



PANASAS® TIERED PARITY™ ARCHITECTURE

Larry Jones, Matt Reid, Marc Unangst,
Garth Gibson, and Brent Welch

White Paper | May 2010



Abstract

Disk drives are approximately 250 times denser today than a decade ago. This is good news for users who are creating, manipulating and storing more data than ever before. It gives them an opportunity to derive more value from their stored data and lowers the capital acquisition and operating expense associated with that data. However, while drive density has increased 250 times, drive interface speeds are only about 10 times faster. The result is that high capacity drives take much longer to reconstruct than the drives in early RAID systems, and the amount of data stored in a single RAID array can take days or weeks of downtime to restore from tape. These new denser disk drives have created substantial challenges for RAID. This white paper discusses how a new architectural approach is addressing these challenges and delivering industry-leading data integrity for large scale storage systems.

Overview

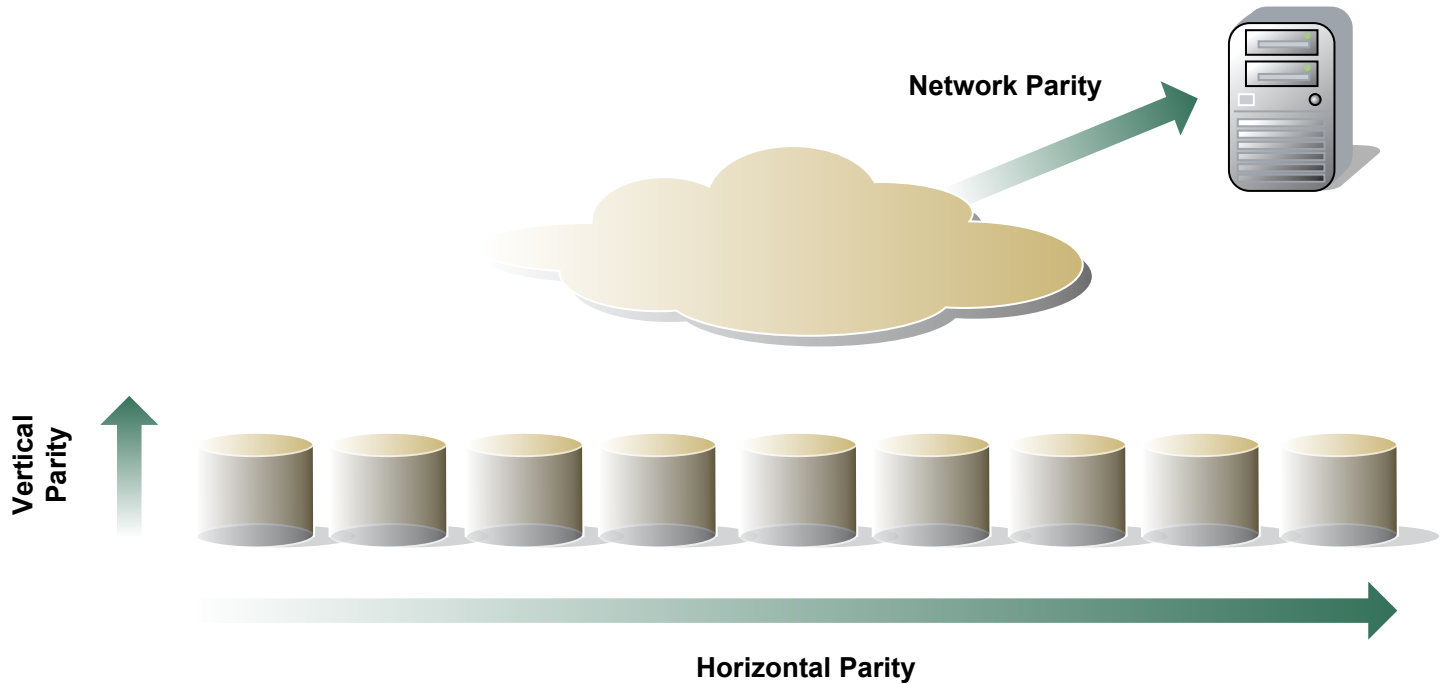
The Panasas Tiered Parity Architecture was designed to solve reliability challenges for large scale storage systems. These reliability challenges are driven by the increasing size and density of network storage systems and the increasing complexity of the data center. Tiered Parity is available in the PanFS storage operating system version 3.4 for the Panasas storage systems.

