Assessing the Reliability of RAID Systems

By Abraham Long, Jr.

To determine the overall reliability of a RAID-based storage system, it is important to accurately assess the reliability of the RAID subsystem. This article presents a method for determining the probability of data loss for different RAID configurations on a Dell® PowerVault® system.

To meet an overall reliability objective, the individual components of a storage system must meet certain reliability requirements. For a system that includes a RAID (redundant array of independent disks) array, the hard disk drives (HDDs) in the array are allocated a reliability expectation. However, the reliability expectation of the RAID array is more than a uniform allocation of the HDD reliability expectation. Each RAID level has different distribution patterns that affect performance and redundancy. A proper assessment of RAID reliability must account for these differences.

Modeling the reliability of RAID sets

This article presents mathematical models of the relationship between HDDs in a RAID set. Using these models, different types of RAID sets for a Dell® PowerVault® 660F/224F are evaluated in terms of their impact on the overall reliability of a disk array. The PowerVault 660F/224F is a low-cost, scalable, external Fibre Channel storage subsystem that supports a total of 14 one-inch HDDs in a three-unit (3U) rack height. Its disk array can be set up at RAID levels 0, 1, 0+1, 3, and 5; these RAID levels, as well as levels 4 and 10, are examined in this article.

Although a single disk array can contain multiple RAID sets of different RAID levels and disk capacities, the examples in this article are based on an array of 14 HDDs dedicated to one RAID level at a time. All examples use a hypothetical 40 GB HDD possessing a reliability figure-of-merit of .90 over three years with a 100 percent duty cycle.

RAID-0: Data striping

RAID-0 involves striping, which is the distribution of data across multiple disk drives in equally sized chunks. For example, a 150 KB file can be striped, or chunked, across ten 15 KB chunks. The RAID set of striped disks appears as a single, logical disk to the operating system.

Striping provides a low-cost method to increase the I/O performance of the disks. However, RAID-0 does not provide any data redundancy; that is, if one drive in the RAID set fails, all data is lost.

Consider an array of six HDDs in a RAID-0 configuration. From the perspective of a reliability block diagram (see Figure 1), the HDDs are considered to be in series. The mathematical relationship that evaluates the reliability of a six-disk RAID-0 array is simply the product of the individual HDD reliabilities:

$$R_{\text{RAIDSET}} = \prod_{i=1}^{6} R_{\text{HDD}_i}$$

Or, for n HDDs,

$$R_{\text{RAIDSET}} = \prod_{i=1}^{n} R_{\text{HDD}_i}$$

Figure 1. Six HDDs in a RAID-0 configuration

For a PowerVault 660F/224F with all 14 HDDs in a RAID-0 disk set, the reliability of the RAID set is

$$R_{\text{RAIDSET}} = \prod_{i=1}^{14} R_{\text{HDD}_i} = \prod_{i=1}^{14} (.9) = .2288$$

This result indicates that the probability of no data loss over three years is 23 percent. Conversely, the probability of losing data over the same period is 77 percent.

RAID-1: Disk mirroring and duplexing

RAID-1 uses mirroring, or shadowing: all data written on a given disk is duplicated on another disk. RAID-1 requires at least two HDDs to implement and consists of paired disks; each pair is considered one RAID set. For example, in a RAID-1 array of three HDDs, HDDs 1 and 2 can mirror data while HDD 3 can be designated as a failover drive (for hot swap in case of failure). A RAID-1 array of four HDDs reduces to two RAID-1 RAID sets, a RAID-1 array of six HDDs reduces to three RAID-1 RAID sets, and so forth.

Mirroring supplies data redundancy and improved read performance. In a RAID-1 configuration, one HDD can fail in a paired set without loss of data. However, if both drives in the same paired set fail, data will be lost. Figure 2 shows the reliability block diagram for six HDDs in a RAID-1 array. The mathematical relationship that evaluates the reliability of this RAID array is

$$R_{\text{array}} = \prod_{i=1}^{\#\ of\ RAID\ sets} [(1 - (1 - R_{\text{HDD}_1})(1 - R_{\text{HDD}_2}))]$$

Note that the number of RAID sets is three in this case. If the HDDs are identical, then the relationship is

$$R_{\text{array}} = \prod_{i=1}^{\#\ of\ RAID\ sets} (2R_{\text{HDD}} - R_{\text{HDD}}^2)$$

Figure 2. Six HDDs in a RAID-1 configuration

For a PowerVault 660F/224F with all 14 HDDs configured as a RAID-1 array (seven RAID sets), the reliability of the array is

$$R_{\text{array}} = \prod_{i=1}^{7} (2(.9) - .9^2) = .9321$$

The result indicates a 93 percent probability of no data loss over three years. Conversely, there is a 7 percent probability of losing data over the same period.

