Advanced SerialRAID Plus Adapter Planning Guide

Note: Before using this information and the product it supports, be sure to read the general information under Appendix A on page 77.

If this document is being viewed on a screen rather than being printed, it is recommended that Adobe Acrobat 4.0 is used. If a level prior to 4.0 is used, some of the lines in the figures may not display unless a magnificantion of >100% is used.

Second Edition (October 2000)

Table of contents

Preface

This document is intended for system designers and administrators who are planning the inclusion of SSA RAID subsystems in their RS/6000 systems. It is written for the Advanced SerialRAID Adapter Plus although sections referring to non-RAID and RAID 5 also apply to previous SSA RAID adapters. It attempts to compare the different RAID or non-RAID configurations for performance and availability to assist in determining which configuration is best suited for the particular environment. It also gives details on how to configure a selected RAID type for optimum performance and availability.

The performance numbers quoted are to be used for comparison between the different types of RAID and are correct only for the environment that was tested. You should not infer that this performance is what would be achieved in your system with your particular applications.

This document provides

- an introduction to Serial Storage Architecture (SSA)
- an introduction to the Advanced SerialRAID adapter
- a description of RAID 1, RAID 0+1, RAID 5 and mirrored logical volumes
- a comparison of the reliability and data availability characteristics of various types of RAID
- some suggestions on how to get the best performance and availability from an SSA subsystem
- some hints and tips relating to the use of the SSA RAID configurator

How This Book is Organised

Section 1: “Introducing SSA and the Advanced SerialRAID Plus adapter”

This describes key characteristics of SSA and of the Advanced SerialRAID adapter

Section 2: “Introducing LVM Mirroring and RAID”

This describes LVM mirroring and the various types of RAID possible and comments on the general advantages and limitations of each type.

Section 3: “Planning for Performance”

This discusses the main factors that affect performance.

Section 4: “Performance Comparison of RAID Types”

This compares the performance of different RAID types for certain configurations. It can be used for comparative purposes of the different types of RAID. It should not be used to predict the exact performance that will be achieved on a given system as the application workload and configuration will be certainly be different.

Section 5: “Planning for Availability”

This discusses how to achieve maximum availability of access to data

Section 6: “Planning your SSA configuration”

This summarises factors that should be considered when planning your subsystem for performance and availability

Section 7: “Configuration Optimization”

This discusses various options available when creating an array that can help to optimize the performance and availability

Section 8: “Comparison of Striping Data Options”

This discusses various options available when striping data across disks when availability of data when a disk fails is not necessary

Section 9: “Split Copy Options”

This discusses various options available to split a copy of the data to separate disks that can then be later backed up to tape or other medium.

Section 10: “Getting More Information About SSA”

This lists where you can obtain further information about SSA and SSA adapters and disks.

Glossary

Availability The ability of a system to continue operation in the presence of failures in one or more components of the system. When applied to disk arrays, it relates to the ability to continue to access and manipulate data in the presence of failures on a disk.

Data Scrubbing A background process that verifies all the data can be read from each array member disk without error. This discovers blocks that cannot be read and that have not been accessed by I/O operations from the host. The RAID manager can then reassign that block and rewrite it with data it has reconstructed from the other array members. In this way when a hot spare is introduced after a disk failure, data can be read successfully from the other member disks to reconstruct the data to be written to the hot spare.

Hot Spare A disk that is not currently being used and has been configured to be available to automatically replace a failed member disk of an array.

Member A disk that has been configured to be part of an array.

Mirroring Two copies of data are held on different disks. This is the means of achieving availability for RAID 1, RAID 10 and LVM mirroring.

RAID 0 Data is distributed across the disks in an array; no redundant information is provided, so data is lost if any disk fails

RAID 1 Data is held on two separate disks. For writes, both disks have to be written. For reads either disk can be read. Availability of data when a disk fails is provided.

RAID 5 Data is distributed across the disks with the exclusive or of the data on each disk held on one of the disks. The disk that holds the exclusive or data is periodically rotated to another disk to even out the disk workloads as the parity data is accessed more than other data. Availability of data when a disk fails is provided.

RAID 0+1 Data is distributed across the disks and two copies of the data are held on separate disks. For writes, two disks have to be written. For reads either disk of mirrored pairs can be read. Availability of data when a disk fails is provided. In the implementation on the Advanced SerialRAID adapter, all the remaining disks of the array can be used when one disk has failed and is not replaced by a hot spare.

Degraded This is the state of an array when one of its disks has failed and write operations have been issued to the array.

Rebuild This is the process of restoring the data that was on a failed disk onto a replacement disk of the array. It involves reading the mirrored disks in RAID 1, RAID 0+1 and mirrored logical volumes and in RAID 5 it involves reading all the other members of the array to reconstruct the data.

Reconstruct The dynamic calculation in RAID 5 of data that cannot be read from a disk by reading data from all the other disks of the array.

Strip A contiguous portion of the array data that is held on one disk of the array before switching to the next disk.

Stripe A portion of the array data starting with one strip on one disk and proceeding with the corresponding strip on each of n-1 disks for a RAID 5 array and for n/2 disks for a RAID 0+1 array where n is the number of member disks. In RAID 5 the nth disk contains the exclusive or of the strip data on the other member disks.

Stretch In RAID 5, the number of stripes that have the parity of the data held on one member disk before it is rotated to the next disk

1 Introducing SSA and the Advanced SerialRAID Plus adapter

This chapter introduces SSA and the key characteristics of the Advanced SerialRAID Plus adapter.

Introducing SSA

Serial Storage Architecture (SSA) is a high performance, serial interconnect technology used to connect disk devices and host adapters. SSA is an open standard, and SSA specifications has been approved by the SSA Industry Association and is an ANSI standard through the ANSI X3T10.1 subcommittee.

SSA subsystems are built up of loops of adapters and disks. A simple example is shown in Figure 1.

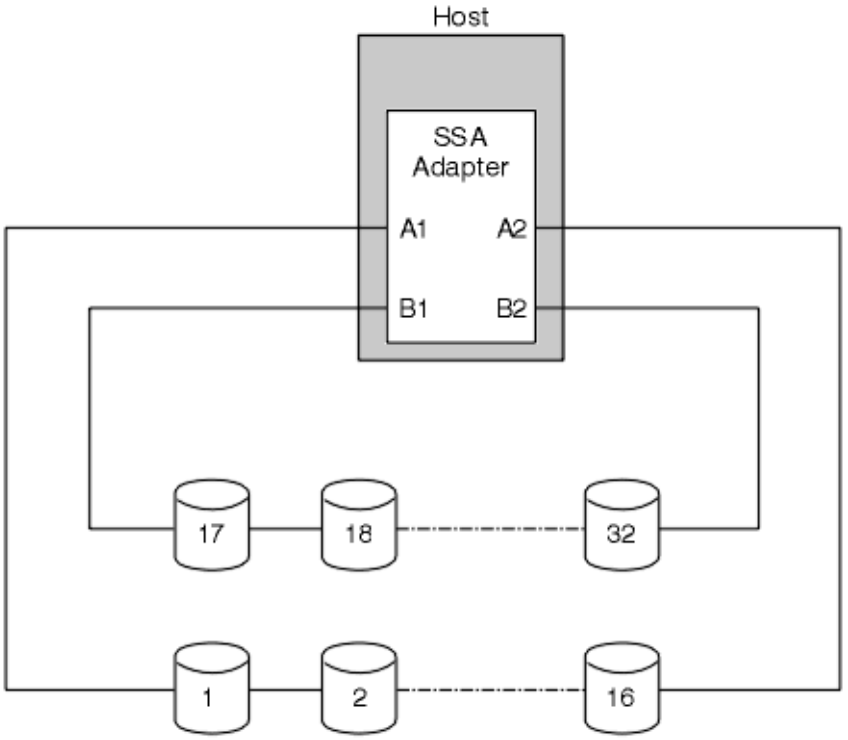

Figure 1: A Simple SSA Disk Subsystem

Here, a single adapter controls two SSA loops, each of 16 disks. Data can be transferred around each loop, in both directions, at 40 MB/sec, consequently the peak transfer rate at the adapter, on each loop, is 160 MB/sec. Since the adapter contains two SSA nodes, one for each loop, the peak SSA data rate at the adapter could be 320 MB/sec, but the adapter and PCI attachment restrict this to about 90 MB/sec. Early disk drives were only capable of operating at 20 MB/sec on the SSA loop. These can still be used with the Advanced SerialRAID Plus adapter even though that is capable of operating at 40 MB/sec. An SSA loop can consist of any combination of devices that can operate at

40 MB/sec or 20 MB/sec. Each node arbitrates with the next node to determine what SSA speed each can operate and these nodes then communicate to each other on that section of the loop at the speed at which they can both operate. A SSA loop may therefore consist of some sections that operate at 40 MB/sec and others that operate at 20 MB/sec.

An SSA node can be either an initiator or a target. An initiator issues commands, while a target responds with data and status. The two SSA nodes in an adapter are therefore initiators while the SSA node in a disk drive is a target. Each SSA node is given a unique address at manufacturing time, known as the UID. This allows the initiators in the loop to determine what other SSA nodes are attached to the loop and to understand the SSA loop topology.

SSA architecture allows more than one initiator to be present in a loop. In that case, commands and data from multiple initiators can be directed to the same or different targets and intermixed freely. The SSA network is managed by one particular initiator, known as the master initiator. This is the initiator with the highest UID. If a new initiator is added to the network with a higher UID than those currently present, then it will take over the master responsibilities for that network. Similarly, if a master initiator is removed from the SSA network, then the initiator with the next highest UID takes over master responsibility. This master takeover occurs automatically without any user intervention. In the SSA configurations used by the Advanced SerialRAID Plus adapter, initiators are adapters and targets are disk drives or other adapters.

The basic unit of data transferred between SSA nodes is a frame, which can contain up to 128 data bytes. SSA nodes are connected together with separate cables or internally within the enclosures to form a loop. The loop has several advantageous characteristics:

- Full duplex support is provided on each link, so that traffic can be present in both directions on the loop simultaneously.
- The loop supports spatial reuse - that is to say, different frames may be transferred between different devices on the loop concurrently. For instance, a frame could be moving from disk 1 to disk 2 at the same time as the adapter is sending a frame to disk 1 (since each disk is dual ported). At the same time a frame may be transmitted from adapter port B1 to disk 17.
- Since a loop topology is used, should a break occur in the loop (e.g. a cable is disconnected, or a disk fails), each adapter on the loop adjusts its routing algorithms, under direction from the master initiator, so that frames are automatically rerouted to avoid this break. This allows devices to be removed or added to the loop while the subsystem continues to operate without interruption.

Summary of Advantages of SSA

- Dual paths to devices
- Cabling simplified - cheaper, smaller cables and connectors, no separate terminators
- Faster interconnect technology
- Full duplex, frame multiplexed serial links
- 40 Mb/sec total per port in each direction, resulting in 160 MB/sec total per node, although a single adapter is not capable of delivering this entire bandwidth.
- Spatial reuse allows concurrent activity in different portions of the same loop
- Hot pluggable cables and disks
- Very high capacity per adapter (up to 127 devices per loop, although most adapter implementations limit this e.g. current IBM SSA adapters limit is 48 devices on each of two loops))
- Large distance between devices (up to 25m with copper cables, 10 km with optical links)
- Auto-configuring - no manual address allocation
- SSA is an open standard

Introducing the Advanced SerialRAID Plus adapter

The Advanced SerialRAID Plus adapter (code above level 7600) supports:

- Non-RAID disks with 8 adapters
- RAID 0 with 1 adapter
- RAID 1 with up to 2 adapters
- RAID 5 with up to 2 adapters
- RAID 0+1 with up to 2 adapters
- Fast Write Cache disk or array with up to 2 adapters

The Advanced SerialRAID adapter (code below level 7600) supports:

- Non-RAID disks with up to 8 adapters
- RAID 0 with 1 adapter
- RAID 5 with up to 2 adapters
- Fast Write Cache disk or array with one adapter

The Advanced SerialRAID Plus adapter Plus has code above level 7600. The Advanced SerialRAID adapter had code below level 7600 and that adapter can be converted to support the additional functions by downloading code at above level 7600.

The Advanced SerialRAID Plus adapter provides:

- Two pairs of SSA ports for the attachment of SSA disk drives conforming to the *Serial Storage Architecture - 1995 Physical and Serial Storage Architecture - 1995 SCSI-2 Protocol.* Each port can operate at 20 or 40 MB/s full duplex up to 25m long using copper cable. Switching between the two SSA speeds is performed dynamically, according to the capabilities of the attached devices.

  The maximum distance between any two SSA devices can be extended to 10 km by using the optical cables. When using optical cables, the data rate possible between the two SSA ports reduces as the distance increases because of the increase in time taken to transmit some control information. The data rate in each direction between any two ports when optical cables are used is:

Distance between ports	**Data rate in each direction**
1 km	12 MB/sec
2 km	6 MB/sec
10 km	1 MB/sec

- Two SSA initiators, one for the A1/A2 port and one for the B1/B2 port.
- Up to 48 disks per loop
- 2 loops, one per initiator
- Non-RAID, RAID 0, RAID 1, RAID 5 and RAID 0+1 operation
- RAID 0 support for:
  - Striping data across from 2 to 16 member disks (with no protection for disk failures)

- RAID 1 support for:
  - Mirroring data on 2 member disks
  - Tolerant to a single disk failure or medium error
  - Tolerant to a multiple medium failures provided not to the same blocks on each member disk
  - Data Scrubbing disks in background for early detection of disk failures
  - Automatic hot spare takeover - hot spares do not need to be pre-allocated to a particular array and are simply taken from a pool as required
  - Up to 2 adapters can access the array
  - The array can be configured so that the total loss of one system and the disks on that site does not prevent access to the array from the other system.
- RAID 5 support for:
  - Arrays may be made up of any mix of between 2+P and 15+P member disks
  - Up to 32 arrays per adapter
  - Tolerant to a single disk failure or medium error
  - Tolerant to a multiple medium failures provided not to the same blocks on multiple member disks
  - Automatic hot spare takeover - hot spares do not need to be pre-allocated to a particular array and are simply taken from a pool as required
  - Delayed writing of the parity data to improve the response time
  - Read cache
  - Data Scrubbing disks in background for early detection of disk failures
  - Up to 2 adapters can access the array
- RAID 0+1 support for:
  - Mirroring data and striping data across disks on 4 to 16 member disks (only even number of disks)
  - Tolerant to a single disk failure or medium error
  - Automatic hot spare takeover - hot spares do not need to be pre-allocated to a particular array and are simply taken from a pool as required
  - Up to 2 adapters can access the array
  - The array can be configured so that the total loss of one system and the disks on that site does not prevent access to the array from the other system.
  - Tolerant to multiple disk failures provided not to the same mirrored pair of disks
  - Tolerant to a multiple medium failures provided not to the same blocks on mirrored pair of disks
  - Data Scrubbing disks in background for early detection of disk failures
- Fast write cache support for:
  - 32 MB NVRAM write cache
  - Can be used with RAID or non-RAID disks
  - Up to 2 adapters can access disks or arrays for which fast write cache is enabled
  - Fast write cache operation can be defined for selected LBA ranges within a disk

- High performance:
  - up to 8,500 non-RAID ops/sec (70/30 read/write, 4k transfers, random seeks)
  - up to 6,500 RAID 0 ops/sec (70/30 read/write, 4k transfers, random seeks with no cache hits)
  - up to 4,700 RAID 1 ops/sec (70/30 read/write, 4k transfers, random seeks with no cache hits)
  - up to 2700 RAID 5 ops/sec (70/30 read/write, 4k transfers, random seeks with no cache hits) and up to 9,000 ops/sec (100% read with cache hits)
  - up to 5,000 RAID 0+1 ops/sec (70/30 read/write, 4k transfers, random seeks with no cache hits)
  - data transfer at up to 90 MB/sec
- Attaches to any mix of 7133 and 7131 disk enclosures

Clusters

The number of adapters supported on the same SSA loop depends on the adapter type, the type of RAID being used on that loop and the level of code in the adapter. The number of adapters supported with the Advanced SerialRAID Plus adapter is:

Array Type	Adapters in loop	Adapter type
Non-RAID	8	Advanced SerialRAID Plus adapter PCI SSA Multi-Initiator/RAID EL Adapter Micro Channel SSA Multi-Initiator/RAID EL Adapter
RAID 0	1	Advanced SerialRAID Plus adapter
RAID 1	2	Advanced SerialRAID Plus adapter at code level above 7600
RAID 5	2	Advanced SerialRAID Plus adapter PCI SSA Multi-Initiator/RAID EL Adapter Micro Channel SSA Multi-Initiator/RAID EL Adapter
RAID 0+1	2	Advanced SerialRAID Plus adapter at code level above 7600
Fast Write	1	Advanced SerialRAID Plus adapter at code level below 7600
	2	Advanced SerialRAID Plus adapter at code level above 7600

Table 1. Adapters supported in a cluster

Host software

The AIX software components for SSA Raid adapters include:

- Adapter and disk device drivers
- Diagnostic and Service packages
- RAID Configurator

The AIX SSA device drivers provide support for the following devices:

- pdisks

  These are devices, associated with each physical disk, on a one to one basis. In non-RAID configurations there is a one to one mapping between pdisks and hdisks. In RAID configurations, more than one pdisk will map onto one hdisk.

- hdisks

  These correspond to either arrays or single disks, and are the resources to which I/O is normally done, or which are included in logical volumes and which are used to store filesystems.

- ssa0, ssa1...

  These are the adapter devices

- ssar

  This is the SSA router device. This is pseudo-device that is not associated with a particular piece of hardware. Its function is to manage the allocation of pdisks and hdisks to individual adapters.

2 Introducing LVM Mirroring and RAID

Users face a wide choice of possibilities when considering methods of enhancing data availability or performance in a disk subsystem. The alternatives are:

- Mirroring at the Logical Volume Manager (LVM) level within the AIX operating system
- RAID 1 functions provided within the Advanced SerialRAID Plus adapter
- RAID 5 functions provided within the Advanced SerialRAID Plus adapter
- RAID 0+1 functions provided within the Advanced SerialRAID Plus adapter

Each of these have advantages and also limitations. This section attempts to compare each of these options in turn so that the user can choose the type of configuration most suitable for the environment.

LVM Mirroring and data availability

The Logical Volume Manager (LVM) provides a method of dividing and allocating storage space on the host's disks. LVM manages disk resources such as single disks or arrays of disks so that they may be collected together in logical groups. This allows data to be striped over multiple disks for performance reasons and/or replicated on multiple disks for improved data availability and performance. LVM manages both physical volumes and volume groups.

A physical volume is a hard disk drive or an array of hard disk drives on which read and write operations can be performed. Each physical volume is given a name of the form hdiskX where X is a number, for example hdisk23. A physical volume is divided into a number of physical partitions. Each of these partitions represents the smallest unit of disk space that can be allocated to a logical volume.

A volume group is a collection of physical volumes of varying size and type. A volume group is divided up into one or more logical volumes, within which filesystems or other data structures may be stored. Logical volumes are divided up into logical partitions. The user may choose to map 1, 2 or 3 physical partitions onto a single logical partition. This in effect means there are 1, 2 or 3 copies of the data stored in the logical volume. These copies are usually on different physical volumes. Physical volumes can be added or removed from the volume group as required and data can be moved from one physical volume to other volumes in the same group. This allows the user to move data off a physical volume that is at high risk of becoming unavailable and remove the physical volume from the group.

The ability to specify more than one copy of the logical volume data through logical partitions is called mirroring. Mirroring reduces data access time by allowing the LVM to choose which copy (physical partition) of the data can be accessed quickest.

Mirroring also increases data availability. Normally, whenever data on a logical volume is updated, all the physical partitions containing the logical volume copies are automatically updated. However, physical partitions containing copies can become stale (no longer containing the most current data) as a result of system malfunctions or because the physical volume was unavailable at the time of an update to a logical volume. The LVM can refresh stale partitions to a consistent state by copying the current data from a valid physical partition to the stale partition. This refresh can take place when the system is restarted, when the physical volume comes back on line, or at the user's request.

With SSA, mirroring can provide assistance in recovering from a disaster, for example loss of power on an entire site. One pair of physical disks can be located up to 10 km (32,808 feet) from the mirrored pair of disks using fiber optic cables and the optional fiber optic extender feature of the 7133 SSA Disk Subsystem. If one of the sites suffers a disaster, the data on the disks at the other site is unaffected and the system can continue to operate. Whilst this is possible with LVM mirroring, the mirrored pair of disks have to be in separate sites and, unlike for RAID 1 or RAID 0+1, no assistance is provided in identifying the physical location of the disks for mirroring under LVM.

LVM mirroring can be used in 2-way systems that share data with HACMP 6000 using concurrent LVM.

It is possible to stripe data across several mirrored pairs of disks entirely by host software to enhance performance. If I/O operations access certain logical volumes more frequently than others or the logical blocks accessed are in close proximity, the performance may be limited by a few disks if the disks are configured as just physical disks or members of a LVM mirrored pair without striping. If data is striped across several LVM mirrored disks, there will be

a more even distribution of I/O operations across all the disks, so enhancing performance by eliminating hot spots. When a disk fails and is not replaced by a hot spare, all the remaining disks of the striped mirrored pairs can still be used. Striping LVM mirrored disks therefore provides the availability benefit of mirroring with the performance benefit of striping.

Summary of the benefits and limitations of using LVM Mirroring:

This highlights the strengths and weaknesses of configuring disks this way. Some of the advantages and disadvantages may also apply to other types of RAID configurations.

Advantages

- Data can be mirrored on three disks rather than having just two copies of data. This provides higher availability in the case of multiple failures, but does require more disks for the three copies.
- The disk types do not have to all be SSA disks. The disks used in the physical volumes could be of mixed attachment types.
- Logical volumes are mirrored. These could be on the same disk that allows mirroring on a single physical disk. This provides protection against medium errors but not against disk failures. It does allow for an odd number of disks to be used and provide protection for disk failures when more than 1 disk is used.
- The disks can be configured so that mirrored pairs are in separate sites or in different power domains. In this case, after a total power failure on one site, operations can continue using the disks on the other site that still has power. No information is displayed on the physical location of each disk when mirrored logical volumes are being created, unlike when creating RAID 1 or RAID 0+1 arrays, so allocating disks on different sites is not easy.
- Mirrored pairs can be on different adapters
- Read performance is good for short length operations as data can be read from either of two disks, so the one with the shortest queue of commands can be used. Write performance requires a write to two disks.
- Mirrored copies can be split for backup purposes and later rebuilt
- Data can be striped across several mirrored disks that avoids hot spots caused by excessive activity to a few disks by evening out the I/O operations across all the member disks.

