

From SCSI to SATA

A Viable, Affordable Alternative for Server and Network Attached Storage Environments

The question “should I switch to SATA from SCSI?” is a broad topic which is best understood in terms of reliability, performance, behavior in a RAID system, and cost.

Reliability

The SATA, SCSI or Fibre Channel (FC) interfaces are comprised of silicon, and the reliability of silicon parts are significantly higher than the reliability of moving parts: the motors, bearings, actuators, heads, platters, and the supporting cooling, filtering, lubricant, enclosure, and quality control (also known as “the platform”). Hard drive reliability is not determined by the interface—it is determined by the reliability of the moving parts that comprise the platform. Any interface (SATA, SCSI or FC) can be offered on a highly reliable hard drive platform. Western Digital’s WD Raptor SATA hard drive rated at 1.2 million hours MTBF (Mean Time Between Failures) and backed with a 5-year warranty.

The market demands affordable, high performance, high capacity, highly reliable server-class hard drives with SATA interfaces. Otherwise stated, the market is demanding a viable, affordable alternative to SCSI in server and network attached storage environments. And these server-class hard drive products with SATA interfaces are shipping now. See for yourself, take a look at the datasheets and compare duty cycle and MTBF ratings, and enterprise features such as time-limited error recovery (TLER) to evaluate if a drive carries enterprise-class reliability.

Performance

Hard drive performance comes from several characteristics: interface speed, rotation speed, and queuing.

Interface Speed

The connection between the computer and the hard drive is commonly referred to as “the interface” (e.g. SCSI, SATA). The SCSI interface is either SCSI Ultra 160 (160 MB/s per channel) or SCSI u320 (320 MB/s per channel). SCSI is a shared bus architecture. In other words, SCSI servers place multiple hard drives on a single SCSI channel. Typical SCSI hard drives (10,000 RPM) can produce a maximum sustained data rate of around 75 MB/s. This data rate is restricted by the speed the data can be transferred from the platters to the heads. Depending on the number of drives on the SCSI bus and the simultaneous workload of all the drives sharing it, there comes a point where the hard drives deliver more data than the SCSI bus can handle and

the SCSI bus becomes a data bottleneck. The normal way to solve this problem is to use more busses with fewer drives per SCSI bus. This works fine, but can be expensive.

Serial ATA is a point-to-point (not a bus) architecture capable of delivering 150 MB/s to each hard drive. Typical 10,000 RPM disks produce 70-75 MB/s and typical 15,000 RPM drives produce 100-105 MB/s, so there is no production hard drive available that will produce more data than the SATA interface can handle.

Therefore, SATA as a server-to-hard drive interface has a performance advantage over SCSI.

Rotation Speed

10,000 RPM rotation speeds are typical for enterprise applications where performance is important. 15,000 RPM hard drives are available today, but the combination of price, performance, and capacity limit the popularity of these drives. At the time of this writing, 15,000 RPM drives account for a very small percentage of the total hard drives for enterprise servers and external storage. 7200 RPM speeds may provide adequate performance for many enterprise applications. And indeed, 7200 RPM drives have been in use successfully for some years in servers. In these cases, performance delivered by 7200 RPM is “good enough” for the application at hand.

Queuing

Random I/O is encountered in servers and networked storage where multiple users simultaneously request random reads or writes. But for a single user desktop system, high levels of random I/O are not typical. Performance of a random I/O workload can be improved through intelligent re-ordering of the I/O requests so they read/write to and from the nearest available sectors and minimize the need for additional disk revolutions or head actuator movement.

SCSI tagged queuing has been available for some years now, and SATA offers the equivalent capability: SATA tagged command queuing (TCQ).

TCQ is an industry standard defined in the T-13 ATA4 spec in the late 1990s. TCQ is well defined, well understood, and thoroughly debugged—www.storagereview.com has published performance benchmarks that show TCQ performance in random I/O workloads is equivalent to SCSI performance.

