

Limitations of Conventional RAID5 on the ATA Platform and the Promise of ATA in the Enterprise

Boon Storage Technologies, Inc.

<http://www.sr5tech.com>

Backdrop

The emergence of ATA drives as a serious alternative to enterprise storage holds the promise of significantly reducing storage acquisition costs. This is amplified by the advent of Serial ATA, which brings features like hot-pluggability, CRC for all communications (including data, commands and status), and thin flexible cabling to further decrease the gap between ATA and more expensive “server” class drive. However, in order to fully realize the advantages of the ATA platform for enterprise storage, new software technologies are required to guarantee the reliability and maximize the performance of the platform.

Specifically, RAID technologies currently used with SCSI and Fibre Channel storage implementations are ill-suited for use in the ATA arena. The pervasive use of write-back caching and the high cost of NVRAM-based board solutions negatively impacts the reliability and price advantages of the ATA platform, introducing the possibility of corruption and data loss and negating much of the cost benefit for the enterprise user. Similarly, the clear attractiveness of RAID Level 5 for large capacity storage is all but eliminated because existing methods for implementing low-cost RAID 5 systems have severe limitations in performance or reliability. On the ATA platform, this results in the undesirable flight to RAID 10 for most types of workloads and directly reduces the cost benefit of the platform.

In order for ATA-based storage to achieve its full potential in the enterprise, it is necessary to understand the limitations of today’s hardware-assisted RAID solutions as these attempt, imperfectly, to address the unique characteristics of the ATA drive platform. Particular attention is placed on

RAID Level 5, which is the most promising RAID type given its natural application to the larger capacity storage applications that will dominate networked ATA adoption.

ATA Characteristics

ATA disk drives emerged in the late 1980s as desktop computers began their ascent into the mainstream of the IT universe. ATA is an acronym which stands for “AT Attachment,” a reference to the IBM PC/AT that served as the de facto reference specification for the desktop since its introduction in the early 1980s. Though synonymous with IDE (Integrated Drive Electronics), the ATA designation is the subject of various ANSI specifications that have evolved the platform over time and is generic to the category.

Since their initial shipments in 1986, ATA drives have grown substantially in volume. Today, ATA drive shipments outnumber SCSI drive shipments by a factor of 6 to 1. And they outnumber Fibre Channel drive shipments by a factor of 10 to 1. Their volume differences are accounted for by the continuing centrality of ATA’s role in the highest volume segment of the PC universe, the desktop computer. Because of their substantial volume advantages, they are subject to far more significant price competition than higher end drive platforms, and on average cost between 3 and 4 times less than SCSI or FC drives. The result has been an increased desire by IT end users to employ ATA drives in enterprise data settings as opposed to using them exclusively in desktop PC devices and workstations.

As engineered products, ATA magnetic disk drives harness the same basic technologies found in higher-end drives that employed different interfaces, most common of which are SCSI and Fibre Channel drives. They employ platters, actuators and a variety of micromotors. As such, ATA drives take advantage of the rapid advances in these component technologies that all disk drive manufacturers are continuously exploiting. Ranging from greater volumetric densities to enhancements in seek performance, ATA drives leverage that same basic technologies as SCSI and FC drives.

However, ATA drives do have significant differences from higher end drive platforms, and these differences must be addressed if the ATA platform is to emerge as a enterprise class storage platform.

The first major difference is that ATA drives are subject to different sorting criteria than higher end platforms. Quality control is relaxed because of the relative tradeoff in profitability and defect rates. Instead of 1% component rejection levels as seen in SCSI drives, ATA drives are typically subject to a less demanding 5% rejection rate. The other differences between ATA and SCSI flow from their different end use targets. Because they are intended for desktop computers, ATA drives use different motors that generate less heat and ambient noise than SCSI. They are also slower than their SCSI counterparts from an RPM basis, given similar design goals to minimize desktop heat and noise but also to maintain SCSI performance advantages at similar capacity levels. That is, drive manufacturers frequently release similar capacity SCSI and ATA drives with higher RPMs available first in the SCSI device.