Disadvantages

- Incurs greater load for writes on the system than when RAID is implemented in the adapter, so performance may be impacted on heavily loaded systems. Host processor utilization may be 25% higher for write operations to a mirrored pair of disks than for writes to RAID arrays. The host processor utilization for read operations is approximately the same for mirrored disks or RAID arrays.
- Data has to be transferred twice across a PCI bus to an adapter for all write operations. If the PCI bus is heavily loaded, this can result in loss of performance for applications that are writing large volumes of data.
- Mirror Write consistency must be set on if the applications require that the two copies of data must always be the same. If Mirror Write Consistency is off, after an error or a power off it is possible that both copies of mirrored data can be different. Running with Mirror Write Consistency on does incur a performance loss that may be significant. Enabling Fast Write Cache for these logical volumes will help to reduce the loss of performance when Mirror Write Consistency has to be on.
- Hot spares are not automatically introduced to replace disks that fail. Scripts are available and must be run to perform this function to automatically introduce hot spares. These scripts can be downloaded from the SSA website at http://www.storage.ibm.com/hardsoft/products/ssa.
- No background data scrubbing of blocks to detect unrecoverable data checks on disks that would result in not being able to restore that block following a disk failure.

RAID arrays

Disk arrays are groups of disk drives that act like one disk as far as the operating system is concerned and which provide better availability or performance than individual drives operating alone. Depending on the particular type of array that is used, it is possible to optimize availability or performance or to select a compromise between both. An array is a set of multiple disk drives plus a specialised controller that keeps track of how the data is distributed across the drives. All the member disks of an array must be on the same SSA loop. Data for a given file may be written in segments to the different drives in the array, rather than being written to a single drive. For some types of array this can provide higher data transfer rates or higher I/O rates than that of a single large drive.

Arrays can provide data redundancy, so that no data is lost if a single drive in the array fails. Depending on the particular organisation of the array, data is either mirrored or striped. Different ways of organising the array are known as different RAID levels. (RAID is an acronym for Redundant Array of Independent Disks). The most commonly available RAID levels are summarised in Table 2 on page 18.

RAID arrays, except RAID 0, contain redundant information to increase data availability. When a disk fails or a block cannot be read, this redundant information can be used to reconstruct the data that was on the failed disk. However, as soon as a disk has failed, there is no further redundancy. This means that a second disk failure may result in data loss. In fact, the level of data protection provided when a disk has failed within a RAID array can be significantly less than that provided by the equivalent single disk. This is because the probability of a subsequent failure in the RAID array is higher than that of the single disk, because of the larger number of disks involved in the RAID array. It is therefore important that when a disk does fail, it is replaced as soon as possible. This can be achieved by the use of a hot spare disk. A hot spare disk waits until a disk in the RAID array fails, and then automatically replaces it.

The steps in restoring RAID protection are:

- A disk fails within a RAID array
- The RAID management firmware in the adapter notes the failure, and marks the array as having one member missing.
- Read operations can still be performed to the array since the RAID management firmware can use the redundant information in the remainder of the array to reconstruct data. Write operations can also be carried out by default although this can be optionally inhibited.
- The RAID management firmware selects a suitable hot spare disk and exchanges it with the failed disk in the array.
- The RAID management firmware now rebuilds the data that should be on this new disk from the redundant information elsewhere in the array. While this rebuild operation goes on, the array still does not have its full complement of members, and could be subject to data loss if a second disk failure occurs.
- When the data on the new disk has been fully rebuilt with the reconstructed data, the array is restored to the full level of RAID protection.

The RAID management firmware in the adapter has thus been able to maintain data availability because of the redundant information in the array, even though one disk in the array has failed. It has also been able to minimise the period that the array is exposed to data loss by using the hot spare mechanism to automatically replace and rebuild the failed disk. On the Advanced SerialRAID Plus adapter, a failed disk in an array is automatically replaced by a hot spare. It is recommended that there is always at least one hot spare disk available to reduce the exposure to data loss.

RAID level	Common Name	Description	Disk Drives	Data Availability	Data transfer capacity	I/O request rate
0	Disk Striping	Data distributed across the disks in the array. No redundant information provided.	N	same as a single disk	very high	high for both read and write. I/O requests distributed evenly across disks.
1	Mirroring	All data replicated on N separate disks. N is most commonly 2.	2N	higher than RAID level 3,5 and single disk	similar to a single disk	higher than that of a single disk for reads, less than a single disk for writes
3 (not supported)	Parallel transfer disks with parity	Each data sector is subdivided and distributed across all data disks. Redundant information normally stored on a dedicated parity disk	N+1	much higher than a single disk;	very high	higher than that of a single disk
5	RAID 5	Data sectors are distributed as with disk striping, redundant information is interspersed with user data.	N+1	much higher than a single disk;	similar to disk striping for read; lower than single disk for writes	similar to Disk Striping for reads; lowest of all types for writes. I/O requests distributed evenly across disks.
0+1	Mirroring with striping	A combination of RAID 1 and RAID 0 in which data is striped across a number of mirrored disks	2N	higher than RAID level 3,5 and single disk	similar to disk striping for reads; lower for writes	higher than Disk Striping for reads, less than Disk Striping for writes. I/O requests distributed evenly across disks.

Table 2: Raid Levels

The Advanced SerialRAID Plus adapter supports RAID 0, RAID 1, RAID 5 and RAID 0+1 types of array. These are controlled entirely within the adapter. Different RAID levels are appropriate for different situations:

RAID 0 in the Advanced SerialRAID Plus adapter

Select RAID 0 for applications that benefit from the increased performance capabilities of this RAID level but, because no data redundancy is provided, do not use RAID 0 for mission critical applications that require high availability.

A RAID 0 array can be built from 2 to 16 member disks. The strip size is fixed at 16 KB. Applications that predominantly transfer large amounts of data, typically sequentially, benefit because data is striped across multiple disks. Applications that predominantly involve random accesses with short, for example 4 KB, data transfers can also benefit because spreading data across disks spreads the disk activity fairly evenly across all the disks. Workloads typically do not issue operations to logical disks randomly. They usually involve frequent accesses to blocks that are in close proximity to previously accessed blocks on the same disk or to certain disks more frequently than other disks. This skewing of activity to certain disks can have the effect that the performance is constrained by a few disks and other disks are not being fully used. Striping data across several disks reduces this imbalance to heavy activity on only a few disks.

Summary of the benefits and limitations of using RAID 0:

This highlights the strengths and weaknesses of configuring disks this way. Some of the advantages and disadvantages may also apply to other types of RAID configurations.

Advantages

- High data transfer rate possible for long data transfers because data is striped across multiple disks
- Good performance for random access, short data transfer because skewing accesses to certain disks is reduced

Disadvantages

- No redundancy, so loss of a disk or a medium error is a data loss condition
- When a disk does fail, the amount of data lost is that of the entire array and not just that of a single disk
- Only supported in a single adapter configuration

RAID 1 in the Advanced SerialRAID Plus adapter

Select RAID 1 for applications where data availability is a key concern, and that have high levels of write operations, such as transaction files, and where the larger number of drives required is not a major consideration.

A RAID 1 array is built from 2 member disks and data is mirrored on each disk. Write operations involve writing the data to both member disks and the write is not considered complete until the data has been written to both disks. The response time to a write is therefore longer than to a non-RAID disk. This disadvantage can be avoided by enabling fast write cache for the RAID 1 array. Read operations can be sent to either of the mirrored pair of disks, so read performance is enhanced over non-RAID disks.

Data is sent only once across the PCI bus for write operations, so PCI bus utilisation is less than when using LVM mirroring. The adapter remembers in non-volatile memory if a write operation has completed to one disk but not to the other disk due to a failure. It ensures that the same data is returned to the system whichever of the mirrored pair of disks is read for the affected blocks. LVM mirroring offers this as an option, (mirror write consistency), but if enabled for this mode a reduction in performance is incurred for all writes.

Member disks of a RAID 1 array can be configured in separate domains, for example in different sites. When configured in this way, an entire site can fail, for example through loss of power, and the system in the other domain can continue operation with the array members in that domain. When the member disks from the domain that was not accessible return to the SSA loop they are rebuilt from the members that have been in operation during their absence. The Advanced SerialRAID Plus adapter ensures that systems do not operate independently on different member disks when both domains cannot communicate yet each domain could still be operational, for example a communication link loss between domains but each domain is still active.

Hot spare disks replace member disks automatically when a member disk fails. These hot spares can be configured into hot spare pools. Each member disk can be assigned a hot spare pool to be searched for replacement when that member disk fails. Using hot spare pools it is possible to arrange that hot spare replacement disks are in the same domain as the disk that is being replaced so that the mirrored pair of disks are in different domains even after a hot spare replacement.

As two disks are used as mirrored pairs, the number of disks required for RAID 1 arrays is twice that of non-RAID and more disks are required than for RAID 5 arrays.

Data is verified on both member disks in the background to normal operation to ensure that no unrecoverable data checks exist on the disks. If one is detected, that block is reassigned and rewritten from the mirrored disk.

RAID 1 arrays can be configured with fast write cache enabled and this reduces the response time for write operations.

RAID 1 arrays can be used in configurations with one or two adapters in the loop. When two adapters are in the loop, each adapter informs the other of its operations to the array. If an adapter fails at any time, the other adapter is able to continue with operations on the disk and the integrity of the data on the array is maintained. Locks are exchanged between adapters to ensure that they do not both change the same area of the disk at the same time.

Summary of the benefits and limitations of using RAID 1:

This highlights the strengths and weaknesses of configuring disks this way. Some of the advantages and disadvantages may also apply to other types of RAID configurations.

Advantages

- High availability of data. No data loss for a medium failure, inability to read a block or a disk failure.
- Read performance is good because data can be read from either of two disks, so the one with the shortest queue of commands can be used.
- The disks can be configured so that mirrored pairs are in separate sites or in different power domains. In that case, after a total power failure on one site, operations can continue using the disks on the other site that still has power. The configurator assists in identifying the physical enclosure that packages each candidate disk.
- The adapter ensures that after any failure or power off that causes a write operation not to be completed successfully, the same data will be returned whichever disk is read after the failure. This function does not incur any performance loss. In LVM mirroring this mirror write consistency is possible, but if it is enabled a reduction of performance is incurred.
- Hot spares are automatically introduced to replace disks that fail.
- Data scrubbing of blocks to detect unrecoverable data checks on disks that would result in not being able to restore that block following a disk failure is performed in the background.

Disadvantages

- Double the number of disks for the required capacity must be provided to ensure availability if a disk fails.
- Write performance is better than for RAID 5, but the response time for writes is higher than for non-RAID disks because two disks have to be written. Enabling for fast write cache significantly reduces the response time to writes.

RAID 5 in the Advanced SerialRAID Plus adapter

Select RAID 5 as a good compromise between high performance, high availability and least number of disks. It requires the least number of disks of all RAID types that provide redundancy. Performance is reduced if the applications involve a high proportion of write operations as these involve multiple disk accesses. Enabling the array for fast write cache may avoid this reduction in write performance. Read performance is good as only one disk access is generally required. It will be more suitable for applications that are predominantly querying a data base rather than those that are frequently updating one.

A RAID 5 array is built with from 3 to 16 member disks. The capacity of the array is the total capacity of all the disks minus the capacity of one disk. All disks are assumed to be the same capacity. The capacity of the smallest disk is used if all the member disks do not have the same capacity. The excess capacity on any other member disk, whose capacity is higher than that of the smallest member, is not used.

Data is organised on the disks in segments called strips. The default strip size is 64 KB, but this can optionally be 32 KB. Data is spread across the disks in strip segments. Each successive disk holds the next strip of data. If there are n disks in the array, the first strip of n-1 disks contain data and the nth disk contains a strip that consists of the parity of each data strip on the other disks. The n-1 strips plus the parity strip is known as a stripe. The same disk is used for the parity data for a number of stripes (known as a stretch) and then the next member disk in the array is used for parity strips. The data allocation employed is that the first disk of the array is used for parity data for the first stretch and then the next member disk is used for parity for the next stretch and so on.

If data on a read operation to a disk fails, or the member disk is missing, the required data is reconstructed by reading all the other member disks for that stripe and, using an exclusive or operation, the required data is reconstructed.

If a member disk fails and is no longer visible to the adapter, the next read or write to that array causes a hot spare disk to be automatically introduced into the array to replace the missing disk. Data that was previously on the disk, now missing, is reconstructed using the data and parity for each stripe and written to the replacement disk. Disks configured to be available for a hot spare can be used for any array on the same loop and are not pre-allocated to a particular array. As the hot spare disk can only be used for arrays on the same loop, if there are arrays on each of the two loops on the adapter, at least one disk on each loop should be configured as a hot spare.

A RAID 5 array can be configured with fast write cache enabled. When fast write cache is enabled, the response time for write operations is reduced because data is held in the adapter's non-volatile cache and no disk activity is required before the host is informed that the operation has completed. The other benefit of configuring with fast write cache is that when data is destaged from the adapter non-volatile cache to the disks, that destage may be for a full stripe of data even though data was received by the adapter from the host as several write operations each of which was less than a full stripe length. This is particularly beneficial for sequential write operations. With fast write cache enabled, sequential write operations to RAID 5 actually require the least I/O operations to disks.

The main disadvantage of RAID 5 is the performance for random write operations, especially without the use of fast write cache. For each write operation, the array normally performs two read operations (to read the old data and the old parity) and two write operations (to write the new data and the new parity) to member disks. This compares with two disk operations required for writes for RAID 1, RAID 0+1 and LVM mirroring. The completion of the write to the host system is indicated when the old data has been read and the new data has been written; the reading and updating of the parity data is handled as a background operation. Only two disk operations are therefore involved in the time the adapter takes to respond to a write operation, but two additional disk operations are required after this response. The utilisation of disks is therefore high which affects performance in heavily loaded environments.

The number of disk operations required is reduced if the length of the write and the alignment of the data is such that one or more complete stripes are written. In this case, the old data and parity do not have to be read so reducing the number of disk accesses required. The higher the number of member disks in a RAID 5 array, the fewer disks are required to provide for parity protection of the data for a given capacity. However, as the number of member disks increases so does the length of the stripe and so write operations are required with more blocks to be written for full stripe writes.

As data is spread across disks with granularity of a strip, applications that access blocks close together now access multiple disks for the data and so avoid skewing accesses to certain disks. If applications do issue I/O operations to blocks close together and data is not striped across disks, for example in non-RAID, RAID 1 or LVM mirroring without striping, certain disks may have more activity than others. This may limit the performance of these disks where data is not striped across several disks.

Data read from the member disk is held in the cache buffer on the adapter. Subsequent reads for data that has already been fetched during a prior read operation can be processed without requiring further accesses to a disk. A subsequent write operation for blocks which have previously been read into the cache buffer do not require the disk access to read the old data to calculate the change to the parity of the data.

Operations are allowed to continue after a member disk fails and is not accessible and no hot spare is available. This option can be inhibited by using ssaraid from the command line. When operations are allowed to continue when the array is in this degraded state, data could be lost if there is a loss of power or a reset of the adapter in the period

between writing data and updating the parity strip. In this case the parity will not be correct for the stripe that was written and that strip of data cannot be rebuilt onto a replacement drive when that is introduced. This could only occur if a hot spare was not available when the member disk failed.

The larger the number of member disks, the more data is affected by a disk failure resulting in reduced performance while rebuilding onto a hot spare and more data is exposed if a second disk fails before the rebuild has completed.

Data is verified on all member disks in the background to ensure that no unrecoverable data checks exist on the disks. If one is detected, that block is reassigned and rewritten by reconstructing the data from the other member disks of the array.

RAID 5 arrays can be used in configurations with one or two adapters in the loop. When two adapters are in the loop, each adapter informs the other of its operations on the array. If any adapter fails at any time, the other adapter is able to continue with operations on the disk and the integrity of the data on the array is maintained. Locks are exchanged between adapters to ensure that they do not both change the same area of the disk at the same time.

Summary of the benefits and limitations of using RAID 5:

This highlights the strengths and weaknesses of configuring disks this way. Some of the advantages and disadvantages may also apply to other types of RAID configurations.

Advantages

- For the same number of disks of equal capacity, RAID 5 array capacity is higher than for all other RAID types that provide redundancy and availability when a disk has failed. The array capacity is (N-1)/N of the total space on the N member disks compared to half the total space for RAID 1, RAID 0+1 and LVM mirroring that all mirror the data on different drives. RAID 5 requires the least number of disks to provide availability of data when disks fail.
- Data striping across the member disks of the array avoids hot spots caused by excessive activity to a few disks by evening out the I/O operations across all the member disks.
- Hot spares are automatically introduced to replace disks that fail.
- Data scrubbing of blocks to detect unrecoverable data checks on disks that would result in not being able to restore that block following a disk failure is performed in the background.

Disadvantages

- Disk and adapter throughput performance for short length write operations is poor as 4 disk accesses are required. This poor write performance can be improved, particularly for sequential write operations, if the RAID 5 array is configured with fast write cache active. Fast write cache coalesces several short length sequential write operations to the array into a single write operation of the length of an entire stripe when the data is destaged from the fast write cache to the disks. This destage to disks does not require so many disk accesses as individual write operations of less than a full stripe as old data and old parity do not have to first be read.
- The availability of data is high, but in some operational modes not as good as RAID 1, RAID 0+1 or LVM mirroring
- When one member disk is missing and a hot spare is not available or rebuilding data onto a hot spare has not yet completed, performance is significantly reduced because all reads and writes to the failed member disk require accesses to all the other member disks of the array.

RAID 0+1 in the Advanced SerialRAID Plus adapter

Select RAID 0+1 for applications where data availability is a key concern, and which have high levels of write operations, such as transaction files, and where the number of disks required is acceptable. It provides better performance than RAID 1 or LVM mirroring without striping because disk accesses are distributed across disks, thereby reducing the skew of the access distribution to a few disks. RAID 0+1 therefore provides the availability benefit of mirroring with the performance benefit of striping.

A RAID 0+1 array is built from 4 to 16 member disks and data is mirrored on pairs of disks. The capacity of the array is half the total capacity of all the member disks. Data is organised on the disks in configurable size segments (16 KB, 32 KB or 64 KB) called strips. When the array is created, the mirrored pairs of disks are defined. Data is spread across half the disks in strip segments each successive disk holding the next strip segment of data. Data is mirrored on the other disks, so that for each disk there is a mirrored copy on another disk.

When one disk fails and cannot immediately be replaced by a hot spare, all the remaining disks of the array can still be used for I/O operations.

Write operations involve writing the data to both member disks of each mirrored pair involved and the write is not considered complete until the data has been written to all disks. The response time to a short write may therefore be longer than to a non-RAID disk. This disadvantage can be avoided by enabling fast write cache for the RAID 1 array. Read operations can be sent to either of the mirrored pair of disks, so enhancing performance of read operations from that of non-RAID disks that are not configured for LVM mirroring.

Data is sent only once across the PCI bus for write operations, so PCI bus utilisation is less than when using LVM mirroring.

The adapter remembers in non-volatile memory that a write operation may have completed to one disk but not to the other disk due to a failure and ensures that the same data is returned to the system whichever disk is read after the failure for the affected blocks.

Member disks of a RAID 0+1 array can be configured in separate domains e.g. on different sites. When configured in this way, an entire site can fail, for example through loss of power, and the system in the other domain can continue operation with the array members in that domain. When the member disks from the domain that was not accessible return to the SSA loop they are rebuilt from the members that have been in operation during their absence. The Advanced SerialRAID Plus adapter ensures that systems do not operate independently on different member disks when both domains cannot communicate yet each domain could still be operational, for example a communication link loss between domains with each domain still being active.

Hot spare disks replace member disks automatically when a disk fails. These hot spares can be configured into hot spare pools. Each member disk can be assigned the hot spare pool to be used for replacement when that member disk fails. Using hot spare pools it is possible to arrange that hot spare replacement disks are in the same domain as the disk that is being replaced. In this case, all mirrored pairs of disks can still be in a different domain even after a hot spare replacement.

As two disks are used as mirrored pairs, twice the number of disks for a given capacity are required for RAID 0+1 arrays than for a non-RAID configuration and more disks are required than for RAID 5 arrays.

Data is verified on both member disks in the background to ensure that no unrecoverable data checks exist on the disks. If one is detected, that block is reassigned and rewritten from the mirrored disk.

A RAID 0+1 array can be configured with fast write cache enabled; this reduces the response time for write operations.

RAID 0+1 arrays can be used in configurations with one or two adapters in the loop. When two adapters are in the loop, each adapter informs the other of its operations to the array. If any adapter fails at any time, the other adapter is able to continue with operations on the disk and the integrity of the data on the array is maintained. Locks are exchanged between adapters to ensure that they do not both change the same area of the disk at the same time.

Summary of the benefits and limitations of using RAID 0+1:

This highlights the strengths and weaknesses of configuring disks this way. Some of the advantages and disadvantages may also apply to other types of RAID configurations.

Advantages

- High availability of data. No data loss for a medium failure or inability to read any block
- Disks can be configured so that mirrored pairs are in separate sites or in different power domains. In that case, after a total power failure on one site, operations can continue using the disks on the other site. The configurator assists in identifying the physical enclosure that packages each candidate disk.
- Read performance is good as data can be read from either of two disks, so the one with the shortest queue of commands can be used. Write performance requires a write to two disks.
- Data striping across the member disks of the array avoids hot spots caused by excessive activity to a few disks by evening out the I/O operations across all the member disks. For long operations, data is read or written to multiple disks, so higher data rates are possible than for LVM mirroring with no striping or RAID 1.
- The adapter ensures that, after any failure or power off, the same data will be returned whichever disk is read after the failure. This function does not incur any performance loss. In LVM mirroring this mirror write consistency is possible, but if it is enabled a reduction in performance is incurred.
- Hot spares are automatically introduced to replace disks that fail.
- Data scrubbing of blocks to detect unrecoverable data checks on disks that would result in not being able to restore a block following a disk failure is performed in the background.

Disadvantages

- Double the required capacity must be provided to ensure availability if a disk fails.
- Write performance is better than for RAID 5, but the response time for writes is higher than to non-RAID disks because two disks have to be written. Enabling for fast write cache significantly reduces the response time to writes.

3 Planning for Performance

The performance achieved by a disk subsystem depends greatly on the particular application being run, the I/O mix (ratio of reads to writes, and seek profiles) that the application generates and the environment that the subsystem is operating in. In this section we will discuss various general aspects of system design that can influence this performance. It should be noted that Engineering/Scientific workloads are usually characterised by long sequential reads and writes, and commercial workloads by short transfers, with random seeks, that are predominantly made up of read operations. This means that optimizing for one application will not necessarily be the best for the other and vice versa.

Performance achieved by the disk subsystem depends on the type of RAID or LVM mirroring being used. “Performance Comparison of RAID Types” on page 33 compares the performance of the different types of RAID and LVM mirroring for some selected configurations and test cases.

Data positioning on disk

The sustained data rate from a disk is greatest on the outside diameter of the disk (the lowest logical block addresses on the disk). Thus data that needs to be read or written at a high data rate is best situated at the outer section of the volume group.

The average access time of a disk is best midway between the outer and inner edges of the disk. Data that is accessed frequently should be placed in this region.

