

WHITE PAPER



# ATRATO DESIGN FOR THREE YEAR ZERO MAINTENANCE

by Sam Siewert, Ph.D

March 2008



**ATRATO**<sup>™</sup>  
ACCESS. THE REVOLUTION.

[www.AtratoInc.com](http://www.AtratoInc.com)



## ATRATO DESIGN FOR THREE YEAR ZERO MAINTENANCE

### Abstract

Many storage RAID (redundant array of inexpensive disks) systems employ data replication or error correction coding to support automatic recovery of data when disk drives fail; but most still require drive maintenance. Most often, maintenance includes hot-plug drive replacement to initiate data migration and restore data from replicated sources or to initiate error correction recoding or recovery after a single fault. Longer rebuild times increase the risk of double-fault occurrence and data loss. To minimize rebuild time and reduce the risk of data loss, replacement disk drives must be kept on hand and arrays need to be closely monitored. Given the cost of stocking replacement disk drives and operator monitoring, Atrato Inc. has researched the concept of building spare capacity into a SAID (self-maintaining array of identical disks) for fully automatic fail-in-place recovery requiring no monitoring and minimizing data loss exposure. This article provides an overview of the Atrato system's unique approach to eliminate drive tending and minimize risk of data loss for a three-year operational lifetime. This design provides superior MTTDL (mean time to data loss), high service availability, lower cost of ownership, minimal spare capacity requirements, and enables deployments with mostly unattended operation.

### Fail-In-Place Design

Fail-in-place design is a fundamental feature of the Atrato system that provides unmatched actuator density, IO performance, storage density, and zero maintenance, which significantly lowers total cost of ownership. Users really don't want to own and care for storage systems. What they really want is worry-free capacity, high performance, and no service or performance loss if components fail in the system. The Atrato system has fail-in-place throughout and takes maximum advantage of this alternative to FRU (field replaceable units). The only FRUs on the Atrato system are fans and cables and both can be over-provisioned so no FRU servicing is immediately critical. Given the relentless increase in drive capacity, very few users operate drives longer than three years and simple end-of-life migration from one Atrato SAID to another makes upgrade planning simple.

### *Testing and calculation confirm three years of high duty cycle operation*

The SATA drives used in the Atrato system have manufacturer specification MTBF of 500K hours with 24/7 power on in a thermally controlled environment with a 50% duty cycle for five years. Duty cycle is defined by the amount of time the drive is actively tracking, reading, or writing with an energized sensor head. Atrato has an ORT (on-going reliability test) with millions of hours of 24/7/365 stress workloads run on a large population of SATA drives installed in Atrato SAID systems. The point of this ORT is to verify expected drive failure rates and to observe drive failure modes and effects with high duty cycle workloads. The ORT testing has indicated that SATA drives can operate in the well-controlled SAID environment at high duty cycle for three years with a modified MTBF model.

Atrato has been testing the assertion that higher duty cycles will decrease MTBF by no more than one third for a three-year operational period. This revised MTBF model is used to compute spare capacity required based on a derated MTBF of 150K hours per drive and three years of operation with no service interruption. This has proven true in the ORT and is, in fact, a conservative estimate.

One very important observation in ORT is that many failures are not total drive failures, but rather partial failures. Partial failure modes for SATA drives sent for manufacturer FA (failure analysis) result in NFF (no failure found) and degraded operation observed in ORT. Latent sector errors (non-recoverable sectors that can't be accessed temporarily or permanently) are a well-known failure mode for SATA drives and can be recovered by AVE (Atrato Virtualization Engine) region remapping features and often prevented by AVE automated background drive maintenance. The AVE error handling is able to remap at a sector level using spare capacity on virtual spare drives so that drives with sector errors are not failed prematurely and can be restored to full performance. Overall, the AVE manages spare regions so that data can be migrated as needed from mirrors or from reconstructed data in the case of RAID-50 or 60 to spare capacity. This migration of data is expected to make full capacity available for at least three years with no service interruption.

## ATRATO DESIGN FOR THREE YEAR ZERO MAINTENANCE

### *Why fail-in-place is more efficient than drive tending*

Fail-in-place is not only fundamentally less expensive than tending hot-plug RAID arrays, it's also safer. Requiring an operator to swap out a failed drive simply delays the start of data migration and can sometimes result in inadvertent data loss due to human error (e.g., pulling the wrong drive that happens to be the remaining good mirror). Spare capacity is kept on hand as shelved drives that may or may not have been recently tested, burnt in, or scrubbed for sector errors. By comparison, spare capacity managed by the AVE is constantly scrubbed at a low background rate, known to be good, and provides hot sparing. Furthermore, given the AVE hot spare management scheme, there is no reason to pull failed drives, again risking human error, so the system can be sealed, vastly simplified and packaged at a lower cost. Either way, 10-15% of the storage capacity must be kept in spares, either on the shelf or, in the case of Atrato, in place and ready to go.

In addition, the failed drives managed by the AVE can be spun down and unlinked to isolate them and reduce power usage. With the fail-in-place strategy, the Atrato SAID has been designed to host higher actuator density and higher overall storage density than any other array ever built. Hundreds of drives are contained in a three RU (rack unit) array, providing up to 50 terabytes of total capacity - more than 10 terabytes per RU. More importantly, hundreds of concurrently operating actuators means that the Atrato system can provide multi-gigabyte IO from this 3RU SAID and tens of thousands of IOs per second with no cache. Fail-in-place not only makes this simplification and unparalleled performance possible, but also lowers total cost of ownership and simplifies administration to the point that the Atrato system can be mostly unattended.

