HP 3PAR Fast RAID: high performance without compromise

Karl L. Swartz

Technical white paper

Table of contents

Introduction

The HP 3PAR® Storage System is a highly virtualized, tightly clustered, dynamically tiered storage array built for utility computing. Each HP 3PAR array implements the HP 3PAR Architecture, a unique approach to the design of a RAID system that offers performance and cost-efficiency advantages unparalleled by traditional storage arrays.

This paper briefly reviews basic RAID terminology and concepts, describes key enhancements in the HP 3PAR Utility Storage implementation of RAID, explains how to match application storage requirements to the right RAID level, and discusses data protection options on HP 3PAR Storage Systems.

RAID basics

Disk drives are mechanical devices, subject to various sorts of failures. These failures could result in data loss if the information is not protected in a suitable manner. RAID, which stands for Redundant Array of Independent Disks, is a method for providing protection by storing data using multiple disks that have sufficient redundancy to survive the failure of a disk drive.[1]

RAID is a collection of methods, referred to as *RAID levels*, which provide tradeoffs between storage efficiency and performance for various workloads. The most widely implemented RAID levels are RAID 1 and RAID 5:

- RAID 1 is the simplest RAID level and consists of writing identical data to two drives, resulting in the drives being mirrors of each other. If one drive fails, the data can still be read from the other drive—and written to a spare or replacement drive, restoring redundancy. Although it is a simple method, RAID 1 requires two blocks of raw storage for every block of usable storage.
- RAID 5 uses a group of three or more drives. Several blocks per drive are combined in a *stripe*, with the blocks of one drive in the stripe being used to store a binary sum (XOR) of the data, called the *parity* of the data, which provides redundancy that can be used to reconstruct data in the event of a single-drive failure. The number of data blocks per parity block can usually be selected to suit given needs; RAID 5 (*n*+1) signifies *n* data blocks per parity block. The next stripe stores its parity on a different drive from the group in order to avoid performance problems that could result from storing all of the parity on one drive. In the event of a drive failure, data blocks can be reconstructed using the remaining data blocks and the parity. Compared to RAID 1, RAID 5 requires less raw storage to provide an equivalent amount of usable storage, with some cost in performance for random-write workloads.

Since Patterson's original RAID paper was published, several new RAID levels have appeared:

- "RAID 0" is a common term used for the concatenation of drives to achieve a larger volume. Applying the term RAID is a misnomer as there is no redundancy, and thus a single drive failure can cause the loss of all data in the RAID group. However, concatenation is commonly used in combination with redundant groups to create larger volumes. For example, concatenating RAID 1 groups results in a larger RAID 1+0 volume.
- RAID 6 is a more recently defined RAID level that is similar to RAID 5 but with additional parity blocks to allow reconstruction of the volume even after the failure of two drives. While disk capacity has benefited from exponential growth, the reliability of disk drives—their Mean Time Between Failure (MTBF)—has not improved substantially. Moreover, the largest disks are generally also the slowest. The result is that the largest disks require a longer time to rebuild a failed disk on a spare—an interval known as the Mean Time To Repair (MTTR)—during which a second disk failure would result in data loss with single parity. RAID 6 was created to address this problem.

RAID on physical drives: random writes

The sequence of events required to process a random write by the host are the basis for the performance disparity between RAID 5 and RAID 1. Assuming that no blocks from the RAID group are in the array's cache, a write to a RAID 1 volume requires two independent disk I/O operations: writing the block twice, once to each disk. The same write to a RAID 5 (3+1) volume requires twice as

[1] D. Patterson, G. Gibson, and R. Katz, "A Case for Redundant Arrays of Inexpensive Disks (RAID)." ACM SIGMOD 88, Chicago, June 1988, pp. 109-116.

many disk I/Os plus the parity calculation, serialized as follows:

1. Read the data block being modified and the old parity block.

2. Compute the new parity: subtract the old data from the old parity, then add the new data (P' = P - D + D').

3. Write the modified data and parity blocks.

HP 3PAR Storage Systems use a large write-back cache to reduce the number of disk I/Os, while wide striping reduces the cost of disk I/Os by distributing them over many disks.

