RAID for the 21st Century

A White Paper Prepared for Panasas
October 2007

Table of Contents

RAID in the 21st Century

Providing users and applications with high-performance, reliable data has always been a challenge, a challenge that RAID in its many forms has been addressing for 20 years. While RAID imposes a price penalty when compared against JBODs, RAID's ability to deliver high-performance I/O, while at the same time adding multiple levels of data protection, have made it an easy choice for IT managers. A new generation of disk drives is now changing the playing field. When RAID was first invented the capacity of disk drives was measured in hundreds of megabytes; today, terabyte-sized drives are beginning to appear. While larger drives offer tremendous advantages in terms of raw storage capacity, when placed in RAID environments they present new challenges. Large drives take much longer to rebuild than was the case with their smaller predecessors. Application performance degradation that occurs during a RAID rebuild is intolerable for any customer-facing or important operational system. Since the number of disk media failures expected during each read over the surface of a disk grows proportionately with the massive increase in density, media defects in these large drives increase the likelihood of a catastrophic RAID failure and loss of all data in the volume.

Fortunately, technologies developed for high performance computing provide a cost-efficient answer to this problem. Commercial sites that may be running thousands of concurrent users or processes can avoid the penalties these large disks impose, while at the same time taking advantage of their performance, scalability, and availability in an easily managed and cost-efficient fashion.

RAID 5 and RAID 6

Network storage systems are all about business, and while there is a need for high-performance data, the greater need is to maintain the integrity of the processes that use the data. To accommodate this, at most sites RAID 5 has been the popular choice. RAID 5 stripes data and parity across all disks in the RAID group at the block level allowing for the failure of a single drive. While some capacity (typically about 10%) is lost due to the parity stripe, RAID 5 provides very good read performance and high reliability at the expense of slower writes due to the calculations associated with the parity data.

RAID 5 systems have been prone to failure due to two principal causes. First, the disk drives now being used are built on much denser media than was available on earlier devices, and the unalterable consequence of more media is more media errors on each of these new larger drives. A secondary concern is the failure of a second device in a RAID 5 set during the reconstruction of a failed drive. While a relatively rare occurrence, losing a second drive within a single RAID 5 set is always catastrophic; when the second drive fails all data within the RAID set is irretrievably lost. RAID 5 devices that have lost a disk are therefore completely exposed.

The vulnerability of RAID 5 implementations that use the new larger disks is what drives interest in RAID 6, which adds a significant amount of protection through its ability to accommodate the failure of a second drive. RAID 6 adds this additional protection by using a second parity stripe. This of course imposes an additional capacity penalty on the RAID group, along with a substantial additional performance penalty. The complexity of the second parity calculation, which only protects against failure of the second disk and offers no protection against the failure of the first disk, slows write operations considerably. Protecting the second disk against the likelihood of failure imposes a full-time penalty on all write operations.

Penalties Associated with RAID 5 and RAID 6

When a drive fails in any RAID 5 or RAID 6 device, the entire drive must be rebuilt. RAID rebuilds have always presented IT managers with a difficult choice: do they pull the system off-line, allowing for comparatively fast rebuilds but making key data unavailable, or are RAID rebuilds done while the devices are still operational, in which case a severe and typically intolerable performance penalty is imposed and the possibility of a second drive failure is increased?

The new larger drives exacerbate the problem because in traditional RAID implementations rebuild times are constant, irrespective of disk drive size, and RAID sets with larger disks always require more time to rebuild than do sets using smaller disks. The fallout from this, in terms of both data availability (when the first choice is made) and I/O performance (in the case of the second choice), may be frightening. Consider for example the case of

storage implementations that use a single system-wide volume. These are vulnerable to failure anytime a RAID group within them fails. Because the entire multi-TB volume must be restored from tape if any RAID group within it fails, restore times at such sites could conceivably take days to correct such a failure.

How the Vendors Compensate

Vendor-specific answers to this problem are varied, and in some cases almost draconian. Isilon, for example, compensates for RAID 5 vulnerability by creating a monolithic RAID 6 failure domain. This provides good protection, but a single node failure might result in a 600 TB reconstruction taking a week to complete. And for large systems Isilon requires even higher RAID levels of N+3 or N+4 (triple or quadruple parity) to maintain acceptable levels of reliability. Network Appliance on the other hand (using RAID 4) has been reducing the size of its RAID groups in order to reduce the probability that they will encounter media errors. By imposing limitations on the sizes of their RAID sets, they can lower the probability of two drives failing in the same RAID group. Unfortunately, IT managers must be willing to accept the penalty of reduced performance and capacity utilization that comes with using smaller RAID 4 groups. Realizing this issue, NetApp has introduced RAID-DP (RAID 6) to address the vulnerability of RAID 4. Both methods provide RAID 6 protection, but because both utilize only traditional RAID architectures this protection comes at the expense of significant capacity, reliability and performance penalties.