### *Virtual sparing scheme*

Virtual sparing is fundamental to the AVE design for managing spare capacity. Instead of sparing at the drive level, the AVE spares regions distributed over multiple drives composing a virtual drive. This allows the AVE to make use of all drive spindles/actuators for high performance. It's as simple as this: Setting aside 15% of the drives would lower performance 15%. But, through virtual sparing, having all spindles/actuators active and performing can be realized. As shown in Figure 1, if disk drive A3 fails, then the 8 regions of data lost on A3 can be recovered from the A3 mirror to Vspare #2. In this example, 2 Vsparm are kept for 16 total drives with mirroring, which is 12.5% capacity sparing. If a physical spare was kept, 14 out of 16 actuators would be active whereas, with the Vsparing, 16 out of 16 are active, so RAID-10 two-way mirroring operations enjoy a 12.5% speed-up compared to physical sparing.

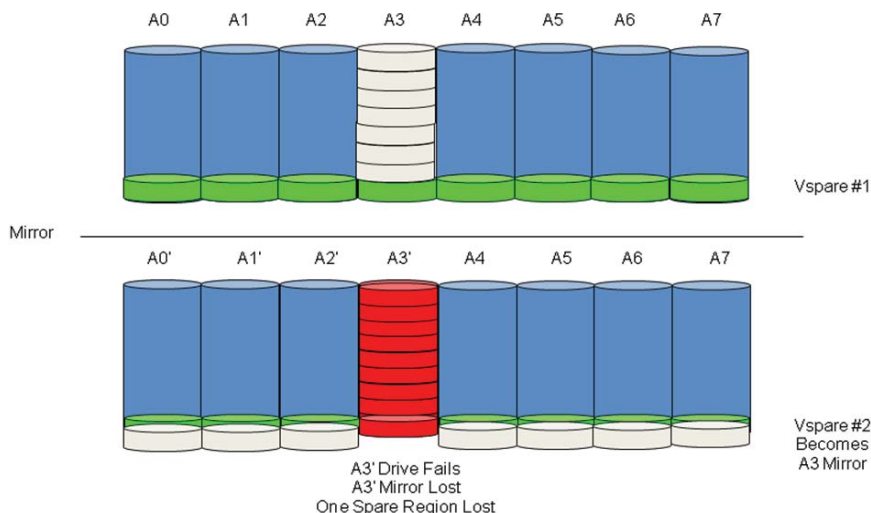


Figure 1. - Example of Virtual Sparing Shown over an Eight Drive Volume



## ATRATO DESIGN FOR THREE YEAR ZERO MAINTENANCE

### *Heuristically guided spare region selection*

When spare regions are selected for a Vspare to replace a failed drive, the selection for the spare regions is heuristically guided so that spares are selected by location, drive rank, paths available, and orientation.

### *Preventative drive maintenance*

Following the old adage for our own health, simple preventative measures are employed in the Atrato system to avoid the much higher cost of failures. While SATA drives do not share the million-hour MTBF 24/7 operation characteristics of enterprise drives, they do survive much better when carefully and regularly maintained by AVE health maintenance automation. This maintenance automation is reliable and not subject to operator omission. And, combined with SAID packaging that keeps drives cool, well isolated mechanically and electrically, and constantly monitored, the drives are most likely going to exceed the worst-case lifetimes assumed in the Atrato design. The SAID environment, background drive scrubbing and testing, and the fail-safe of duty cycle control combine to significantly increase MTTDL in the Atrato system.

### *Drive environment*

The Atrato system fail-in-place design has enabled superior SAID drive environment management. The SAID is permanently sealed with no internal FRUs, therefore minimizing risk of hot-plug related damage (e.g., ESD, improper insertion, incorrect drive types, etc.) and ensuring a well controlled thermal, mechanical, and electrical environment. The SAID takes maximum advantage of fail-in-place by laying out drives to minimize rotation and vibration, and maximize cooling efficiency, providing isolation and multi-pathing from the controller to each drive connect expander.

Superior environmental control requires active monitoring, which is provided by SAID firmware and AVE software in partnership with SMART, SES, and SCSI sense data to monitor drive health, enclosure health and status, and drive IO interface errors. These monitor inputs are used to drive preventative maintenance automation, spare region management, and predictive rebuild heuristics. For example, a drive may be taken off-line and its data migrated to spare capacity so that the AVE can perform short and long duration SMART tests, full scrubbing, and diagnostics. Then, the drive can be fully failed or returned to the spare pool if NFF is determined in place. Likewise, any enclosure issues, such as elevated enclosure temperature will be detected early, before drives are allowed to get hot, so that fan speeds can be adjusted and/or the user can be notified of the need for fan replacement. To ensure that the SAID drives are always monitored and history is complete, the SAID can only be powered on by an Atrato controller and history is stored both on the SAID and on the controller. In extreme cases, where a data center is allowed to over-heat, a SAID may suffer out-of-range operational temperatures, but the controller will know this and can alert the user and provide updated lifetime projections.

### *Latent sector error scrubbing*

SATA drives are known to suffer from latent sector errors (non-recoverable sector errors) that are not detected by the built-in drive controller. The Atrato AVE controller provides background scrubbing to discover and remap those sectors as early as possible. This has the same distinct advantage as enterprise drive controllers, but with centralized global management of spare sectors virtualized over all drives rather than within a single drive. Ferreting out sector errors provides earlier detection of high-growth counts (increasing numbers of non-recoverable sectors and/or drive controller remapped sectors on a single drive). The growth count is a leading indicator in SMART for determining which drives are most likely to fail next. This provides regular updates of drive rankings, from the drive most likely to fail next to the least likely, which is information the AVE uses to its advantage in heuristically guided spare selection and predictive rebuilds or data migration.