RAID 6 adds further I/O cost to an uncached, random write because a data block participates in at least two parity calculations. The 6+2 configuration requires an average of 6.67 disk I/Os per random write in HP 3PAR's implementation; a 14+2 configuration requires an average of 6.28 disk I/Os per random write.

RAID on physical drives: sequential writes

Sequential writes are much simpler than random writes. The number of disk I/Os required when the host writes *n* blocks, again ignoring caching, depends on the RAID level:

- **RAID 1**: 2*n* disk I/Os
- **RAID 5 (*n*+1)**: *n*+1 disk I/Os
- **RAID 6 (*n*+1)**: *n*+2 disk I/Os

For example, consider a host write of 6 blocks of data. RAID 1 will require 12 disk I/Os, while the default configurations for parity RAID (3+1 for RAID 5, 6+2 for RAID 6) will require only 8 disk I/Os. The disk I/O cost of RAID 1 in this case is 50% more than either form of parity RAID.

RAID on physical drives: reads

Read performance (uncached) requires one disk I/O per host read, whether random or sequential. RAID 5 and RAID 6 store less redundant data than RAID 1, and thus require shorter seeks for random reads if the same number of physical disks are used. Storage efficiency—the ratio of usable data to raw data—is a useful approximation of the relative cost of seeks between the different RAID levels.

Improved implementation of RAID

HP 3PAR Storage Systems incorporate several enhancements over conventional RAID arrays. By making more effective use of all drive resources in the array, these enhancements allow higher performance with less hardware, which in turn leads to cost reduction.

Chunklet-based RAID

HP 3PAR InForm® Operating System Software divides physical drives into 256-MB *chunklets*, each of which can be viewed as its own small disk. RAID groups are constructed from chunklets on separate drives. Several of these RAID groups are then combined into a Logical Disk (LD), with LD data striped over the RAID groups. LDs using RAID groups from as many physical drives as possible are then concatenated to form virtual volumes. By widely striping each virtual volume over chunklets on many physical drives, much higher performance is achieved than if RAID groups were constructed from whole disks as is done in traditional storage arrays.

Wide striping for performance

In a traditional storage array, small volumes either suffer from poor performance by using few drives or waste expensive resources by using more drives than required for capacity in order to obtain sufficient performance.

On an HP 3PAR array, even modest-sized volumes will be widely striped using chunklets spread over all drives of the same type. Wide striping provides the full performance capabilities of the array to

small volumes without provisioning excess capacity and without creating hot spots on a subset of physical drives. (Figure 1) Other chunklets on the drives are available for other volumes.

Figure 1: wide striping on HP 3PAR arrays compared to traditional RAID

Wide striping: physical drives operate within a few percent of the same IOPS and service time. No hot spots.

Traditional RAID: mix of hot and cold drive-based RAID groups, fluctuating over time.

Physical drives can hold a mix of RAID levels

On an HP 3PAR array, RAID groups are constructed from chunklets, not from whole drives. Different chunklets on a physical drive can be used for volumes with different RAID levels. On a traditional array, a storage administrator might be forced to use RAID 1 for an archival volume in order to use space that is available on a RAID 1 disk even though RAID 5 would deliver adequate performance with less overhead. The chunklet-based approach deployed by HP 3PAR Utility Storage allows all RAID levels to coexist on the same physical drives, using the optimal RAID level for each volume.

ASIC speeds data movement and parity calculations

Each HP 3PAR array contains up to four pairs of controller nodes. Each controller node contains at least two Intel® processor cores[2] to handle control tasks, but all bulk data movement is efficiently performed by the proprietary HP 3PAR Gen3 ASIC in each node. This ASIC is also used to perform the XOR parity calculation.

In addition, the Gen3 ASIC uses built-in zero-detection capability to recognize blocks of zeros on-the-fly. Zero-detection enables conversion of "fat" legacy storage to thin-provisioned storage using HP 3PAR Thin Conversion Software and recovery of deleted blocks from file systems using HP 3PAR Thin Persistence Software.

