DriveGuard
The Choice in Availability

Garth Gehlbach

Technical Report 18
June 1997

Abstract

This report introduces DriveGuard™, Auspex's RAID 5 product which delivers an exceptional balance of availability, performance and economy. With the addition of DriveGuard, Auspex broadens its impressive portfolio of high availability offerings for enterprise-class NFS service. DriveGuard benefits align with key themes of configuration flexibility, protection, and performance, offering system administrators an unparalleled level of choice with which to address the availability and performance requirements of their user communities.

Flexibility: Realizing that applications and data may have different protection and performance requirements, Auspex enables NetServer™ configurations to be tailored filesystem by filesystem to the desired levels of availability and performance, thus allowing for optimization according to the needs of the data, not the peculiarities of a server architecture. DriveGuard RAID 5 arrays can be configured concurrently with arrays that are configured as mirrors and/or striped such that the right balance of protection, performance and economy can be selected for each filesystem and application.

Protection: Auspex's robust implementation of RAID 5 augments basic data protection from disk failure with an innovative, industry-first Hot-Move feature, which reduces the impact of rebuilds in half, and provides speedy rebuilds which make Auspex NetServers as much as 10,000 times safer in protecting against data loss as compared to other products.

Performance: DriveGuard's embedded hardware-based design and underlying Functional Multi-processing® (FMP®) architecture enhances performance and addresses the challenges inherent with write operations in a RAID 5 scheme. Auspex leverages hardware acceleration and a high degree of parallelism to provide high performance in normal, degraded and rebuild modes.

DriveGuard is yet another advancement in Auspex's commitment to continuous data access. DriveGuard features include:

- Embedded hardware design
- "On the fly" hardware parity generation
- Intelligent, battery-backed write cache and parity cache
- Industry unique "Hot-Move" function
- Concurrent support of RAID 0, 1, 0+1, 5 and individual disks
- Arrays of 3-6 disk drives
- Up to 7 arrays per Storage Processor (SP) subsystem
- Floating and dedicated hot-spares
- Configurable stripe size

Document 300-TC046

Auspex Systems Inc.
5200 Great America Parkway
Santa Clara, California 95054 USA
Phone: 408/986-2000 • Fax: 408/986-2020
Email: Info@auspex.com

WWW: http://www.auspex.com

Table of Contents

Tables and Figures

1. Introduction

In the world of Enterprise Computing, all data is not equal. The service requirements of data vary with application, frequency of access, criticality to the business, and impact on productivity. Since price is usually a function of availability as well as performance, the goal of Systems Administrators generally becomes matching the needs of the data with the characteristics of the solutions. If this were not true, System Administrators would either overspend and manage all their data on the most sophisticated high-availability systems, or suffer the pains of putting critical data on inexpensive, unreliable systems.

2. The Spectrum of Fault Resiliency

In looking at servers and file service, there are a number of points of failure which can lead to downtime. These sources include Operating Systems, applications, disk drives, power and cooling subsystems, network failures and potentially a variety of system hardware and software elements. Auspex has systematically introduced products which address the availability issues associated with the components of failure listed above, and has prioritized its efforts on those that have the most impact. From the fundamental Functional Multi-processing (FMP) architecture, whith its streamlined code focused on quick and reliable data movement, to DataGuard™ which isolates UNIX administrative functions and applications, and from disk mirroring for protection against drive failures, to ServerGuard™ which offers resiliency to disasters, Auspex has delivered high-availability to mission critical environments.

With the introduction of DriveGuard, Auspex provides another choice for protection against disk failures. DriveGuard offers performance similar to mirrored configurations at a significantly lower cost. As disk drive manufacturers accelerate the introduction of new technologies, with gains in capacity and performance, and speed the end of life of the more reliable and mature product lines, disks are shipping at lower points on the Mean Time Between Failure (MTBF) curves. This makes it increasingly important to have protection against disk failures, for a greater portion of storage. Thus, for data that requires disk failure protection, but not to the level of mirroring, DriveGuard is the appropriate choice.

3. RAID Review

A description of the various RAID (Redundant Array of Independent Disks) levels can be found in Table 1 below. Also included in the table are attributes associated with the RAID levels, such as relative protection, performance, and cost. Since there are tradeoffs to be considered with each scheme, it is important to understand relative positioning of each RAID level. Striping, or RAID 0, tends to be a good choice for environments where performance is the key objective. Mirroring, or RAID 1, is the choice for highest availability. RAID 3 is often preferred for highly sequential I/O environments, and RAID 5 is best for I/O intensive environments.