The Tiered Parity Architecture consists of three components: Horizontal Parity, Vertical Parity and Network Parity. Horizontal Parity is traditional RAID which was invented by Panasas, Inc. founder and CTO Garth Gibson and two colleagues at UC Berkeley in 1988. By striping data across an array of disks and calculating one or more parity blocks across that stripe, RAID is able to protect against the failure of one or more components in the array, providing very high levels of reliability.

The second component of protection as part of the Tiered Parity Architecture is Vertical Parity. Vertical Parity provides the ability to protect the data on a single disk from problems with the disk media itself. This component was designed to protect against reliability issues with the newest generations of ultra-dense disk drives (1TB per drive or greater). A way to conceptualize Vertical Parity is to think of it as providing RAID across sectors of each individual disk.

The third component of protection is Network Parity which extends the protection of your data from the disk drive right into the client system.

A storage system that has a comprehensive approach to managing data integrity provides a lot of benefit to end users. First is the ability to build larger storage systems without compromising reliability, allowing server consolidation and lowering the operational cost of providing storage. It also provides better availability of the data, which improves the productivity and competitiveness of the organization, and in general provides better service to their users.



Innovative end-to-end data integrity architecture addresses the root cause of disk reliability problems

RAID Worked Well in 1996, and Even Better Today

RAID was designed to keep a set of relatively inexpensive disks from losing data due to disk failures. By creating a parity block associated with the set of blocks striped across an array, the array controller could recreate on the fly the data lost due to a disk failure, recover the entire disk and provide full performance again in a matter of hours. If a second failure occurred, the worst case in 1996 was that a RAID failure meant recovering approximately 50GB of data from tape which could take a few hours to achieve.

Since RAID became the standard for data protection, much has improved. In particular, disk drive reliability has dramatically improved. In 1996, when RAID 5 data protection was the “gold standard,” the typical disk drive had an MTBF of 100,000 hours as specified by the manufacturer’s warranty. Today’s enterprise-class disks have a typical MTBF specification of more than 1.2 million hours, making them 10 times more reliable. From a purely statistical basis, RAID reliability should be 10x greater today.

Denser Disk Drives are Creating New Challenges for RAID

However, disk drives are approximately 250 times denser today than a decade ago. This is good news for users who are creating, manipulating and storing more data than ever before. It gives them an opportunity to derive more value from their stored data and lowers the capital acquisition and operating expense associated with that data. However, while drive density has increased 250 times, drive interface speeds have only gotten about 10 times faster. The result is that high capacity drives take much longer to reconstruct than the drives in early RAID systems, and the amount of data stored in a single RAID array can take days or weeks of downtime to restore from tape. These new denser disk drives have created substantial challenges for RAID.



Disk Media Errors are Now the Challenge

Disk drives are 250 times denser than a decade ago, providing great value to storage users. However the media error rates on those drives have remained constant over that decade at about 1 failed bit in 12.5 Terabytes. A decade ago, all the data on the drive would need to be read from end to end 3,000 times before it was likely to produce a media error. On today's 1TB drives, reading a drive just 12 times moves enough data that a media error is likely to occur. In addition, a decade ago it was typical for a disk drive to go its entire lifetime without a single grown media error. Today, it is common and expected for a disk drive to develop a small number of media errors over its lifetime.