To compensate for decreased performance, ATA drive manufacturers have employed a variety of techniques to enhance the ATA platform. The most important of these techniques is called Write Back Caching. Write Back Caching involves the use of small memory chips contained in the drive electronics that buffer data transfers to the ATA disk. By using these memory modules, which are typically deployed in 2MB to 8MB configurations, the ATA drive can signal the completion of writes more quickly than if it had to wait until that data was completely transferred to the disk media. However, even as write back caching provides a performance boost, it introduces a series of reliability concerns that contribute to the failure of the drive platform to achieve enterprise-class acceptance. These and other obstacles to reliability in the ATA drive platform will be discussed in detail below.

One of the most significant developments in the ATA world has been the evolution of the platform from a parallel bus architecture to a serial one. This evolution was undertaken to accelerate the use of ATA in networked storage environments and it has proven to be a crucial step in raising the awareness of the platform in multi-drive configurations. Technically, the Serial ATA drive is a seven-wire replacement for the physical ribbon of parallel ATA with a variety of benefits for denser storage implementations. The most important of these includes the cabling change (which facilitates better airflow and easier assembly) as well as the addition of capabilities like hot-pluggability and a point-to-point topology that enables full datapath switching. The first Serial ATA specification was completed in 2000 and drives supporting serial ATA begin initial production runs in the second half

of 2002. Major research houses like IDC predict that Serial ATA will dominate the ATA platform within three years, rising to a 95% to 99% share of new drive shipments by the mid 2000's. In the area of networked storage, IDC further predicts the possibility of Serial ATA commanding at least 20% of entry-level servers by 2004.

Serial ATA Characteristics
Narrower Cabling
Supports Lower Power Requirements
Lower Pin Counts
10-Year Roadmap
Higher Performance
Improved Connectivity (No Master-Slave)
Longer Cabling
PC Economies of Scale

Several years ago the ANSI Parallel ATA specification was amended with the Ultra DMA protocol, which brought advanced CRC algorithms into the ATA world. These have been carried into the Serial product. While this inclusion has addressed low-level data transfer integrity issues, a new series of problems have surfaced that stand to pose the largest obstacle to ATA acceptance in the enterprise storage world. These problems center around the use of RAID technologies that have been largely tailored and refined through their application to multi-drive SCSI and Fibre Channel storage. As ATA begins to enter the multi-drive network storage world, enterprising vendors are attempting to apply legacy RAID strategies to multi-drive ATA installations but are achieving mixed results. Today, all hardware-assisted RAID technologies native to the ATA platform—as well as ascendant software RAID packages—fail to address key performance and reliability concerns that are unique to the ATA market. By failing to address these problems, it is unlikely that the ATA platform will break beyond the entry-level category that IDC and others envision for it.

RAID5

RAID5 is one of the methods for achieving higher performance and greater resilience to drive component failure that was originally developed by the U.C. Berkeley RAID team in the late 1980s

and early 1990s under the auspices of principal investigators David Patterson, Randy Katz and their students. RAID is an acronym that refers to Redundant Array of Inexpensive Disks, and the original RAID project was conceived as a way to exploit the benefits of high volume magnetic disk drives by using strings of lower cost drives together in order to achieve the same benefits as more expensive storage configurations popular in the high end systems of the day. The groundbreaking work of the RAID team and the industry acceptance that shortly followed have made RAID strategies and resultant technologies the ascendant paradigm for dealing with magnetic disk storage today.

RAID5 specifically is a methodology for achieving redundancy of data on a group of drives without sacrificing $\frac{1}{2}$ of the available capacity as mirroring (RAID1) and its variations (i.e., RAID 10) do. RAID5 achieves this storage efficiency by performing a parity calculation on the data written to disk and storing this parity information on an additional drive. Should a disk drive fail, the data can be recovered by computing the missing data using the parity and data blocks in the remaining drives. RAID5 is an especially popular methodology for achieving redundancy because it is more economical than RAID1 insofar as more disk drive capacity can be rendered usable from a group of active drives. It has been estimated that RAID5 accounts for 70% of all drive volumes shipped into RAID configurations (the actual percentage of RAID5 per discrete RAID configuration is lower, given the popularity of striping and mirroring with OLTP). This would be sensible given that RAID5 is typically associated with file serving and similar workloads, which account for significantly more capacity usage on a global basis than higher intensity OLTP workloads, for which RAID5 is rarely used.