Fast RAID reconstruction

Chunklets, wide-striping, and the Gen3 ASIC combine to provide extremely fast RAID reconstruction after a drive failure, with minimal impact to performance for ongoing activity. Fast rebuilds reduce the window during which an additional drive could lead to data loss, allowing the use of RAID 5 to provide a level of data protection that would otherwise require the additional storage cost of RAID 6 in other arrays.

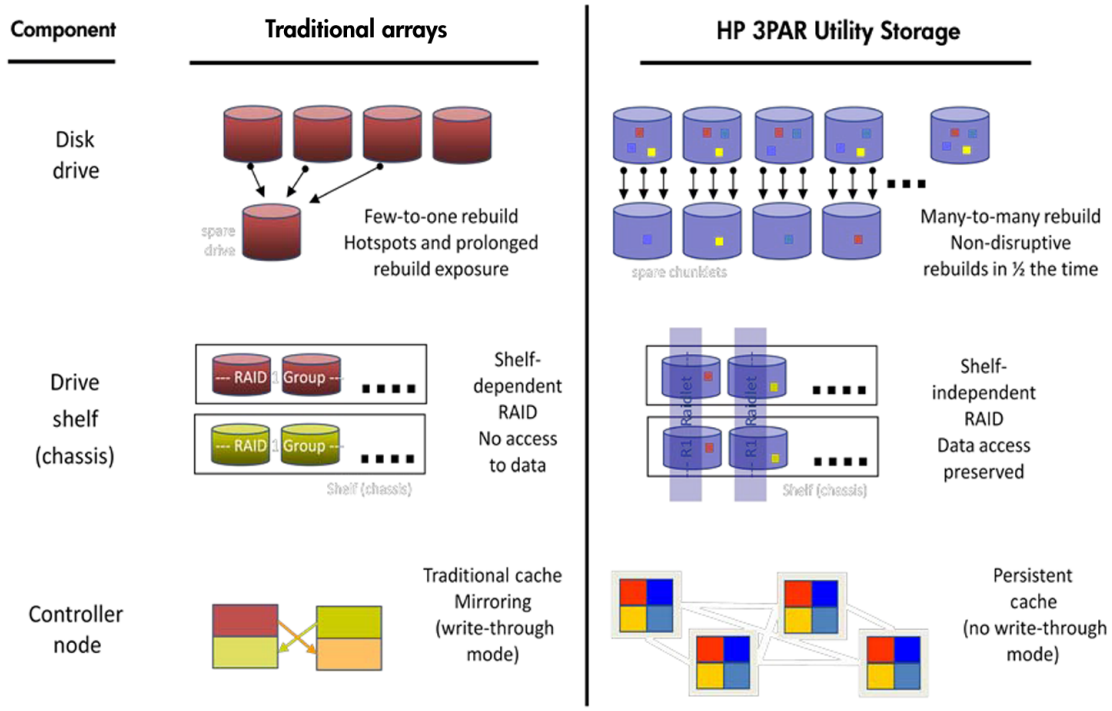
First, only allocated chunklets need to be reconstructed, not the entire drive. The platform's unique Thin Technologies help by not allocating physical storage for unwritten or zeroed data. The data required to reconstruct allocated chunklets comes from many other drives in the array, even with simple RAID 1 mirror pairs. This allows the wide striping that aids normal performance to speed reconstruction without causing hot spots on other drives. Spare chunklets used during reconstruction are also wide-striped, so the bottleneck of a single spare drive is avoided. (Figure 2.)

Secondly, the Gen3 ASIC helps to speed reconstruction by efficiently moving data and by accelerating parity calculations.

[2] Dual processors and/or multi-core processors depending on the array model.

Together, these features enable fast RAID reconstruction that minimizes the time interval during which data is unprotected (the MTTR), all without compromising the performance of ongoing application requests for the array. After the failed drive has been replaced, its contents are rapidly restored using data that was already reconstructed in spare chunklets wide-striped throughout the array.

Figure 2: wide striping enables faster, non-disruptive RAID rebuilds**.**



Choosing the right RAID level

In traditional arrays, RAID 1 is used to increase performance, despite the cost it adds by using two raw drives for every drive's worth of user data. RAID 5 is used to improve storage utilization where performance is less important. RAID 6 can be used to provide adequate data protection for large, slow disks.