### *Duty cycle management*

Many users won't keep the SAID busy 24/7/365 and duty cycle is monitored along with all other SMART/SES/SCSI-sense data. But strict duty cycle enforcement can be imposed by the AVE as a fail-safe in extreme cases where data loss is imminent prior to end-of-life. Likewise, the AVE can lower duty cycle during periods of low demand by customer workloads. Overall, the system has been designed to operate at very high duty cycle workloads verified with ORT at Atrato Inc. So, duty cycle control provides users with extra assurance that lifetime will be maximized and that, in accelerated failure scenarios, the AVE will attempt to prevent early data loss through fail-safe duty cycle control. For example, if the Atrato system is deployed in an inadequately cooled data center, duty cycle control would be invoked as a fail-safe, degrading performance, but preventing data loss. The extent to which duty cycle can be controlled is determined by RAID mapping and user configuration for fault handling. Users can specify the degree to which performance is degraded to prevent data loss in scenarios where spares are being over-consumed, most likely due to operation out of thermal specifications.

### *RAID-10 mirror duty cycle management modes*

A 50% duty cycle can be enforced for RAID-10 two-way mirror operation that is predominately a read workload with strict alternation and a single input queue to each mirror pair, compared to use of dual-read queues. This increases performance by keeping all drives active all the time and results in a 100% duty cycle. Likewise, for three-way RAID-10, duty cycle can be reduced to 33% for read-only operation. For ingest applications that are predominately write workloads, RAID-10 experiences full duty cycle as well. As a fail-safe, the RAID-10 two-way and three-way mirroring modes for Atrato VLUNs (virtual logical units) provide duty cycle control for predominately read workloads (e.g., media streaming) that can be throttled down to or below 50% duty cycle. The RAID-10 duty cycle management mode is a fail-safe to prevent data loss while maintaining operation at a degraded level. This mode would only be entered if specified upon user configuration and if spare consumption is significantly exceeding expectation or operating drive temperatures are out of range for extended periods of time.

Figure 2 shows a scenario where spare consumption is exceeding expectation and duty cycle management is triggered to prevent early data loss. When spare consumption trends start to vastly exceed a linear projection over 3 years, this can be detected and correlated with other key parameters, like temperatures and sector error rates, to determine that duty cycle management mode is needed to prevent early data loss. Once these parameters are back on a safe trend, duty cycle management would be exited to restore full performance operation.

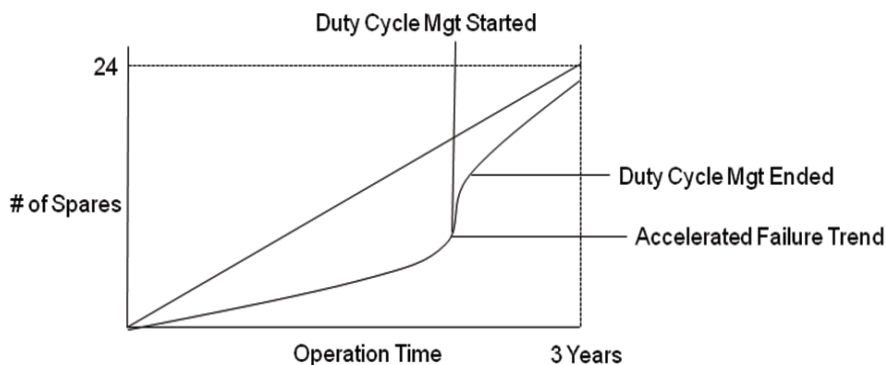


Figure 2. - Duty Cycle Management Fail-Safe

The RAID-10 duty cycle management mode commands mirror drives into a non-tracking idle mode when active. Figure 3 shows the normal full performance mode of operation with dual input queues that are both active (and load balanced) interfaced to mirrors that are both active. By comparison, with input queue control, one of the two drives can be idled for an extended period of time and one of the mirrors can handle workload at 50% performance for reads - also shown in Figure 3. The duty cycle management alternates the active drive so that each drive has workload reduced to 50% over time. During duty cycle management, the VLUN is not only operating in degraded performance, but will reject writes or, if configured, will exit duty cycle management to accommodate writes. As a fail-safe, this mode allows for data migration and continued service in read-only-mode with no risk of data loss. In many cases, temperature overloads would be handled by user corrective action which allows for full operation to be restored quickly; for example, when cooling capability in the data center is increased and drive temperatures and failure rates are restored to expectation. The increased failure rates are most often symptoms of a root cause such as elevated drive temperatures. So the FDIR correlates increased failure rates to root cause and is able to restore normal operation once those root causes are corrected. So, in Figure 2, the duration of duty cycle management would be expected to be short, based upon user corrective action such as lowering data center temperatures.