Frame multiplexing

The SSA adapter card is limited at any one instant to transferring data over 4 duplex links (2 per loop). There is a physical limit of 8 data transfer operations at any one instant. The adapter data bandwidth capability supports multiple concurrent transfers on the 4 SSA ports.

Transfers from each disk are broken up into 128 byte frames. Each frame is individually addressed. It is possible to interleave frames for different transfers, so that the frames from many different data transfers can be intermixed and sent down the same link one after the other. This is called frame multiplexing. The SSA hardware automatically manages this, deciding when to insert a new frame, or when to allow existing traffic to flow through. Fairness is implemented to ensure that no single node blocks transfers through it from other nodes. Similarly, the hardware automatically interprets the frame address, and de-multiplexes the data stream, separating it out into its individual logical transfers.

The number of data streams that SSA can support on a port in theory is very large. The SSA adapter supports 11 data transfers plus one control message frame transfers per port. Hence 44 data transfers could be streaming between an adapter card and the SSA network at any point in time. This flexibility in the SSA architecture means it can be quite difficult to state categorically the exact number of disks that an adapter can keep busy, since this is often highly dependent on the particular workload being executed at any particular time. “Performance Comparison of RAID Types” on page 33 discusses the maximum disk drives that can be fully supported for the different types of RAID for long sequential data transfers.

Adapter and host bandwidth limitations and adapter processor capability

Several factors can determine the performance limit of a subsystem that include the data transfer bandwidth, the processor capability in the adapter and the response time to each I/O operation. When many small I/O operations are being performed, the time taken by the adapter processor to set up, monitor and complete data transfers dominates the time taken to transfer the actual data. In this case, the clock speed or computational capability of the processor on the adapter card is the limiting factor, and this results in an I/O operations limit of 8,500 non-RAID ops/sec for a 70/30 mix of 4 Kbyte, read/write operations. (The SSA network or the disk capability does not impose a limit here, since if there are enough disks attached, each disk will only be transferring data for a very short period. During the rest of the time it will be performing a seek or waiting for a new command, and will consequently consume few resources associated with data transfers. In particular, it will not consume many of the 44 available 'channels' above).

If the adapter has an I/O workload that matches the random seek pattern with 4 KB transfers above, then we might expect each disk to execute up to 80 ops/sec. This means, at the 8,500 ops/second limit we would need about 106 disks to saturate the adapter capability. The maximum number of disks possible on two SSA loops attaching to an adapter is 96, so the adapter processing limit for non-RAID operations is not the limiting factor. The limit will either be the number of disks or the ability of the host system to issue I/O operations at the high rate. When RAID is being used, the maximum transaction processing performance possible is more likely to be limited by the adapter than for non-RAID or LVM mirroring environments.

At the other end of the scale, when very long data transfers are being performed, the adapter processor is not utilised very highly and the performance limit is then the available bandwidth in the adapter internal buses or the host bus bandwidth. For the Advanced SerialRAID Plus adapter, with 64 KB read or write data transfers, the maximum data transfer rate can be up to 85 MB/sec for non-RAID operations. This maximum depends on the host bus implementation details and can vary from system to system. Disks are currently available that can read and write at over 20 MB/sec, so this data rate can be generated by relatively few disks.

If the adapter bandwidth does limit the number of disks that can be kept busy by a single adapter then further disks may be attached by adding a second adapter. In this case, the bandwidth of the host bus can become a factor. Adding a second adapter does not necessarily double the available bandwidth if the second adapter is on the same PCI bus as the first adapter.

The adapter processing and bandwidth limitations when using RAID arrays are discussed later in this section and in detail in “Performance Comparison of RAID Types” on page 33.

Number of disks in the subsystem

In any disk subsystem, the larger the number of independent actuators per Mbyte, the smaller is the impact of queuing operations to disks. This effect is more pronounced as the number of I/O operations per second increases and the transfer length of each operation decreases. When approaching limiting cases performance is better when a larger number of small capacity drives (e.g. 4N x 9.1 Gbyte disk drives) is used rather than a few high capacity disk drives (N x 36.4 Gbyte drives). For lower cost it might be preferable to use the higher capacity disks where these deliver satisfactory performance.

Different capacity disk drives are available for the same disk drive technology. These different capacities are achieved by using more disk platters and heads on the higher capacity disk drives. The data rate and latency of each of these different capacity drives are the same and the access times vary only slightly. Whilst the individual disk drive performance for any individual I/O operation is approximately the same for all the different capacity disk drives, the system performance will be much better for configurations that have more disks as there are then more independent operations that can take place concurrently.

Number of disks per SSA loop or per SSA adapter

For maximum data transfer capability, there should be an equal number of disks on each of the two SSA loops of each adapter.

If there are only a small number of disks, the subsystem performance is limited by the disks rather than the adapter. As the number of disks per adapter increases the adapter becomes the limiting factor for performance. Table 3 shows the maximum number of highly active disks that should be attached to a single adapter for random 4 KB operations for each type of RAID. More disks than this will not result in increased performance.

Type of RAID	Maximum Number of Highly Active Disks for Performance (one adapter)
Non-RAID	96
RAID 5	48
RAID 0+1	88
RAID 1	72
LVM Mirroring	88

Table 3: Maximum number of highly active disks per adapter for transaction processing

The performance measured in operations per second achievable increases linearly as the number of disks increases to about 2/3 of the number of disks in Table 3. There is some increase in performance as the number of disks are increased above this up to the maximum number of disks in Table 3, but the benefit is not as significant as when there are less than 2/3 of the maximum number of disks. If you require more than 2/3 of the disks in Table 3and they are all highly active, you could get better performance by using two adapters and using more SSA loops.

More disks can be supported than is shown in Table 3 before the performance is limited by the number of disks if two adapters are used and both adapters are operating to disks on the loops. In this case, the number of disks in Table 3 is increased by 50% for RAID 1, RAID 5 and RAID 0+1 (up to the 96 disk maximum). This assumes that the I/O operations from each adapter on separate systems are split across arrays or disks, or both adapters are on the same system (in which case the device driver ensures access to arrays or disks is evenly distributed between the adapters). If there are two adapters in the loops, the adapter executing the I/O operations has to perform extra processing to keep the partner adapter in step and this results in the maximum number of disks for performance being reduced by 25% for each adapter from the numbers in Table 3. For this reason the number of disks for performance when operations are sent through both adapters is 1.5 times and not 2 times that of a single adapter.

If fast write cache is enabled, extra processing is required in the adapter, and this reduces the maximum number of disks by 25% from the numbers in Table 3 assuming 30% of operations are writes.

These limits assume that all the disks are being used to their maximum all the time. Real benchmarks will not be totally random and not all the disks will be used to their maximum all the time. You can probably safely attach more disks to the adapter than have been indicated in Table 3 without losing performance, but this table is provided to suggest limits for peak workloads involving large numbers of disks.

The maximum number of disks that can be on the SSA loops for long sequential operations depends on how the operations are being issued. If the operations are synchronous and 64 KB sequential operations and to the whole array or LVM mirrored disks, the limit is the number of arrays rather than the number of disks. For this type of operation, the adapter and PCI bus bandwidth is reached with approximately the same number of disks as in Table 3 for RAID 5 arrays with 7 member disks and RAID 0+1 arrays with 8 member disks. For RAID 1 arrays and mirrored logical volumes, the maximum number of disks should be 25-30. If the operations were not synchronous or they were being sent to multiple logical volumes for each array, the bandwidth limit of the adapter and host attachment would be met with fewer disks, particularly for RAID 5 and RAID 0+1.

Again, if two adapters are in the loops and operations are being issued concurrently through each adapter, the number of disks that can be supported for long operations is increased because of the characteristic of SSA that data can be transferred simultaneously to multiple adapters. It is unlikely that the bandwidth of the SSA loop will be the limit to performance.

Effect of distribution of data over disks

The way in which data is distributed between the SSA disks on a loop can have an effect on subsystem performance. SSA can transfer data in several different portions of a SSA loop concurrently, provided these portions do not overlap. This is known as spatial reuse. The adapter port that is used to access a particular disk is the one that is closest to it around the loop. Thus in the one adapter, single loop, 6 drive configuration shown in Figure 2, the adapter accesses drives A,B,C along portion X of the loop, and accesses drives D,E,F along portion Y of the loop.

Since the SSA link is full duplex, reads and writes in each portion of the loop can also be processed concurrently. This means that if conditions are such in portion X that approximately half the data transfers correspond to reads and half to writes at any one time, then the best use is being made of the available bandwidth. If however all reads were performed in portion X and all writes in portion Y, then only half the available bandwidth would be being utilised, and the maximum number of drives that could be supported without saturation would be correspondingly halved.

If the application is arranged such that its I/O is distributed equally between the X portion of the loop and the Y portion, then this will balance the load within the loop and thereby ensure that the maximum number of disk drives can be included in the loop before the available bandwidth is saturated. When arrays are built, the member disks should ideally be located so that half are accessed from one port and the other half from the other port of the adapter. This applies for all the arrays located on the SSA loop.

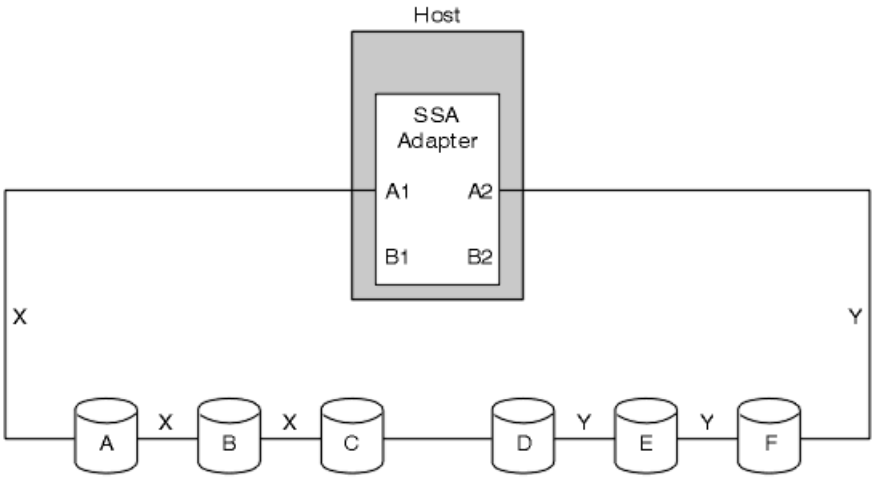

Figure 2: Spatial Reuse in a SSA Disk Subsystem

Distance of disk from adapter

In multiple adapter configurations, data can be transferred to several adapters simultaneously due to the spatial reuse characteristic of SSA. When data is being transferred from several disks to the same adapter, SSA disks that are positioned further away from the adapter can occupy more of the available SSA bandwidth than disks that are close to the adapter. This only occurs in heavily loaded systems. This is because the frames from distant disks have to pass through the frame buffers in nearby disks, thus reducing the opportunity of these close by disks to send or receive frames of their own. Thus in Figure 2, disk C and disk D can operate at somewhat higher data rates than disks A and F. This effect only becomes significant when the loop is heavily loaded or under data intensive workloads involving sequential read operations. A fairness algorithm is implemented on SSA loops to ensure that no disk can use all the available bandwidth, and that the available bandwidth is shared amongst all the disks that need to transfer data, but the general rule is that the disks midway around the loop from the adapter can use slightly more of the bandwidth than those close to the adapter.

Mixing different types of disk on the same loop

SSA80 and SSA160 disks can be mixed on the same SSA loop. Each node arbitrates with the adjacent node to determine the speed to be used between these nodes. However, if there is a mixture of SSA80 and SSA160 disks on the loop, care should be taken in the placement of these disks to ensure that the performance benefit of the SSA160 disks is not inhibited by the data for these disks having to be passed through SSA80 disks that will reduce the data rate. If there is a mixture of SSA160 and SSA80 disks, the SSA160 disks should be placed closer to the ports on the adapters using those disks than the SSA80 disks. Each adapter will access disks to halfway round the loop from each of its ports. If the SSA80 disks are located nearest this halfway point around the loop, data to or from the SSA160 drives will not have to pass through the SSA80 disks and will operate at the maximum data rate.

When using LVM mirroring, RAID 1 or RAID 0+1, mirrored pairs of disks should operate at the same motor speed. If they are mixed motor speeds, the write performance is limited by the speed of the slowest disk as both disks have to be written before an operation is completed.

Number of member disks in an array

In a RAID 5 array short read operations may be processed in parallel as the operations may involve different disks. Long read operations map into several disk operations to separate member disks and these are processed in parallel. This striping effect means that arrays with a large number of members can offer better read performance, than those with small numbers of members for long operations. This only applies, of course, if the read request is long enough to span across all the striped members. It could be argued that if 6 disks worth of data are required, then this could be provided as 2 arrays of 3+P or one array of 6+P. If the read operations are sufficiently long to span all 6 members of the large array, then response time for those operations will be better than that in the 3+P configuration. However, the 3+P configuration will be able to provide comparable throughput since it can still involve all 6 disks, provided that

two commands are in process at any one time. The 3+P configuration requires an extra parity disk. It also may require some manual tuning or use of the logical volume manager to ensure that the load is distributed evenly across the two arrays.

Write operations to a RAID 5 array that involve writing the entire stripe, (i.e. 256 KB for a 4+P array), are much faster than writes of shorter length because disk operations to read the old data are not required. The larger the number of member disks in an array, the longer has to be the length of the write operation to obtain full stripe writes and this may not be achieved with large numbers of member disks. Use of fast write cache results in full stripe writes for sequential operations from the host, each of which is not for a full stripe, because fast write coalesces several write operations before destaging data to the disks.

The larger the number of members in a given array, the longer the rebuild time required when a member disks fails, and the poorer is the performance of the array while the disk is missing.

The larger the number of disks in an array, fewer disks are required to provide availability for a given capacity (1 per array).

In RAID 0+1 arrays, twice the number of disks for the required available capacity are required whatever number of member disks per array are used. There is no rebuild time disadvantage involved in using more member disks. There is a performance benefit obtained by using a larger rather than a smaller number of disks as for long operations data is transferred to several members rather than to a single disk. Striping data across several disks also reduces the limitations that would be caused by operations that were skewed to a narrow range of block addresses and would, if striping was not in use, result in heavy activity to a small number of disks.

The use of striping to improve performance is not restricted to RAID systems. The AIX LVM offers striping across hdisks for instance, and in general, any system that allows parallel access to the data will outperform systems that impose a sequential method.

Performance Implications of Disk Mirroring

If LVM mirroring is being used with non-raid disks and Mirror Write Consistency is on (the default), you may want to locate the copies in the outer region of the disk, since the Mirror Write Consistency information is always written in Cylinder 0. This region corresponds to the lower logical block addresses near 0. From a performance standpoint, mirroring write operations is costly, mirroring with Write-Verify is costlier still (extra disk rotation per write), and mirroring with both Write Verify and Mirror Write Consistency is costliest of all (disk rotation plus a seek to Cylinder 0). To avoid confusion, it is important to remember that although the *lslv* command will usually show Mirror Write Consistency to be on for non-mirrored logical volumes, no actual processing is incurred unless the COPIES value is greater than one. Write Verify, on the other hand, defaults to off, since it does have meaning for non-mirrored logical volumes.

The performance degradation due to Mirror Write Consistency is greatest for random writes and least for short block sequential writes. Enabling for fast write cache can mask the performance reduction of Mirror Write Consistency.

It should be remembered that for read operations, mirroring can have significant performance advantages since the system has a choice of two disk from which it can read the desired data and it can therefore choose whichever disk has the smaller queue or seek.

Note that these comments are not specific to SSA. SCSI disks have the same effects.

Response Time Considerations

A fast write cache can significantly improve the response time for write operations, however care must be taken not to flood the cache with write requests faster than the rate at which the cache can destage its data.

A fast write cache typically provides significant advantages in specialised workloads, for example copying a database onto a new set of disks or a database log file written with short sequential writes. Applications where the next I/O operation is not issued until the previous one has completed and that have a high percentage of write operations will benefit by using fast write cache. If multiple adapters, each with a fast write cache, are used, then this will multiply the benefit.

A fast write cache may also coalesce adjacent write operations prior to actually moving the data to disk. This can be a significant performance benefit particularly for RAID 5 arrays.

A fast write cache can adversely affect the maximum throughput, since additional processing is required in the adapter card to determine if the data that is being transferred is in the cache or not. This only has an effect if there are enough disks attached that the throughput is limited by the adapter. This increases if there are two adapters in the loop. Fast write should only be used in a two adapter configuration when operations are sent to an array enabled for fast write cache from either but not both of the adapters. It can be used, therefore, for failover protection.

Performance Recommendations for SSA Raid Adapters

Non-RAID operation

- Distribute the disks evenly over the available SSA loops.
- If possible, distribute read and write data evenly throughout the SSA loops.
- For high throughput applications, use logical volumes made up of small disks. (The adapter can sustain a 70/30 mix of read/write operations, with 4Kbyte transfer lengths, at up to 8,500 operations/sec. This I/O operation rate can be produced by two fully populated SSA loops of disks.
- When using LVM mirroring arrange to have the mirror copies on different disks on different loops.
- When using LVM mirroring with Mirror Write Consistency on, throughput is increased for write operations if fast write cache is enabled.

RAID 5 operation

- For transaction processing applications, where a large number of small, unrelated I/O requests are made:
  1. Use smaller physical disks rather than larger disks.
  2. If a database log file is being used enable fast write cache for the array.
- For data intensive applications, where a small number of large, possibly related I/O requests are made:
  1. It is tempting to use a large number of members in each array (to maximise the effect of striping) However, when the array is operating with one member missing, it must read every other member to reconstruct the missing data. This means that when reconstructing data, performance reduces as the number of members in the array increases. It is better to limit the number of members in the array to a relatively small number, to make, say, a (4+P). In fact, (6+P) is a popular size, since together with a spare it occupies exactly half the 7133 enclosure capacity.
  2. Use equal numbers of arrays on each loop.
  3. Locate member disks on the loop so that approximately half the disks are accessed by one port of the adapter and the other half by the other port.
- Enabling fast write cache for a RAID 5 array is beneficial as it does not just improve the response time for an I/O operation but also results in a higher data rate for sequential write operations. Writes are coalesced in the fast write cache and the destage to the disks is for an entire stripe length. No read operations are then required for RAID 5 to read the old data as is required if only a partial stripe is written.

RAID 1 operation

- For transaction processing applications, where a large number of small, unrelated I/O requests are made, use smaller rather than larger disks
- Locate each member disk on the loop so that they are normally accessed by different ports on the adapter

RAID 0+1 operation

- For transaction processing applications, where a large number of small, unrelated I/O requests are made use smaller rather than larger disks.
- For data intensive applications, where a small number of large, possibly related I/O requests are made:
  1. Use equal numbers of arrays on each loop.
  2. Locate member disks on the loop so that half the disks are accessed by one port of the adapter and the other half by the other port

Having selected a particular SSA Raid configuration it is important not to neglect the conventional LVM tuning techniques (e.g. spreading a Logical Volume across multiple disks, using striped Logical Volumes, multiple JFS logs). As ever, it is also important to avoid file system fragmentation.

4 Performance Comparison of RAID Types

The performance numbers quoted are correct and should only be used for comparison between the different types of RAID and LVM mirroring for the environment that was tested. You should not infer that this performance is what would be achieved in your system with your particular applications. The performance actually achieved depends on:

Function	Assumptions use in this section of the document
System	RS/6000 7015 model S70 was used. This had 8 GB memory and 8 processors
Disks	A mixture of 10,000 rpm and 7,200 rpm disks were used. All disks were capable of operating at 40 MB/sec SSA speed.
Read/Write ratio of I/O operations	A spread of performance with different read/write ratios are included. Read/write ratio significantly affects performance.
Transfer size	4K transfers were assumed for transaction processing and 64 KB transfers were assumed for sequential operations.
Randomness of operations	All transaction comparisons assumed totally random I/O operations. In practice operations will skew towards subsequent accesses to block addresses that are close to the previous address. This skewing of accesses will reduce the transaction performance of RAID 1 and mirrored logical volumes that are not striped where the performance may be limited by the excessive use of a few disks. RAID 5 and RAID 0+1 will not be affected by skewing of operations to localised block addresses because consecutive strips of data are located on different disks.
Number of member disks of arrays	RAID 5 assumed 6+P (7 member disks); RAID 0+1 assumed 4+4 (8 member disks). More member disks per array results in more I/O operations per second possible to that array but the initial build and rebuild times are higher.
Location of array members on the SSA loop	Array members were positioned on the loop as recommended for optimum performance, that is half the member disk were closest to one adapter port and the other half were closest to the other adapter port.
Synchronous operations for sequential workloads	Synchronous operations were assumed with the next operation not being issued until the previous one had completed. If a queue of sequential operations had been issued, the data rate possible would increase because revolutions would not be missed on the disks.
Command Queue depth	For the transaction processing test cases, a queue of 2 times the number of array member disks or a queue of 2 for LVM mirrored disks was maintained.
Number of adapters on the loop	A single adapter was assumed for most comparisons. A second adapter increases the availability of the system by continued operation when one adapter fails, but there is a decrease in the performance of all types of RAID.
Mix of array types	For the performance comparisons, all the disks were assumed to be of the same type

With your particular workload and configuration, you may see better or worse performance than is referred to in this section.

The performance available from a disk subsystem can be limited by any or all of the following:

- Maximum adapter throughput capability
- Maximum throughput from a fixed number of disks that is not dependent on the response time for each I/O operation
- Response time to each I/O operation

One of these may be much more significant than others for your application. This section compares the effect of each of these for each type of RAID array and LVM mirroring. Unless there are a large number of disks attached, (at least 48), the adapter throughput limitation is not relevant as the performance is limited by the number of disks. If the response time for each I/O operation is critical to the overall performance, then the response time comparison for different RAID types with and without fast write cache is more relevant than the maximum throughput comparison.

The arrays and LVM mirrored logical disks used for the performance results shown in this section consisted of the following member disks (unless otherwise stated):

RAID 5	6 + P
RAID 0+1	4+4
RAID 1	1+1
Mirrored logical volume	1 + 1

All the SSA disks were capable of 40 MB/sec operation and were a mixture of those operating at 7200 rpm and 10,000 rpm.

Transaction Throughput Performance

Transaction processing is characterised by I/O operations of short length to random locations on the disks. For the throughput performance comparisons for transaction processing in this section, all I/O operations are assumed to be 4 Kbytes long and to a random logical block address. It is also assumed that copper rather than optical cables are used for the SSA links. If long optical cables are used, the SSA data rate possible is reduced (see page 11). This may have an effect on the operations per second possible for transaction processing. The throughput varies according to the ratio of Reads and Writes, so the spread from 100% Writes to 100% Reads is shown. The following abbreviations have been used in the figures:

MLV	Mirrored logical volume
MWC	Mirrored write consistency

Transaction adapter throughput comparison - maximum number of disks

Figure 3 shows the throughput comparison for random 4KB I/O operations for a single adapter with the maximum number of disks attached. A raw logical volume that spanned each array was used. A queue depth of twice the number of member disks in the array was maintained by the test cases for all transaction processing comparisons.