Disk duplexing

Some RAID-1 configurations use disk duplexing, where each HDD is connected to its own SCSI channel. This setup provides additional redundancy and increases the speed of read/write operations. However, the SCSI channel is now in series with the HDD to which it is connected (see Figure 3), so the reliability of the SCSI channel becomes a factor in the equation. Therefore, in this configuration, $R_{\text{HDD}} = R_{\text{SCSI}} \times R_{\text{HDD}}$. Assuming that the SCSI reliability is .99, then $R_{\text{HDD}} = (.99)(.9) = .891$.

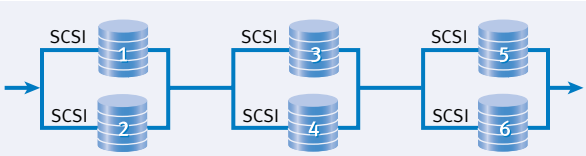

Figure 3. SCSI channels in series with connected HDDs

For a PowerVault 660F/224F with all 14 HDDs configured as a RAID-1 array (seven RAID sets) with disk duplexing, the reliability of the array is

$$R_{\text{array}} = \prod_{i=1}^{7} (2(.891) - .891^2) = .9197$$

The result indicates a 92 percent probability of no data loss over three years. Conversely, there is an 8 percent probability of losing data during the same period.

RAID-0+1: Mirror of stripes

In RAID-0+1, data is striped to one disk set and then mirrored to another disk set, resulting in good I/O performance and reliability. If a drive in one disk set fails, that disk set is lost, but all data will remain available from the mirrored disk set. However, if any HDD in the remaining disk set (the mirror) fails before the first disk set is restored, all data is lost.

RAID-0+1 requires a minimum of four HDDs to implement. Figure 4 shows the reliability block diagram for eight HDDs in a RAID-0+1 array with two RAID sets.

Figure 4. Eight HDDs in a RAID-0+1 configuration

Assuming all HDDs are identical, the mathematical relationship that evaluates the reliability of an array of HDDs in a RAID-0+1 configuration is

$$R_{\text{array}} = \prod_{i=1}^{\#\ of\ RAID\ sets} [(1 - (1 - R_{\text{HDD}}^2)(1 - R_{\text{HDD}}^2))]$$

For a PowerVault 660F/224F with 14 HDDs, one possible configuration is 12 HDDs dedicated to a RAID-0+1 array (three RAID sets) with the remaining two HDDs available for failover. The reliability of this configuration is

$$R_{\text{array}} = \prod_{i=1}^{3} [(1 - (1 - .9^2)(1 - .9^2))] = .8956$$

The result indicates a 90 percent probability of no data loss over three years. Conversely, there is a 10 percent probability of losing data during the same period.

RAID-3: Bit-level data striping with dedicated parity

In RAID-3, the RAID controller calculates parity (error correction) information and stores it to a dedicated parity HDD. Data is striped in byte- or bit-sized chunks to the remaining HDDs.

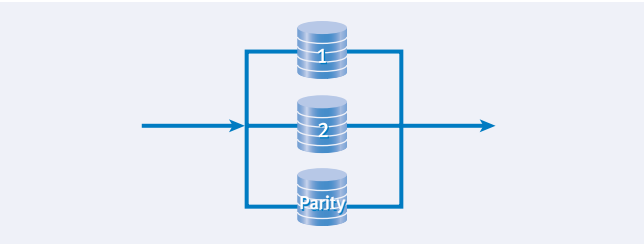

Figure 5. Three HDDs in a RAID-3 configuration

RAID-3 provides a high data-transfer rate; however, write times are slower because parity information needs to be written on the single parity drive each time a write occurs.

RAID-3 can tolerate one HDD failure in an array of *n* HDDs. For example, if the parity HDD fails, the remaining data HDDs are not affected, but redundancy is lost. If a data HDD fails, the RAID controller uses the remaining data HDDs and the parity HDD to recalculate the missing data on the fly. System performance slightly degrades until the failed HDD is replaced; however, no data is lost. All data in the RAID set will be lost if another HDD fails before the failed HDD is restored.