Figure 1 is representative of the SATA-to-SCSI benchmarking for random I/O workloads, in this case a Web server.

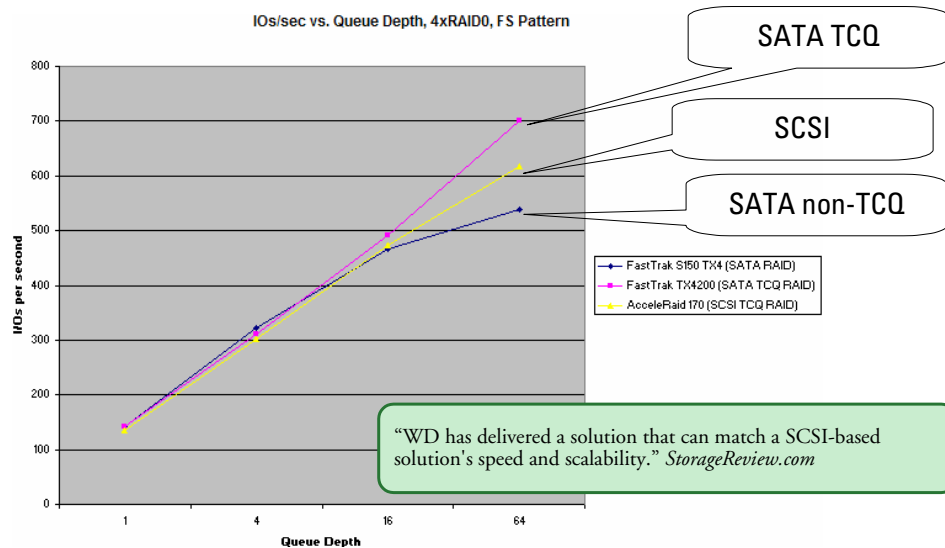


Figure 1. Independent Performance Testing Storage Review - Fileserver 4xRAID0

More Info on Performance and Tagged Command Queuing (TCQ)

- support.wdc.com - SATA Tagged Command Queuing info sheet.
- www.storagereview.com/articles/200406/20040625TCQ_1.html the performance benchmark between SATA and SCSI.
- www.esata.com – A listing of TCQ-capable RAID and motherboard products shipping today.

SATA Native Command Queuing

Most manufacturers of SATA hard drives, SATA controllers, and SATA RAID adapters are working on a technology to solve the SATA Native Command Queuing (NCQ) problem. This is expected to ship later in 2004, perhaps 2005. At the time of this writing, NCQ has not yet shipped, so a precise assessment on NCQ performance is not possible now. However, NCQ performance is expected to be roughly equal to TCQ. An NCQ white paper is available at www.serialata.org.

Performance in Server Environments with Vibration

Several years ago, hard drives had lower capacity and lower spin speeds, so rotational vibration was not a problem. Now drives have much higher capacities, narrower tracks, and faster spin speeds. The vibration generated by one hard drive can impact the tracking of an adjacent drive and cause read/write misses and retries. This problem can be overcome with technologies that sense vibration and correct the drive tracking. WD has a technology available on selected drives called Rotational Accelerometer Feed Forward (RAFF). RAFF senses vibration and actually corrects the head tracking. The result is performance degradation in server environments with vibration considerably reduced. Figure 2 illustrates the impact of rotational vibration on WD Raptor 74 GB and three other SCSI 10,000 RPM drives.