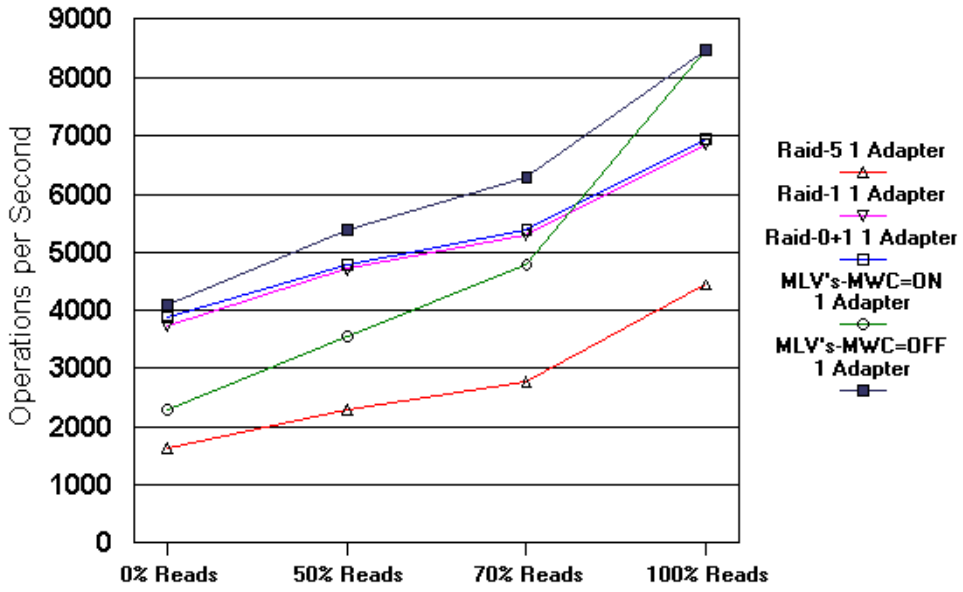

Figure 3: Transaction throughput comparison of RAID types (96 disks, single adapter)

The throughput in terms of operations per second possible is different for the different array types and depends on several factors. One key consideration is the predominance of write activity. For RAID 5 arrays, each 4KB write involves 2 read operations and 2 write operations to read and write the data and parity disks. If the application writing data has previously already read the data and it is still held in the cache on the adapter, then this will be reduced to 1 read and 2 write operations. The benchmark shown in Figure 3 assumed all operations were to random logical blocks and so 4 disk operations were required for each write I/O operation. The throughput for write operations for RAID 5 is the lowest of all RAID type because of the disk activity required.

When using mirrored logical volumes, RAID 1 or RAID 0+1 arrays, each write operation requires 2 disk accesses to write the data on each mirrored pair of disks. When using mirrored logical volumes, the set up can enable or disable mirror write consistency. If enabled, it is guaranteed that after a write is not completed successfully all further accesses to the data that had been attempted to be written the same data is returned whichever disk of the mirrored pair is read. If it is not enabled, subsequent reads to data that was not successfully written may return different data depending on which disk has been read. This is not allowed in some environments. The default setting is that mirror write consistency is enabled. RAID 1 and RAID 0+1 always return the same data read after a write operations was not completed successfully. Even though writes for RAID 1, RAID 0+1 and mirrored logical volumes (with mirror write consistency disabled) require write operations to two disks it can be seen that the throughput for writes is much better than for writes to RAID 5 arrays. Mirror write consistency requires extra disk accesses for writes to ensure synchronism of the data and it can be seen that enabling this introduces a significant reduction in throughput for writes and is only slightly better than RAID 5 write throughput. This would have improved if fast write cache had been enabled.

Read throughput is better for RAID 1 and RAID 0+1 arrays than for RAID 5 arrays. The best throughput performance limited by the adapter is for mirrored logical volumes as for this environment the adapter processor is doing less work than for RAID 1, RAID 0+1 or RAID 5.

Write operations to LVM mirrored logical volumes require up to 25% more host processor utilization than that required for all types of RAID that are implemented in the adapter. This may be a consideration if the applications have a high percentage of writes and the host processor limits the throughput available.

This throughput performance comparison shows that RAID 5 arrays provide the least throughput of all array types when the adapter is limiting the throughput. This is more noticeable for workloads that involve heavy write activity. The throughput performance of RAID 0+1 is 2.5 times better than RAID 5 if all the operations are writes, but is only 1.5 times better if all the operations are reads. However, RAID 5 requires the least number of disks to provide availability when a disk fails (only 1 disk for the width of the array) compared to 100% for RAID 1, RAID 0+1 or mirrored logical volumes.

Another way of looking at this is that for the same number of disks as RAID 1, RAID 0+1 or mirrored logical volumes, RAID 5 provides almost twice the usable capacity and better throughput performance than any alternative that uses mirroring for read intensive workloads and only slightly lower throughput performance for write intensive workloads.

Transaction throughput performance comparison - 6 disks capacity

Figure 3 shows the transaction throughput performance comparison for different array types when the adapter is fully configured with 96 disks and throughput is limited by the adapter. These 96 disks, however, provide different usable capacity for RAID 5 compared to using mirroring so it may not be a meaningful comparison for throughput for a given usable capacity. Also if the configuration does not have such a high number of disks, the throughput may be limited by the disks rather than be the adapter. Figure 4 shows the transaction throughput performance comparison for the different array types for a fixed user capacity (6 disks of data). For this comparison the RAID 5 array is 6+P, RAID 0+1 array is 6+6, there are 6 RAID 1 arrays and 6 LVM mirrored logical volumes.

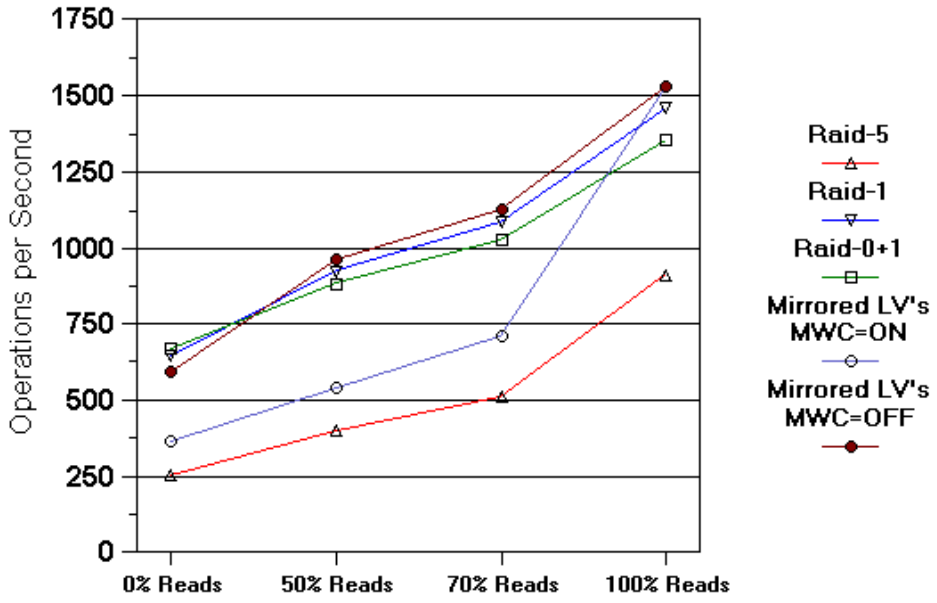

Figure 4: Transaction throughput comparison of RAID types (6 disks capacity, single adapter)

Figure 4 shows similar throughput performance comparisons between the arrays when the comparison is made with the same usable data capacity as was shown in Figure 3.

When a disk fails in a RAID 1, RAID 0+1 or RAID 5 array, a hot spare is automatically introduced and the data on this new member is rebuilt from the other member disks of the array. Whilst rebuilding, transaction throughput performance is reduced. Figure 5 shows the performance comparison when different RAID types are rebuilding.

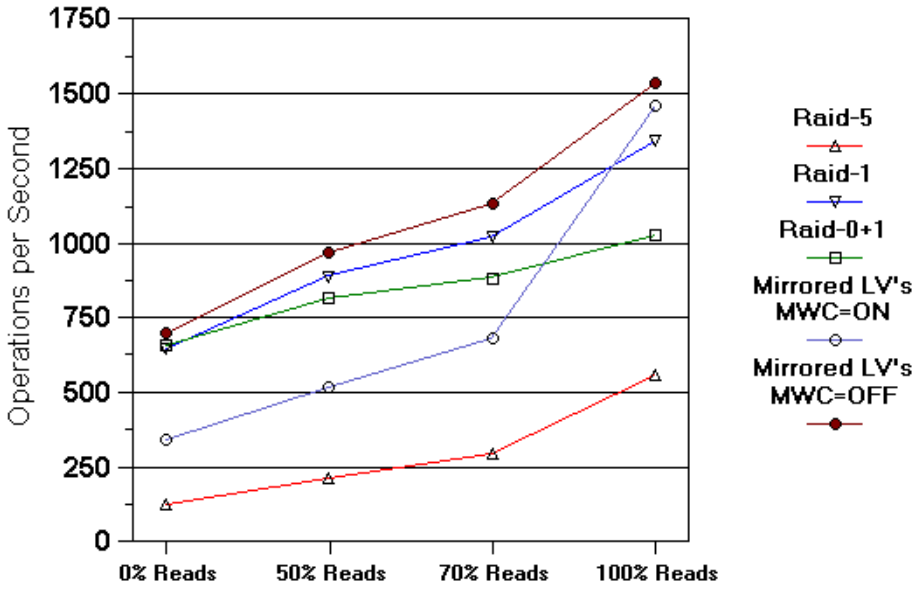

Figure 5: Transaction performance comparison of rebuilding arrays (6 disks capacity, single adapter)

Comparing Figure 5 with Figure 4 shows that there is very little performance reduction when rebuilding one disk for RAID 1 or mirrored logical volumes of 6 disk of usable data. That is because only one disk out of the 12 disks is being used for the background rebuild and only one of the 6 arrays is limited to reading from a single disk during the rebuild. In RAID 5 arrays, all the disks of the array are used to rebuild data and the performance during the rebuild is consequently reduced.

Transaction throughput comparison - increasing number of arrays

The transaction throughput performance available for any given type of array depends on how many disks are used. For RAID 5 you can expect twice the transaction throughput performance for a 12+P array compared to a 6+P array. The following figures (Figure 6 to Figure 10) show how the throughput increases as the number of arrays increases for each array type.

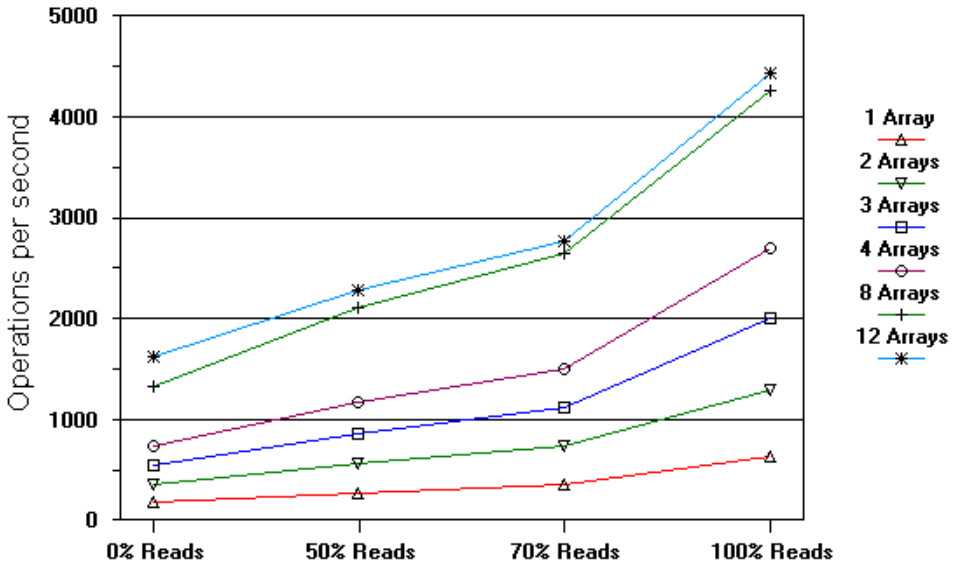

Figure 6: Transaction processing throughput performance, increasing number of RAID 5 arrays (6+P)

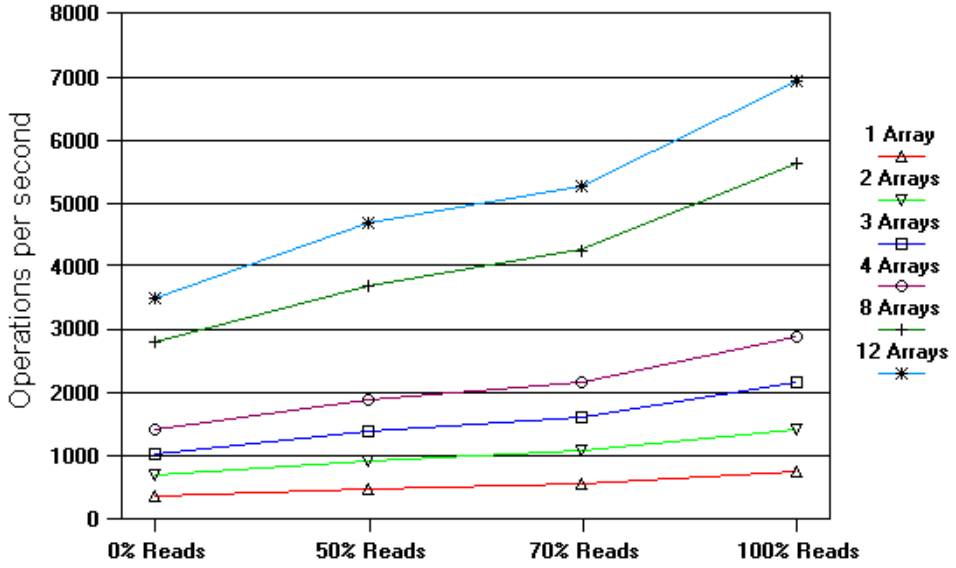

Figure 7: Transaction processing throughput performance, increasing number of RAID 0+1 arrays (4+4)

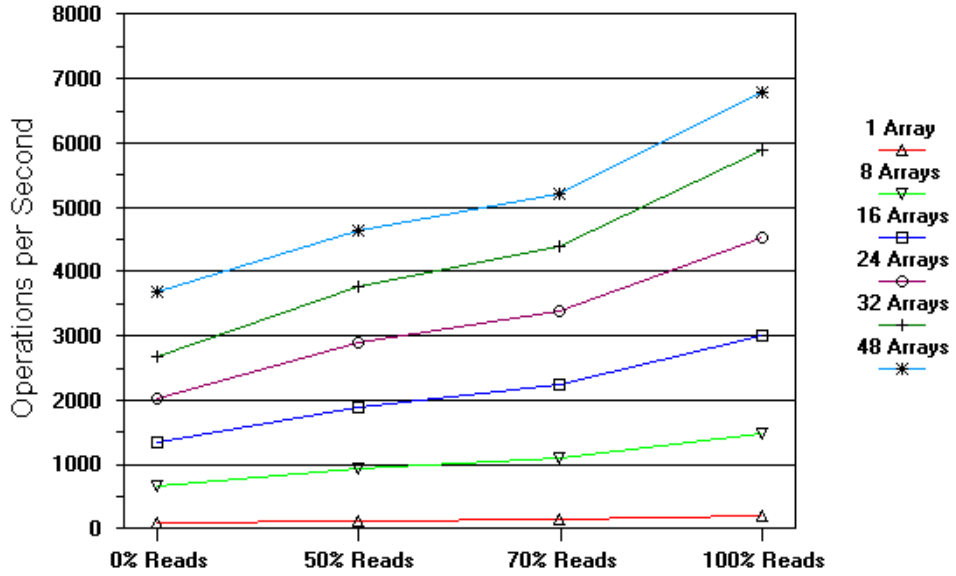

Figure 8: Transaction processing throughput performance, increasing number of RAID 1 arrays

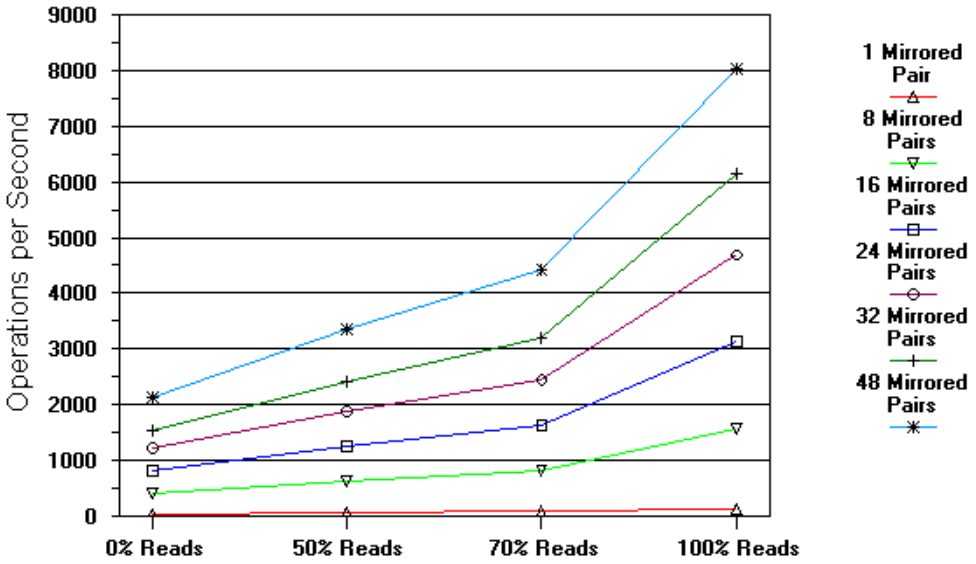

Figure 9: Transaction processing throughput performance, increasing number of mirrored logical volumes (MWC on)

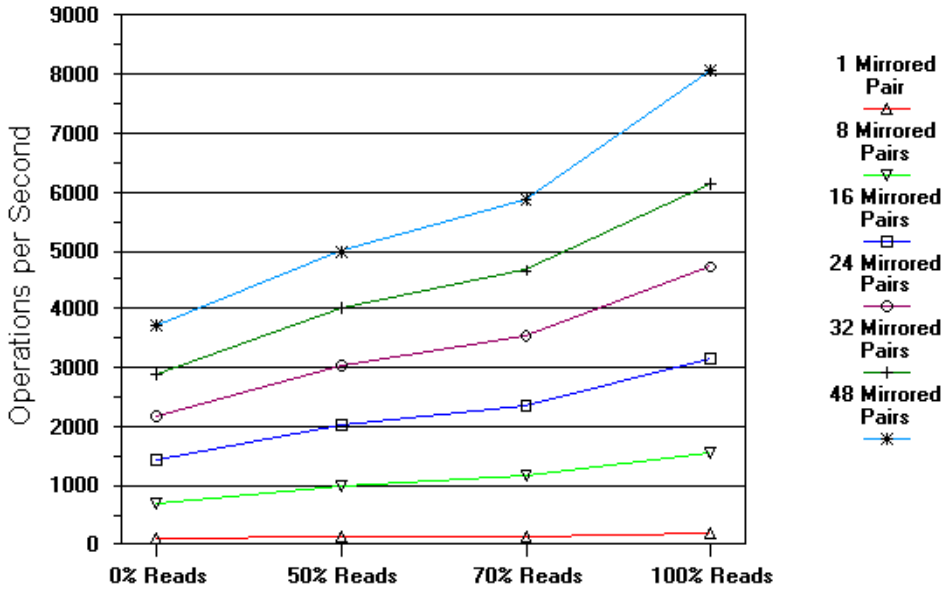

Figure 10: Transaction processing throughput performance, increasing number of mirrored logical volumes (MWC off)

As the number of disks increases for each array type, the transaction throughput performance increases linearly with the number of disks initially. The figures show this effect by increasing the number of arrays that each have the same number of member disks. The same effect could have been shown by increasing the number of member disks of RAID 5 or RAID 0+1 arrays rather than increasing the number of arrays. The transaction throughput of a 15+P RAID 5 array is about the same as the throughput possible of two 7+P RAID 5 arrays. The choice of array size should be determined by other factors, such as sequential performance, build and rebuild times and usable capacity requirements.

If applications are able to submit I/O operations at high rates there comes a point where the performance is limited by the adapter rather than by the number of disks. This limit is different for each RAID type.

Transaction adapter throughput comparison - effect of a second adapter

If there are sufficient disks and applications are able to submit I/O requests at a level where the adapter becomes the limiting factor, throughput capability can be increased by adding another adapter and spreading the disks over both adapters. Figure 11 shows the throughput achievable by using a second adapter for a maximum configuration of 96 disks.

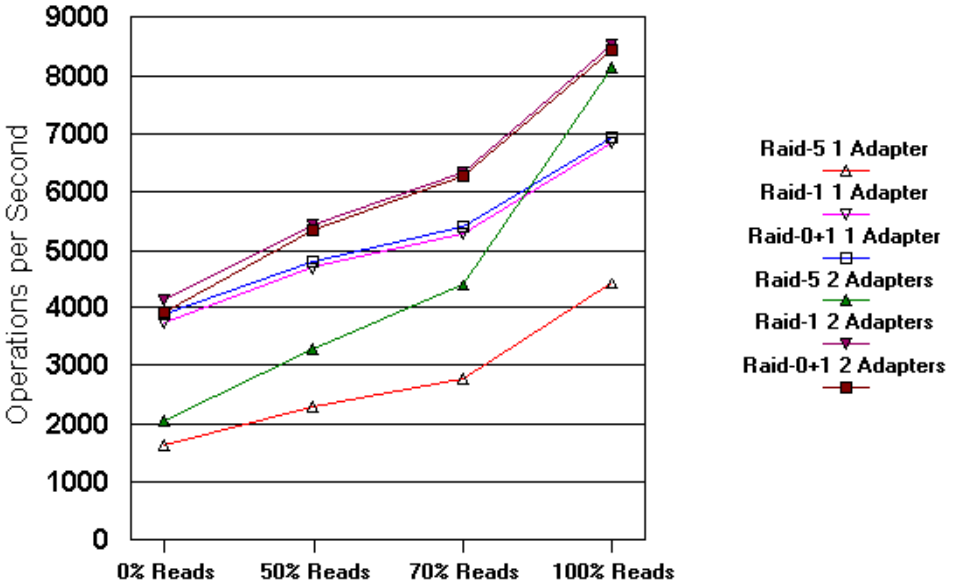

Figure 11: Transaction throughput performance for array types with 1 or 2 adapters

The effect of adding a second adapter is most noticeable if RAID 5 arrays are used. This is because for RAID 5 the transaction throughput performance becomes adapter limited and adding a second adapter almost doubles the throughput possible for a maximum configuration of disks for read intensive workloads. For the RAID types that use mirroring, the throughput possible is improved by adding a second adapter but the improvement is not as noticeable as for RAID 5.