Table 1. Summary Comparison of RAID levels[1]

RAID Level	Common Name	Description	Cost (Disks req'd)	Reliability	Data Transfer (MB/s)	I/O Rate (IOPs)
0	Striping	Data distributed across the disks in the array. No redundancy or protection from disk failure.	N (all storage is usable)	Lower than single disk (consider MTBF of all drives in array)	Very high	Very high
1	Mirroring	All data is duplicated, stored on two distinct disks	2N	Higher than RAID 0, 3, 5	Similar to single disk, may be higher for reads, may be lower for writes	Similar to single disk for writes, higher for reads[2], may be lower for writes
3	Parallel transfer disks with parity	Each data sector is subdivided and distributed across all data disk. Redundant information is stored on a dedicated parity disk	N+1 (N is the number of data disks in the array)	Higher than single disk	Highest of all RAID levels (normal mode)	Similar to single disk (normal mode)
5	Parity RAID	Data sectors are distributed as with disk striping; redundant information is interspersed with user data	N+1	Higher than single disk	Similar to striping for reads; lower than single disk for write[3] (normal mode)	Similar to striping for reads; generally lower than single disk for write (normal mode)
0+1	Mirrored & Striped	Combination of striping and mirroring	2N	Similar to mirroring	Similar to striping	Similar to striping

Figure 1 below illustrates how a block of data is segmented and distributed in a RAID 5 array[4]. Notice both the distribution or striping of the data and the placement of the parity. RAID 5 is often viewed as the RAID level of choice for I/O intensive environments, where multiple read operations can be performed on the array (logical drive) simultaneously. This is because each read usually only requires one physical disk to satisfy the request, and multiple disks within the array can be operating in parallel.

[1] Much of the table contents have been extracted from the RAID Advisory Board RAIDbook.
[2] Mirrored reads may be faster if implementation allows read to be satisfied by whichever drive can satisfy the request faster.
[3] Performance of RAID 5 write is commonly referred to as the RAID 5 write penalty and can be mitigated by employing cache in the design of RAID storage subsystems.
[4] An array is defined as a group of drives configured into a single logical drive, with data organized according to the RAID level and parameters such as stripe size.

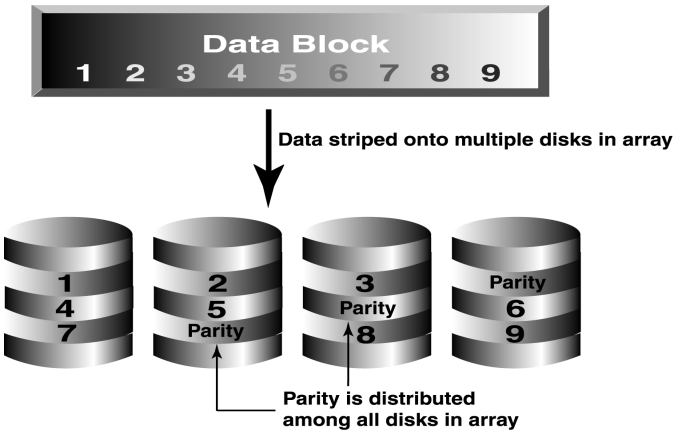

Figure 1: RAID 5 format

4. DriveGuard: Auspex RAID 5

DriveGuard is a RAID 5[5] solution whose benefits extend beyond economical data protection against drive failures. In keeping with Auspex's commitment to continuous data access, superior availability, performance and flexibility are cornerstones to the design.

DriveGuard is a hardware-based RAID implementation where the most demanding tasks are quickly performed in dedicated hardware, thus relieving the host CPU/OS from the burden of calculating parity or maintaining arrays. Auspex emphasizes availability through its protection against drive failures, battery-backed cache, and quick rebuilds to minimize the risk of data loss due to multiple failures within arrays.

Another key theme to DriveGuard is flexibility, which affords the capability to tailor NetServers to best meet the service requirements of NFS clients. Just as it is important in golf to have a variety of clubs from which to select for a given shot, the flexibility to select the right balance of protection, performance and cost for each filesystem produces an optimum server configuration. This flexibility comes in a variety of features, including concurrent support for RAID 0, 1, 5, 0+1, and individual drives within an SP subsystem. This lets System Administrators select different levels of protection for different filesystems, even within an SP, based on the importance associated with the data.