The attractiveness of RAID5 to the ATA storage opportunity is even more pronounced. Given the great volumetric density advantages of the ATA platform versus SCSI and Fibre Channel, ATA is ideally suited for larger capacity storage installations. The capacity efficient RAID Level 5 is functionally allied with this focus on maximum capacity per dollar of storage cost. Though some have speculated that the high density advantage of the ATA platform will result in a willingness of end users to employ mirroring given a surplus of raw capacity, the fundamental laws of technology would seem to argue against this. The sharp and continuous rise in the processing power of the Intel chip, for instance, has not been accompanied by an increase in the sales of 4-way or 8-way servers—quite the reverse is true, with one and two-way processor servers today dominating most application

usages on the market. In the storage market, given its long evidenced storage elasticity, greater volumetric densities will be accompanied by a growth in the desire to maximize capacity as well as prevent disruption from drive failure. In this view data protection based on parity strategies, as opposed to redundancy ones, will be maximally appealing—provided that they pose no crippling obstacles in their implementation.

Today, even for expensive solutions on SCSI and Fibre Channel platforms, there are obstacles to the universal ascendance of RAID Level 5 and the foremost among these is speed. For instance, one reason that RAID5 is rarely used for OLTP application storage is because of its low performance for such workloads. As a tradeoff to its storage efficiency benefits, RAID5 imposes additional computational as well as I/O burdens on the underlying magnetic disk storage. These additional burdens in many cases result in the general characterization that RAID5 is slower than other types of RAID. And, in fact, with many commercial RAID controller technology—both hardware and software—RAID5 is often the slowest performing configuration, especially when compared to straight striping (RAID0), mirroring (RAID1) or striping + mirroring (RAID 10). In some cases, for instance software RAID from vendors like VERITAS, the difference in performance between RAID5 and RAID0 is as much as 10X.

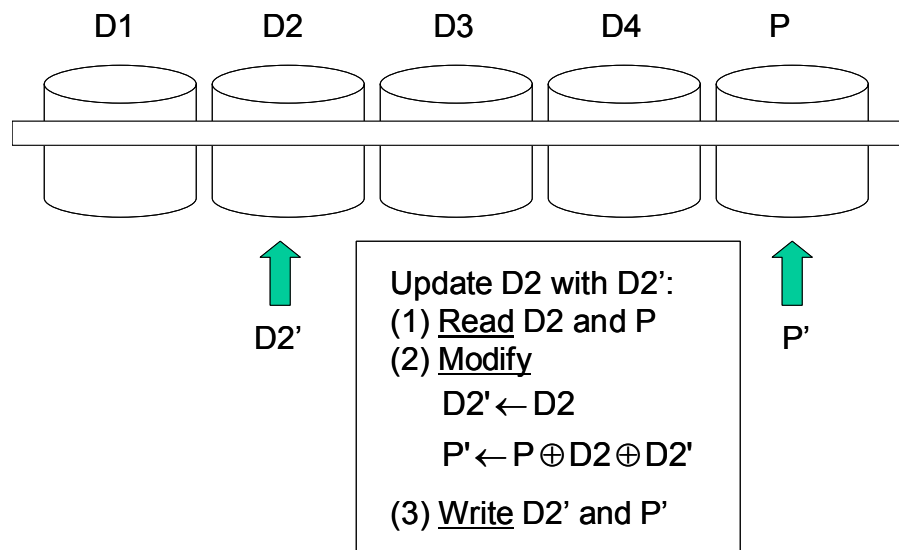


Figure 1: Read-modify-write process

Conventional RAID5 Performance Penalties

The reason that RAID5 imposes performance penalties when compared to other methods of RAID is due to two principal and related requirements. The first is the calculation of the parity itself, which requires computational resources and takes place in real time. This calculation can be accelerated by the use of specialized hardware such as an XOR engine, and most hardware RAID controllers employ this type of component to assist performance. The second performance cost, by far the most extensive, is due to the way that RAID5 typically conducts its writes. This process is called Read-Modify-Write.