Fast RAID 5, described in section 4.1, allows more cost-effective RAID 5 to be used on HP 3PAR arrays instead of RAID 1. Testing of OLTP throughput performed by Oracle® showed that Fast RAID 5 (3+1) delivered 91% of the performance of RAID 1[3] while using 25% less raw storage for the same amount of usable storage.

The Fast RAID reconstruction performed by HP 3PAR arrays, described in section 3.3, allows volumes on large, slow disks to use RAID 5 instead of RAID 6. Traditional arrays need RAID 6 because their slower rebuild times (i.e., their MTTR) dramatically increase their exposure to data loss due to multiple disk failures while using 25% less raw storage for the same amount of usable storage.

HP 3PAR Utility Storage arrays also offer high availability options that allow access to data even after the failure of an entire drive chassis.

Drive type, RAID level, and high availability options are chosen when a virtual volume is created. Should the need arise to change any of these attributes later, HP 3PAR Dynamic Optimization and Adaptive Optimization Software can be used to make this change on active volumes without any

[3] Simplified Database Storage Management That Lowers Management Costs And Yields High Storage Utilization, An Oracle & 3PAR White Paper. April 2008.

downtime or interruption. Dynamic Optimization tunes entire volumes with minimal performance impact to the array. Adaptive Optimization provides even finer-grained tuning, moving sub-volume regions between storage tiers in response to changing performance.

Fast RAID 5

Fast RAID 5 combines the HP 3PAR Gen3 ASIC, a large, battery-backed memory cache, and wide striping for reducing spindle contention to offer performance that approaches that of RAID 1, minimizing the performance impact typical of RAID 5 on legacy storage architectures. For certain workloads, RAID 5 can provide higher performance than RAID 1. The write-back cache in the HP 3PAR Storage System allows sequential writes (as generated by transaction journals, logs, and similar performance-sensitive workloads) to be collected until a full parity group can be written, reducing disk I/O traffic and possible back-end bottlenecks. RAID 5 is also appropriate for volumes that are dominated by read activity, since RAID 1 provides no benefit for reads as detailed in section 2.3.

HP 3PAR Storage Systems allow selection of the number of data blocks per parity block (*n*+1) to suit different needs. For RAID 5, 3+1 is the default, but any value from 2+1 to 8+1 can be selected. Higher values of *n* result in higher storage efficiency but can reduce the performance of random writes.

HP customers using HP 3PAR arrays typically choose HP 3PAR Fast RAID 5 for most or all volumes, as Fast RAID 5 minimizes the performance disadvantages associated with traditional RAID 1 while providing greater storage efficiency.

SATA disks and Fast RAID 6

Exponential growth in disk capacity without commensurate improvements in reliability or performance results in greater risk of data loss. For example, consider the 300-GB Fibre Channel (FC) disks and 2-TB Nearline (Enterprise SATA) disks available on HP 3PAR Storage Systems. The capacity difference alone implies that reconstruction of a failed disk on a replacement can be expected to take more than six times as long with the 2-TB disk. The Nearline disks are slower, too, which further increases the MTTR relative to smaller, faster FC disks. A longer MTTR creates a larger window during which a second disk failure could cause data loss when using RAID 1 or RAID 5.

RAID 6 was created to address this problem. Like RAID 5, RAID 6 uses distributed parity, but it stores two different parity values, calculated from different parts of the stripe in a manner that allows the data to be reconstructed, even in the event of two disk failures.

HP 3PAR RAID MP (multiple, distributed parity) initially supports dual parity, equivalent to RAID 6, but is capable of supporting higher parity levels in the future. However, as noted earlier, even the extra protection of RAID 6 relative to RAID 5 is less important for Nearline disks on HP 3PAR arrays than on traditional storage arrays where slower rebuilds make RAID 6 crucial.

Environments such as highly consolidated virtual host environments tend to have unusually high data protection requirements due to the large number of users that could be affected by data loss, and so demand the highest level of data protection. High I/O loads make RAID 6 problematic on traditional arrays; the implementation of RAID 6 on HP 3PAR arrays is the only choice that provides the extra increment of data protection without compromising I/O performance.