5. Performance and Availability

Auspex's DriveGuard implementation includes a number of performance oriented features, many of which also improve availability by reducing the window of time where an array is rebuilding and is not resilient to a second disk failure. A custom ASIC − called the Auspex FIFO Controller (AFC) − is at the heart of the design and resides in the data path, calculating RAID parity on the fly. The AFC thus delivers the performance advantages of generating parity (XOR function) in hardware without requiring an intermediate location to store and forward the resulting data, and without draining valuable cycles from the Storage Processor (SP) or Host Processor (HP) CPUs.

The Auspex write-back cache further boosts NFS performance to clients by immediately acknowledging write completion and later executing the write to disk, thus reducing write acknowledgment latency. In addition, the intelligent cache efficiently caches data, metadata and parity, eliminating some disk operations and greatly mitigating the RAID 5 write penalty. The cache is battery-backed and provides safe, stable storage, protecting against the unlikely event of a server failure or reboot before data is written to disk.

---

[5] For a more detailed explanation of RAID, see the RAID Advisory Board RAIDbook. Information on obtaining a copy can be found on the web site *www.raid-advisory.com*.

6. DriveGuard Write Operation

Figures 2-4 demonstrate the benefits of cache on DriveGuard write operations. In the Figure 2, the cache does not already have the data or parity resident and is required to read old data and old parity from disk. Performance benefits are realized, however, in that write acknowledgment is sent back to the client immediately, without waiting for the subsequent disk operations. In Figure 3, the write operation involves data from the *same parity group or portion of the RAID 5 stripe* as the previous write, and thus benefits from having parity information in cache and not requiring a disk operation to read parity into the cache. In Figure 4, the write operation further benefits from having both the old data and old parity already in cache, thus all the inputs required to complete the RAID 5 write operation are available in cache and no disk operations are required at this time.

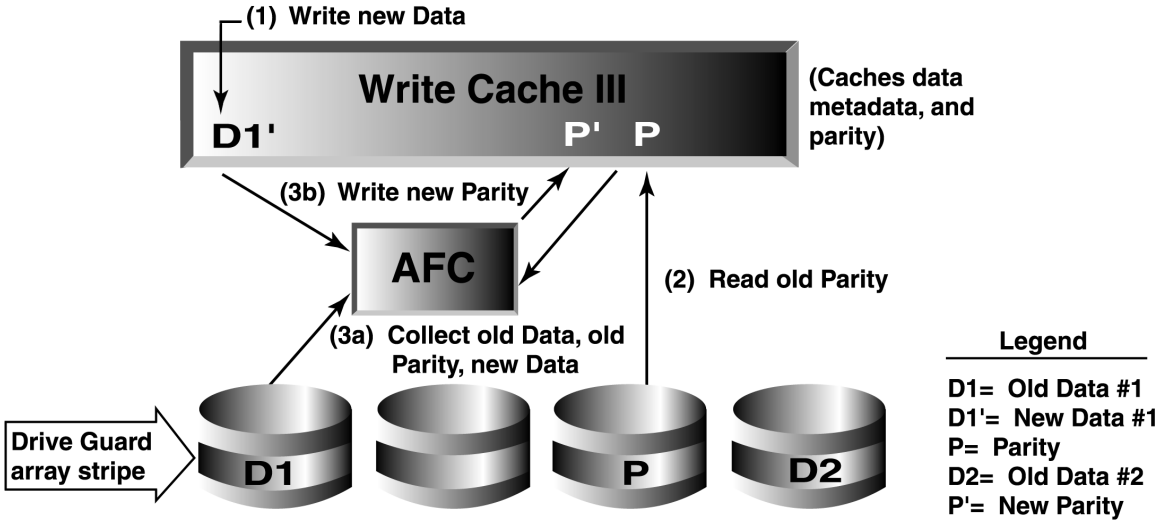

Figure 2. DriveGuard Write Operation

The first scenario of a DriveGuard write operation is shown in Figure 2. The first step in the diagram is where the new data (D1') enters the SP write cache. Immediately after the new data (D1') is cached an acknowledgment of completion of the write is sent to the client (step 1b). Step 2 reads the old parity (P) into the cache. For demonstration purposes, step 3 is segregated into two parts, 3a shows the AFC collection of old data (D1) being read from disk, and new data (D1') and old parity (P) being read from cache. Step 3b finishes the process with the AFC calculating the new parity (P') and writing the result into cache. Since the AFC calculates parity on the fly, steps 3a and 3b are essentially one uninterrupted flow. Subsequent writes can further benefit from data and/or parity overwrites (see Figures 2 and 3 below). The cache is flushed, or written to disk, according to a Least Recently Used (LRU) algorithm, maximizing the probability of cache hits and the associated performance benefits (see Figure 4).

