a] HostNVSRAMByte [0x09,0x24=0x00;
set controller[a] HostNVSRAMByte [0x0a,0x24=0x00;
set controller[a] HostNVSRAMByte [0x0b,0x24=0x00;
set controller[a] HostNVSRAMByte [0x0c,0x24=0x00;
set controller[a] HostNVSRAMByte [0x0d,0x24=0x00;
set controller[a] HostNVSRAMByte [0x0e,0x24=0x00;
set controller[a] HostNVSRAMByte [0x0f,0x24=0x00;
show “Complete”;
show "Disabling AVT on Controller B...";
```

set controller[b] HostNVSRAMByte [0x00,0x24=0x00;

set controller[b] HostNVSRAMByte [0x01,0x24=0x00;

set controller[b] HostNVSRAMByte [0x02,0x24=0x00;

set controller[b] HostNVSRAMByte [0x03,0x24=0x00;

set controller[b] HostNVSRAMByte [0x04,0x24=0x00;

set controller[b] HostNVSRAMByte [0x05,0x24=0x00;

set controller[b] HostNVSRAMByte [0x06,0x24=0x00;

set controller[b] HostNVSRAMByte [0x07,0x24=0x00;

set controller[b] HostNVSRAMByte [0x08,0x24=0x00;

set controller[b] HostNVSRAMByte [0x09,0x24=0x00;

set controller[b] HostNVSRAMByte [0x0a,0x24=0x00;

set controller[b] HostNVSRAMByte [0x0b,0x24=0x00;

set controller[b] HostNVSRAMByte [0x0c,0x24=0x00;

set controller[b] HostNVSRAMByte [0x0d,0x24=0x00;

set controller[b] HostNVSRAMByte [0x0e,0x24=0x00;

set controller[b] HostNVSRAMByte [0x0f,0x24=0x00;

```
show "Complete";

show "You must now reboot both controllers for these changes to take
effect!";
```

Legal Notices

Document Description

This document describes the best practices for using VMware with an LSI 2600 storage system.

Ownership of Materials

The Document is provided as a courtesy to customers and potential customers of LSI Corporation ("LSI"). LSI assumes no obligation to correct any errors contained herein or to advise any user of liability for the accuracy or correctness of information provided herein to a user. LSI makes no commitment to update the Document. LSI reserves the right to change these legal terms and conditions from time to time at its sole discretion. In the case of any violation of these rules and regulations, LSI reserves the right to seek all remedies available by law and in equity for such violations. Except as expressly provided herein, LSI and its suppliers do not grant any express or implied right to you under any patents, copyrights, trademarks, or trade secret information. Other rights may be granted to you by LSI in writing or incorporated elsewhere in the Document.

Performance Information

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of LSI products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they want to purchase.

Disclaimer

LSI has provided this Document to enable a user to gain an understanding of LSI CTS2600 Configurable Storage Components . This document is designed to assist a user in making a general decision as to whether an LSI CTS2600 storage system configuration is appropriate for such user's objectives. Neither this Document nor the Tool are designed or intended to guarantee that the configuration a user chooses will work in a specific manner. While the guidance provided by this Document can help a user to choose an appropriate configuration (or avoid a configuration that is not appropriate), there is no way LSI can guarantee the exact performance and/or results of the Information contained in this Document. Accordingly, LSI assumes no obligation whatsoever for the use of the Information provided in this Document, AND UNDER NO CIRCUMSTANCES WILL LSI OR ITS AFFILIATES BE LIABLE UNDER ANY CONTRACT, STRICT LIABILITY, NEGLIGENCE OR OTHER LEGAL OR EQUITABLE THEORY, FOR ANY SPECIAL, INDIRECT, INCIDENTAL, PUNITIVE OR CONSEQUENTIAL DAMAGES OR LOST PROFITS IN CONNECTION WITH THIS DOCUMENT.

THE INFORMATION AND MATERIALS PROVIDED IN THIS DOCUMENT IS PROVIDED "AS IS" AND LSI MAKES NO WARRANTIES EXPRESS, IMPLIED OR STATUTORY, INCLUDING, WITHOUT LIMITATION, THE IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE WITH RESPECT TO THE SAME. LSI EXPRESSLY DISCLAIMS ANY WARRANTY WITH RESPECT TO ANY TITLE OR NON-INFRINGEMENT OF ANY THIRD

PARTY INTELLECTUAL PROPERTY RIGHTS, OR AS TO THE ABSENCE OF COMPETING CLAIMS, OR AS TO INTERFERENCE WITH USER'S QUIET ENJOYMENT.