In enterprise storage systems, RAID is typically used to recover from these errors. When an error is detected, the RAID controller uses data and parity from the other drives in the RAID set to rebuild the data in the bad sector, usually on-the-fly, and without requiring a complete reconstruction. But if a media error occurs when the controller is reconstructing a failed drive, the controller no longer has enough data and parity to rebuild both the failed drive and the media error. With a typical RAID 5 block-based controller, this will cause a catastrophic failure of the RAID set and result in the loss of that volume's data. A 10+1 RAID set requires that 10TB of data be read in order to reconstruct the failed drive, which is very close to the specified error rate of 1 media error in every 12TB read. The probability of a media error during a reconstruction causing a catastrophic failure is approaching 100%, and with the next generations of drives this problem will only get worse.

More Media = More Media Errors Per Drive

Panasas is not the first or only storage vendor to see the problems with increasing disk densities. When NetApp introduced their Double Parity feature, they noted:

“With modern larger disk media, the ability of traditional single-parity RAID to protect data is stretched past its limits”

“These factors – larger disks, unimproved reliability, and increased bit error rates with larger media – all have serious consequences for the ability of single-parity RAID to protect data.”

RAID-DP™: NetApp Implementation of RAID
Double Parity for Data Protection

Chris Lueth, Network Appliance, December 2005, TR 3298

Panasas Tiered Parity Architecture

Panasas has developed Tiered Parity architecture, a comprehensive architecture to deliver superior reliability and data integrity. This architecture provides three tiers of error detection and correction:

1. **Horizontal Parity**, also known as Panasas object-based RAID, is an innovative protection method in which the system computes parity for each file (object) and writes the data and parity across multiple disks. Thus, if there is a failed disk, the files can be reconstructed from the remaining disks. Panasas Object RAID uses two innovative techniques to make rebuilds faster and thus more robust. First,



Object RAID uses multiple RAID controllers to cooperate to rebuild a failed drive in parallel (“Parallel Reconstruction”) which provides a linear speedup of reconstruction performance resulting in dramatic improvements in rebuild time of 10X or more in larger storage systems. Second, Panasas Object RAID is more efficient, by reconstructing only user data rather than every sector (allocated or free), as is done in a typical block-based RAID controller. Object RAID rebuilds the lost file, not the entire drive. As a real-world example these two innovations allow a Panasas system at Los Alamos National Laboratory to rebuild a failed 800GB Panasas StorageBlade in about 30 minutes which might take a traditional RAID over 12 hours to complete! This means that a Panasas storage system is back at full performance with complete data protection faster than any other system available.

2. **Vertical Parity** is a new and unique-to-Panasas technology that solves the media error problem regardless of drive density. It provides a second level of RAID within an individual StorageBlade allowing the StorageBlade to independently and transparently recover from media errors without using the Horizontal Parity information. In effect, Vertical Parity improves on the internal ECC capabilities of the drive itself by at least 10X, in a way that is independent of and complementary to horizontal array-based parity schemes like Object RAID. When a media error does occur, the Vertical Parity capability is used to provide seamless detection and recovery. When no media errors occur, the system still verifies the Vertical Parity in order to catch cases where a drive returns incorrect data. Otherwise silent errors are detected by the Vertical Parity encoding, and repaired using redundant information in the horizontal RAID stripe. Other vendors detect parity or CRC errors in data as it is being read back from the drive; the Panasas architecture goes the extra step to correct any errors that are found before they are transmitted to the user.
3. **Network Parity** is another unique-to-Panasas data protection system that extends error detection across the full data path between storage and the client or server node. The optional Network Parity feature in the DirectFLOW® client software examines the data that it has received over the network and compares it against the stored on-disk parity to ensure correctness. This enables end-to-end validation of data integrity. Other storage systems can detect errors within their system at best and are powerless to detect errors introduced in the network or on the client node. The Panasas Network Parity takes the next step to protect user data from errors introduced by disks, firmware, server hardware, server software, network and transmission components. The application on the client either receives valid data or an error notification allowing the “read” to be retried and if necessary, the data path corrected to eliminate the errors.