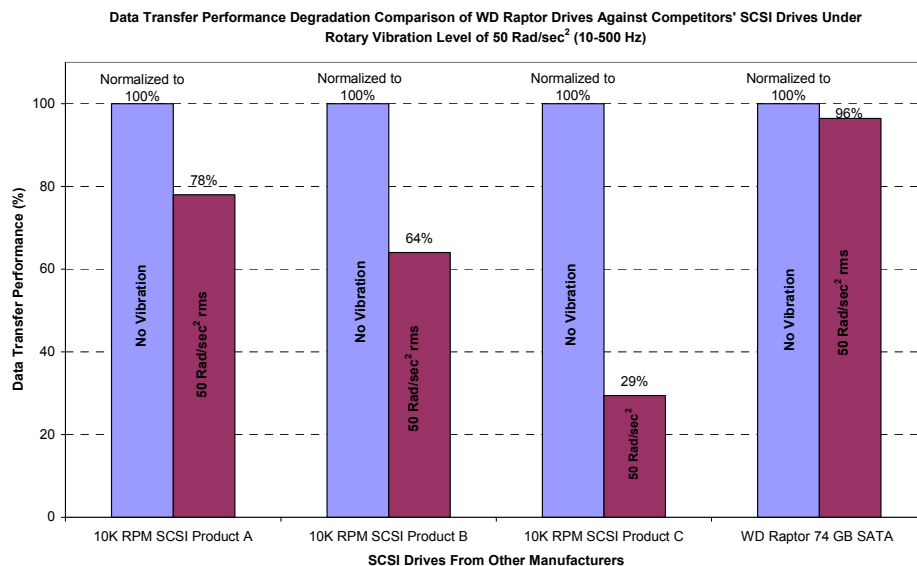


Figure 2. Data Transfer Performance Degradation Comparison of WD Raptor vs. Competitors' SCSI Drives

A RAFF info sheet is available at support.wdc.com.

Behavior in RAID system

Time-Limited Error Recovery (TLER)

Without TLER, desktop drives can, and do, get into a prolonged error recovery phase, and the drive may take time to reallocate around bad blocks or similar prolonged error recovery activity. During that time, the drive may not respond to the RAID adapter. RAID adapters will pass a threshold (typically 8 seconds) and if the drive does not respond, the RAID adapter will drop the drive from the RAID array and run in degraded mode. After the drive is replaced, a RAID recovery will commence. Now, with 200+ GB drives, the RAID recovery can take several hours to several days. If another drive fails (or engages in another prolonged error recovery) during that time, the data on the entire RAID volume is lost. This is like driving on a long trip without a spare tire. Risk is increased because of the long RAID rebuild times associated with very large hard drives, and this risk is further magnified without TLER.

The benefit of TLER is the substantially reduced risk of a second failure that will result in the loss of all data on the RAID volume. *With* TLER, the drive will send an error to the RAID controller before the RAID timeout period expires. RAID adapters then respond to an error message from the hard drive. The hard drive is allowed to proceed without the drive being dropped from the RAID array. The RAID array operates in degraded mode for a short time, then the drives re-synch from the RAID journal. The extended performance degradation and RAID rebuild (the typical RAID error recovery process) are thus avoided.

Hot Plug

The Serial ATA standard is designed for hot pluggability. The connector/receptacle is designed with alignment ears and a 3-stage connection scheme. So, the interoperability standard for hot pluggability is in place and many products make good use of the standard and are hot pluggable. All WD SATA drives are hot pluggable.

LED

In RAID arrays, LED activity/status indicator is an important feature. If there is a failure, LEDs make it obvious which drive has failed. Without LEDs, there is some guesswork about which drive to replace, and pulling the wrong drive usually results in losing all data on the RAID volume. SATA drives can engage an activity LED through two methods: via pin-11 on the hard drive power connector, or via activity LED capability in many RAID controllers.

Thermal Sensor

Heat is the enemy of hard drive reliability. Enclosure thermal sensors are necessary to provide warning in time to take steps to avoid a failure. WD Raptor drives go one step further by providing a thermal sensor on each hard drive; the sensor data is available through the S.M.A.R.T. interface.

