Configuring RAID for Optimal Performance

- Intel® RAID Controller SRCSASJV
- Intel® RAID Controller SRCSASRB
- Intel® RAID Controller SRCSASBB8I
- Intel® RAID Controller SRCSASLS4I
- Intel® RAID Controller SRCSATAWB
- Intel® RAID Controller SRCSAS18E
- Intel® RAID Controller SRCSAS144E
- Intel® Server System SR2500ALLX (Intel® Integrated RAID)
- Intel® Server System SR1550ALSAS (Intel® Integrated RAID)
- Intel® Server Board S5000PSLROMB (Intel® Integrated RAID)
- Intel® Server System S7000FC4UR (Intel® Integrated RAID)

Revision 1.1

September, 2008

Enterprise Platforms and Services Division - Marketing

Revision History

Date	Revision Number	Modifications
April, 2008	1.0	Initial revision
September, 2008	1.1	Minor corrections

Disclaimers

Information in this document is provided in connection with Intel® products. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted by this document. Except as provided in Intel's Terms and Conditions of Sale for such products, Intel assumes no liability whatsoever, and Intel disclaims any express or implied warranty, relating to sale and/or use of Intel products including liability or warranties relating to fitness for a particular purpose, merchantability, or infringement of any patent, copyright or other intellectual property right. Intel products are not intended for use in medical, life saving, or life sustaining applications. Intel may make changes to specifications and product descriptions at any time, without notice.

Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them.

Intel® RAID Controllers may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel, Pentium, Itanium, and Xeon are trademarks or registered trademarks of Intel Corporation.

*Other brands and names may be claimed as the property of others.

Table of Contents

1. Overview

The target audience for this guide includes users, technical support personnel, and pre-sales personnel who work with Intel® Server RAID products. It is assumed that the reader has a basic understanding of hard-drive operation, RAID operation, and RAID levels.

This guide provides information to help achieve optimal performance of a RAID array depending on the RAID level and performance measurement tool used. It does not include techniques for fine tuning RAID performance for a specific application.

This guide covers Intel® RAID Controllers and Intel® Integrated RAID only. This guide does not cover Intel® Embedded Server RAID Technology II.

2. Performance Measurement Tools

When measuring RAID performance it is important to understand how the measurement tool works, the capabilities of the tool, and its limitations. It is better to avoid using a tool with unknown or unclear data access patterns, which can make the results difficult to interpret.

2.1 Copying large files

Copying files with the Microsoft Windows* and Linux* standard OS utilities generates primarily sequential stream of data with 64 KB blocks and no queuing. Read Ahead (Always or Adaptive) and Write Back modes must be used to achieve good performance with this type of operations.

2.2 Tools without queuing

There are many tools available for measuring storage performance, but many of them do not have queuing capabilities and they can measure sequential throughput only.

Without queuing, Read Ahead and Write Back caching modes must be used to achieve good throughput. Tools without queuing should not be used for measuring the maximum throughput capabilities of a RAID controller because the RAID cache causes a throughput bottleneck. This is especially important for read operations.

Some tools allow the access block size to be changed. If a large block size can be selected, this can partially mitigate the lack of queuing. For example, sequential throughput with 2 MB block size and no queuing may be similar to performance with 64 KB block size and a queue depth of 32 IOs.

2.3 IOmeter*

IOmeter* (www.iometer.org) is a sophisticated tool that can measure RAID performance, including sequential, random and mixed workloads, adjustable block sizes, and queuing. This tool requires certain level of proficiency to use it.

Unfortunately, queuing does not work under Linux with the current IOmeter version 2006.07.27.

One of the common mistakes made with IOmeter is setting the access type to mixed (random + sequential). This selection will have big impact on performance measured in MB/s.

For achieving maximum MB/s numbers, 100% sequential (0% random) access pattern must be used with a large enough block size (between 64KB and1MB). Also the number of Outstanding I/Os (queue depth) typically should be in the 8-128 range.

The following formula provides guidance for selecting optimal outstanding I/Os and block size values:

> *Outstanding_IOs x Block_Size = 2 x Strip_Size x Number_of_HDDs_in_Stripe*
>
> The number of HDDs in stripe does not include parity HDDs. For example, RAID5 array with 9 HDDs will have 8 HDDs in the stripe.

2.4 Measuring Performance under Linux

The Linux* OS has a substantially different I/O queuing model (asynchronous I/O) than the Microsoft Windows* OS. We are not aware of any performance measurement tools that support I/O queuing for read operations under Linux. When measuring sequential read throughput under Linux, it is important to enable Read Ahead mode (either Adaptive or Always).