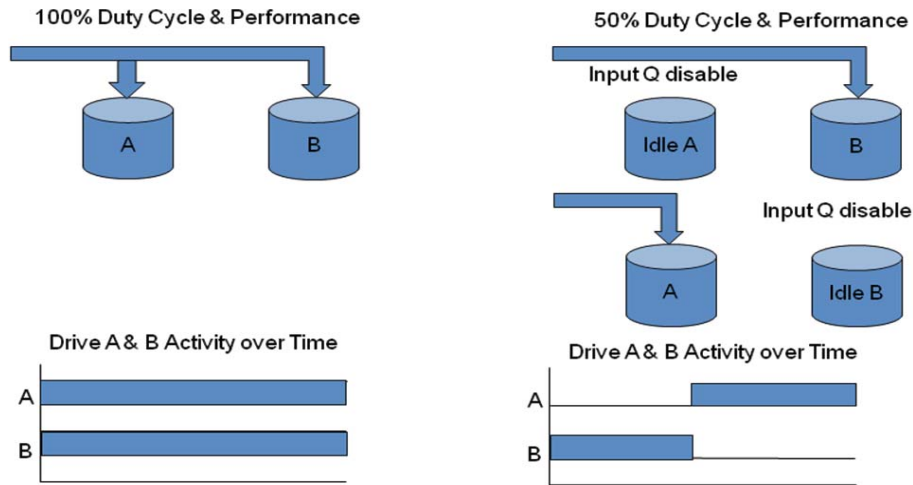


Figure 3. - Normal Full Performance Mode vs. Duty Cycle Management for RAID-10 two-way Mirrors

**RAID-50/60\* duty cycle management modes**

For RAID-50/60 VLUNs, the ability to manage duty cycle is limited to 80%. The same scheme can be used like RAID-10, but must be spread over drives in the RAID-5 or RAID-6 set. Like RAID-10, this mode works for read-only workload and writes must either be rejected in this mode or the mode must be exited to accommodate writes. In the case of RAID-5/6, the drive idled is round-robin selected from all the drives in the set, so in RAID-5, with five drives in the set, one of five drives is idled at any time, leading to 80% duty cycle seen for all in each set. When a read request is made, the drive that is idle is treated like a failed drive and data requested is rebuilt to satisfy the read request. RAID-6 similarly can rest up to two drives in a set of seven for 71% duty cycle on each drive in the set.

\*RAID-60 available only with custom controller

## ATRATO DESIGN FOR THREE YEAR ZERO MAINTENANCE

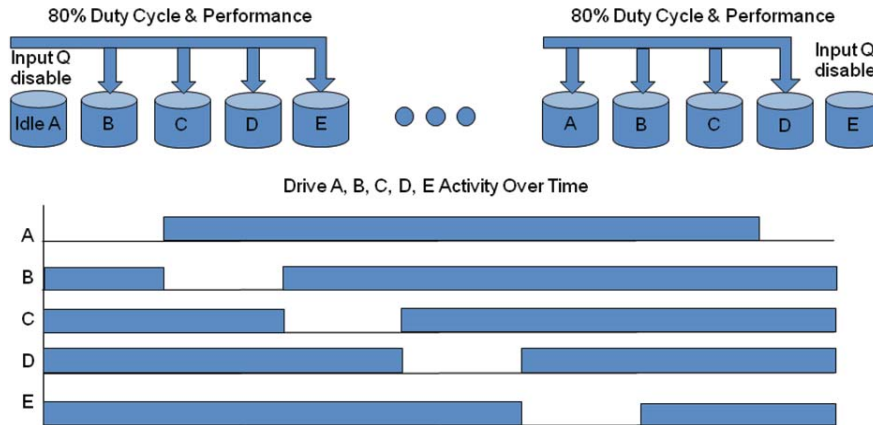


Figure 4. - RAID-5 Full Performance and Duty Cycle Management Scheme

### *Lowering duty cycle for decreased customer workload*

Since customer workloads will likely include natural quiescent periods for mirror pairs, the AVE has the ability to detect empty input queues and will place one of the two drives for two-way or two of the three drives for three-way mirroring into an idled state, always leaving one drive ready to respond when requests come in until the other resting drives can come back on-line. Likewise, the AVE can do the same for RAID-5 or RAID-6 sets in RAID-50 or RAID-60 VLUNs. The duty cycle controls are user configurable so that appropriate tradeoff between lifetime and performance can be selected or adjusted over time. In general RAID-10, RAID-50, and RAID-60 can all take advantage of read-only periods of operation with idle time distributed over the RAID set. The wake-up latency is, in general, less than a second when writes are resumed or workload increases to the point that full performance is required. This may be acceptable to some users and not others.

The lifetime models for sparing, adjusted MTBF, and fail-safe duty cycle modes of operation are being tested in ORT at Atrato Inc. The ORT drive population will always have far more hours on drives in testing than any shipped system. Not only has this lifetime model proven itself over the past two years of testing, any deviation from this would be detected first by Atrato Inc. and not by users. This allows Atrato Inc. to notify customers if for some reason a particular drive type or serial lot were to suffer systematic early failures. It should be noted again that, given the highly controlled operating environment of the SAID, drives are expected to be stressed only by workload, not thermally, with rotational vibration, shock loading, or any of the many stresses SATA drives often see when deployed in personal computers or laptops.

### **FDIR Design**

Almost as fundamental to the Atrato system as fail-in-place is FDIR (fault detection, isolation, and recovery). This terminology was coined by the aerospace industry to describe automation for highly reliable and available systems. Deep space probes must employ FDIR automation simply because operator attendance is either not possible or practical. Central to FDIR design is over-provisioning and redundancy. When faults are detected, spare resources and redundant components are employed to prevent service interruption. The space environment is much harsher than the data center environment, so FDIR is simpler and lower cost for RAID arrays like the Atrato system. But, surprisingly, many FDIR concepts are not employed in other RAID arrays. The Atrato system design enables cost effective and powerful FDIR features that differentiate the product from all other SAN/NAS devices.



## ATRATO DESIGN FOR THREE YEAR ZERO MAINTENANCE

### *Detection - SMART, SES, SCSI sense data*

Reliable detection is the most important part of FDIR. False alarms that trigger unnecessary isolation and recovery waste resources and time. In extreme cases, false alarms can themselves cause service interruption. The AVE is designed with multiple high-quality detection sensors that provide high integrity health and state estimation from numerous samples and sources. Critical parameters such as temperatures, voltages, fan speeds, and more, are measured. And multiple sensors minimize the possibility of both missed and false alarms (both false positives and negatives).