EMA's Perspective

Fortunately, affordable high performance/high reliability RAID storage is also available using a technology that eliminates the downtime penalty. Aggregate performance, viewed here as a mix of traditional operational I/O, rebuild performance, and I/O during that rebuild, will be the key issue at commercial sites which cannot afford to have business-critical operations offline even briefly.

Just any RAID implementation will hardly be satisfactory however. Traditional RAID rebuild technologies require full disk reads and writes for each rebuild. Worse yet, should even one failed read operation prove to be unrecoverable during the rebuild, the entire rebuild operation can fail catastrophically. Few commercial sites can accept such a penalty. They need a method of compensating for the increased number and size of disks, which cause higher probability of encountering disk failure.

The Panasas approach and technology differs from that of their competitors in several key areas. Because of this Panasas storage is able to provide scalability, performance, reliability, rapid rebuilds, and significant ease-of-use. Fundamental to this is an approach that EMA views as *protecting data rather than the disks that hold the data*. Here is what we mean.

Architectural differentiators. Two key points that need to be understood about the Panasas solution are that it provides a two-tiered parity structure, and the fact that it provides RAID over objects (files and their metadata). This is in contrast to the traditional method of providing RAID over disks.

Panasas Tiered-Parity provides two tiers of data protection – horizontal and vertical. The horizontal parity tier provides protection across the array in a way that is similar in many respects to the more traditional RAID schemes currently in use. It is Panasas' vertical parity tier that provides significant differentiation from other vendors' approaches. The vertical parity tier provides what is essentially *RAID across sectors within individual disks*. This sector-based RAID within each disk eliminates the vast majority of unrecoverable read errors (UREs), and thereby removes nearly all the risk associated with secondary failures during a rebuild.

The Panasas approach is object-based, with each object holding both data and metadata. RAID protection levels are assigned on a per-object basis rather than on a per-disk basis, which will always deliver reduced reconstruction times should a failure occur. Why? Because while traditional systems must read and rewrite the entire disk in order to rebuild an array, the Panasas controller need only read the part of the disk that actually contains the invalidated data. The two-tiered parity scheme provides all the metadata necessary to reconstruct the array.

Why a storage administrator should care. Ninety-seven percent of RAID 4 and RAID 5 double disk failures occur due to unrecoverable read errors appearing during reconstructions. The Panasas vertical parity technology eliminates the likelihood of UREs, essentially reducing the likelihood of secondary disk failures to zero while

imposing no performance overhead. This protection against UREs does not vary when disk size increases.

The clustered architecture also provides clear performance benefits: throughput will exceed 20 GB per second and should scale in an essentially linear fashion as additional nodes are added. This contrasts with what is seen in most data centers today: administrators are used to seeing a RAID 6 performance penalty of 50% on Read/Modify/Write operations. There is no requirement for any dedicated spares in the Panasas approach.

Administrators can tune the system to provide specific protection levels for each object within the system, choosing RAID 1, RAID 5 or RAID 10 to optimize both protection and workload for each object within the system. Administrators also may now sharply define failure domains, which in turn will deliver dramatically shorter periods of degraded performance whenever a rebuild must be undertaken.

The bottom line. Enterprise Management Associates identifies the following as key benefits of this architecture:

- Scalability: linear in terms of both capacity and performance
- Performance: extreme bandwidth, with high random I/O capabilities
- Reliability: UREs, or media errors, are eliminated; the scope of each failure is limited to the object itself, and does not impact the whole disk
- No performance-reliability trade-off: object-based rebuilds take seconds rather than hours, while providing the same level of protection
- Scalable Reliability: Vertical Parity means larger disks are no more failure prone than smaller disks, allowing systems to scale without compromising data integrity and rebuild times
- Ease-of-use: the global namespace results in simplified management, which should limit operational expenditures

About Panasas

Panasas, Inc., the global leader in parallel storage solutions, helps commercial, government and academic organizations accelerate their time to results leading to real world breakthroughs that improve people's lives. Panasas' high-performance storage systems enable customers to maximize the benefits of Linux clusters by eliminating the storage bottleneck created by legacy network storage technologies. The Panasas ActiveStor Parallel Storage Clusters, in conjunction with the ActiveScale® Operating Environment and PanFS™ parallel file system, offer the most comprehensive portfolio of storage solutions for High Performance Computing (HPC) environments. Panasas is headquartered in Fremont, California. For more information, please visit www.panasas.com.

About Enterprise Management Associates, Inc.

Enterprise Management Associates is an advisory and research firm providing market insight to solution providers and technology guidance to Fortune 1000 companies. The EMA team is composed of industry respected analysts who deliver strategic awareness about computing and communications infrastructure. Coupling this team of experts with an ever-expanding knowledge repository gives EMA clients an unparalleled advantage against their competition. The firm has published hundreds of articles and books on technology management topics and is frequently requested to share their observations at management forums worldwide.

Corporate Headquarters:
5777 Central Avenue, Suite 105
Boulder, CO 80301
Phone: +1 303.543.9500
Fax: +1 303.543.7687
www.enterprisemanagement.com

1452.101507