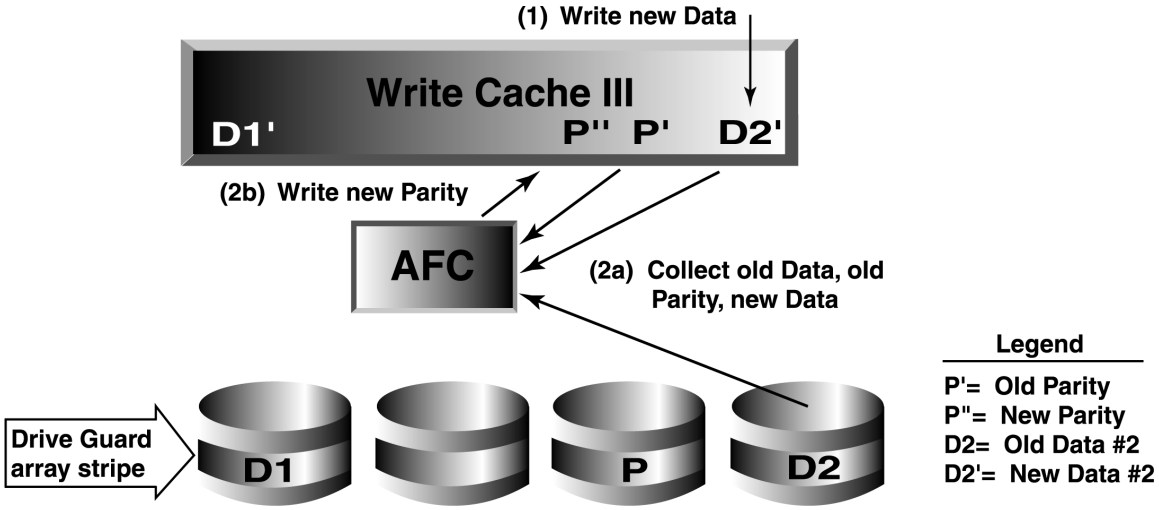

Figure 3. DriveGuard Write Operation - Parity cache overwrite

In Figure 3, the DriveGuard Write is similar to that described in Figure 2; specifically where new data (D2') enters the cache and the AFC collects the old parity (P') and new data (D2') from cache, and the old data (D2) from disk, generates new parity (P'') and writes this to cache. The difference is that old parity is already in cache and does not need to be read from disk, thus eliminating a disk operation and improving performance.