### *Isolation - active SAS link management and drive state management*

Failed components, including disk drives, SAID SAS (serial attached SCSI) expanders, SAID interface controllers, or front-end customer initiator interfaces, can be failed and then isolated so that partially functioning components will not interfere with normal Atrato system operation. During ORT, Atrato Inc. has observed drive failure modes where drives may interfere with other drives on common SAS expanders. So, it is important that once a fatal drive fault is detected, the AVE FDIR software will quickly isolate that failed drive by disabling its link and powering it down. Non-fatal errors and partial failures such as non-recoverable sectors will, of course, not trigger isolation but will trigger error handling and continued use of the degraded component. The isolation handling underscores the importance of reliable detection and the ability of the AVE software to distinguish fatal and non-fatal faults. Isolation itself is, for all components, a simple process of disabling, unlinking, or removing power. So it is the decision logic based on years of ORT and millions of hours of stress and FMEA that is most important. Isolation has been designed to be quick and effective.

### *Recovery - Error handling levels with RAID recovery and data migration support*

Recovery for failed drives or sectors is handled by AVE data migration. In the case of individual sector failures, this is a simple region migration that is triggered by a read/write failure, SCSI sense data, and/or a timeout retry. In the case of fatal drive faults, migration is much more time-consuming, since, in a worst-case scenario, it may involve migrating all regions on a drive. This risk is mitigated by AVE predictive rebuild. Since spares regions are always in place, rather than on a shelf, they can be rebuilt ahead of time, guided by heuristic prediction of drives most likely to fail. The heuristic prediction sounds tricky at first. Analysis has shown, however, that even the most misguided predictions cause no harm and, in fact, even with random selection, provide faster recovery.

### *Abstracted spare consumption monitoring*

One interesting side effect of the Atrato system design and AVE FDIR is that it leads to a very simple high-level capability to manage Atrato storage by exception. The single-most important monitoring parameter is spare capacity consumption over time. Failures of the few FRUs in the system are also unlikely exceptions. So, the Atrato system has been designed to provide notification through its management interface so that regular monitoring or tending is not necessary. The AVE can provide remote monitoring via its remote CLI (command line interface) or web browser interface. Furthermore, it can be set up to send e-mail or provide text messaging. This single parameter monitoring (spare capacity and remaining projected lifetime) makes scaling of data center operations much less costly. In fact, monitoring of many Atrato systems can be aggregated via SNMP (simple network monitoring protocol) with a small number of MIB (management information base) objects.

### *SATA drive failure modes and effects analysis*

Atrato Inc. started ORT shortly after the bring-up of the very first SAID and has continued ORT non-stop since that day in spring of 2006. Focus has been on a limited number of drive types and families, with all drives identical - matching systems that are planned for shipping. The one exception for all identical drives would be a mixed install of SSD (solid state disk) drives along with traditional disk drives. Solid state drives have different failure modes, but very little effect on other drives (most often flash fails to erase and becomes read-only). Identical drives offer the



## ATRATO DESIGN FOR THREE YEAR ZERO MAINTENANCE

advantage of failure modes and effects that are mostly identical. The only real difference for any drive is location and original manufacturing differences. This is tightly controlled by the Atrato Inc. manufacturing process, which ensures all drives have identical burn-in times, scrubbing, test workloads, and handling. The same procedures have been used in ORT, but ORT has always been run with a 24/7 stress workload to accelerate failure rates and to provide conservative validation of lifetime and spare capacity estimates. Given the accelerated stress testing in ORT and the significant two year or more head start that ORT has on shipping systems, it is more probable that all unique failure modes and effects will be first observed in ORT and not in customer systems.

### *Ongoing reliability testing and failure modes and effects analysis*

In ORT to date, failures have never exceeded the available levels of sparing; furthermore, numerous failure modes have been observed and analyzed by Atrato Inc. and partner drive manufacturers. In early ORT, cases of mishandling were found to be responsible for some drive failures, which fortunately allowed Atrato Inc. to improve handling procedures to eliminate this failure mode. Initial manufacturing SMART baseline scans, performance tests, as well as improved shipping, in house handling, and tracking, have all contributed to the elimination of these failures. Likewise, ORT has revealed that, while the manufacturer provides quality control, Atrato Inc. burn-in and testing of drives in the SAIDs has provided the opportunity to weed out drives with performance issues before they are shipped to customers. Likewise, the manufacturing process of scrubbing all drives has eliminated latent sector errors from being present at time of SAID deployment. Operational background scrubbing ensures that latent sector errors do not develop after these manufacturing processes are completed.

Failure elimination and prevention are always best, but drives are still expected to suffer head burn-outs, crashes, media errors, and drive controller malfunctions. In the case of a single head burn-out on multi-platter drives, in some cases a partial migration is possible. But, in all other cases, the drive itself must be failed in place and isolated. Drive failures and non-recoverable sector failures are not surprising; they are expected, often predicted, and handled. Only accelerated failure rates would raise concern. However, given sufficient margin and time to migrate data to a full replacement or system level Atrato mirror, this is not even cause for alarm since failures may not be uniform over time. As long as the failure rate does not remain accelerated and spare margin is not too low for customer data migration, this situation can still be monitored by exception.

