Infortrend®

Using RAID6 for Advanced Data Protection

Table of Contents

The Challenge of Fault Tolerance

RAID technology is used to protect data as well as provide more efficient data access and capacity utilization with disk-based subsystems. By placing data on multiple disks, I/O (input/output) operations can overlap in a balanced way, improving performance. Using multiple disk drives increases the Mean Time Between Failure (MTBF), while storing data redundantly and this increases the fault-tolerance of the entire system. To meet requirements such as utilization, performance, and data protection, there are several types of RAID implementation including RAID0, RAID1, RAID5, and combinations such as RAID10 and RAID50. RAID0 provides the highest performance but no redundancy while RAID1 mirrors the data stored in one hard drive to another and can only be performed with two hard drives. With RAID5, parity data is not stored in a dedicated hard drive and in the event of a drive failure, the controller can recover/regenerate the lost data of the failed drive by comparing and re-calculating data on the remaining drives. RAID preferences greatly depend on the user's needs and budget.

New technologies have led to the availability of inexpensive, high capacity disk drives in SAS (Serial Attached SCSI) and SATA (Serial ATA) formats, allowing storage managers to build low cost, high capacity arrays to protect their data. However, these affordable options also pose their own risks as SATA drives fail more often than Fibre Channel (FC) or SCSI drives. In a typical RAID5 implementation, one drive fails and another will take over. Failure of two drives at once would cause loss of data and system downtime. Therefore, when these less reliable drives are common, advanced protection is needed to guard against multiple drive failures and provide fault tolerance and high availability of data.

A Compelling Technology: RAID6

RAID6 is essentially an extension of RAID5 that allows for additional fault tolerance by using a second independent distributed parity scheme (dual parity).

Data is striped on a block level across a set of drives, just like in RAID5, except a second set of parity is calculated and written across all the drives. Because RAID5 uses only one set of parity codes, data will be permanently lost if another disk fails or access errors occur during a long rebuild time. With the protection of RAID6, data can be reconstructed even if a second drive fails. With minimal performance impact and capacity consumption, RAID6 provides for extremely high data fault tolerance and can sustain multiple simultaneous drive failures.

RAID5 and RAID6 Comparison

	RAID5	RAID6
Parity	Single	Dual
Protection	One drive	Two drives
Implementation	Requires N+1 drives, minimum of 3	Requires N+2 drives, minimum of 4
Performance	Medium impact on write operation and rebuild	More impact on sequential write operation vs. RAID5

Why Use RAID6

As described above, RAID5 protects against a single drive failure without downtime; however, if a second drive goes down, data will be lost. While two drive failures are less likely than one failure, the probability is an increasing concern based on the factors listed below:

1. *Growing adoption of less reliable drives*: The benefits of SATA drives include lower prices and high capacity; however, their MTBF is lower than FC or SCSI drives. The increased use of these drives increases the probability of two drives failing at the same time.

2. *High capacity means longer rebuilding time:* Greater capacity on a single drive means more time is needed to rebuild data if one drive fails. Subsystems suffer heavy loading during the rebuild process and it's therefore more likely to damage another drive or for a second drive failure to occur during this longer rebuilding time.

3. *Human error:* When one drive fails, someone has to replace the damaged drive with a new one. An error, such as removing the wrong drive, can create other drive failures and data would be lost.

4. *The failure rate significantly rises as the number of hard drives increases:* Increasing the number of hard drives in an array effectively raises the expected failure rate of the first hard drive. During system recovery using a spare drive, the failure rate of the second hard drive is also multiplied. A subsystem composed of multiple hard drives needs added protection to guarantee data availability in case two drives fail simultaneously.

Given the greater chance that two drives could fail simultaneously, it is clear to see the appeal of implementing RAID6 in a storage array.

How RAID6 Works

RAID6 algorithms perform with two parity data, P and Q, using two linear independent equations. The first parity data P, the same as the parity data in RAID5, is performed by the equation:

$$P = D_0 \oplus D_1 \oplus D_2 \oplus \cdots \oplus D_{n-1}$$

The second parity data Q is generated by the equation:

$$Q = (A_0 \otimes D_0) \oplus (A_1 \otimes D_1) \oplus (A_2 \otimes D_2) \oplus \cdots \oplus (A_{n-1} \otimes D_{n-1})$$

Where $\oplus$ means XOR operation,

$\otimes$ means RAID6 multiplication,

Di is the data block, and

Ai is a coefficient.

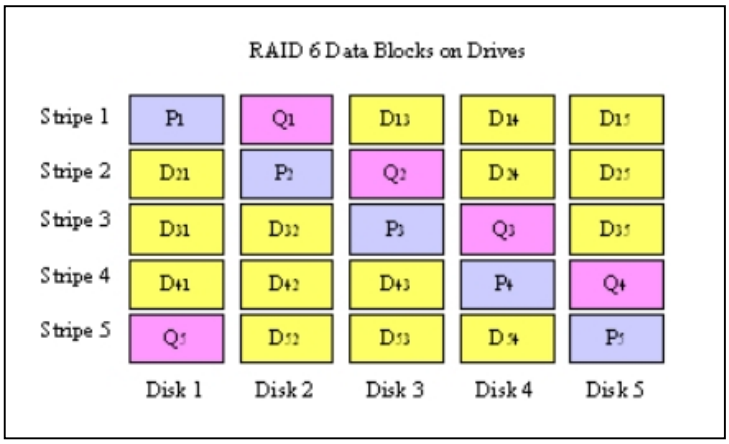

Figure 1

For example, according to ***Figure 1***, parity data P1 *(1)* will be:

$$P_1 = D_{13} \oplus D_{14} \oplus D_{15} \quad (1)$$

and parity Q1 *(2)* will be:

$$Q_1 = (\alpha \otimes D_{13}) \oplus (\beta \otimes D_{14}) \oplus (\gamma \otimes D_{15}) \quad (2)$$

where α, β, γ are specific numbers.