High availability options

Physical drives in an HP 3PAR array are mounted on magazines that are contained within drive chassis. Each magazine on a T-Class system contains four drives, with up to ten magazines (40 drives) per 4U drive chassis. The midrange F-Class systems use single-drive magazines with up to 16 magazines per 3U drive chassis. Each drive chassis has redundant access paths via two Fibre Channel connections, one to each controller node in a pair.[4]

With RAID 1 and RAID 5, virtual volumes default to a configuration in which access to the data will survive the failure of an entire *cage* or *drive chassis*: In this configuration, the default value for the *createcpg* command's *–ha* option is *-ha cage,* which causes each chunklet in a RAID group to be allocated from a physical drive in a different chassis. For RAID 6, *-ha cage* means the CPG or virtual

[4] F-Class drive chassis may be daisy chained in pairs, with one chassis in a pair connected to the second chassis which is then dual-attached to the controller nodes.

volume will tolerate the failure of two physical drives in the same chassis. Thus, a RAID 6 (6+2) volume can be created on an array with as few as four drive chassis. The 14+2 configuration requires eight drive chassis per node pair.

In some cases it might be desirable to create volumes that do not meet the *-ha cage* requirements, but will survive the failure of two drive magazines. Examples include a small array with only two or three drive chassis, or RAID 6 (14+2) on an array with fewer than eight drive chassis. The *-ha mag* option can be used to specify that the volume must survive the failure of any two magazines, though not necessarily the failure of even one drive chassis.

Dynamic Optimization: tuning volume configurations

HP 3PAR Dynamic Optimization Software is an optional product that allows the underlying characteristics of a volume to be changed transparently and without disruption of service. Drive type, RAID level and configuration, and high availability options can all be changed simply, easily, and non-disruptively. Unlike traditional arrays, where a poor initial choice can be difficult and/or costly to change, Dynamic Optimization allows changes to be made easily.

Dynamic Optimization also makes it easy for IT personnel to adapt to the changing needs of a modern, dynamic computing environment. For example, a previously high-priority project that used RAID 1 on high-performance Fibre Channel disks could be moved to more cost-effective RAID 5 storage on SATA disks.

Another use of Dynamic Optimization is to redistribute volumes after adding drives to an HP 3PAR Utility Storage array. Using Dynamic Optimization, existing volumes are autonomically striped across existing and new drives for optimal volume performance after capacity expansions. The increase in the total disks for the provisioned volume contributes to higher performance.

Adaptive Optimization: autonomic sub-volume tuning

HP 3PAR Adaptive Optimization Software is a fine-grained, policy-driven, autonomic storage software solution that delivers service level optimization for enterprises and cloud datacenters at the lowest possible cost while increasing agility and minimizing risk. Adaptive Optimization analyzes performance (access rates) for sub-volume regions, then selects the most active regions (those with the highest I/O rates) and uses the proven sub-volume data movement engine built into HP 3PAR InForm Operating System Software to autonomically move those regions to the fastest storage tier. It also moves less active regions to slower tiers to ensure space availability for newly-active regions.

Traditional storage arrays require the storage administrator to choose between slow, inexpensive storage and fast, expensive storage for each volume—a process that depends on the storage administrator's knowledge of the application's storage access patterns. Moreover, volumes tend to have hot spots rather than evenly-distributed accesses, and these hot spots can move over time.

Using Adaptive Optimization, an HP 3PAR array configured with SATA disks plus a small number of Solid State Drives (SSDs) can approach the performance of an all-SSD configuration at little more than the cost per megabyte of SATA-based storage, adapting autonomically as access patterns change.

Summary

HP 3PAR Storage Systems synergistically combine the unique features of the HP 3PAR Architecture to enable higher performance than traditional storage arrays with less hardware. Fast RAID allows more cost-efficient RAID levels to be used without sacrificing performance. Fast RAID reconstruction after a hardware failure reduces the risk to data, which also allows the use of more cost-efficient RAID levels without disruption or performance impact. HP 3PAR Dynamic Optimization and Adaptive Optimization Software allow RAID levels and other storage attributes of volumes to be altered in response to changing needs, without interruption.

For more information

Visit www.hp.com and www.hp.com/go/3PAR.

4AA3-4815ENW, Created June 2011