### **RAID Mappings**

RAID mappings in the AVE are not only integrated with IO to provide detection of corrupted data and to handle failures, but also to provide improved performance, error handling, and overall system reliability. The RAID replicated regions for mirrors, RAID-5 parity, or RAID-6\* parity and error correction codes are mapped so they are striped, integrated with error handling, and plug into a common overall FDIR framework. The methods of data corruption detection, recovery, and encoding/replication are specific, but detection, isolation, and heuristically guided recovery is common to all RAID mappings. This is enabled in part by the observation that RAID-10 two-way and RAID-50 have common single-fault characteristics, as do RAID-10 three-way and RAID-60 for double faults. Likewise, RAID-10 two-way and RAID-50 can be paired with striped data digests to strengthen silent data corruption protection. The RAID-60 mapping provides 12-bit Galois Field (Reed Solomon 2-bit) detection and correction ¶ much stronger than simple parity checks. Likewise, full data compare for RAID-10 two-way is not feasible, so data digests are provided as an option for reliable accelerated detection of corruption between mirrors. Overall, MTDL (mean time to data loss) is the important figure of merit. The tight integration of RAID, FDIR, and data protection mechanisms (data digest options) provide simple but very scalable mechanisms to maximize MTDL even in very high bandwidth, high capacity, high workload usage scenarios. Finally, unique mapping algorithms, rebuild strategies, and continued operation during rebuilds provides best-in-class performance.

\*Available only with custom controller

### *RAID-10 two-way and three-way mirroring*

RAID-10 two-way mirroring has the advantage of simple immediate recovery for data corruption or loss using the replicated mirror data to satisfy IO requests while mirroring is restored with rebuild data migration. RAID-10 three-way mirroring is very costly in terms of capacity, but provides the maximum performance reliability protection.

### *Two-way mirror design for single-fault protection*

The RAID-10 two-way mirror provides options such that mirrored regions can be mapped to have equal performance criteria (in similar bands and closer to outer diameter than inner). For many users, the 50% capacity given up in exchange for simple high performance reliable IO may be a worthwhile trade-off. Use of mirrors has fringe benefits such as remote mirror data replication acceleration (for disaster recovery or backup). With RAID-10 mirrors, the current idle mirror is always used along with copy on write to migrate data from the non-busy device. The RAID-10 striping over mirror pairs has significant advantage over RAID 0+1 where striped volumes are then mirrored at a system level because it makes maximum use of concurrent actuation and common controller resources, provides rapid recovery for transient errors, and, in the AVE design, directly integrates extended features like SDC (silent data corruption) protection. The main disadvantage of two-way mirroring is duplication on writes, so ingestion of content is slowed down to half that of read performance.

### *Three-way mirror design for double-fault protection*

RAID-10 three-way mirroring provides double-fault protection to prevent data loss (or corruption with SDC protection on) and also provides single-fault performance protection. The performance protection is less obvious and is implemented using a permuted mapping unique to the AVE which ensures that first fault leaves a copy of data in the same diameter band, but allows double-fault data to be mapped into a lower performance band. This allows for more outermost band data mapping, but provides latency bound guarantees at the same time. This mapping, called the triplex is shown in Figure 5. The triple mirror lowers duty cycle significantly below manufacturer limits and provides ample opportunity to overlap backup and remote mirror operations with ongoing IO services.

### *Three-way mirror triplex permuted region mapping*

As shown in Figure 5, the triplex mapping feature of RAID-10 three-way mirroring not only provides double-fault protection against data loss, but also preserves data banding in single faults to prevent any performance degradation. Banding data results in higher performance access due to short stroking, where a VLUN defined with a particular band will enjoy lower latency due to areal density and limited motion of the drive actuator. When a failure occurs, if the mirrored data is out of band with the primary, then performance would degrade. In Figure 5, regions 1,3,5 are all in one band as are 2,4,6. If a drive fails, there are still two mirrors which have 1,3,5 in similar, but permuted bands. In a random workload, the banding with permutation provides speed-up by statistically keeping actuators in bands of operation. Tests with random workloads and no permutation, just exact copies of regions on each mirror, compared to the permuted mapping, have shown increases in performance up to 24%. This can be seen in Figure 5 where head position on drive A might be accessing region 3, so region 5 is the minimum distance from 3 on drives A and C. Likewise, head position on drive B might be 3, so region 1 is the minimum distance from 3 on drives B and C. The queue interface has been designed to place requests into queues which result in minimum head movement and the permutation mapping ensures that there is always an optimal choice for minimum motion regardless of data access patterns. The triplex mapping of RAID-10 three-way mirroring preserves this even when one drive is in recovery. When two drives are in recovery, data is still not lost, but the banding can't be controlled as tightly, so performance would drop.