Enabling fast write cache for any array type significantly reduces the response time for all write operations (see “Response Time Performance” on page 47). It does, however, require extra processing on the adapter card with the result that the maximum operations per second that the adapter can support is reduced as shown in Table 4 when fast write cache is enabled:

Array Type	Single adapter throughput (operations/sec, 30% writes)		2 adapters throughput (operations/sec, 30% writes)	
	No Fast Write	Fast Write Enabled	No Fast Write	Fast Write Enabled
RAID 5	2,700	2,000	4,200	3,000
RAID 1	4,800	3,300	6,100	4,000
RAID 0+1	5,200	3,500	6,200	4,400

Table 4. Effect of 2 adapters / Fast Write Cache on transaction processing adapter capability

This reduction in transaction processing capability will only have an effect if the throughput possible is not limited by the number of disks and the host system is able to generate sufficient I/O operations. If two adapters are in the loops, fast write cache can be enabled if an array is only accessed by one and not both of the adapters. In this case the adapter transaction processing capability is reduced by 25% to keep the other adapter synchronized with operations.

Sequential Performance

Performance of long sequential operations is limited by the bandwidth available rather than operations per second that can be sustained. Figure 12 shows the data rate that can be sustained for a single adapter with 96 disks attached using a workload of synchronous sequential operations with a queue depth of 1 each operation reading or writing 64 Kbytes. It is assumed that copper rather than optical cables are used for the SSA links. If long optical cables are used, the SSA data rate possible is reduced (see page 11).

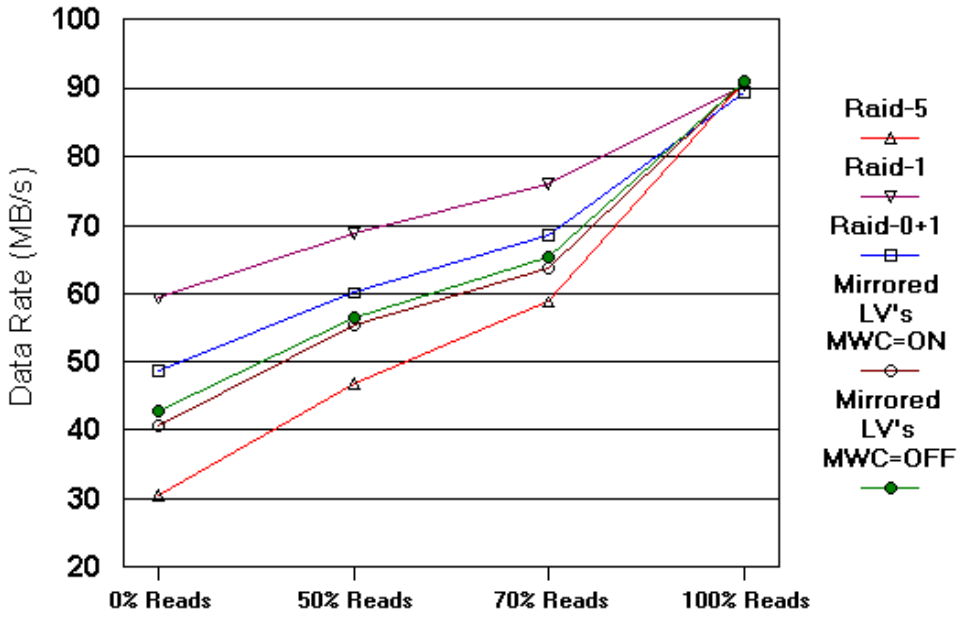

Figure 12: Sequential bandwidth comparisons for different array types (96 disks, single adapter)

Figure 12 shows that the data rate possible when all operations are sequential reads is about 90 MB/sec for all types of RAID. This data rate was measured on a S70 system and it may be slightly different when the adapter is used on other systems as the PCI bandwidth available differs across systems.

The data rate possible when there is a high degree of write operations differs for the different types of RAID. RAID 1 followed by RAID 0+1 is highest because data is only transferred once from the system into the adapter. Mirrored logical volumes require two transfers of data to the adapter and this means that the data rate possible for write operations is less. RAID 5 has the lowest data rate because 4 transfers of data to or from the disks are required. The data rate for writes to RAID 5 would be improved to 70 MB/sec if the length of the writes were longer so that 64 KB had been written to each member disk for a single write operation. In this case, RAID 5 does not have to read the old data or parity and the number of write operations is then less than for any of the mirroring options which always involve writing to both mirrored pairs of disks. When a RAID 5 array is enabled with fast write cache, sequential write operations, each of which may be short, will be coalesced by the fast write cache before destaging to the disks into writes that are the length of the full stripe and therefore do not require the old data to be read before data is written. The bandwidth for sequential write operations to RAID 5 arrays is therefore much higher if the array is enabled for fast write cache.

The results shown were obtained for synchronous write operations with a block size of 64 Kbyte. In a sequence of synchronous writes, the application issues a write command and waits for completion of that operation before issuing another write command. This waiting causes revolutions of the disks to be missed for write operations which reduces the overall data rate. Not all applications require synchronous writes and in that case the data rate possible would be higher than shown for writes. Use of synchronous operations for reads has little effect on the data rate as the disks read ahead 128 Kbyte of data into the buffer on the disk after the read has completed, so even if there is a delay between issuing the next read to the disk, there is no extra revolution of the disk as the required data is already held in the read ahead buffer on the disk.

If all the operations are reads, the same data rate is sustained for all RAID types. This data rate is limited by the adapter to system interface and may vary when the adapter is attached to different systems.

Figure 12 shows the data rate limit of the adapter with the maximum number (96) of disks. Figure 13 shows the data rate comparison for the different array types for a user capacity of 6 disks.

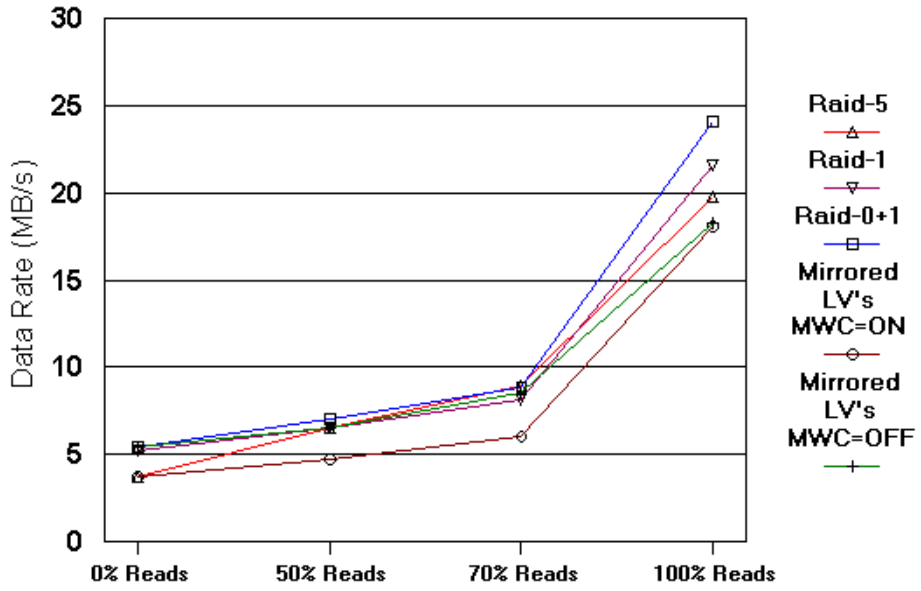

Figure 13: Sequential bandwidth comparisons for different array types (6 disks, single adapter)

In Figure 13, the data rate was measured with synchronous 64 Kbyte sequential operations to the array. In order to sensibly compare between array types with single synchronous operation threads, data was striped across the 6 pairs of RAID 1 and mirrored logical volume disks into a single logical volume with 64 Kbyte strip length.

The bandwidth for sequential operations shown for RAID 5 is for arrays consisting of 7 member disks. The effect on the bandwidth for sequential operations of changing the number of disks in an array is shown in Figure 14.

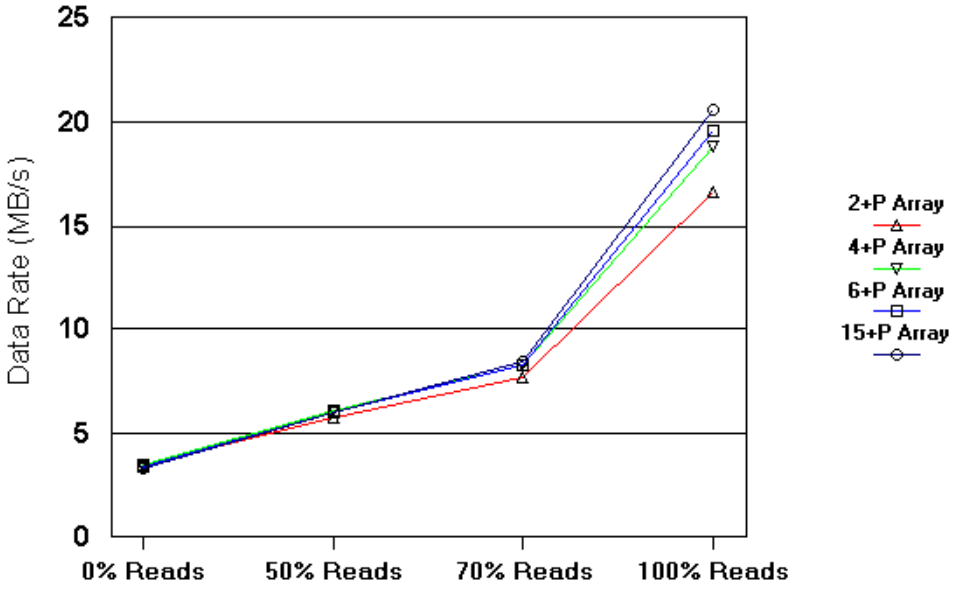

Figure 14: Sequential bandwidth comparison with increasing number of member disks for a single RAID 5 array

Figure 14 shows that for synchronous 64 KB sequential operations there is little difference in the possible bandwidth to a single array. There would be a benefit if operations were issued with a queue greater than 1. The main benefit of increasing the number of disks in an array is to reduce the number of disks required to provide availability for a given

capacity. Transaction throughput possible for an array is increased if there are more member disks in the array, assuming queue depths > 1. The higher the number of disks in the RAID 5 array, the longer is the time to rebuild a replacement disk after a disk failure and the longer the period of exposure to data loss if a second disk fails in this period.

If fast write cache is enabled for a RAID 5 array, the data rate that is possible for synchronous sequential 100% write operations for the 6 disk capacity array is increased from about 4 Mbyte/sec to about 24 Mbyte/sec. This is because fast write coalesces the writes to disks into full stripe writes that write to all member disks and eliminate the need to precede the writes with reads. The increase in data rate for synchronous sequential write operations of different lengths to a single 6+P RAID 5 array is shown in Figure 15.

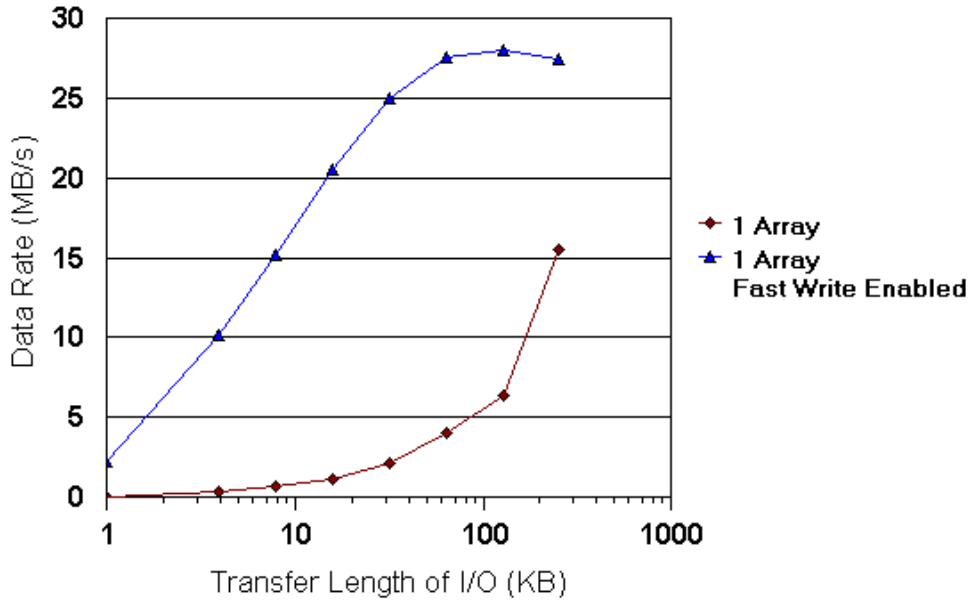

Figure 15: Sequential bandwidth comparison for a single RAID 5 array with fast write cache enabled

The data rate shown in Figure 15 is when the disks are on loop A of the adapter. Fast write cache and SSA transfers on loop A share the bandwidth available internally within the adapter. If the disks are on SSA loop B, the fast write cache operations do not compete for the same bandwidth and the data rate for one fast write RAID 5 array increases to 37 MB/sec. It is therefore better for write operations for arrays and disks that are enabled for fast write to locate these on SSA loop B rather than loop A. However, if all the disks are located on loop B, read performance is reduced as only one loop is used so reducing the available bandwidth. Disks should therefore be approximately evenly distributed across the two loops, but if not all are enabled for fast write cache, it is advisable to locate the fast write enabled disks on SSA loop B.

Figure 16 to Figure 20 show the data rate possible for synchronous 64 KB operations for increasing numbers of arrays for different array types. These show that for small number of arrays the data rate possible increases linearly as the number of arrays increases, but after a certain number of arrays the data rate still increases but at a lower rate.

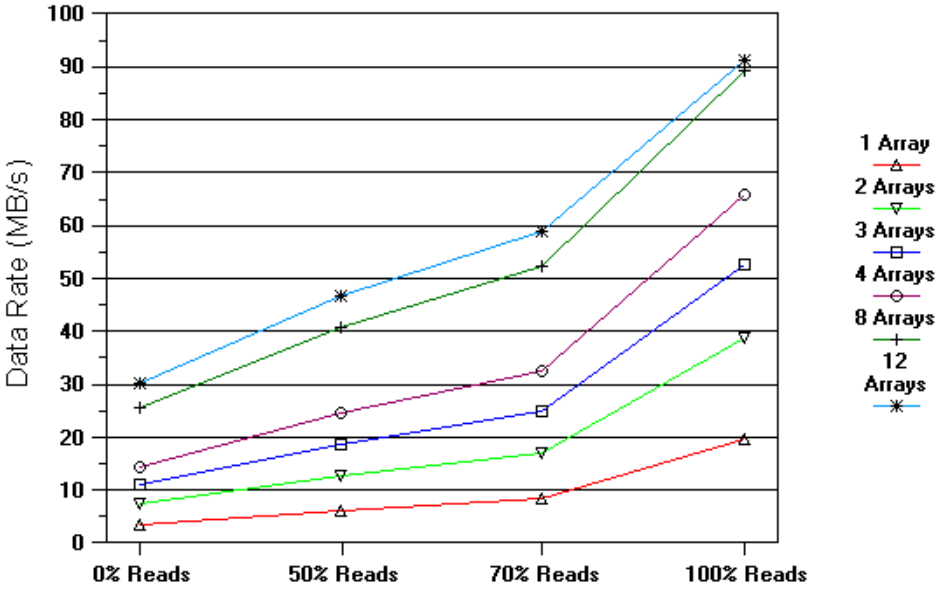

Figure 16: Sequential bandwidth for RAID 5 arrays with increasing number of arrays

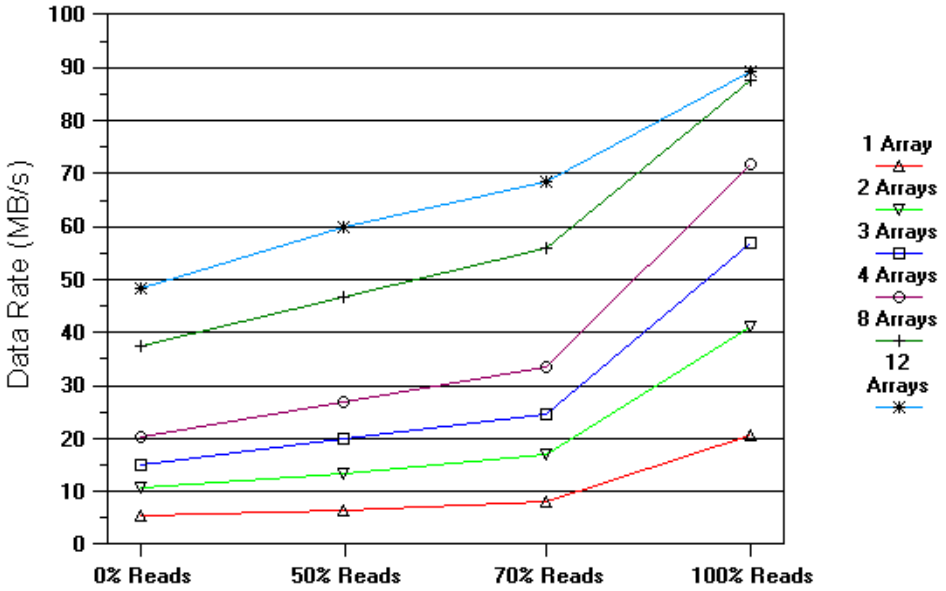

Figure 17: Sequential bandwidth for RAID 0+1 arrays with increasing number of arrays

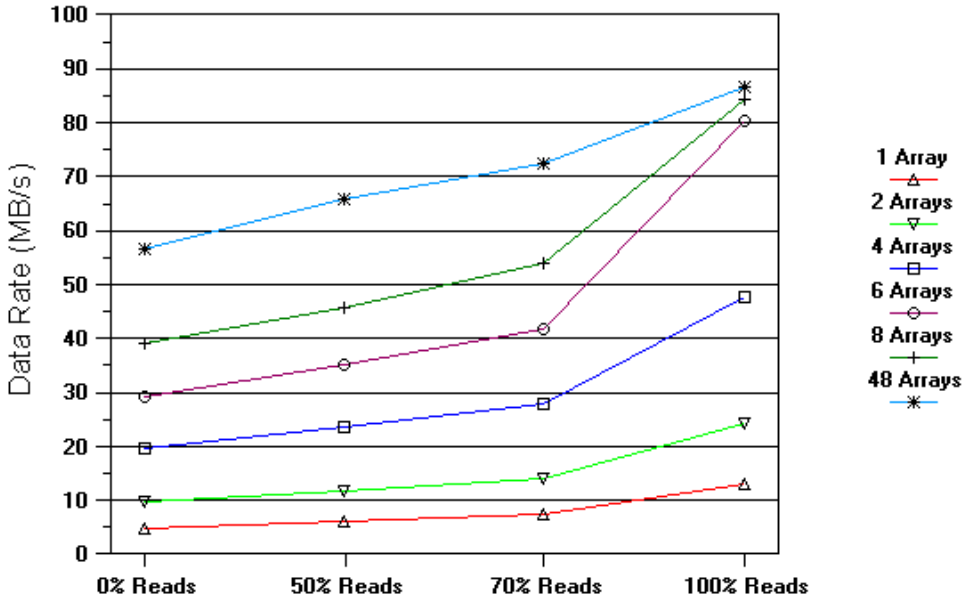

Figure 18: Sequential bandwidth for RAID 1 arrays with increasing number of arrays

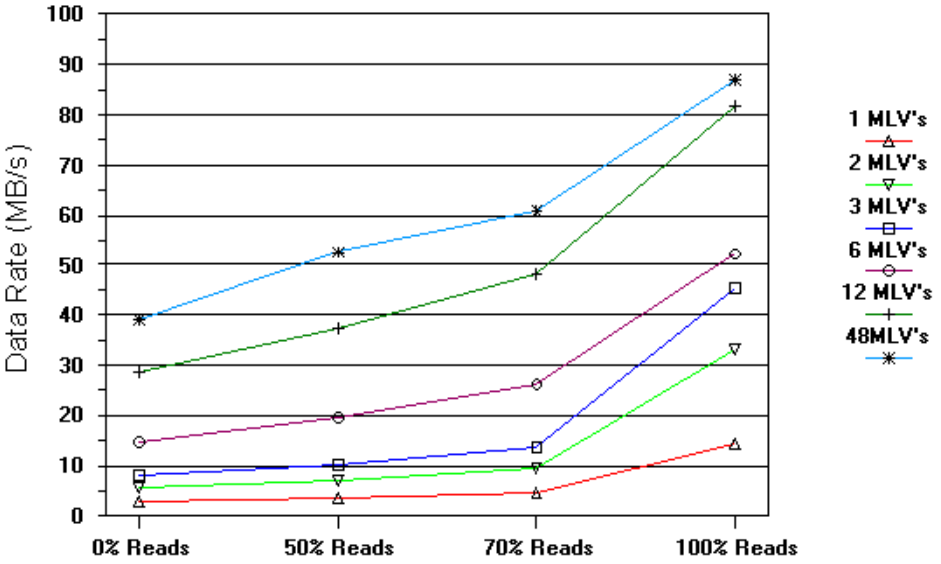

Figure 19: Sequential bandwidth for mirrored logical volumes with increasing number of disks (MWC on)

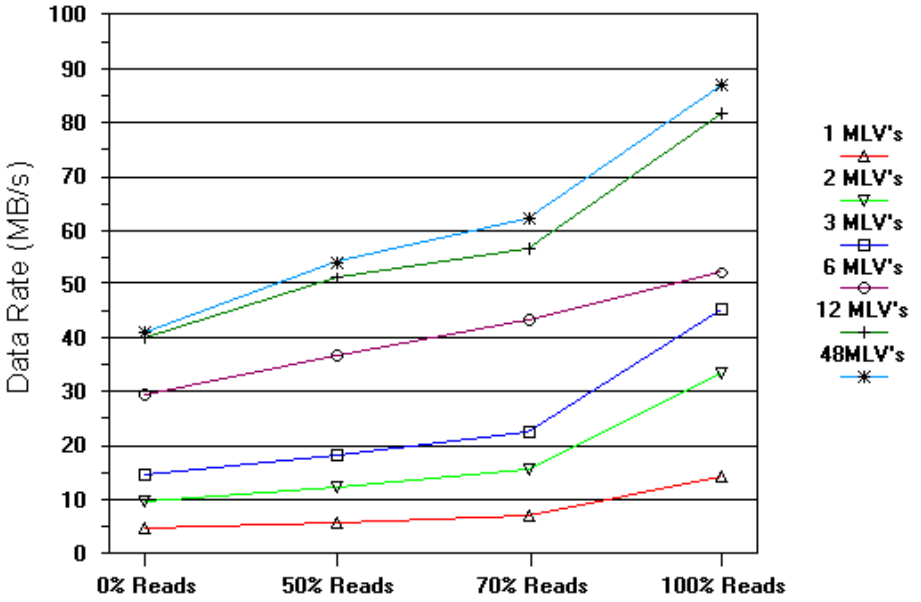

Figure 20: Sequential bandwidth for mirrored logical volumes with increasing number of disks (MWC off)

In summary, for sequential operations, RAID 1 and RAID 0+1 provide the best bandwidth and RAID 5 the least but they but they do require more disks than RAID 5. The bandwidth possible of all RAID types depends on the read/write ratio. When fast write cache is enabled, the sequential write performance of RAID 5 increases dramatically. This is because fast write coalesces the write operations to disks. As a result, writes are issued to all the member disks of the stripe, thus avoiding the read operations to the disks.