RAID-3 requires a minimum of three HDDs to implement. Figure 5 shows the reliability block diagram for three HDDs in a RAID-3 disk set.

The mathematical relationship that evaluates the reliability of *n* HDDs in a RAID-3 configuration is

$$R_{\text{array}} = \sum_{j=k}^{n} \binom{n}{j} R_{\text{HDD}}^{j} (1 - R_{\text{HDD}})^{(n-j)}$$

where *k* is the number of HDDs that must operate out of *n* HDDs.

For a PowerVault 660F/224F with 14 HDDs, one possible configuration is 13 HDDs dedicated to RAID-3 with the remaining HDD available for failover. Figure 6 shows the reliability calculation for this configuration, in which 12 of 13 HDDs must operate.

The result indicates a 62 percent probability of no data loss over three years. Conversely, the probability of losing data during the same period is 38 percent.

RAID-4: Data striping with dedicated parity

RAID-4 is identical to RAID-3, except that RAID-4 accommodates larger chunks. Here again, one HDD failure is tolerated; that is, when one HDD fails, all data is still fully available.

RAID-4 requires a minimum of three HDDs to implement. The reliability block diagram (see Figure 5) and the mathematical relationship that evaluates the reliability of *n* HDDs in a RAID-4 configuration are identical to the ones for RAID-3.

For a PowerVault 660F/224F with 14 HDDs, one possible configuration is 13 HDDs dedicated to RAID-4 with the remaining

$$R_{\text{array}} = \sum_{j=12}^{13} \binom{n}{j} R^{j}_{\text{HDD}} (1 - R_{\text{HDD}})^{(n-j)} = \frac{13!}{12!\,(13-12)!} .9^{12} (1 - .9)^{(13-12)} + \frac{13!}{13!\,(13-13)!} .9^{13} (1 - .9)^{(13-13)}$$

$$R_{\text{array}} = .3672 + .2542 = .6213$$

Figure 6. Reliability calculation for a RAID-3 configuration

HDD available for failover. The reliability calculation for this configuration, in which 12 of 13 HDDs must operate, is the same as the one for RAID-3. The result indicates a 62 percent probability of no data loss over three years. Conversely, the probability of losing data during the same period is 38 percent.

RAID-5: Data striping with striped parity

RAID-5 is similar to RAID-4 except that the parity data is striped across all HDDs instead of written on a dedicated HDD, eliminating the single parity drive as a bottleneck. Here again, when one HDD fails, all data is still available; the missing data is recalculated from the remaining HDDs and parity information. RAID-5 requires a minimum of three HDDs to implement. The reliability block diagram (see Figure 5) and the mathematical relationship that evaluates the reliability of *n* HDDs in a RAID-5 configuration are identical to the ones for RAID-3.

For a PowerVault 660F/224F with 14 HDDs, one possible configuration is 13 HDDs dedicated to RAID-5 with the remaining HDD available for failover. The reliability calculation for this configuration, in which 12 of 13 HDDs must operate, is the same as the one for RAID-3. The result indicates a 62 percent probability of no data loss over three years. Conversely, there is a 38 percent probability of losing data during the same period.

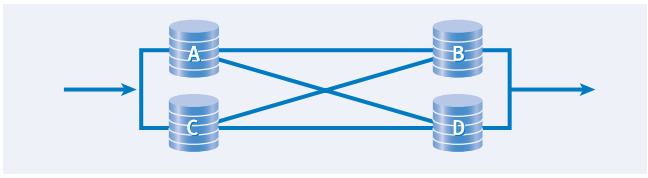

Figure 7. Four HDDs in a RAID-10 configuration

In a RAID-1 configuration, one HDD can fail in a paired set without loss of data.

RAID-10: Stripe of mirrors

RAID-10 combines the mirroring of RAID-1 with the data striping of RAID-0; specifically, data is striped across mirrored sets of drives. During an HDD rebuild, a RAID-10 system performs better than a parity-based RAID system because the missing data is not regenerated from parity information; instead, the data is copied from a surviving drive.

Only one HDD in a mirrored set of a RAID-10 array can fail without any data loss. As long as each mirrored set contains one active drive, all data is still available. However, if both HDDs in a mirrored set fail, all data is lost.