3. Optimal RAID Settings

The following table provides a simplified quick reference for RAID settings for achieving optimal performance depending on the type of application or test. Please refer to the following sections for more details on the meaning and impact of each setting.

	Maximum Sequential Throughput		Recommended settings for most real-world applications with sequential + random storage access
	Under Linux* or Microsoft Windows* without I/O queuing (Including copying files and using simple tools like HDSpeed*)	IOmeter* under Microsoft Windows*	
Strip Size	512K	512K	256K
Read Cache Policy	Direct I/O	Direct I/O	Direct I/O
Read Ahead Policy	Adaptive Read Ahead	No Read Ahead	Adaptive Read Ahead
Write Cache Policy	Write Back	RAID 0/10: Write Thru RAID 1/5/6/50/60: Write Back	Write Back
Disk Cache Policy**	Enabled	Enabled	Disabled
Virtual Drive Initialization	Full Initialization Completed	Full Initialization Completed	Full Initialization Completed
Consistency Check	Disabled	Disabled	Scheduled
Patrol Read	Disabled	Disabled	Scheduled

***Note**: Enabling disk cache in Write-Back mode provides little or no performance improvement, while the risk of data loss due to power failure increases. See Section 4.2 for more information.*

4. Impact of RAID Settings on Performance

4.1 Write Policy

The Write Policy can have a very big impact on write performance. There are two modes available – Write Back and Write Thru.

4.1.1 Write Back Mode

This mode provides better performance in most cases. In Write-Back mode, the RAID controller acknowledges write I/O requests immediately after the data loads into the controller cache. The application can continue working without waiting for the data to be physically written to the hard drives.

If a power loss occurs in write-back mode, there is a risk of losing data in the RAID cache. The data loss may be fatal and may require restoring data from a backup device. It is critical to have protection against power failures. Using a UPS with redundant system power supplies is highly recommended. RAID Backup Battery Unit can provide additional protection.

4.1.2 Write Thru Mode

This mode does not utilize the RAID cache for accelerating write I/O requests. In most cases it will be slower than Write-Back mode. However, Write Thru mode allows achieving the highest sequential write bandwidth with RAID 0 or RAID 10.

4.2 Disk Cache Policy

Disk Cache Policy determines whether the hard-drive write cache is enabled or disabled. When Write Policy is set to Write Thru mode, Disk Cache Policy can have very big impact on write performance. When Write Policy is set to Write Back mode, impact of Disk Cache Policy is much smaller and in many cases negligible.

When Disk Cache Policy is enabled, there is a risk of losing data in the hard drive cache if a power failure occurs. The data loss may be fatal and may require restoring the data from a backup device. It is critical to have protection against power failures. Using a UPS with redundant system power supplies is highly recommended. RAID Backup Battery Unit can provide additional protection.

***Note**: A RAID Backup Battery Unit does not protect the hard drive cache.*

4.3 Read Ahead Policy

The Read Ahead Policy determines whether the RAID controller will read just a block of data that an application has requested, or whether it will read the whole stripe from the hard-drives. This setting can have big impact on read performance.

4.3.1 No Read Ahead (Normal)

The RAID controller will read only the block of data that the application has requested. This mode is preferred when read requests are primarily random. Also this mode is recommended when measuring sequential read throughput with IOmeter* under Windows.

4.3.2 Always Read Ahead

The RAID controller will read the whole stripe containing the requested data block and will keep it in cache. Each read operation will consume more hard drive resources, but if the read requests are primarily sequential it can substantially reduce the amount of read requests to the hard drives and can substantially increase performance.

Note*: This setting will only make difference if the typical read request size is smaller than the stripe width.*

4.3.3 Adaptive Read Ahead

The RAID controller automatically adjusts the read policy based on the current pattern of read requests. It combines the benefits of No Read Ahead and Always Read Ahead modes. This mode is recommended if the workload has mixed sequential and random patterns, or if the pattern is unknown.

4.4 I/O Policy

The I/O Policy determines whether the RAID controller will keep data in the cache, which can reduce the access time if subsequent read requests are made to the same data blocks.

4.4.1 Direct I/O

Direct IO mode is recommended in most cases. Most file systems and many applications have their own cache and do not require caching data at the RAID controller level.

4.4.2 Cached I/O

In Cached I/O mode the controller caches both read and write requests. If there are subsequent read requests to the same data blocks, they are read from the RAID cache instead of the hard drives. This mode may be required if the application or file system does not cache read requests.