During the process of a sequential write, the RAID5 implementation will attempt to write data in full stripes corresponding to the number of drives in the RAID group. However at the end of any sequential write process and during any modification of data in place, it is not possible to write a complete stripe and the technique of Read-Modify-Write must be employed. The Read-Modify-Write process is the prototypical RAID5 process and it is responsible for much of the performance limitations seen in most implementations of RAID5.

RAID 5 Performance Limitations
Multiple I/Os in Read-Modify-Write
Parity Calculation Overhead
Fixed Stripe Size
RAID Group Initialization
RAID Group Rebuilding

In a typical Read-Modify-Write operation, multiple I/Os must be executed for each logical write request. The first I/O involves reading an existing block or sequence of blocks on the disk. The second I/O involves reading the parity associated with the block(s) that will be modified. The third I/O involves writing the new data blocks, and the fourth I/O involves updating the parity associated with the relevant block(s) corresponding to the new data that is being written. No matter how small the set of drives that comprise the RAID group, the minimum number of I/Os required in a single write operation that involves the standard Read-Modify-Write approach is four, with an even greater number of I/Os associated with multiple data block writes in larger RAID sets. Furthermore, certain

approaches to ensuring reliability in RAID5 implementations (see section below) involve additional I/O activity such as logging atomic parity updates separately which increases the minimum number of Read-Modify-Write I/Os to six or higher. Figure 1 shows a typical read-modify-write process. In this figure, it is desired to update block D2 with D2'. It is also necessary to update the parity P to P'. Two reads are needed to obtain block D2 and P. D2' and P' are then computed. Finally, two writes are performed to write D2' and P' to disks.

Because of the multiple I/Os required in existing RAID5 implementations, write performance is characteristically poor, often 5X – 10X slower than mirroring or striping alternatives. There are hardware limits to the performance that is achievable given the amount of I/O activity that is generated upon each write.

In addition to low write performance, conventional RAID5 implementations have other performance limitations that are unique to its RAID flavor. Two of the most common are RAID group initialization and RAID group rebuilding. In RAID5 group initialization, the RAID solution needs to perform a scan of every data sector on each disk in the RAID set and initialize the corresponding parity. This initialization process is time consuming, the magnitude of which is directly related to the size of the RAID set and the capacity of each drive in the group.

RAID5 rebuilding is a process that must occur after a RAID5 set experiences a disk failure. When a disk fails in a RAID5 set, the missing data and parity contained on the failed drive must be regenerated on a replacement drive once the new working drive is inserted into the set or an existing hot spare is activated as the replacement drive target. Similar to initialization, the process of rebuilding requires that each data block on the system is read and the XOR computations are performed in order to obtain the absent data and parity blocks, which are then written onto the new disk. Often, during the process of reading all data from the disk to recompute the missing data and parity, bad sectors may be encountered, and it is no longer possible to rebuild the array. Depending on the size of the RAID group and the capacity of each drive, the rebuilding process is time consuming and may degrade the use of the drives in the RAID5 set for normal activity. Both the initialization and the rebuild processes are additional performance and reliability penalties of conventional RAID5 implementations that will occur as a matter of normal operation.

Conventional RAID5 Reliability Penalties

Based on the dominant approach to implementing RAID5 at present, there are several discrete reliability problems that arise in common implementations. Many of these reliability concerns are generated by events like power failure, which can often set in motion a cascade of correlated failures. For instance, a power failure not only interrupts active writes, which can invalidate any parity that is in the process of being updated, but can also burn out disks with aging components. As a result, power failures can often cause data loss in many types of RAID implementations by destroying both the parity and data associated with a “parity stripe.” Part of this is due to characteristics of the ATA platform itself, such as differences in assembly line quality control processes that have more tolerance for production variability. However a large part of the quality differential is due to ineffective strategies employed by the ATA RAID community using legacy RAID methodologies.