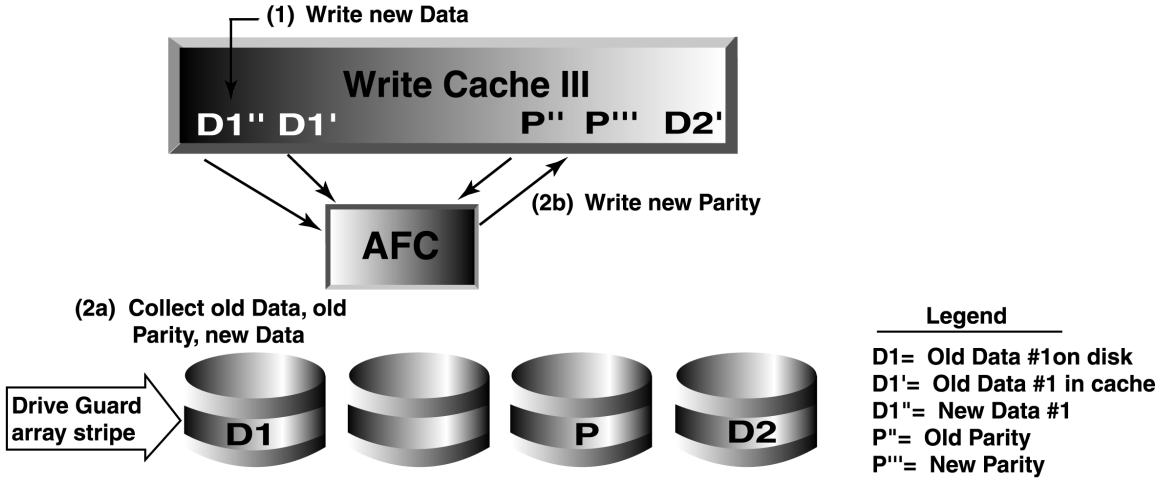

Figure 4. DriveGuard Write Operation - Data and parity cache overwrite

Figure 4 illustrates the ultimate benefit of write cache where the write operation was completed without requiring a disk operation. New data (D1'') entered the cache and the AFC received the necessary inputs of old parity (P''), old data (D1') and new data (D1'') from cache. The new parity (P''') was calculated by the AFC and sent to cache.

7. Degraded Mode and Rebuild Mode Performance

When considering performance issues with RAID systems, there are two essential topics that generally get neglected: degraded mode[6] performance and rebuild mode[7] performance. The reason these two performance dimensions are critical in evaluating a RAID system is that a key criterion for purchasing RAID is continuous data access. If a RAID system cannot satisfy I/O demand when in degraded mode or rebuild mode, or loses data because a rebuild took days and a second disk failure occurred before the failed array could be restored, continuous data access is not realized.

As with all parity RAID implementations, a failed drive will cause reads of the failed drive to be satisfied via a calculation of the data on the remaining drives in the RAID array. This extra processing and disk accesses obviously decreases performance, but for implementations such as Auspex's, where the XOR calculation is done in hardware, the performance impact is minimized. In fact, there is no discernable impact to the throughput of a Storage Processor subsystem with a failed drive.

Rebuild performance can be analyzed from two perspectives: rebuild duration and NFS service during rebuild mode. In the case of rebuild duration, a 6-disk DriveGuard array can reconstruct a 4.29 GB drive in as little as 90 minutes. Under light to moderate workloads, rebuild of a 4.29 GB drive requires 2 hours with little impact to NFS service for the affected Storage Processor (SP); under heavy workloads, rebuild takes 6 hours. While 6 hours may seem a long time, lesser solutions, particularly those that are software-based RAID solutions, can require 24 hours or more to complete a rebuild while serving a significant workload.

Rebuild duration directly affects the window of vulnerability where an array is not redundant and a second disk failure will cause data to be lost. The embedded hardware-based Auspex design provides for continued data service (even under heavy loads), a short rebuild duration, and no impact to other SPs. By comparison, alternative designs suffer from a single CPU and software based implementation that significantly impact performance of normal I/O and may require many hours or days to complete a rebuild.

Table 2. Summary of DriveGuard Performance: Normal, Degraded and Rebuild Modes

Workload	**Degraded Mode** Impact to SP throughput vs. Normal mode	**Rebuild Mode** Impact to SP throughput vs. Normal mode	**Rebuild duration**
Light	none	none	90 minutes
Moderate	none	none	2 hours
Heavy	none	little	6 hours

Table 2 above summarizes the performance impact of the various modes of DriveGuard. As outlined, there is little or no impact to SP throughput of an array in degraded or rebuild mode; however, depending on the workload, there may be a noticeable effect on the throughput of the particular array itself. It is important to note that a NetServer supports up to five Storage Processors (SPs) and any performance impact associated with a degraded array or an array being reconstructed will not impact the performance of the other SPs. Multiple concurrent rebuilds are supported within an SP as well as for multiple SPs within a NetServer.

[6] Degraded mode is defined as an array state where one drive has failed and I/O continues to be performed on the remaining drives.

[7] Rebuild mode is defined as an array state where one drive has failed, I/O continues to be performed on the remaining drives and the contents of the failed drive are being reconstructed to a designated spare drive using the content of the remaining drives in the array and XOR logic.

8. Hot-spares: Floating or Dedicated

Auspex has implemented spares in a powerful and flexible way. Spares can be configured as dedicated to a particular array, or floating to serve as substitutes for failed drives anywhere in the subsystem (within an SP). Furthermore, spares can be designated to rebuild a failed drive's contents automatically or upon manual command, and the priority of the rebuild can be specified by the System Administrator. A higher priority for the rebuild devotes more SP cycles to the task of rebuild, thereby increasing the speed of the rebuild and possibly decreasing the performance of normal I/O. Alternatively, a lower priority of the rebuild will put emphasis on I/O performance and if the SP is heavily taxed, the rebuild duration will be longer. This degree of flexibility allows administrators to tailor fault recovery policies appropriate for their operations.