If sequential performance is important to applications and you require least number of disks, use RAID 5. It is recommended that you also use fast write cache.

Response Time Performance

Figure 21 to Figure 26 show how the response time varies for the different types of RAID for different read/write ratios. For all these comparisons the operations were 4 KB in length to random block addresses.

RAID Configurations:

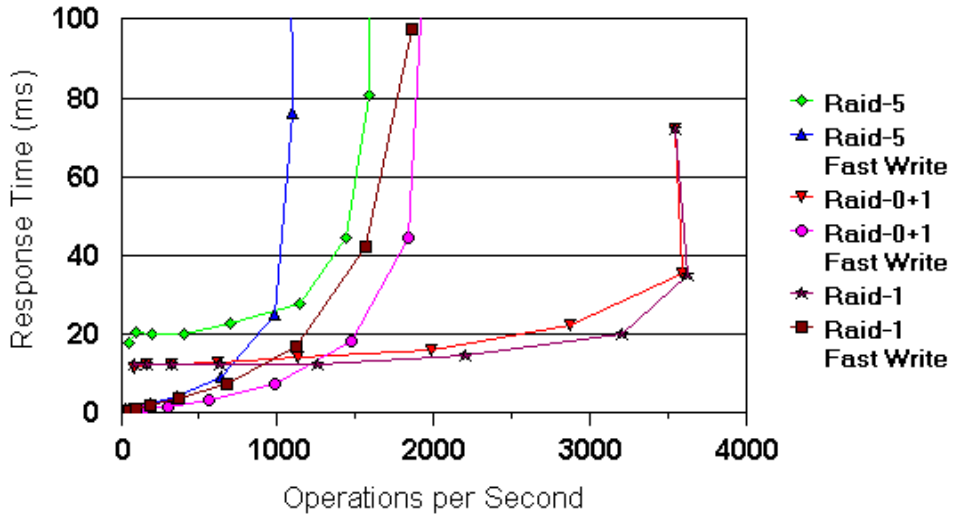

Figure 21: Response time comparison of arrays for 100% write operations (single adapter, 96 disks, random 4KB operations)

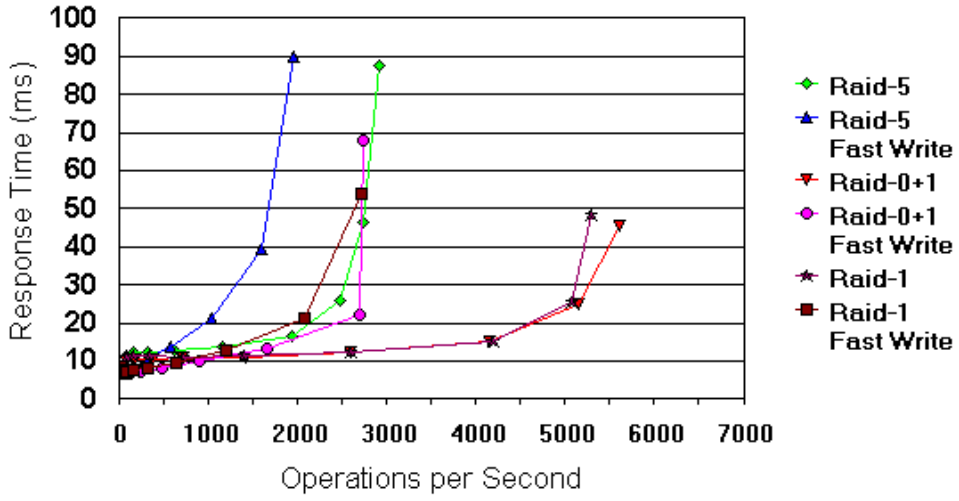

Figure 22: Response time comparison of arrays for 70% read operations (single adapter, 96 disks, random 4KB operations)

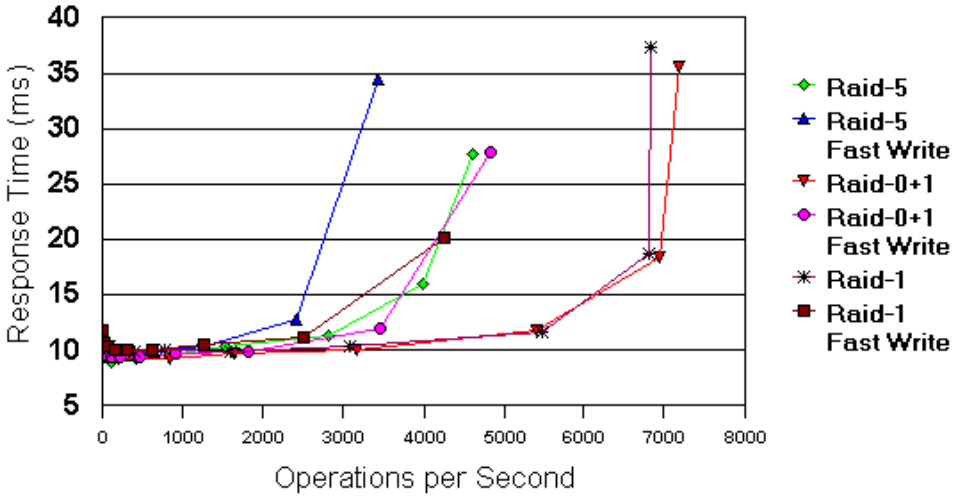

Figure 23: Response time comparison of arrays for 100% read operations (single adapter, 96 disks, random 4KB operations)

Non-RAID Configurations:

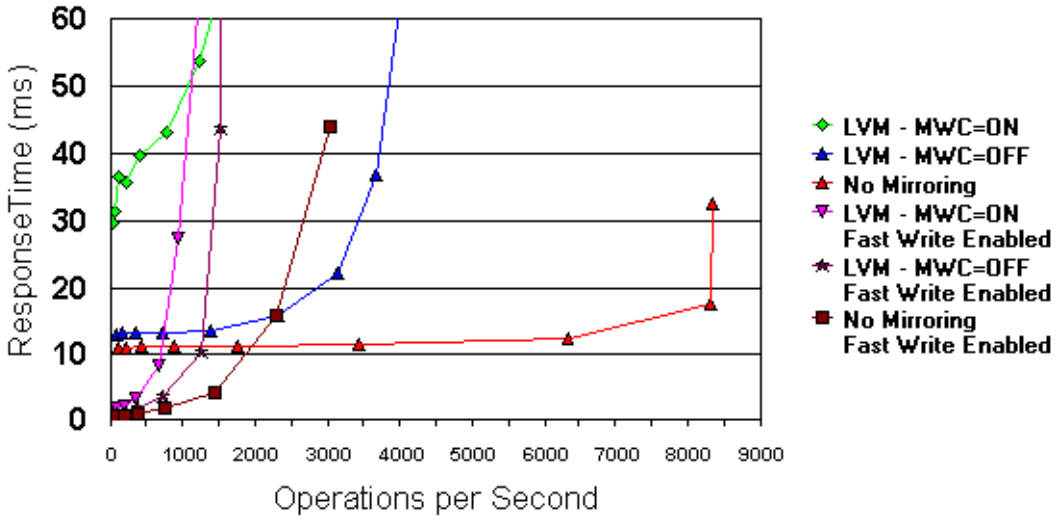

Figure 24: Non-RAID response time comparison of for 100% write operations (one adapter, 96 disks, random 4KB operations)

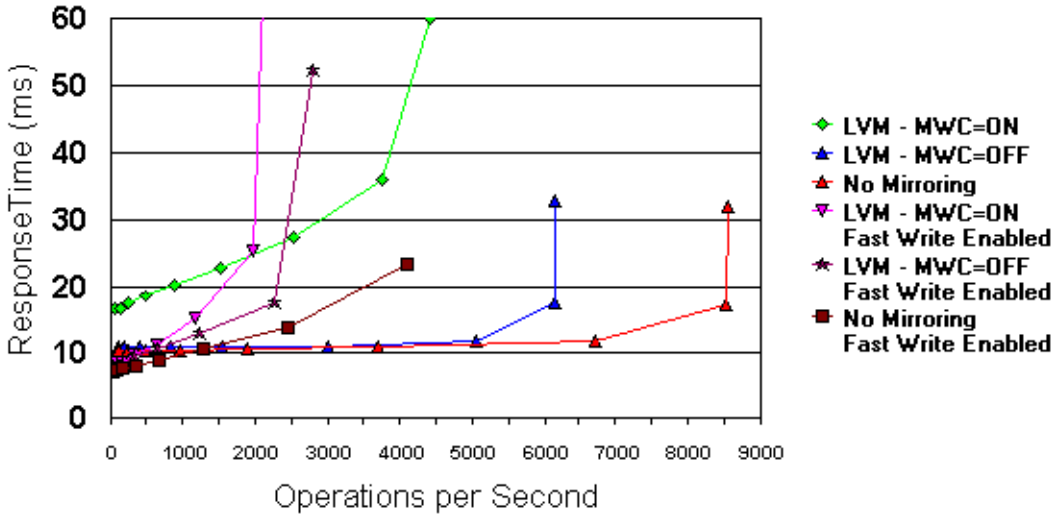

Figure 25: Non-RAID response time comparison of for 70% read operations (one adapter, 96 disks, random 4KB operations)

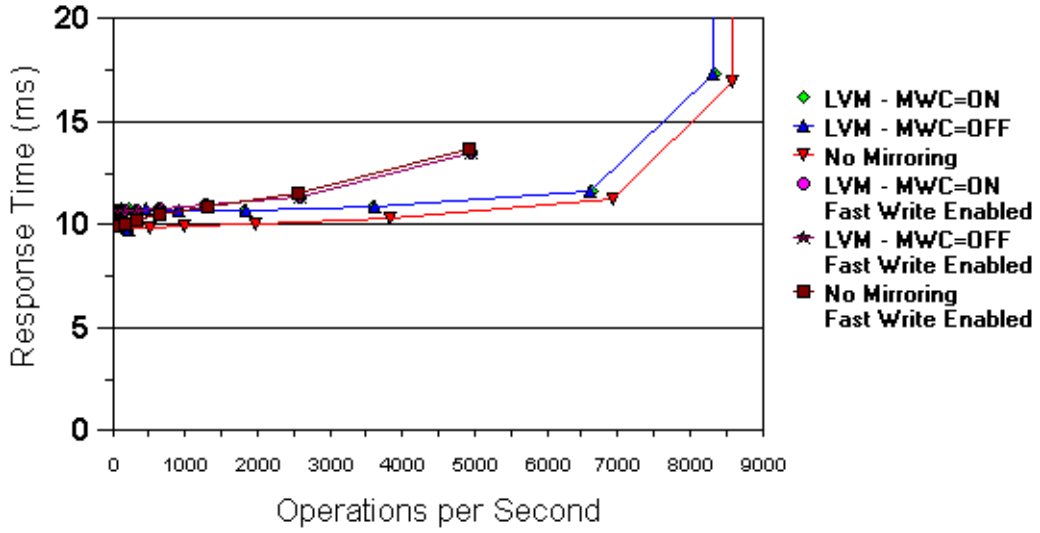

Figure 26: Non-RAID response time comparison of for 100% read operations (one adapter, 96 disks, random 4KB operations)

These figures show the effect of the array or adapter becoming saturated; that is when the array receives too many requests for it to handle immediately and therefore queues them. This causes the response time to increase significantly. If the application is waiting for that I/O request to complete before preceding (for example when performing synchronous writes), that application sees high I/O wait times and the overall system performance may be reduced.

The figures show that the response time of a given array type is approximately constant for different operation rates provided the array is not driven to the point on the response curve where saturation begins. To ensure an application does not experience excessive I/O response time, you should ensure that the disks are operating in the flat section of the figures.

Figure 3 on page 35 showed the maximum operations per second possible for the given environment for the different array types. Figure 21 to Figure 26 show that the response times increase dramatically for each RAID and non-RAID type as the operations per second approach these limits. The more operations per second with relatively constant response time for different arrays depends on the ratio of reads to writes:

- If the operations are 100% reads, LVM mirrored logical volumes provide the maximum range of operations per second with a flat response time and RAID 5 provides the least range. The actual response time during this flat region is approximately the same for all RAID and non-RAID types as they all involve reading a single disk.
- If the operations per second are 100% writes, RAID 1, RAID 0+1 and LVM mirrored logical volume all provide the maximum range of operations per second with a flat response time. However, if the application requires that mirror write consistency has to be on for mirrored logical volumes, the flat response time range of operations is reduced to the same range as for RAID 5. The actual response time during this flat region is best for RAID 1, RAID 0+1 and LVM mirrored logical volumes with write consistency off. The response time for mirrored logical volumes with write consistency on is higher than the response time for RAID 5.

Figure 21 to Figure 26 also show the effect on response time of using fast write cache. When fast write cache is used, the write operation is completed as soon as the data is stored in the non-volatile cache and no disk access is required before completion of the operation. The figures show that when fast write cache is used, the response time is reduced significantly for write operations and this can have a significant effect on synchronous write performance. This response time reduction is independent of which type of RAID is being used. Fast write cache, however, does incur extra processing within the adapter and so it can be seen that the maximum number of operations per second possible and hence the flat response time section for increasing operations per second is reduced by using fast write cache. Provided the throughput is not so high that the adapter is becoming saturated because of the extra processing required for the fast write cache, the response time is much improved if fast write cache is enabled for all types of RAID and non-RAID configurations.

Array Build and Rebuild Times

After a RAID 5 array has been created, there is a time taken before the parity has been built. Until this has completed, use of the array is not protected against the failure of one of its member disks. After the parity for the entire RAID 5 array has been built, the array is now protected against disk failure and a failure of one of its member disks causes the failed disk to be replaced by a hot spare immediately. The data on this replacement disk is rebuilt from the data on the other member disks. Until this rebuild has completed, the array is not protected against the failure of another member disk.

For RAID 1 and RAID 0+1 arrays the array is protected immediately after array creation against a disk failure. A failure one of its member disks causes the failed disk to be replaced by a hot spare immediately and the data on this replacement disk is rebuilt from the mirrored pair disk. Until this rebuild has completed, the array is not protected against the failure of another disk that is part of the mirrored pair.

For RAID 5, the build and rebuild times increase as the number of member disks increases and as the size of each member disk increases. If the array consists of 6+P 9.1 GB disks, the following times are taken to initially build and to rebuild after a failure if there are no concurrent I/O operations to the array:

Initial build time	32 min.
Rebuild time	49 min.

In practice, applications may need to issue I/O operations to the array whilst it is being rebuilt after a failure of a disk. In this case, the rebuild time is longer as it is performed in the background to the I/O operations. The priority of the rebuild can be increased to shorten the rebuild time at the expense of creating more impact to concurrent I/O operations to the array. The default rebuild priority is a value of 50% which can be increased to 100%. The rebuild time for the 6+P array of 9.1 GB disks with 3 operations per second per Gbyte of data being issued concurrent with the rebuild can be reduced as follows:

Rebuild Priority	Rebuild Time
50%	265 min.
85%	85 min.

For RAID 1 and RAID 0+1, the rebuild time increases as the disk size is increased but is not affected by how many member disks are in the array. If the array consists of 9.1 GB disks with a 16 KB strip size, the following times are taken to rebuild after a failure if there are no concurrent I/O operations to the array:

Rebuild time	50 min.

The rebuild time is increased if I/O operations are taking place concurrently with the background rebuild. The rebuild time for the RAID 1 or RAID 0+1 array can be reduced by increasing the rebuild priority setting.

For Mirrored Logical Volumes, the rebuild time of a 9.1 GB disk with no concurrent I/O operations taking place is 25 minutes.

Performance Conclusions

The performance numbers listed in this section are for guidance to help in determining which type of RAID is best suited for your particular environment, workload and cost requirements. It can be seen that there are many different options to consider from a performance perspective with the Advanced SerialRAID Plus adapter. It would be wrong to say one type of RAID is always best as that oversimplifies the situation. The following guidelines can be developed based on the performance information in this chapter:

- If you want the minimum number of disks and your application has a low percentage of write activity, you should use RAID 5. RAID 5 becomes less attractive for applications that are heavily write intensive.
- If you want the best transaction throughput performance possible, particularly if there are a high percentage of writes, you should use one of the mirroring options. Mirrored logical volumes provides the best throughput performance only if the applications permit mirror write consistency to be disabled. If this has to be enabled, RAID 1 and RAID 0+1 provide better transaction performance when writes are more than 25%. Alternatively you should consider enabling fast write cache when using LVM mirroring with Mirror Write Consistency on.
- Of the mirroring options, RAID 0+1 has the advantage over RAID 1 and LVM mirrored logical volumes of reducing the skew to certain disks by spreading the data across disks. This does not show on the benchmarks described in this section as they were to totally random block addresses, but real applications may access certain disks and certain ranges of blocks more often.
- The maximum data rate possible through the adapter for long sequential operations is independent of RAID type if all the operations are reads and the maximum is best for RAID 1 and RAID 0+1 arrays if there is a high percentage of write activity.
- If you require the minimum response time, you should use fast write cache that significantly reduces the response time for write operations for all RAID types. If fast write cache is used with RAID 5, the performance of short sequential write operations will also be improved because fast write coalesces writes to the member disks into ones that write to all the members for a stripe and avoid having to read the old data.

  Use of fast write cache does, however, require extra processing by the adapter resulting in the adapter limit of short transaction processing type operations being reduced. Your applications are probably not limited by this adapter limit, so this reduction may not cause any performance loss.
- For each RAID 5 array one disk is required to provide availability. The total number of disks in the system required to provide availability is reduced as the number of member disks per array is increased. The rebuild time of a replacement disk after a failure of a member disk is, however, increased as the number of member disks is increased
- If using RAID 5 for intense processing of long operations, it may be preferred to attach the required drives to two adapters rather than a single adapter if the number of drives causes performance to be limited by the adapter or host PCI bus capabilities. You can also use multiple independent adapters with separate loops to reduce the load on any adapter. This is only relevant if there is a large number of disks.
- If using more than one adapter in the same system, it is advisable to install them on separate PCI busses as the Advanced SerialRAID Plus adapter can use more than half the available bandwidth of a PCI bus, far higher than most other adapters.
- To avoid any single point of failure, two systems should be attached to the SSA loop and this is possible for all RAID types. When 2 adapters are used, the total transaction processing possible is increased by 50% compared to that possible with a single adapter. Processing is required between the adapters so that they both are aware of write activity to arrays and fast write cache so that if one adapter fails, the other adapter can continue to keep the parity and mirrored copies valid.

- You should use both SSA loops. You should also position array member disks on each loop so that half are closest to one adapter port and half are closest to the other adapter port for maximum bandwidth. Also, if two host systems are being used and one host uses one array and the other host another array, the member disks of the arrays should be positioned close to their using systems. This enables the spatial reuse ability of the SSA loop to be used and data can be transferred to both adapters simultaneously.

5 Planning for Availability

Data availability is dependent on various characteristics of a subsystem. The following scenarios need to be considered, where relevant:

- Host failure
- Adapter failure
- Drawer failure
- Disk failure
- Cable or connector failure
- Miscellaneous other types of failure (for example: power down, or deliberate maintenance action)

Terminology

In this section, the following terms are used:

host

An RS/6000 workstation containing a disk subsystem

loop

A closed SSA network, that is, one in which you can trace the path around the network and eventually return to the start point. So, there are two paths round the network to each device. Adapter 1 and disks 1, 2, and 3 in Figure 27 form a loop. The loop would be converted to a string if disk 2 was removed.

string

An SSA network with an open link. Adapters 1 and 2 and disks 4 and 5 in Figure 27 form a string. The string could be converted into a loop by connecting disk 5 to port B2 on adapter 1.

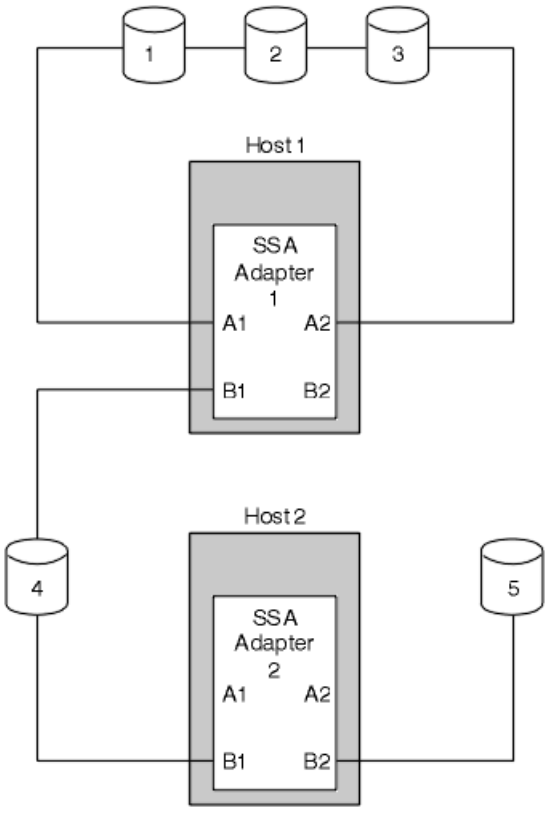

Figure 27: Examples of SSA networks

Availability Characteristics of a Non-RAID SSA Subsystem

In a single-host non-RAID subsystem where disks are not configured with LVM mirroring, a single failure of a disk or host causes data to be unavailable.

In a typical multi-host SSA subsystem, each host is connected to one or more drawers of 7133 disks. Each disk can be accessed by up to 8 hosts. The connection is in a loop as illustrated in Figure 28 for a 4 host configuration. The adapters can be PCI SSA Multi-Initiator /RAID EL adapters, Micro Channel SSA Multi-Initiator /RAID EL adapters or Advanced SerialRAID Plus adapters in any combination.