Panasas Tiered Parity Architecture Protects Data Better Than RAID 6

Other vendors typically use some flavor of RAID 6 (RAID-DP, Reed-Solomon, etc.) to protect against a media error during rebuild. However, RAID 6 imposes high costs both during normal operation and during a rebuild. RAID 6 doubles the parity overhead of the system, reducing usable capacity. It also significantly increases the effective RAID stripe size, making higher-overhead “small write” updates more frequently and reducing performance for many workloads. The algorithms used for many RAID 6 implementations are more complex than XOR-based parity, slowing writes and rebuilds.



In addition, although correcting simultaneous drive failures is often cited as an important advantage by advocates of RAID 6, the amount of data that must be read to rebuild from this type of failure virtually guarantees that a media error will be encountered on one of the surviving drives, causing the rebuild to fail. Therefore, although RAID 6 is theoretically capable of handling a double catastrophic failure, in practice it only protects against a media error during the rebuild of a single failed drive. This weakness will only continue to grow as drives become larger (3 TB and 4 TB drives), the probability of a triple failure will double as compared to the 2 TB drives available today.

The Tiered Parity protection technique (Horizontal, Vertical and Network Parity) provides equivalent protection to RAID 6 from the combination of a disk failure and a media error - which is by far the most common form of double failure in a RAID system. The probability of suffering two simultaneous drive mechanism failures continues to diminish as drives have become 10 times more reliable in the last decade. Object RAID further diminishes the window of vulnerability to failure using Parallel Reconstruction to achieve dramatically times faster rebuild.

By seamlessly correcting media failures in place, Tiered Parity architecture delivers several important advantages over RAID 6 or other multiple RAID parity approaches:

- Provides the only scalable architecture that delivers full protection against media problems, even as drive densities continue to increase.
- Provides an additional level of assurance that the data read is exactly what was written.
- Ensures any errors encountered during data being read from the disk are automatically and seamlessly corrected.
- Eliminates the 50% performance impact to small write performance of RAID 6.
- Eliminates the 12 to 24 hour performance impact caused by reconstruction of a second drive.
- Cuts in half the number of drives that must be replaced, lowering administrative costs and TCO.

Panasas Tiered Parity Architecture Summary

Tiered Parity architecture is a comprehensive approach to improving reliability & data integrity. It is an architectural approach to addressing every layer of the storage system to deliver industry-leading data integrity.

Disk Drive – Vertical Parity

- Parity to detect whether the data on the drive is correct when it is read back
- Automatic and seamless correction of errors caused by media defects on the drive

RAID Array – Horizontal Parity

- Parallel reconstruction to improve RAID rebuild performance times dramatically (>10X) which improves reliability and application performance
- Per file RAID eliminates the rebuild of unused disk space, improving RAID rebuild times and thus reliability and system performance



Data path – Network Parity

- Detects errors induced by any component between the disk drives and the end-system at any time between when the data was written and when it is read back

Other systems have only one tool (RAID6) to solve integrity issues and attempt to use it to solve all issues; a strategy which is incomplete today and will become more inadequate over time. Tiered Parity architecture addresses the real challenge in building reliable large scale storage systems – media errors that cause reconstruction to fail.

Tiered Parity architecture is the only comprehensive solution to data integrity for storage systems.

It is only available with Panasas Storage Systems.

Glossary

MEDIA ERROR

A disk drive failure condition that causes a persistent error when reading back a previously written sector, but does not cause catastrophic failure of the drive. May be caused by any of a number of faults, including latent damage to the platter surface, a prior “high-fly write”, or a head positioning error.

RAID

Redundant Array of Independent Disks: A methodology that involves storing redundant parity information for use in recovering data in the event that a disk fails.

RECONSTRUCTION

The process of recreating lost data after a disk fails.

About Panasas

Panasas, Inc., the leader in high-performance scale-out NAS storage solutions, enables enterprise customers to rapidly solve complex computing problems, speed innovation and bring new products to market faster. All Panasas solutions leverage the patented PanFS™ storage operating system to deliver exceptional performance, scalability and manageability.



| Phone: 1-888-PANASAS

| www.panasas.com