The benefits of choices in configuring spares enable System Administrators to guarantee a level of protection for filesystems. For instance, a filesystem with extremely critical data may justify a dedicated spare for the filesystem, plus a floating spare for the entire SP subsystem. Such a configuration would provide each array with access to the protection of the floating spare, but also ensure that at least one spare drive was devoted to a drive failure in the array with the sensitive filesystem and that drive failures in other arrays could not access the dedicated spare.

9. Hot-move

Auspex's unique "Hot Move" feature improves system availability and performance by allowing disks within an array to be moved to restore optimized configurations, typically one disk per SCSI bus[8]. This operation can be performed without taking down the system or forcing a second rebuild – actions that would impact availability and performance. In addition, from a physical management perspective, Hot-Move allows the NetServers to be returned to a sensible physical configuration – highly desirable with the large disk farms that the NetServers support.

Figure 5 below shows a failed disk automatically initiating a rebuild to a spare at a different location. Although this spare disk is now part of the RAID array, it may be undesirable to leave the configuration "unbalanced," where two disks are on one SCSI bus and another SCSI bus is unused by the array. In other products, a replacement drive is inserted in the failed drive slot and the rebuilt spare is removed from the array, causing a rebuild to the new drive to restore the original configuration. This scheme requires two rebuilds, a process that degrades system performance, and is an availability exposure since redundancy is not provided until the array is rebuilt again.

[8] A common RAID configuration is to establish arrays where each array has one drive on each SCSI bus. This generally provides for maximum performance (high degree of parallelism, no contention for SCSI bus within an array) and maximum availability (data is protected against a SCSI bus failure, since such an event would only cause one drive in each of the arrays to be lost).

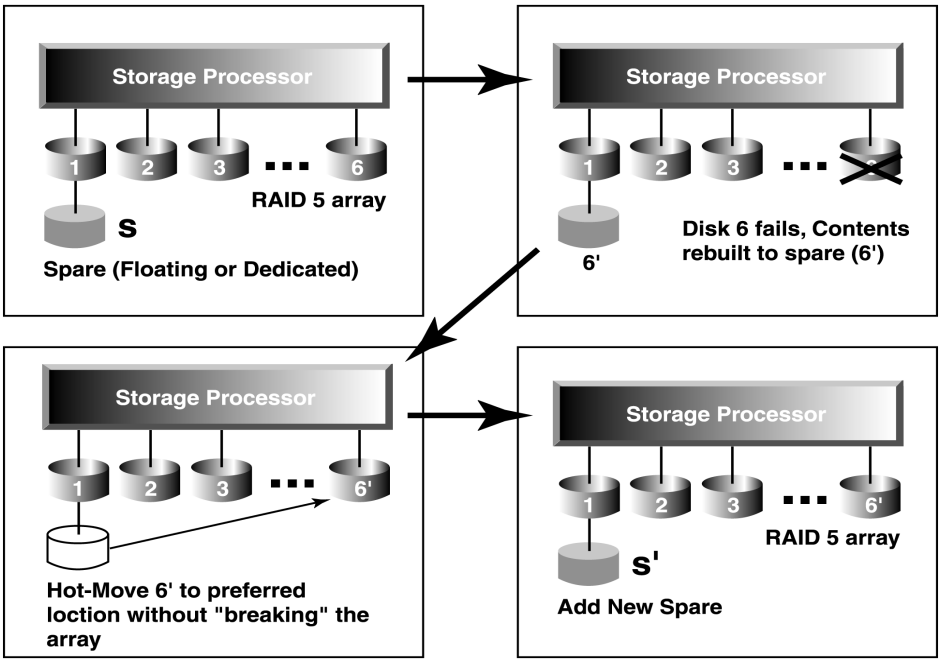

Figure 5. Hot-Move Process

10. Risk of Data Loss: The Poisson Model

The Poisson process is a statistical model for independent, identically distributed exponential random variables. This model is appropriate for the case of a subsystem of disk drives if the events, namely disk failures, are independent of each other. If we rule out the possibility of a subsystem induced catastrophic failure (such as a power surge to several disks), this model is reasonable: specifically, that one drive failure is independent of any other drive failure. The formula for the Poisson process is

$$P(k) = \frac{(\lambda t)^{k} e^{-\lambda t}}{k!}$$

where **P(k)** is the probability of k (the number of events or failures), λ is the arrival rate, in this case defined as the number of drives in the array divided by the drive MTBF, **t** is the time window, in this case the rebuild time, and **k** is the number of events (failures).