4.5 Strip Size

Strip size determines how data is distributed across hard drives. It also determines how many drives are accessed to service a single I/O request. Strip size can have big impact on performance. Typically, sequential workloads benefit from using large strip sizes (512 KB or 1 MB).

With random types of access, the strip size depends on the typical access block size and on data alignment. For example, if a database is using 16 KB records with 16 KB alignment, the optimal strip size can be 16KB. For file- or web-server a large (512 KB or 1 MB) strip size can be optimal. Software vendor documentation often provides recommendations on how to select RAID strip sizes.

***Note**: Matching the strip size to the file system cluster size does not usually provide any benefit. Data block or file sizes used by the application are usually more important. However, setting the strip size smaller than the cluster size is not recommended.*

5. Other Performance Factors

When measuring performance of a RAID subsystem, it is important to remove factors that can limit the performance or cause variations in the performance.

5.1 Backup Battery Status

When doing write performance measurements in Write Back mode, it is important to check the status of the battery and the Current Write Policy. When the battery is not fully charged or is in the process of relearning, Write Policy will be automatically switched to Write Thru. This will have big impact on write performance.

You can disable the *Write Thru for a failed or missing battery* option to make sure that the Write Back mode is used regardless of the current battery status. In a production environment disabling this option may cause data loss if power failure occurs when the battery does not have sufficient level of charge.

5.2 Virtual Drive Initialization

For RAID 5/6/50/60 volumes it is important to perform Full Initialization of the volume before doing performance measurements. On large virtual disks full initialization may take many hours to complete. For performance measurements you can create smaller virtual disks.

For RAID 0/1/10 there is no need to perform Full Initialization before measuring performance. However, it is important to disable Background Initialization (BGI) during measurements if Full Initialization was not performed. Unless it is disabled, BGI will start automatically on RAID 1/10 volumes and will substantially reduce the RAID performance.

5.3 Patrol Read

Patrol Read helps to detect and reallocate bad blocks on hard drives and to prevent possible data loss. Patrol Read generates substantial number of disk requests that may reduce the RAID performance.

You should disable or enable Patrol Read depending on the purpose of your performance measurements. You can also adjust the Patrol Read rate to reduce or to increase the priority of the requests. Patrol Read settings can be changed in Adapter Properties in RAID BIOS Console or in RAID Web Console.

5.4 Consistency Check

Consistency Check is an important function that helps to detect inconsistencies in data stored across hard drives in redundant RAID arrays and to identify possible sources of silent data corruption.

Consistency Check generates substantial number of disk requests that may reduce the RAID performance. You should disable or enable Consistency Check depending on the purpose of your performance measurements. You can also adjust the Consistency Check rate to reduce or increase the priority of the Consistency Check requests. Consistency Check settings can be changed in Adapter Properties in RAID BIOS Console or in RAID Web Console.

5.5 Data Location on Physical Drives

When measuring sequential throughput, it is important to remember that the speed of the hard drives depends on the location of the data.

The maximum sequential speed is achieved when reading or writing data in the beginning of the hard-drive. When data is read or written to the end of the hard-drive, the sequential speed will be 30%-40% lower.

If only a part of the hard drive capacity is occupied by the virtual disk or by the partition used for the performance measurements, the random performance will be higher than if the whole hard drive capacity was used. This happens because the hard drive heads make shorter movements. It is important to remember about this factor affecting random performance when using several partitions or several virtual disks in one array.

5.6 PCI Express* Slot

In configurations with many hard drives, the sequential throughput of the RAID may exceed the bandwidth of the PCI Express* link. When using a RAID controller with a PCI Express x8 interface, make sure it is installed into a PCI Express* x8 (or higher) slot.

Typical bandwidth of a PCI Express Gen1 link is approximately 800 MB/s for a x4 interface, and 1600 MB/s for a x8 interface. In some systems it may be up to 10% lower depending on chipset capabilities and BIOS settings.

5.7 Vibration

Vibration in the chassis can substantially impact hard drive performance. Vibration can be caused by chassis fans or by other hard drives. Vibration is usually higher in rack-mount systems, because smaller and higher speed fans are used. SATA drives are typically more sensitive to vibration than SAS drives. Different models and even different sizes of the same hard drive family can have substantially different sensitivity to vibration.

5.8 Overheating

Overheating can impact hard drive performance. When measuring performance, especially in a lab environment, make sure that the chassis lid is closed and the air flow is sufficient to cool the hard drives.