RAID-10 requires a minimum of four HDDs to implement. Figure 7 shows the reliability block diagram for a RAID-10 disk set consisting of four HDDs.

The mathematical relationship that evaluates the reliability of a RAID-10 configuration is

$$R_{\text{array}} = \prod_{i=1}^{\#\ of\ RAID\ sets} (R_{\text{disk set}_i})$$

$$\text{Let } R_{\text{disk set}} = P(AB \cup CD \cup AD \cup CB)$$

The reliability of one RAID-10 disk set $R_{\text{disk set}}$ is .98, as calculated in Figure 8.

For a PowerVault 660F/224F with 14 HDDs, one possible configuration is 12 HDDs dedicated to RAID-10 with the two remaining HDDs dedicated to failover. The reliability of this array, which contains three RAID-10 disks sets, is

$$R_{\text{array}} = (.98)^3 = .94$$

The result indicates a 94 percent probability of no data loss over three years. Conversely, the probability of losing data during the same period is 6 percent.

(+) One term at a time: $+ P(AB) + P(CD) + P(AD) + P(CB)$
(–) Two terms at a time: $- P(ABCD) - P(ABD) - P(ABC) - P(BCD) - P(ACD) - P(ABCD)$
(+) Three terms at a time: $+ P(ABCD) + P(ABCD) + P(ABCD) + P(ABCD)$
(–) Four terms at a time: $- P(ABCD)$

$$R_{\text{disk set}} = [P(AB) + P(CD) + P(AD) + P(CB) - P(ABCD) - P(ABD) - P(ABC) - P(BCD) - P(ACD) - P(ABCD) + P(ABCD) + P(ABCD) + P(ABCD) + P(ABCD) - P(ABCD)]$$

which reduces to

$$R_{\text{disk set}} = [P(AB) + P(CD) + P(AD) + P(CB) - P(ABD) - P(ABC) - P(BCD) - P(ACD) + P(ABCD)]$$
$$R_{\text{disk set}} = [.81 + .81 + .81 + .81 - .729 - .729 - .729 - .729 + .6561] = .98$$

Figure 8. Reliability calculation for a RAID-10 disk set

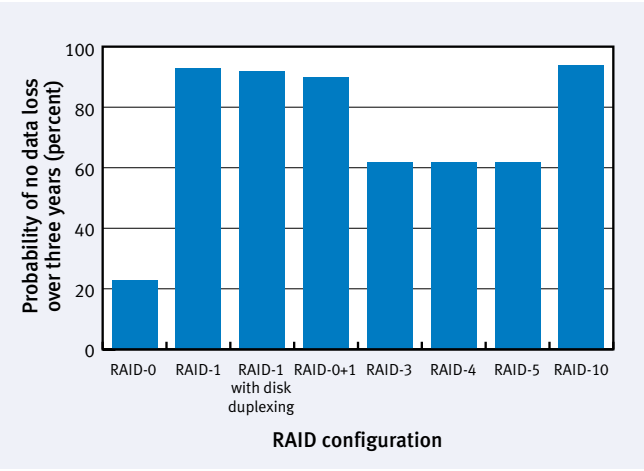

Figure 9. Probability of no data loss in a three-year period

Mirroring provides the most reliable RAID system

Figure 9 summarizes the reliability expectations calculated for RAID configurations on a PowerVault 660F/224 with 14 HDDs. These reliability expectations show that RAID levels using mirroring are less likely to fail than those using parity. RAID-10, a combination of striping and mirroring, has the highest probability (94 percent) of no data loss over three years. Because RAID-0 does not use mirroring, it has the highest probability (77 percent) of data loss.

From a reliability engineering point of view, the RAID probabilities should be combined with the probabilities of success from other system elements to develop figures-of-merit for overall system availability. In this manner, the influence of different RAID levels on the overall product platform can be evaluated against various marketing strategies.

Abraham Long, Jr. *(abraham_long@dell.com) is a senior engineering consultant on the Reliability Engineering Storage Products team in the Dell Enterprise Systems Group (ESG). Abraham is currently working on Linear Tape-Open™ (LTO™) and super digital linear tape (SDLT) tape-drive storage solutions and ESG power supplies for a variety of platforms. He has a B.S. in Industrial Engineering from New Mexico State University and an M.S. in Systems Engineering from California State University, Northridge. He is also an American Society for Quality Certified Reliability and Quality Engineer.*