The critical measure of the Poisson model for a storage subsystem is the risk of data loss, or with a parity RAID implementation, the risk of losing two drives within an array before a rebuild can complete. This can be accomplished by calculating the probability of exactly zero failures and the probability of exactly one failure during the rebuild window, and subtracting the sum of these probabilities from 1 to yield the probability of 2 or more failures.

An example may reinforce the point that choices in configuring a RAID subsystem are important in realizing the desired level of availability in a system. Consider the impact of RAID arrays of different sizes and with different rebuild times. An array of 6 drives with a rebuild time of 2 hours is over ***10,780 times less likely to lose data*** than an array of 52 drives with a rebuild time of 24 hours.[9] Table 3 provides additional data points for perspective.

[9] This example assumes disk drive MTBF of 300,000 hours. Other values for MTBF yield essentially the same relative results.

Table 3. Risk of Data Loss Comparison

Array size	Rebuild Time	Drive MTBF	Probability of data loss vs. baseline
6 drives	2 hours	300,000 hours	baseline (approximately 8 x 10^{-10})
6 drives	6 hours	300,000 hours	**9** x baseline
26 drives	12 hours	300,000 hours	**676** x baseline
52 drives	12 hours	300,000 hours	**2,700** x baseline
52 drives	24 hours	300,000 hours	**10,786** x baseline
6 drives	2 hours	100,000 hours	**9** x baseline

Another key aspect in comparing solutions on dimensions of risk of data loss, and degraded performance due to rebuilding arrays, is the necessity of other solutions to perform a second rebuild to restore the desired configuration. This *doubles the impact of degraded performance and more than doubles the risk of data loss*. With the combination of configuration flexibility and Hot-Move, Auspex maximizes availability and minimizes risk and impact of degraded performance.

11. Flexibility: Tailoring Configurations

Auspex's implementation of DriveGuard has a number of valuable configuration choices, allowing NetServers to be fine tuned to meet the needs of the data and the environment. Central to this flexibility is the concurrent support of striping, mirroring, striping and mirroring, and parity RAID (RAID levels 0, 1, 0+1 and 5) all within the same Storage Processor (SP). For performance-oriented applications where protection against drive failures is not a concern, striping the data in a RAID 0 virtual partition is appropriate. For the most critical data where the highest degree of protection and availability are the absolute objectives, mirroring or mirroring and striping are the best choices. For applications that require protection against drive failures but not to the degree of mirroring, and deliver a high ratio of usable storage to total storage, DriveGuard is the solution.

DriveGuard arrays can co-exist with mirrored and striped virtual partitions such that a combination of RAID storage can be defined within an SP. Furthermore, drives of different capacities, such as 4.29 GB and 9.1 GB drives can be mixed within the SP subsystem, but not within a DriveGuard array.[10]

Other parameters such as DriveGuard array stripe size can also be defined by the System Administrator. Stripe size is the depth, or amount of storage, that is spread over the drives in the array in a single stripe. Stripe sizes generally should be sufficiently large such that the typical I/O can be satisfied by the data on a single drive's stripe (avoid "disk boundary crossings" where more than one disk holds the data for the I/O operation) and not so large as to create hot spots (diminish the advantages of spreading data). The default stripe size of 128 KB generally accomplishes the objectives previously stated; other options include 64 KB and 256 KB.

[10] Drive capacity of hot-spares must match the capacity of the drives in the array for a successful reconstruction.

12. Ease of Administration

Auspex provides management tools for defining and maintaining DriveGuard arrays that are similar to the tools for managing the NetServer in general, promoting a rapid learning curve that speeds time to productivity and keeps training costs to a minimum. Included in the tool set are options for event notification. Administrators can extract event information from log files and choose to send email to designated users as well as be notified via a pop-up window, which itself can be customized.

13. Summary

Auspex's DriveGuard RAID 5 implementation provides an excellent balance of availability, performance and economy, and delivers valuable choices in configuring NetServers to match the needs of the data, filesystem by filesystem. With its embedded hardware-based design, DriveGuard can deliver performance with protection, enabling continuous data access.