ATA RAID 5 Reliability Penalties
Platform requires write-back-caching
Data loss on power failure
Data out-of-sequence failure
Parity recalculation failure
File system corruption
NVRAM is not an economic answer
Single drive failure problem

The most salient reliability problem in the ATA RAID arena is the nearly universal use of write back caching in all ATA implementations, even those driven by hardware RAID solutions. Write back caching is a function that is enabled by the inclusion of small cache memory components within the disk drive electronics. By providing this additional memory, the drive is able to commit to write commands by buffering bursts of data in memory prior to the full completion of writing data onto the disk platter. When the drive signals that a write has been completed, the application moves on to its subsequent operation even if the data in question remains in the drive’s write back cache. Quicker completion of writes leads to faster application performance when disk latency is the primary

performance limitation. Because of this, the logic behind making write back caching a default strategy is straightforward: to increase the performance of the disk platform.

This performance enhancement is understandable given ATA's traditional role as a desktop device with most target implementations limited to one or two drives. Drive manufacturers have sought to differentiate the high-volume ATA offering from the higher margin SCSI and Fibre Channel drive business by limiting rotational speed thresholds on the platform. This gives pressure to optimize for performance gains like those presented by write back caching, and for the most part the industry benchmarks the ATA platform with write back caching enabled. It is possible that this will change in the future, but at the present moment this strategy is so pervasive that drive manufacturers presume write back caching to be enabled when certifying their ATA products.

Though performance enhancement is helpful, the use of write back caching in ATA RAID implementations presents at least two severe reliability drawbacks. The first involves the integrity of the data in the write back cache during a power failure event. When power is suddenly lost in the drive bays, the data located in the cache memories of the drives is also lost. In fact, in addition to data loss, the drive may also have reordered any pending writes in its write back cache. Because this data has been already committed as a write from the standpoint of the application, this may make it impossible for the application to perform consistent crash recovery. When this type of corruption occurs, it not only causes data loss to specific applications at specific places on the drive but can frequently corrupt filesystems and effectively cause the loss of all data on the "damaged" disk.

The reason that this more global type of corruption occurs is due to another problem with using a write back cache. This second problem involves the sequencing of data that enters and exits the write back cache. That is, ATA drives are free to reorder any pending writes in its write back cache. This allows the write back cache to obtain additional performance improvements. Instead of issuing sector commitments and then initiating rotational seeks for each sector in the exact sequence that commits were made, the drive places data on sectors that it encounters as platters rotate through an increasing or decreasing sector path. This reduces seek times and speeds up cache throughput. However, if a power or component failure occurs during a write process, the identity of sectors that make it to disk will not correspond to the sequence in which they were written. This causes

corruption as applications are unable to recover from drive failures because they have no way of resolving the order in which data made it to the disk media versus which data was lost in cache. Even if individual drives did not reorder writes, there is no convenient way of preventing the reordering of writes that are striped across multiple drives that use write back caching, since any individual drive is unaware of the writes being serviced by another drive.

These write back cache problems are a common cause of data corruption. In fact the weakness of the write back cache is even a relatively well understood problem, and in higher end drive platforms RAID devices and sophisticated storage administrators will default to a policy of prohibiting the use of the SCSI write back cache. However, in the ATA RAID arena, the write back cache is usually enabled by default, and performance measurement is conducted with the caching enabled, which is misleading given that the reliability implicit in RAID is compromised by the use of write-back-caching.

Deactivation of write-back caching prevents the most severe of the ATA RAID corruption problems. The tradeoff for RAID5, however, involves even lower performance. As discussed in the previous section, the legacy methodologies for RAID5 impose a significant performance limitation on this type of RAID, one that is partially addressed by vendors through the default use of write-back caching. Unfortunately, deactivating write back caching usually has a dire effect on performance.

And yet, there is a further dilemma. Since ATA vendors are not currently certifying the recovery of drives that deactivate write-back caching, it is possible that drives operating without this function will have greater failure rates. So, while vendors do achieve the goal of preventing an obvious source of data corruption, they run the risk of increasing drive failure.