If a host is powered down, circuitry in the 7133 causes the SSA connection to that host to be bypassed and the

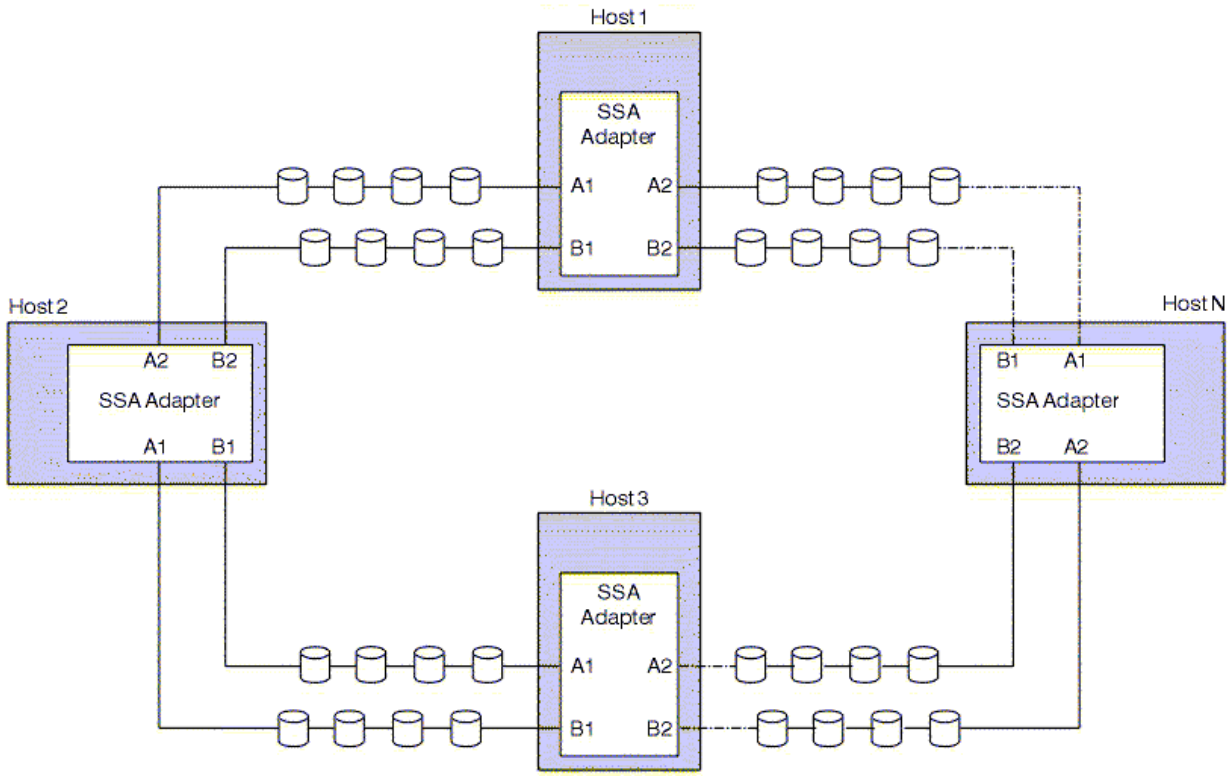

Figure 28: Example of a Multi-host Non-RAID SSA Configuration

continuity of the loop without that adapter is maintained. If a disk drive fails, the SSA loop continuity may be broken and the SSA network becomes a string rather than a loop. In this case, all SSA traffic that used to pass through the failed disk is now rerouted around the network to use the alternate path so that all adapters can continue to access all the operational disks.

Whilst the SSA network is operating as a string, the configuration is exposed to the possibility of a second failure (examples are a disk or cable failure). This second failure would make some disks on the loop unavailable to some or all hosts.

It is possible, by careful selection and placement of disks within a loop, to minimise the problems that a second failure can cause. To do this, you must consider such factors as volume-group takeover, ownership of disks and which hosts need to access which disks. Your local RS/6000 HACMP expert has details of these options and can provide more advice on their selection.

Availability characteristics of SSA RAID Subsystems

In a typical SSA RAID subsystem, one or two hosts are connected to one or more drawers of 7133 disks. The connection is in a loop, as in Figure 28 for SSA non-RAID systems, but there can be only one or two hosts in each loop. The comments above relating to availability after a host powers down or fails and the rerouting of access to disks after a disk or SSA cable failure apply equally to SSA RAID systems as for SSA non-RAID systems. Not all the disks in fact have to be configured as members of arrays, and the SSA loop can support any combination of RAID type or non-RAID disk provided there are no more than two adapters on the SSA loop. The main enhancement of using RAID is that access to data is still available after a member disk has failed.

It is a characteristic of all types of RAID that access to data is still made available after one disk has failed. Hot spare disks should be available and this will automatically replace the failed disk and the data rebuilt on this new member after the failure for RAID 1, RAID 0+1 and RAID 5. Special scripts are necessary to provide this function on mirrored logical volumes and can be obtained from http://www.storage.ibm.com/hardsoft/products/ssa. Once the data on the replacement disk has been built, all arrays can tolerate a second disk failure.

Until the replacement disk has been rebuilt, a RAID 5 array is exposed to loss of data due to any second member disk failure. RAID 1, RAID 0+1 and mirrored logical volumes are only exposed to loss of data due to a second member disk failure before rebuild has completed if the second failure is to the other disk of a mirrored pair. If the second disk failure is to a different mirrored pair of disks, then this second failure does not cause any data loss.

RAID 1 and RAID 0+1 arrays can be configured to still provide access to data after an entire site loses power or connection to an entire site is lost from another site. The RAID configurator provides information about the location of each candidate disk from which the RAID 1 or RAID 0+1 array is to be built and the array can be built so that mirrored pairs of disks reside in different sites. Sites can be separated by use of up to 10,000 m of optical fiber. If the RAID 1 or RAID 0+1 array is configured so that the mirrored pairs of disks are in separate sites and there is a host on each site, operations to the array continues automatically when the primary site loses access to the secondary site (due to a power loss or loss of communication). Operations can continue to the array if the secondary site loses access to the primary site but only after user intervention. Hot spare disks can be assigned to pools along with selected array member disks, so that if a member disk fails, it is replaced by a hot spare from the same site as the failed member disk. The subsystem is then still tolerant to the loss of an entire site even after a hot spare has been introduced to an array. Mirrored logical volumes could also be configured to achieve mirror pairs being located on different sites, but not easily. When RAID 1 and RAID 0+1 arrays and hot spares are configured, the configurator identifies the physical enclosure that holds each disk, but this information is not presented when LVM mirrored disks are configured.

There can be up to 2 adapters per loop if disks are configured as members of RAID 1, RAID 0+1 or RAID 5 arrays. If both of these adapters are in the same host system, operations will failover to the remaining adapter when one adapter fails. If there are two systems and one SSA adapter in each system, a failure of one SSA adapter will cause HACMP failover to the other system only if the SSA loop is used for the only heartbeat communication. Using SSA as the only path for heartbeats is not recommended. It is recommended that HACMP is set up to failover when particular error log entries are cut. Then, either the entry that indicates loss of adapter, or the entry that indicates loss of the volume group could be used to trigger HACMP to move the application to the takeover node. In this way availability of data can be maintained when an SSA adapter fails.

Summary of Availability Characteristics

Host failure

All SSA configurations permit multiple hosts to be connected to an SSA loop and if there is more than one host per loop, a failure of a host can be tolerated. Powering off a host does not impact the availability possible due to SSA cable failures as the 7133 bypasses the powered off host so the network is still an SSA loop.

Adapter failure

SSA supports two adapters in the same host in the same loop. This means that if the adapter electronics fail, the second adapter still provides the host with access to the loop. The failover mechanism when both adapters are in the same host system is architected into the SSA device driver layers and needs no application intervention. If any of the

disks are configured as members of an array, the maximum number of adapters that can be in the SSA loop is two. Having both of these in the same host system would mean that if the single system rather than one of the adapters fails, operations could not continue.

If there are two SSA adapters in separate systems on the SSA loop, it is recommended that HACMP failover is initiated by an error log entry that indicates an adapter failure or loss of a volume group in addition to the heartbeat failure to maintain access to data when an adapter fails.

Disk failure

RAID 5, RAID 1, RAID 0+1 or mirrored logical volumes can be used to allow the system to tolerate a single disk failure without loss of data availability. All RAID types or LVM mirrored disks can tolerate a second disk failure if that disk is not part of the same array or mirrored logical volume. RAID 0+1 and LVM Mirroring with striping can tolerate multiple disk failures within the array or LVM striping set provided both disks of any mirrored pair do not fail.

Failure of an entire site

RAID 1 and RAID 0+1 permit the array to be configured from disks on separate sites to allow one system to tolerate the loss of the system and disks on the other site without the loss of data availability. Hot spares and array member disks can be assigned to pools so that the loss of a site can still be tolerated after a hot spare has been introduced.

Drawer failure

The 7133 drawer has redundancy for power supplies and fans so that it continues to operate when members fail.

Backup of Data

One of the most common causes of data loss is accidental erasure by users. The use of RAID arrays to enhance availability does not reduce the need to take regular backup copies of data. This can be automated by using the IBM product ADSM/6000, or an equivalent.

6 Planning your SSA configuration

When planning your installation you need to make a number of decisions related to:

- Number of disks required

  This is largely determined by the total amount of storage required, the number and type of arrays (and hence parity or mirrored disks) required, the number of hot spares required and whether any non-RAID disks will be mirrored. Some consideration should be given to avoid the situation where there are 16 disks in a configuration and an additional hot spare might be allocated in future. Since this would require a 17th disk, purchase of an additional 7133 would also be required, making this additional disk rather expensive.

- Number of adapters required

  Each adapter can sustain a certain I/O rate or data rate. If a single adapter cannot supply adequate I/O bandwidth then additional adapters need to be considered. Bearing in mind the relatively small cost of adapters versus disks, additional adapters can add significant performance at little additional cost. Increasing the number of adapters to improve performance is only relevant if there is a large number of disks on each loop.

- Size and arrangement of SSA loops

  This has been discussed in some detail already in “Planning for Performance” on page 25.

- Array sizes

  Once created, arrays cannot easily be increased or decreased in size. To transfer data to a larger or smaller array, requires that the data be backed up, the old array deleted, the new array created, and the data restored onto the new array. Since each array corresponds to a physical volume and this is the unit from which logical volumes are built, there is an argument for using relatively small numbers of members in each array. However the more member disks in the RAID 0+1 or RAID 5 array, the more chances of parallel reads being able to take place. When using RAID 5, the more member disks in the array, fewer disks are required to provide availability for the total capacity required (1 per array). If you are unsure of what RAID 5 array size to you, it is suggested you choose 6+P arrays.

- Positioning of array candidates

  Member disks of RAID 1, RAID 0+1 or RAID 5 should be located on the loop so that half the member disks are closest to one port of the adapter and half are closest to the other port to achieve maximum performance.

  Mirrored pairs of disks of a RAID 1 or RAID 0+1 array should be located on different sites if it is required to continue operations when an entire site is powered off or access to it is lost. If this is required, hot spares and array member disks should be assigned to pools such that hot spares are used to replace failed member disks from the same site.

- Use of hot spares

  Here the strategy that should be employed depends somewhat on the number of arrays that are attached to the adapter. If there are only few (say 1 or 2) arrays attached then it may be appropriate to have one hot spare per array. Having a hot spare per array is expensive however, and since hot spare disks are held in a pool before they are needed, there is little value in having many hot spares since the chances of them all being used at once is small. Since up to 96 disks can be attached to a single adapter, having 2 hot spares per loop may be a sensible number.

  The size of the hot spare disks must be at least that of the smallest member of the array it is associated with. It is preferable that the hot spare disk should have at least the capacity of any member disks in any array on the loop. This makes the hot spares usable for any array.

  Because of the possibility of data loss due to a second disks failure or when a RAID 5 array is running in

degraded mode and power is lost, it is strongly recommended that the system is never run without hot spares being available.

Another important consideration is that the adapter requires a hot spare to be in the same loop as the array in which it may be used. If the installation consists of only a small number of arrays per adapter, it may make more sense therefore to keep all the arrays (and hot spares) on a single loop. Provided the number of arrays is relatively small, there will be no performance impact in restricting operation to a single loop.

- Loop configuration rules
  1. The A1/A2 and B1/B2 ports of the adapter cannot be in the same loop.
  2. The Advanced SerialRAID Plus adapter can coexist on the same SSA loop as the following adapters:

Feature code	Adapter type	Maximum adapters per loop
6215	PCI SSA Multi-Initiator/RAID EL Adapter	8 Non-RAID
		2 RAID 5
		0 Fast Write Cache
6219	Micro Channel Multi-Initiator/RAID EL Adapter	8 Non-RAID
		2 RAID 5
		0 Fast Write Cache
6230	Advanced SerialRAID Plus adapter	8 Non-RAID
		2 RAID 5, RAID 1, RAID 0+1
		2 Fast Write Cache
		0 RAID 0

  3. Between 0 and 48 disks may appear in a loop, together with as many adapters as are permitted by the above rules.
  4. Each slot in the 7133 drive enclosure must contain either a disk or a blank carrier.
  5. No more than 3 blank carriers may appear adjacent to one another in a 7133 enclosure.
  6. Maximum copper cable length is 25m, (7 Cable lengths: 0.18, 0.6, 1, 2.5, 5, 10, & 25m are available).
  7. Maximum fiber optic extender length is 10 km.
  8. Disks do not have to be cabled in any particular sequence, although as described elsewhere there may be some performance benefits by positioning member disks of arrays such that half are accessed by one port of the adapter and half by the other port.
  9. All array members for any particular RAID 0, RAID 1, RAID 0+1 or RAID 5 array are all in the same loop (and also that any hot spare for that array is also in the same loop).
  10. When two adapters from the same host appear within the same loop that also contains adapters in other hosts, the two adapters should be placed next to one another without any intervening disk drives. This is because the disks may become isolated if the adapters in that host are reset or held at a reset level. The Bypass Circuits in the 7133-020 and 7133-600 enclosure help to resolve the problem of a host being powered down, but will not operate if the adapter is simply held reset. In many host systems, an adapter may be held reset for a long period during the boot sequence or if the host fails in certain ways. This point only applies to multi-host loops. In a single host application, in which two adapters appear in the loop, disks should be placed equally between the adapters, all other things being equal. This is because access to those disks from other hosts is not an issue in that particular configuration.
  11. Each host may contain a maximum of two adapters in the same loop. These are the primary and secondary adapters for the hdisks that are configured on that loop and in normal operation the load on the loop will be dynamically balanced between these adapters. The hdisks could be non-RAID disks or RAID arrays. The SSA device drivers provide an automatic failover mechanism between the primary and secondary adapter should one or other adapter fail. If only one adapter is in the loop then no failover is possible.
  12. More than two adapters may be installed in a host, but they must be in different loops.

13. The user should refer to the 'PCI adapter Placement Reference' to determine how many adapters are permitted per PCI bus.

- Selection of operation modes:
  - Hot Spares Enable

    This is the recommended mode of operation. Sufficient hot spares should be allocated to cover the possibility that multiple disks may fail within the period it takes to rebuild an array. If no hot spare is available when a member disk fails and a second disk fails or if power is lost when writes are in progress to a RAID 5 array, data could be lost. It may be preferable, depending on the application that is being run, to prevent any writes when a RAID 5 array goes degraded. This can be done by setting the array attribute Read_Only_When_Exposed to True. This means that any write operation that is requested when the array is in a exposed state (when one member is missing) will be failed back to the application with an I/O error. This means that data availability is not risked, and the application can continue to read the data on the array. The application cannot update the data on the array until the failed disk has been replaced. Operating with Read_Only_When_Exposed=TRUE is a highly specialised mode of operation and should only be employed after careful investigation of the consequences of writes being failed back to the application.
  - Fast write cache enabled

    The fast write cache facility provides a means of substantially reducing the response time of write operations. This is done by caching write data on the adapter card and returning good status to the host prior to writing the data to disk. Since the write cache is a non-volatile memory, the data can be safely held on the adapter card until a suitable opportunity to destage the data to disk arises. Early return of completion to the write operation avoids the need to wait for the disk rotation and seek to complete and can result in each I/O operation being completed within a few milliseconds.

    As using the fast write cache improves the response time of individual writes, synchronous write streams are significantly speeded up. By saving up a lot of writes and destaging them together, it allows the disk to re-order the commands for minimal seek time. It also improves the throughput of a sequential write stream by coalescing short transfers into longer ones which will not miss revolutions when writing to disk. This coalescing also results in significant performance improvement for sequential RAID 5 writes as writing to a full stripe avoids read accesses.

    These gains have to be traded off against the loss in maximum adapter throughput due to extra internal overheads which reduce both the maximum operations/second and the maximum data rate.

    Generally, the adapter is not near its maximum capability, so using the fast write cache gives a noticeable performance gain to customer applications involving significant numbers of write operations.

    Since the write cache is relatively small compared to the actual disk capacity, if the best performance is to be obtained, it is important that only the key write operations are directed through the write cache. For example:

1. In database applications, where a journal log is being kept, that log should be positioned on a fast write cache disk.
2. In applications where a number of disks are being written to in a synchronous fashion, these disks are candidates for fast write cache operation as the use of fast write cache means that the completion of the overall operation need not wait for the write to the last disk to complete.

Note that the above comments regarding Fast Write Caching are not specific to the SSA adapters, but apply to Fast Write Caches in general.

7 Configuration Optimization

When an array is created using the ssaraid configurator, there are several options that can be chosen to provide optimum performance and availability. This section discusses these options for each type of RAID array and explains how they can be used for the maximum benefit for your particular environment.

RAID 5 ssaraid options:

Function	Default	Comments
Member disks		The number of member disks can be from 3 to 16 and they must all be on the same SSA loop. The larger the number of disks, fewer disks are required to provide availability for the same total capacity (1 per array). However, more disks means a longer time is taken to rebuild a replacement disk after a disk failure, and during this period a second disk failure will cause a loss of all array data. Large numbers of disks in the array, particularly with high capacity disks, may result in the physical volume of the array being unacceptably high. More data is affected by a disk failure for larger number of member disks resulting in reduced performance while rebuilding onto a hot spare and more data being exposed if a second disk fails before the rebuild has completed. The smaller the number of disks, the more likely it is that write operations span an entire stripe (strip size x number of members minus one). In this case, write performance is improved because then disk writes do not have to be preceded by disk reads. The number of disk drives required to provide availability may be unacceptable if arrays are too small. If in doubt, arrays of 6+P (7 member disks) is recommended.
Strip size	64 Kbyte	This can be 32 Kbyte or 64 Kbyte. This defines the amount of contiguous data that is assigned to one member disk before assigning the next strip of data to the next member disk. If 32 Kbyte strip size is selected rather than the default 64 Kbyte, there will be twice the number of strips in the array and the rebuild time of a replacement disk after a disk has failed will be longer as rebuild is performed on a strip granularity. The number of arrays that can be supported for the large 36 Gbyte disks is reduced from 32 if the strip size is 64 Kbyte to 17 if the strip size is 32 Kbyte unless the adapter memory is increased to 128 Mbyte. Using 32 Kbyte strip sizes may make it more likely that long write operations are for a full stripe, resulting in improved performance, but this also depends on the number of member disks.
Enable Hot Spares	Yes	It is strongly recommended that there is always a Hot Spare available and that this option is enabled.

Hot spare only from Preferred Pool	No	Hot spares can be assigned to specified pools and array members can also be assigned to these pools so that when a disk fails, only hot spare disks are used from the specified pool. This has limited use for RAID 5 arrays, so the default is to disable this option. Allocating hot spares and arrays to specified pools would allow you to ensure that a hot spare was used from a selected physical drawer of disks if, for example, you wanted all the member disks of an array to always be in specific drawers.
Allow page splits	Yes	Data is allocated to member disks in blocks of the length of the strip (default 64 Kbyte). Write operations that write to blocks on one or more strips involved writes to one or more disks. These write operations are normally performed in parallel (allow page splits = yes) to maximise the performance. This does, however, mean that if a write operation that spans two strips of data fails to complete, the end of the data may have been written and also the start of the data, but some blocks in between may not have been written. If your application requires that data must be written in a contiguous manner, such that on a failure to complete the write operation the end of the data is only written if all the preceding data has been written, you should set allow page splits = No. This will incur a reduction in performance for write operations that straddle strip boundaries.
Enable fast write	No	Enabling fast write will significantly reduce the response time for write operations. It will also improve the performance of sequential write operations as the writes to disk will be coalesced in the adapter before being issued to disks, and will then be for a full stripe (to all the member disks). This avoids having to read disks before writing. Extra processing is required in the adapter for fast write cache operations, so reducing the maximum operations per second possible by the adapter, but this may not limit the performance which is probably limited by the number of disks attached to the adapter, or the ability of the system to issue operations at a high rate.
Rebuild Priority	50%	Unlike all the other options, Rebuild Priority can only be changed through the SSA command line. The syntax for this command line to change the rebuild priority to 80% is: ssaraid -I ssaX -H -n array_name -a rebuild_priority=80 where ssaX is the identification of the adapter (for example ssa0, ssa1) and array_name is the 15 character identification of the array. Rebuild priority can be any percentage up to 100; the higher the value, the shorter is the rebuild time when a hot spare is introduced, but also the higher is the performance degradation to processing I/O operations concurrent with the rebuild.

RAID 0+1 ssaraid options:

Primary Disks		The total number of disks in a RAID 0+1 array can be any even number between 4 and 16. Half of these are to be primary disks and half to be secondary disks. All the disks must be on the same SSA loop. A fundamental characteristic of any mirroring implementation is that there are two copies of the data available and either copy can be used for reads. Care has to be taken to ensure that if the network is broken, or for some other reason, one system can only see one copy of the data and another system can only see only the other copy of the data, then both systems must not be allowed to continue to operate on the copy of data they can access. This would result in data becoming unsynchronized and difficult from which to recover. Logical volume mirroring has a concept of a quorum that creates an imbalance between the two copies to ensure operations only continue by the system can access the quorum. RAID 0+1 and RAID 1 prevent erroneous operations after a network failure by defining half the disks to be primary disks and half to be secondary disks. When access to both primary and secondary disks is not possible and the system cannot access the other system, the system is only allowed to continue to operate to the array if it has access to the primary disks by default. This can be changed to allow access to secondary disks instead of the primary disks in this situation by the Split Array Resolution option. When selecting disks from the list of candidates for primary disks, the drawer identification is displayed for each disk. If you require that one of the systems can continue to operate when an entire site fails, for example due to loss of power, you can define all the primary disks to be in drawers on one of the sites. Disks can then be selected for secondary disks that are in drawers located on the other site.
Secondary Disks		The total number of disks in a RAID 0+1 array can be any even number between 4 and 16. Half of these are to be primary disks and half to be secondary disks. All the disks must be on the same SSA loop. The differences between primary and secondary disks is described for the primary disks option above.
Strip Size	16 Kbyte	This can be 16 Kbyte, 32 Kbyte or 64 Kbyte. This defines the amount of contiguous data that is assigned to one member disk before assigning the next strip of data to the next member disk. The smaller the strip size, the more disks are used to share data transfers for long operations, so maximising the data rate performance. Rebuild of a replacement disk is performed a strip at a time, so the smaller the strip size the longer is the time taken to rebuild a replacement disk after one disk member fails.