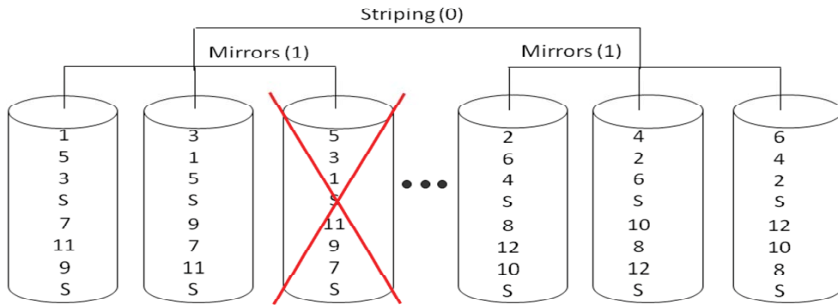


Figure 5. - Triplex mapping of Regions for RAID-10 Three-way Mirroring

**RAID-50 and RAID-60**

RAID-50 and RAID-60 provides significant capacity savings. For example, RAID-50, striped over 10 drives per RAID-5 set and then striped over multiple sets in the array, provides great striped performance and uses 90% of available capacity. In general, the larger the number of drives in a RAID-5 set, the more useable capacity approaches 1-1/n. This amounts to 90% for 10 drives, but has the drawback of requiring larger and larger XOR computations for parity encode and rebuild from parity. Simply mapping RAID-5 over all drives rather than striping sets (RAID-50) would lead to maximum capacity. For example, capacity would be 1-1/160 (99.375%) in the extreme case for a full box SAID, but the parity calculation would be 160 LBAs (logical block addresses) and not feasible. Another interesting note about RAID-5 is that sets of five drives work nicely with tens of drives overall in the array and block sizes that are multiples of four. File systems most often deal in block reads and writes that are, in fact, multiples of four such as 4K, 8K, and 64K. Figure 6 shows this relationship of diminishing returns for larger RAID-5 sets.

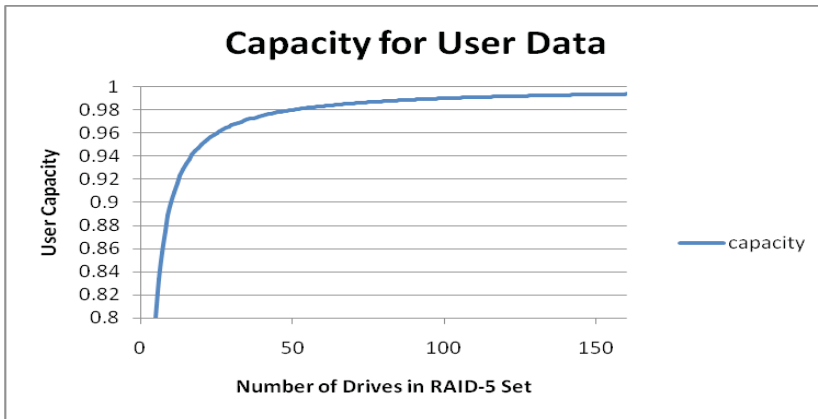


Figure 6. - Capacity Tradeoff with RAID-5 Set Size

Given the diminishing utility of larger RAID-5 sets, it is recommended that a tradeoff such as 5, 9, 17, or 33 drives per set be used, along with striping over the RAID-5 sets, for RAID-50 with a balance between size of the parity calculation and user capacity. The number of drives in each RAID-5 set must also be a factor of the total number of drives used in the physical drive pool for the VLUN being created.

**RAID-5 parity design for single-fault protection**

The RAID-5 uses simple XOR parity computation for a parity LBA (logical block address) striped into the set along with data LBAs. When a drive fails or LBA is lost due to sector failure or corruption, the lost LBA is rebuilt by simply restoring parity to that originally computed. If the parity LBA is itself lost, it is simply recomputed. The parity LBA is never on the same drive as the LBAs it encodes so that parity and data can never be lost with a single fault. The RAID-5 mapping over five drives in each RAID-5 set with striping over multiple sets for RAID-50 is done as shown in Figure 7. Figure 7 shows only two RAID-5 sets, but this can be generalized to any number or size of set to achieve higher capacity, but with larger RAID-5 set extents. The AVE has a minimum RAID-5 set of five for 80% capacity and a maximum of nine for 89% capacity.

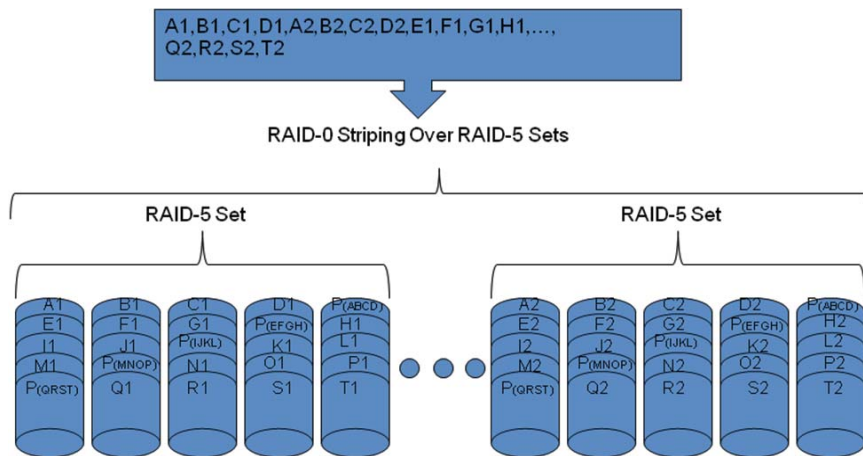


Figure 7. - Striping of RAID-5 Sets for RAID-50 Mapping

The reason that RAID-5 sets are NOT made as large as possible - e.g., over all drives in one SAID - is that the cost of the RAID-5 parity and rebuild times goes up with set size and the sets span so many resources that the isolation advantages of independent sets are lost. Instead, a trade-off is made with RAID-5 sets of 5 or 9 drives with striping over 32 5 drive sets for example in a RAID-50 mapping that spans the entire 160 drive SAID. Furthermore, as Figure 6 shows, the advantage of larger and larger RAID-5 sets diminishes and the advantages compared to added complexity are less significant beyond a double RAID-5 set (10 drives). RAID-5 is often noted as slower than RAID-10 for ingest, which is generally true. However, for content ingest that is large block, the AVE makes use of a write cache so that all data and parity LBAs are operated on in memory. For example, if a 128K block write is done to a nine-drive RAID-5 set, 32 eight-LBA and parity LBA updates can be made in memory and written out to the nine-drive set. Odd-sized blocks require read, modify, and write operations to read the current parity block, update it with the odd number of block updates, and then send both back to the drives.