The other showstopper problem posed by disk failure in ATA RAID5 solutions is the parity recalculation problem. If the system crashes during the middle of a write process, the parity calculation that applied to the active data write may be inconsistent. As a result, when the system is powered back on, it is necessary to regenerate this parity and write it to disk. Since the system will not be able to determine where the last active write was in progress, one solution is to recalculate all of the parity on the RAID5 group. This recalculation process takes time and every sector of each

participating RAID group must be scanned. Based on various leading system implementations currently available, the parity recalculation process can take between forty-five minutes for a standard RAID5 group of five or six drives to several hours for larger sets.

Currently, the parity recalculation problem is a significant drawback of software RAID5 solutions. There is no easy way to avoid this penalty when using the traditional read-modify-write approach to RAID5. Some RAID5 solutions in the ATA universe do avoid this limitation, however, through the use of “pointers” that records the positions of the in-place updates. These pointers are stored either on another disk or within a small NVRAM component. This technique is called “dirty region logging.” If the pointer is stored on another disk, it generates an additional I/O step that will further degrade performance. Nonetheless, it will deliver a performance benefit by avoiding the need to recalculate all parity upon power failure; however, it does not eliminate the associated reliability problem since, in the event of a crash, some parity will still be left in an inconsistent state until recovery can be performed. If dirty region logging is combined with write-back-caching, the original reliability problem caused by a power failure or power spike event will result in inconsistent or corrupt data. Another solution is to log the data and parity to a separate portion of the disks before responding to the write request; the logged data and parity are then copied to the actual RAID stripe. In the event of a failure, the data and parity can be copied back to the RAID stripe. This approach, while much more reliable than dirty region logging, imposes additional disk latency and makes RAID5 writes significantly slower.

A complete, high-performance way around these parity update problems in RAID5 is to use significant quantities of NVRAM with reliable battery backup. Unfortunately, the use of NVRAM will tend to degrade RAID5 performance for streaming where throughput rather than latency is important. NVRAM is often employed in higher-end SCSI and Fibre Channel RAID controllers because it improves performance for many applications and confers reliability benefits in the face of power failure. Nevertheless, it is undesirable for the ATA world to move to this type of solution. One of the most important aspects of the ATA storage opportunity involves its cost savings over alternative drive platforms. Given this, vendors do not have the luxury to equip ATA RAID solutions with a lot of expensive hardware components. Moreover, there is some expectation within the ATA community that the widespread adoption of serial ATA will result in an increase of drive counts

within standard rackmount servers. In many of these scenarios, the real estate required for additional board-level components will not be readily available on motherboards or easily addressable through the use of expansion boards. This means that the ATA world will continue to have relatively few options available for addressing reliability concerns associated with RAID5 implementations simply by applying more hardware.

Conclusion

The advent of Serial ATA drive technology holds the promise for radically altering the economics of networked storage. However, the ATA drive platform is largely unsuitable for enterprise class storage because of severe reliability problems in RAID solutions addressing the ATA universe. These reliability problems are exacerbated in the case of RAID Level 5, which amplifies susceptibility to drive failures and imposes crippling performance limitations. While RAID Level 5 has great popularity and should top demand for the overwhelming bulk of drive shipments that address mass storage, it fails to confer these advantages in the ATA world where expensive NVRAM-based hardware is economically infeasible and performance limitations make it impractical. As a result data protection must be achieved through mirroring rather than parity, which is wasteful for many applications and reduces the cost savings advantage of the ATA platform.

A new methodology to conduct RAID 5 is required if its promise in an era of low cost drive platforms is to be realized. Such a methodology would provide enterprise-class reliability without NVRAM and would deliver near-wirespeed write performance within existing ATA rotational speed frameworks. If this type of solution were available, RAID Level 5 ATA-based storage would achieve rapid and ready acceptance throughout the enterprise-class universe.

Boon Storage Technologies, Inc. has a breakthrough RAID5 technology called SR5 that overcomes the limitation of existing ATA RAID5 solution. SR5 truly makes ATA drives enterprise quality; it delivers the ultimate cost benefit to ATA drives while delivering high reliability and high performance to ATA RAID5.

Contact Boon Storage Technologies at sr5@sr5tech.com for more information.