Split Array Resolution	Primary	Split Array Resolution controls whether access is required to the primary or secondary disks for operations to be allowed when access to both primary and secondary disks is not possible and when the systems cannot access each other. This avoids both systems being able to continue operations independently to the copies of the array they can both access when the network is broken into two sections each of which contains an operational system and half the disks. The default is that access to the primary disks is required. The system that cannot access the primary disks cannot continue operation to the array. If the reason for the failure is, say, a failure or power outage of one system and its local disks, you may chose to continue operation from the system that can still access the secondary disks. This is possible by using ssaraid to set the split array resolution to Secondary. The system that can access the secondary disks can now operate to the array. When the primary disks next become accessible, those primary disks will be rebuilt from the secondary copies before being used. When the rebuild is complete, the split array resolution option will revert back to primary.
Hot Spare only from Preferred Pool	No	Primary and secondary disks can be allocated to disk enclosures in different power domains to permit operations to continue when one power domain fails. If this is done, it is therefore desirable to ensure that if any disk fails, it is replaced by a hot spare disk from the same power domain. Hot spare pools can be created using the ssaraid Add Hot Spare Pool function. These create for a specified pool a list of array member disks and a list of hot spare disks. If the Hot Spare only from Preferred Pool is set to Yes, then when a disk that is a member of a hot spare pool fails, it will only be replaced by a hot spare that is available from the same pool. If a hot spare is not available from this pool, no hot spare will be introduced. The array will continue to operate, but in a degraded mode and it is exposed to data loss if the other member disk of the mirrored pair fails. If the Hot Spare only from Preferred Pool is set to No (default setting) and a member disk fails that has been assigned to a hot spare pool, the failed disk will be replaced by a hot spare disk from the specified pool if such a hot spare disk is available. If such a hot spare is not available, the failed disk will be replaced by a hot spare from another pool. An error log will be reported to alert you that this has taken place, and you may then need to add a new disk to the appropriate domain and change the disk member that was replaced to again ensure that all primary and all secondary disks are in separate domains.
Allow Hot Spare Splits	No	When this is No, if a failure, for example a power loss of one domain, results in a system being able to access exactly half the disks, that is all the primary or all the secondary disks, the adapter will not attempt to replace any disks not accessible by hot spares. This option is provided to avoid hot spares being used to replace disks that are not accessible due to a loss of power in a domain.

Allow Page Splits	Yes	Data is allocated to member disks in blocks of the length of the strip (default 16 Kbyte). Write operations that write to blocks on one or more strips involved writes to one or more disks. These write operations are normally performed in parallel (allow page splits = yes) to maximise the performance. This does, however, mean that if a write operation that spans two strips of data fails to complete, the end of the data may have been written and also the start of the data, but some blocks in between may not have been written. If your application requires that data must be written in a contiguous manner such that on a failure to complete a write operation the end of the data is only written if all the preceding data has been written, you should set allow page splits = No. This will incur a reduction in performance for write operations that straddle strip boundaries.
Initial Rebuild	No	After an array is created, each mirrored pair of disks contain different data. When Initial Rebuild = No (default) each mirrored pair of disks will continue to have different data except for all blocks that have subsequently been written by write operations to the array. This default setting avoids the adapter having to take processing time and disk utilization initially to ensure that both mirrored copies of data are the same for data that has never been written. If your applications require that when blocks of data that have never been written are read the same data is always returned for those unwritten blocks, you should set the Initial Rebuild option to Yes. This will incur some processing overhead by the adapter when the array is first created that will reduce the performance available from the adapter to process I/O operations.
Enable Fast-Write	No	Enabling fast write will significantly reduce the response time for write operations. Extra processing is required in the adapter for fast write cache operations, so reducing the maximum operations per second possible by the adapter, but this may not limit the performance which is probably limited by the number of disks attached to the adapter, or the ability of the system to issue operations at a high rate.
Rebuild Priority	50%	Unlike all the other options, Rebuild Priority can only be changed through the SSA command line. The syntax for this command line to change the rebuild priority to 80% is: ssaraid -I ssaX -H -n array_name -a rebuild_priority=80 where ssaX is the identification of the adapter (for example ssa0, ssa1) and array_name is the 15 character identification of the array. Rebuild priority can be any percentage up to 100; the higher the value, the shorter is the rebuild time when a hot spare is introduced, but also the higher is the performance degradation to processing I/O operations concurrent with the rebuild.

RAID 1 ssaraid options:

Primary Disks		For each RAID 1 array there will be 1 primary disk. The reason for defining the disk as a primary or secondary is for the actions taken after loss of access to the other system and half the disks as described for RAID 0+1 above.
Secondary Disks		For each RAID 1 array there will be 1 secondary disk. The reason for defining the disk as a primary or secondary is for the actions taken after loss of access to the other system and half the disks as described for RAID 0+1 above.
Split Array Resolution	Primary	Split Array Resolution controls whether access is required to the primary or secondary disks for operations to be allowed when access to both primary and secondary disks is not possible and when the systems cannot access each other. This is as described for RAID 0+1 arrays above.
Hot Spare only from Preferred Pool	No	This controls whether a hot spare can only be taken from a designated hot spare pool or any other hot spare can be used when a member disk fails. This is as described for RAID 0+1 above.
Allow Hot Spare Splits	No	This controls whether a hot spare disk is used to replace a member disk when access is lost to one of the member disks and also to the other system. This is as described for RAID 0+1 above.
Allow Page Splits	No	When Allow Page Splits is no, a write operation that straddles a 64 Kbyte address can result in the end of data being written before the start of data as long writes are broken down into multiple 64 Kbyte writes to the disks and these may be executed in any order to maximise performance. This does mean that if a write operation that spans a 64 Kbyte address fails to complete, the end of the data may have been written and also the start of the data, but some blocks in between may not have been written. If your application requires that data must be written in a contiguous manner, such that on a failure to complete the end of the data is only written if all the preceding data has been written, you should set allow page splits = No.
Initial Rebuild	No	This controls whether both member disks are made to contain the same data initially or not as described for RAID 0+1 above.
Enable Fast-Write	No	Enabling fast write will significantly reduce the response time for write operations. Extra processing is required in the adapter for fast write cache operations, so reducing the maximum operations per second possible by the adapter, but this may not limit the performance which is probably limited by the number of disks attached to the adapter, or the ability of the system to issue operations at a high rate.
Rebuild Priority	50%	Unlike all the other options, Rebuild Priority can only be changed through the SSA command line. The syntax for this command line to change the rebuild priority to 80% is: ssaraid -I ssaX -H -n array_name -a rebuild_priority=80 where ssaX is the identification of the adapter (for example ssa0, ssa1) and array_name is the 15 character identification of the array. Rebuild priority can be any percentage up to 100; the higher the value, the shorter is the rebuild time when a hot spare is introduced, but also the higher is the performance degradation to processing I/O operations concurrent with the rebuild.

RAID 0 ssaraid options:

Function	Default	Comments
Member disks		The number of member disks can be from 2 to 16 and they must all be on the same SSA loop. Data is held on each disk in 16 Kbyte strips. If operations to the array are typically 128 Kbyte long and there are 8 member disks, these operations cause concurrent operations to all the 8 disks. Larger number of member disks than this would only show a performance increase if operations to the array were longer than 128 Kbyte. An advantage of RAID 0 for short transaction processing type of operations is that as data is spread across the member disks in 16 Kbyte strips. If the application accesses closely addresses data blocks frequently, these access different disks rather than the same disk as in the non-RAID situation, and so performance will not be limited by excessive accessing to a few disks. 4 member disks are sufficient to benefit from this reduction in skew for short operations A disadvantage of large number of member disks is that as no redundancy is provided for RAID 0 arrays, when a disk fails all the data of the array is lost and not just that of the failed disk and more data would be lost the larger is the number of member disks.
Allow page splits	Yes	Data is allocated to member disks in blocks of the length of the strip (default 64 Kbyte). Write operations that write to blocks on one or more strips involved writes to one or more disks. These write operations are normally performed in parallel (allow page splits = Yes) to maximise the performance. This does, however, mean that if a write operation that spans two strips of data fails to complete, the end of the data may have been written and also the start of the data, but some blocks in between may not have been written. If your application requires that data must be written in a contiguous manner, such that on a failure to complete the end of the data is only written if all the preceding data has been written, you should set allow page splits = No. This will incur a reduction in performance for write operations that straddle strip boundaries.
Enable fast write	No	Enabling fast write will significantly reduce the response time for write operations. It will also reduce the write operations to disks as write operations from the host are coalesced into fewer and longer operations to disks when data is being destaged.

Fast Write Cache options:

Fast write cache can be enabled or disabled for any array or non-RAID logical SSA disk by using smitty and the ‘Change/Show Characteristics of an SSA Logical Disk’ screen obtained by selecting from the smitty System Management the following:

Devices

SSA Disks

SSA Logical Disks

Change/Show Characteristics of an SSA Logical Disk

Function	Default	Comments
Enable Fast Write	No	When Fast Write is enabled, the response time for write operations is significantly reduced. It also has the benefit of increasing the performance for RAID 5 arrays for sequential write operations. Some processing overhead is required in the adapter for all operations when fast write is enabled and this reduces the maximum operations per second that the adapter can process.
Bypass Cache in 1 way Fast Write Network	No	If this is set to Yes and one adapter fails in a 2 adapter configuration, data in the fast write cache of the remaining adapter is destaged to disk and the fast write cache is bypassed until the other adapter becomes available. This option is provided so that operations can continue without risk of data loss due to a failure of the remaining adapter before a failed adapter has been replaced.

8 Comparison of Striping Data Options

The performance comparisons previously described have compared performance characteristics when availability of data is required when a disk or other failure occurs. If no availability of data is necessary when a disk fails, performance can be increased if data is striped across several disks in strips. This can be achieved by configuring disks as members of a RAID-0 array or by using the striping option of the Logical Volume Manager. This section compares the performance of these two mechanisms.

When striping data across several disks, performance is enhanced in two ways:

1. If operations are not evenly distributed across all disks, as is most likely, some disks will have more activity than others and the performance of the system may be limited by the performance of a few disks. If data is striped across disks, then all the disks will be subjected to approximately the same rate of receiving requests. Striping data across disks, therefore, has the effect of eliminating a skew of operations to a few disks and showing a benefit to throughput for short non-sequential operations. This benefit of eliminating skew by striping data is not shown in the following performance comparisons, or probably in a performance benchmark test, but may be the most significant performance benefit of striping data. This benefit is seen for short as well as long operations.
2. For long operations, multiple disks are used to fetch or store the data rather than using a single disk when data is not striped across disks. This has the benefit of increasing the data rate to transfer data for long operations.

The bandwidth comparison of RAID-0 and LVM striping is shown in Figure 29. The configuration for this comparison was 8 disks configured as either two RAID-0 arrays each having 4 member disks or two logical volumes each striped across 4 disks. In the case of LVM striping the disks of each logical volume were on both SSA loops, each disk on each loop being closest to a different adapter. In the case of RAID-0, all the disks in an array have to be on the same loop, so one array was located on one SSA loop with 2 member disks closest to one adapter port and the other two closest to the other port and the other array was located on the other loop. This locating of disks enabled the maximum performance for each type of configuration and is important if striping is used to increase bandwidth of long operations. If all the disks that have data striped across them are all located closest to the same adapter port, then the bandwidth will be limited by the bandwidth possible from one adapter port and the benefit of striping to achieve high bandwidth will be lost.

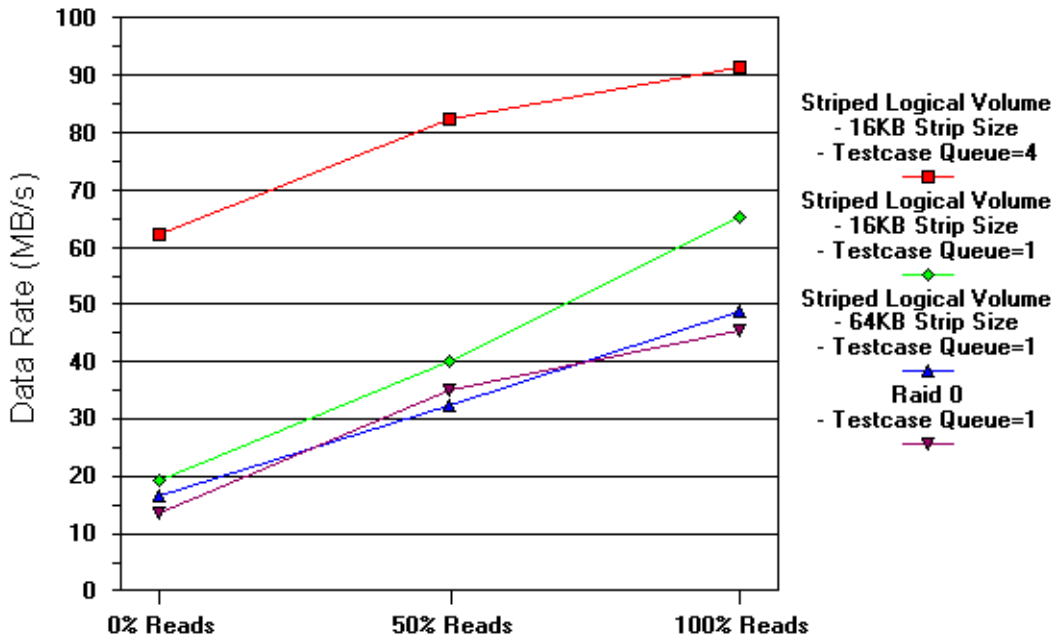

Figure 29: Bandwidth comparison of RAID-0 and LVM Striping (8 disks, 64 KB sequential operations)

In Figure 29 it can be seen that for 64 Kbyte sequential synchronous operations (queue depth = 1) the bandwidth provided by striping the data using LVM is higher than that achieved by using RAID-0 when the same strip size is used. For RAID-0 the only strip size supported is 16 Kbyte. The bandwidth for reads is much higher than for writes as the disk reads ahead into its own buffer when not responding to a request. For synchronous operations, a revolution of the disk is required for each write operation but for reads the data for the next operation is fetched from the read ahead buffer on the disk, so an extra revolution is not required.

For LVM striping the strip size can be increased to larger than 16 Kbyte. The bandwidth achieved with a 64 KByte strip size is shown in Figure 29 and it can be seen that this is less than when the strip size is 16 Kbyte. When the strip size is 16 Kbyte, data for each 64 KByte operation involves operations to 4 disks, but when the strip size is 64 Kbyte only one disk is involved. When using striping, the strip size should always be much smaller than the average length of the operations to achieve a performance benefit.

Figure 29 also shows the effect of increasing the queue depth of operations. The bandwidth is increased significantly if operations can be queued, particularly for write operations. When operations are queued, there is less idle time waiting for the next request and so the bandwidth is increased. For write operations this is important because consecutive write operations may be executed within the same revolution and that is not possible if there is a delay before the next request is issued to the disk.

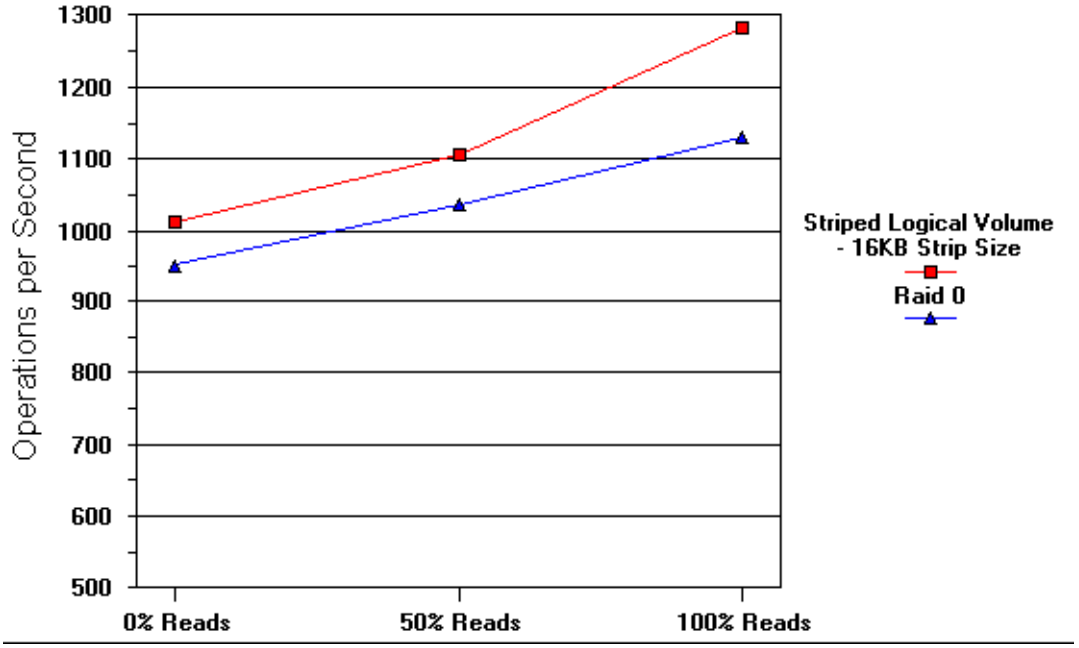

Figure 30: Throughput comparison of RAID-0 and LVM Striping (8 disks, 4 KB random operations)

Figure 30 shows the throughput comparison of operations for 8 disks configured as two RAID-0 arrays or as two logical volumes with data striped across 4 disks for each volume using LVM. It shows that there is a small benefit in throughput if LVM striping is used. There may however be a small increase in the processor utilization in the system when LVM striping is used, but this should be small.

If RAID-0 is used, there can be only a single adapter on the loop.

Striping data across disks is used by RAID-0+1 and RAID-5 and the benefits of striping apply to those configurations as well as for RAID-0 and LVM striping.

If availability of data is required when a disk fails as well as the performance benefit of data striping, the following configurations could be used:

1. RAID-0+1
2. LVM striping and LVM mirroring
3. RAID-5

The benefits of striping data across several disks is that for short operations disk accesses are more evenly distributed across all the disks and hot spots can be avoided. The data rate possible is increased for long operations as data is accessed from several disks in parallel and not from a single disk. When comparing striping using LVM or by using RAID-0, LVM provides more flexibility and some performance advantages.

9 Split Copy Options

There are two alternative ways that data can be copied:

1. When using Logical Volume Manager, logical volumes can be copied, split using Split Copy and later accessed separately from the original volume. This can be performed on any logical volume under LVM.
2. For RAID-1 and RAID-0+1 SSA disks, whole volume groups or physical volumes can be copied to a separate disk or disks, split and later accessed separately from the original volume using 3-Way Copy. This can only be used on RAID-1 or RAID-0+1 disks. It can be used if these disks are not managed by LVM.

Each method has its advantages and limitations that should be considered when deciding which of the methods is best for your applications. The advantages of each method are as follows:

Split Copy	3-Way Copy
Copies logical volumes. Copying entire volume groups requires copies of all the appropriate logical volumes in the volume group	Copies whole volume groups or physical volumes
Can only be used to copy individual LVM logical volumes.	Can be used to copy any volume group or physical volume whether or not it is configured for LVM
Copies any logical volume configured for LVM (e.g. non-RAID, RAID-0, RAID-1, RAID-0+1, RAID-5, LVM mirrored pair)	Only copies RAID-1 or RAID-0+1 volume groups or physical volumes.
Data is copied and synchronized to the new disk(s) at a rate of 7 MBytes/sec although this depends on the type of disks being used. This means synchronizing and copying an 18 GByte volume takes about 43 minutes when Splitlvcopy is used. If Chlvcopy is used, only the data changed since the last synchronization has to be copied to the copy disk when synchnonizing and this takes less time than Splitlvcopy. The copy must however be for the same volumes as were used for the previous copy for this reduction in synchronizing time. The copy disks cannot be used for copying different volumes on each synchronization of the copy and still require only the changes to be copied	Data is copied and synchronized to the new disk(s) at a rate of 16 MBytes/sec for RAID-1 and 29 MBytes/sec for RAID-0+1although these depend on the type of disks being used. This means synchronizing and copying an 18 GByte volume takes about 19 minutes for RAID-1 and 10.5 minutes for RAID-0+1. When copying and synchronizing there is less system interaction than for split copy, so performance may be better on heavily loaded systems.
Copying a volume group that consists of 200 logical volumes requires 600 steps as 3 steps are required for each logical volume (add, synchronize and split)	Copying a volume group requires only 2 steps
If only a single or a few logical volumes are to be copied, then this is a short operation	The smallest entity that can be copied is the physical volume. If this contains more logical volumes than are required to be copied, then logical volumes may be copied that are not needed.

Split Copy	3-Way Copy
The copied disks are under control of the system that initiated the copy.	The copied disks can be accessed from another system to the system from which the copy was made without affecting that system.

Copying a whole volume group or physical volume is much faster if 3-Way Copy can be used and the other benefits may desirable. It is, however, limited to only copying RAID-1 or RAID-0+1 disks.

10 Getting More Information About SSA

The following SSA manuals are available:

SA33-3285	Advanced SerialRAID Plus adapter; User's Guide and Maintenance Information
SA33-3287	Advanced SerialRAID Plus adapter; Installation Guide
SA33-3286	Advanced SerialRAID Plus adapter; Technical Reference
SA33-3278	7133 Models D40 and T40; Serial Disk Systems; Operator Guide
SA33-3281	7133 Models D40 and T40; Serial Disk Systems; Hardware Technical Information
GY33-0192	7133 Models D40 and T40; Serial Disk Systems; Service Guide
GA33-3279	7133 Models D40; Installation Guide
GA33-3280	7133 Models T40; Installation Guide

There is online information available on the AIX device drivers and subsystem interface within AIX. Refer to the AIX 4 information pages and search for SSA. This will bring up sections describing the SSA device drivers, their IOCTL interfaces and how they are used within an AIX system.

The IBM Hursley SSA Online Customer Support page is at:

http://www.storage.ibm.com/hardsoft/products/ssa

This contains adapter and disk drive microcode download packages as well as technical data and product descriptions for IBM SSA products developed at Hursley.

The Open Systems Disk Home Page is at:

http://www.storage.ibm.com/hardsoft/diskdrud.htm

The Open Systems Disk Storage Products page is at:

http://www.storage.ibm.com/hardsoft/tech/tech.htm

This contains technical information on SSA products.

The SSA Redbooks page relating to storage is at:

http://w3.austin.ibm.com/:/projects/itso/itso_web/areas/ST.html

Appendix A: Notices

References in this book to IBM products, programs, or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Subject to IBM's valid intellectual property or other legally protectable rights, any functionally equivalent product, program, or service may be used instead of the IBM product, program, or service. The evaluation and verification of operation in conjunction with other products, except those expressly designated by IBM, are the responsibility of the user.

Any examples of parameters or definitions are for guidance only. Some details may differ from the requirements in your environment. Contact your IBM representative if you need further assistance.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license enquiries, in writing, to the IBM Director of Licensing, IBM Corporation, 500 Columbus Avenue, Thornwood, New York 10594, U.S.A.

Trademarks

The following terms are trademarks of International Business Machines Corporation in the United States, or other countries, or both:
IBM
RS/6000
Micro Channel

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