RAID Rebuild Times and RAID 5 vs. RAID 0/1

In the case of a drive failure in a RAID 5 volume, the array will operate in degraded mode. In other words, all the reads and writes will include the overhead and performance impact of a parity calculation until the failed drive is replaced. Once the failed drive is replaced, the RAID 5 rebuild stage will commence. During that stage, the RAID adapter uses the parity protection data to re-populate the replaced hard drive with user data and parity protection data. With very large hard drives, these RAID 5 rebuilds can take several hours to several days. SATA RAID 1 (mirroring) offers an affordable, viable alternative. A SCSI RAID 5 (including adapters and drives) costs approximately the same as a SATA RAID 1 mirroring. After a drive failure in a mirrored system, the system continues to run at the same performance level. When a failed drive is replaced in a mirrored system, the rebuild happens very rapidly and without excessive burden to the I/O bus or I/O processor. In fact, the principal benefit of an I/O processor is for RAID 5 parity processing. There is now an option to buy RAID 0,1 RAID cards without an I/O processor (often referred to as hardware RAID 5). So, why not take advantage of lower priced SATA hard drives to take advantage of mirroring, which in every way is simpler and better than RAID 5? All hard drives are electromechanical devices, and they are remarkably reliable; however they are still subject to failure. Why not plan for the optimal failure recovery scenario with mirroring?

Figure 3 illustrates the differences in RAID 5 recovery and RAID 1 (mirrored) recovery.

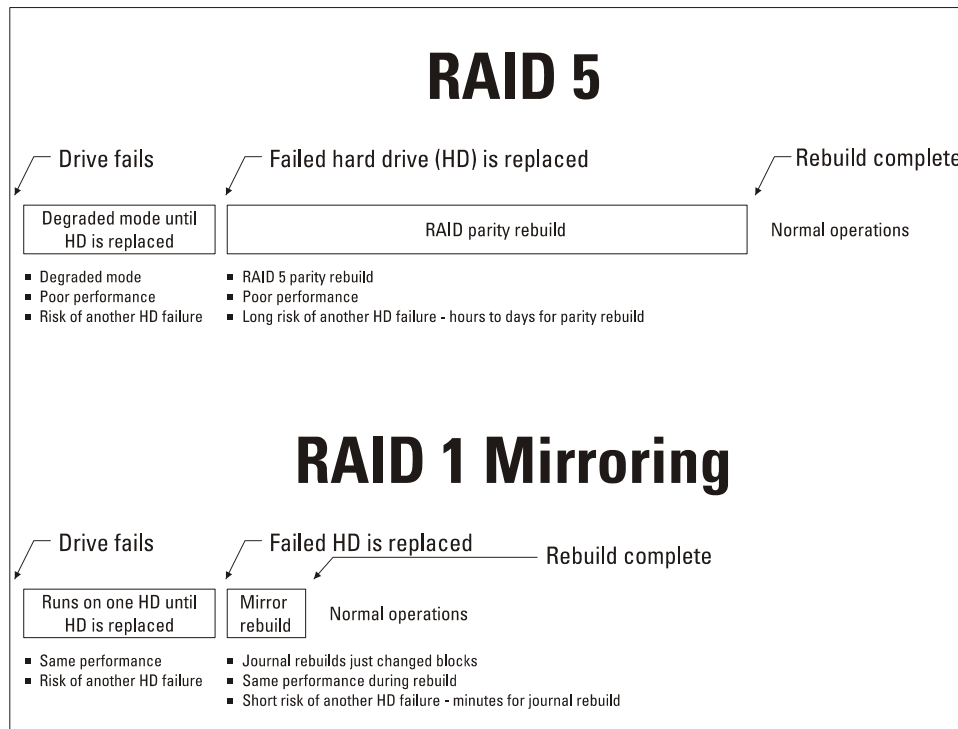


Figure 3. RAID 5 vs. RAID 1 (Mirrored) Recovery

Scalability

Just a bunch of disks (JBOD)—is an enclosure with drives, traditionally SCSI hard drives, a SCSI backplane, and a SCSI I/O controller that provides a connection to rack servers. SATA offers similar types of products either using either SCSI-to-SATA controllers, FC-to-SATA controllers or SATA port multiplier (please refer to www.sata.org, use case diagrams). These mechanisms allow storage expansion and scalability at a very affordable price compared to SCSI-based storage.