**RAID-6 Reed-Solomon design for double-fault protection**

The AVE supports RAID-6 sets arranged in a RAID-60 mapping for double-fault protection with similar characteristics

of RAID-10 three-way, but with higher capacity. The RAID-6 set includes seven drives with a parity (P) and error correction code (Q) sector for every five sectors of data striped over all seven drives. This provides double-fault protection with 71% of capacity available for user data. This RAID mapping is very computationally complex and therefore employs hardware acceleration to keep up with the high data rates of the Atrato system. Otherwise, it provides the same advantages of RAID-50, but with double-fault protection and slightly lower user capacity.

### Capacity and performance trade-offs

The RAID-50/60 mappings have the advantage of 80% or greater capacity efficiency compared to 50% for two-way and 33% for three-way mirroring. Users are provided a clear choice to trade capacity for lifetime performance and fault tolerance; RAID-10 two-way and RAID-50 providing single-fault tolerance and RAID-10 three-way and RAID-60 providing double-fault tolerance. This can be useful tailoring for different types of applications and workloads. For example, a user wanting maximum capacity, single-fault tolerance, and mostly unattended operating of a media streamer would likely choose RAID-50. For a streaming application where fault protection, service reliability, and performance QoS must be exceptionally high, RAID-10 three-way mirroring could be selected and provides the best overall service reliability. Table 1 provides an overall comparison of the RAID mappings available and their performance and reliability. These mappings can be made on a VLUN basis so that Atrato customers can choose capacity, reliability, performance, and data loss risk trade-offs to best fit their application needs.

Mapping	Fault Protection	User Capacity	Performance Protection	Recovery Time
RAID-0	none	100%	Data loss on any fault	N/A
RAID-10(2)	single	50%	Latency increase on single fault	low with predictive
RAID-10(3)	double	33%	No latency increase on single fault	low with predictive
RAID-50(5-drives)	single	80%	Minimal latency increase on single fault	medium w/ predictive
RAID-50(9-drives)	single	89%	Minimal latency increase on single fault	medium w/ predictive
RAID-60(7-drives)	double	71%	Minimal latency increase on single/double fault	high, no predictive

Table 1. - Comparison of Atrato RAID Mappings and Integrated FDIR

### Snapshots

Snapshots are supported by the AVE with copy-on-write so that they are instantaneous and, in the case of RAID-10, backed by mirrored data. Figure 8 shows the copy-on-write for an instant snapshot in time using RAID-10. The meta-data includes date and time and any differences stored to the VLUN after that snapshot was taken with original data and address information. Snapshots provide an instant copy in this way and RAID-10 two-way or three-way allows for backup with less interruption to continued operation.

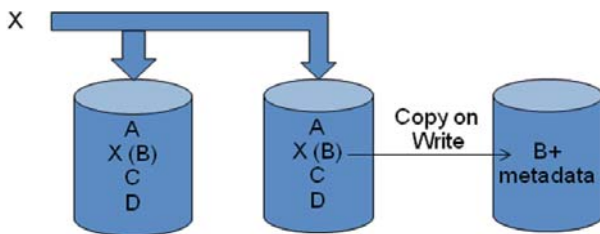


Figure 8. - Copy-on-Write for Snapshots

With the RAID-10 mapping, if a user initiates backup from the snapshot, the backup operation can copy unchanged data from that point in time from the inactive mirror and data that was updated from the snapshot capacity set up



## ATRATO DESIGN FOR THREE YEAR ZERO MAINTENANCE

by the user. By comparison, RAID-50/60 requires the continued operation IOs and backup IOs to access the same drives, so RAID-10 offers an advantage for backup.

### New Operating Paradigm

Overall, the Atrato system design for fail-in-place with FDIR provides for management of storage by exception. Storage now appears as capacity, with known access performance, and a maintenance-free lifetime that requires only occasional management for exceptions and end-of-life replacement planning. This makes the Atrato system better suited for remote operation and to assist with upgrade planning. Customers can focus more on deploying capacity, access, and quality of service, and less on specific drive types, drive tending, servicing rebuilds, managing spares, and tending to their storage.

### References

1. "Improved Disk Drive Failure Warnings", IEEE Transactions on Reliability, September 2002.
2. "Failure Trends in a Large Disk Drive Population", Proceedings of 5th USENIX Conference on File and Storage Tech (FAST'07), February 2007.
3. "An Analysis of Latent Sector Errors in Disk Drives", ACM SIGMETRICS'07, June 2007.
4. "Intelligent RAID 6 Theory Overview and Implementation", White Paper, Intel SCD, [www.intel.com/design/storage/intelligent\\_raid.htm](http://www.intel.com/design/storage/intelligent_raid.htm), 2005.
5. Storage Networks Explained, Ulf Troppens, Rainer Erkens, Wolfgang Mueller, SNIA, John Wiley & Sons, Inc., 2004.

### About Atrato

Atrato, Inc. is revolutionizing the data access and storage markets by challenging the traditional thinking on how to rapidly access stored data. Atrato's system provides unparalleled performance, linear scalability, and significantly reduced operational costs in a self-maintaining architecture for the Entertainment, Web 2.0, IPTV, and VOD markets. For more information about Atrato and its products, visit [www.AtratoInc.com](http://www.AtratoInc.com).

Atrato Inc.  
1120 West 122nd Ave., Suite 300  
Westminster, CO 80234  
P 303.245.1045  
F 303.479.9385