RAID6 Recovery

Recovery of P and Q data failure

The rebuild procedure will regenerate parity data P and Q from the data blocks.

Recovery of P and single data block failure

The rebuild procedure will use parity data Q and all remaining data blocks to rebuild the failed data block. After the failed data block is rebuilt, parity data P can be regenerated from all data blocks, the same as the generating method of parity data P.

For example, assume parity data P1 and data block D13 in ***Figure 1*** have failed.

$$Q_1 = (\alpha \otimes D_{13}) \oplus (\beta \otimes D_{14}) \oplus (\gamma \otimes D_{15})$$

$$D_{13} = \alpha^{-1} \otimes [Q_1 \oplus (\beta \otimes D_{14}) \oplus (\gamma \otimes D_{15})]$$

Parity data can be recomputed using the normal operation:

$$P_1 = D_{13} \oplus D_{14} \oplus D_{15}$$

Recovery of Q and single data block failure

The rebuild process is similar to a RAID5 recovery. The failed data block will be rebuilt from parity data P and all remaining data blocks. This process is like the RAID5 rebuild process. After the data block has been rebuilt, parity data Q can be regenerated.

For example, assume parity data Q1 and data block D13 in ***Figure 1*** have failed. The data in D13 can be rebuilt from parity data P1 and other data blocks.

$$P_1 = D_{13} \oplus D_{14} \oplus D_{15}$$
$$D_{13} = P_1 \oplus D_{14} \oplus D_{15}$$

Then, parity data Q1 can be regenerated by the normal operation.

$$Q_1 = (\alpha \otimes D_{13}) \oplus (\beta \otimes D_{14}) \oplus (\gamma \otimes D_{15})$$

Recovery of two data block failures

The rebuild procedure to recover data from two failed data drives is the most complicated case. From the parity data generating equations, there are two equations and two unknowns. Using matrix inversion however, the two unknowns can be solved and the lost data can be rebuilt.

Assume data blocks D13 and D14 have failed.

From the parity data generating equations:

$$P_1 = D_{13} \oplus D_{14} \oplus D_{15}$$
$$Q_1 = (\alpha \otimes D_{13}) \oplus (\beta \otimes D_{14}) \oplus (\gamma \otimes D_{15})$$

come the following two equations:

$$D_{13} \oplus D_{14} = P_1 \oplus D_{15} \quad \textit{(3)}$$

and

$$(\alpha \otimes D_{13}) \oplus (\beta \otimes D_{14}) = Q_1 \oplus (\gamma \otimes D_{15}) \quad \textit{(4)}$$

The following is calculated from equations *(3)* and *(4)*:

$$\begin{bmatrix} 1 & 1 \\ \alpha & \beta \end{bmatrix} \begin{bmatrix} D_{13} \\ D_{14} \end{bmatrix} = \begin{bmatrix} P_1 \oplus D_{15} \\ Q_1 \oplus (\gamma \otimes D_{15}) \end{bmatrix}$$

Using the matrix inversion, the two unknowns, D13 and D14, are solved.

$$\begin{bmatrix} D_{13} \\ D_{14} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \alpha & \beta \end{bmatrix}^{-1} \begin{bmatrix} P_1 \oplus D_{15} \\ Q_1 \oplus (\gamma \otimes D_{15}) \end{bmatrix}$$

RAID Configuration Tradeoffs

When configuring a storage array, system managers must balance their needs for high performance and data security. For example, if high performance is a priority, RAID10 is the best option. For increased data protection, RAID5 is an excellent choice, but as described in this paper, RAID6 provides valuable extra protection against the loss of data due to a second drive failure in large capacity disk drives, especially for SATA drives used in enterprise environments. There is a performance loss compared to RAID5 but this can be an acceptable tradeoff for the improved data safety.

A Balanced Solution: Infortrend RAID6

Infortrend has developed an efficient new RAID6 scheme for its RAID subsystems that provides the highest fault tolerance with minimal loss of performance. The RAID6 array can tolerate the failure of more than one disk drive; or, in the degraded mode, one drive failure and bad blocks on the other. In the event of disk drive failure, the controller can recover/regenerate the lost data of the failed drive(s) without interruption to I/O access.

Compared to existing RAID6 solutions, Infortrend's RAID6 technology delivers significant performance improvements and features. Users will get twice the parity protection with only a 10 to 15 percent performance loss compared to RAID5.

Summary

RAID6 provides an extremely high level of fault tolerance and can sustain two simultaneous drive failures without downtime or data loss. This is a perfect solution for the rigorous fault tolerant requirements of mission-critical data.