Cost

WD Raptor SATA hard drives for the enterprise are available and priced significantly below equivalent SCSI drives. These SATA drives offer 10,000 RPM performance, best-in-class benchmarks, and excellent reliability.

Recently, an independent authority published an in-depth performance comparison between SATA enterprise to SCSI. Please refer to www.storagereview.com/articles/200406/20040625TCQ_1.html for the complete review. The findings of the *StorageReview.com* article include:

- “SATA command queuing and SATA RAID have the potential to deliver benefits to the server market just as great as that of SCSI command queuing and SCSI RAID.”
- “The cost difference between the two arrays works out to 40 percent, significant indeed.”
- “SATA’s performance is competitive with, and in some cases exceeds, that of a comparable SCSI solution.”

Additionally, Serial ATA RAID adapters are considerably less expensive than equivalent SCSI RAID adapters, often half the price for equal performance and features. Furthermore, Serial ATA RAID on motherboard solutions are substantially less expensive than equal SCSI on motherboard solutions.

Applications Best Suited for SATA

In terms of enterprise reliability, availability, and performance, a key consideration is business impact of systems decisions. When comparing hard drive products across the board, one observes the tradeoff between performance and capacity. High capacity drives are available at slower spin speeds:

- 7200 RPM drives – typical 160 GB to maximum 400 GB
- 10,000 RPM drives – typical 74 GB to maximum 300 GB
- 15,000 RPM drives – typical 36 GB to maximum 74 GB

So, this begs the question: “which is more important, performance or capacity?” Otherwise stated, will your business be impacted more by running out of space or by the performance difference between 7200 and 10,000 RPM or between 10,000 and 15,000 RPM? As you ask yourself these questions the answer will likely be: “depends on the application, the size of the dataset, and user demand on the system.”

Example Applications for Serial ATA

Non-Random<----->Random I/O (queuing)

Video Surveillance	General File Servers	Database Servers
Video streaming	Medical records	E-mail
Nearline storage	Insurance image data mgmt.	e-Commerce
Scientific/Engineering	Engineering data mgmt.	Web server

Conclusion

Why switch from SCSI to SATA?

- **Reliability** – SATA hard drives are available today with reliability that is equivalent to SCSI in high duty cycle environments. It is important to distinguish between desktop SATA products and enterprise SATA products. RAID 1 (mirroring) is a considerable improvement on RAID 5.
- **Performance** – Serial ATA 10,000 RPM queuing-capable hard drives have demonstrated equivalent performance to SCSI in server workloads.
- **Performance in a RAID environment** – TCQ provides performance which equals SCSI performance in random I/O benchmarks. TLER serves to allow RAID adapters and RAID on motherboard to properly handle error conditions. RAFF vibration compensation serves to correct head tracking and keep I/O performance high in rotary vibration environments commonly found in servers and networked storage systems. And with SATA affordability, RAID 1 (mirroring) becomes a much more attractive option than SCSI RAID 5.
- **Cost** – SATA hard drives, SATA RAID adapters, and SATA on motherboard offer a compelling alternative to SCSI because SATA offers the same reliability, performance, and features, but at a considerably lower price.

And now to the market realities: every major server/storage system vendor including IBM, HP, Dell, EMC, Network Appliance, and others offer SATA-based storage. These vendors place high value on customer satisfaction, cost of return, reputations, and public perception. Would they offer products with reliability risk? Driven by an underserved need—affordable, high performance, high capacity storage for servers—the adoption of SATA in the enterprise has already happened.

My suggestion to the readers is simple: you be the judge! Discuss this with your team, suppliers, and your professional associates. Pick an application and evaluate SATA. Reconsider your preference for RAID 5 or RAID 1. You be the judge of the reliability, performance, and price.

For additional enterprise storage systems employing SATA drives, please visit www.esata.com.

For service and literature: