Technical Report

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

Jay White & Chris Lueth, NetApp

May 2010 | TR-3298

ABSTRACT

This document provides an in-depth overview of the NetApp® RAID-DP® implementation for double parity–based data protection. Several aspects of RAID-DP are covered, including how it works, resiliency versus other RAID technologies, performance, and capacity utilization.

TABLE OF CONTENTS

1 INTRODUCTION

NetApp introduced double-parity RAID, named RAID-DP, starting with Data ONTAP® 6.5 in 2003, and since then it has become the default RAID group type used on NetApp storage. This document provides an overview of RAID-DP and how it dramatically increases data fault tolerance from various disk drive failure scenarios. Other key areas covered include how much RAID-DP costs (it's free), special hardware requirements (none), converting from existing RAID 4–based volumes to RAID-DP (it's easy), and performance characteristics. This document also presents a double–disk drive failure recovery scenario to show how RAID-DP both allows the volume to continue serving data while recreating data lost on the two failed disk drives.

2 WHAT IS THE NEED FOR RAID-DP?

As mentioned earlier, traditional single-parity RAID offers adequate protection against a single event, which could be either a complete disk failure or a bit error during a read. In either event, data is recreated using both parity and data remaining on unaffected disks in the array or volume. If the event is a read error, then recreating data happens almost instantaneously, and the array or volume remains in an online mode. However, if a disk fails, then all data lost on it has to be recreated, and the array or volume will remain in a vulnerable degraded mode until data has been reconstructed onto a spare disk. It is in reconstruct or degraded modes that traditional single-parity RAID shows that its protection capabilities have not kept up with modern disk architectures.

1.1 THE EFFECT OF MODERN LARGER DISK SIZES ON RAID

Modern disk architectures have continued to evolve, as have other computer-related technologies. Disk drives are orders of magnitude larger than they were when RAID was first introduced: most recently NetApp introduced a 2TB-sized disk in 2009. As disk drives have gotten larger, their reliability has not improved, and, more importantly, the bit error likelihood per drive has increased proportionally with the larger media. These three factors—larger disks, unimproved reliability, and increased bit errors with larger media—all have serious consequences for the ability of single-parity RAID to protect data.

Given that disks are as likely to fail now as when RAID technology was first introduced to protect data from such an event, RAID is still as vital now as it was then. When one disk fails, RAID simply recreates data from both parity and the remaining disks in the array or volume onto a spare disk. However, since RAID was introduced, the significant increases in disk size have resulted in much longer reconstruct times for data lost on the failed disk. Simply put, given the same rotational speed, it takes much longer to recreate data lost when a 274GB disk fails than when a 36GB disk fails. Compounding the longer reconstruct times is the fact that the larger disk drives in production use today tend to be SATA, which reconstruct more slowly and are slightly less reliable than smaller FC or SAS ones.

1.2 PROTECTION SCHEMES WITH SINGLE-PARITY RAID USING LARGER DISKS

The various options to extend the ability of single-parity RAID to protect data as disks continue to get larger are not attractive. The first is to continue to buy and implement storage using the smallest disk sizes possible so that reconstruction after a failed disk completes more quickly. However, this approach isn't practical from any point of view. Capacity density is critical in space-constrained data centers, and smaller disks result in less capacity per square foot. Moreover, storage vendors are forced to offer products based on what disk manufacturers are supplying, and smaller disks aren't readily available, if at all. The second way to protect data on larger disks with single-parity RAID is slightly more practical but, with the introduction of RAID-DP, a less attractive approach for various reasons. Namely, by keeping the size of RAID arrays small, the time to reconstruct is reduced. Continuing the analogy about a larger disk taking longer to reconstruct than a smaller one, a RAID array built with more disks takes longer to reconstruct data from one failed disk than one built with fewer disks. However, smaller RAID arrays have two steep costs that cannot be overcome. The first cost is that additional disks will be lost to parity, affecting usable capacity and total cost of ownership (TCO). The second cost is the extra overhead many smaller RAID groups cause on the controller, which in turn affects business and users.

The most reliable protection offered by single-parity RAID is RAID 1, or mirroring. In RAID 1, the mirroring process replicates an exact copy of all data on an array to a second array. While RAID 1 mirroring affords maximum fault tolerance from disk failure, the cost of the implementation is severe, since it takes twice the disk capacity to store the same amount of data. Earlier it was mentioned that using smaller RAID arrays to improve fault tolerance increases the total cost of ownership of storage due to less usable capacity per dollar spent. Continuing this approach, RAID 1 mirroring, with its unpleasant requirement for double the amount of capacity, is the most expensive type of storage solution with the highest total cost of ownership.

1.3 RAID-DP DATA PROTECTION

In short, given the rapid adoption of larger disk drives creating data protection challenges, customers and analysts demanded a better story about affordably improving RAID reliability from storage vendors. To meet this demand, NetApp released a new type of RAID protection named RAID-DP. RAID-DP stands for RAID double parity, and it significantly increases the fault tolerance from failed disk drives over traditional single-parity RAID. When all relevant numbers are plugged into the standard mean time to data loss (MTTDL) formula for RAID-DP versus single-parity RAID, RAID-DP is thousands of times more reliable on the same underlying disk drives. With this reliability, RAID-DP exceeds even RAID 10 mirroring for fault tolerance, but at RAID 4 pricing. RAID-DP offers businesses the most compelling total cost of ownership storage option without putting their data at an increased risk.

2 HOW RAID-DP WORKS

Traditional levels of existing RAID technology offer data protection through various approaches. The RAID used by NetApp, a modified RAID 4, stores data in horizontal rows, calculates parity for data in the row, then stores the parity in a separate row parity disk. Please see A Storage Networking Appliance for a more in-depth overview on how traditional RAID 4 works on NetApp storage. However, a constant across the different RAID levels, including the modified NetApp RAID 4, was that they used a single-parity scheme, which in turn limits their ability to protect past a single disk failure.

2.1 RAID-DP WITH DOUBLE PARITY

It is well known that parity generally improves fault tolerance and that single-parity RAID improves data protection. Given that traditional single-parity RAID has established a very good track record to date, the concept of double-parity RAID should certainly sound like a better protection scheme. But what exactly is RAID-DP with its double parity?

At the most basic layer, RAID-DP adds a second parity disk to each RAID group in an aggregate or traditional volume. A RAID group is an underlying construct that aggregates and traditional volumes are built upon. Each traditional NetApp RAID 4 group has some number of data disks and one parity disk, with aggregates and volumes containing one or more RAID 4 groups. Whereas the parity disk in a RAID 4 volume stores row parity across the disks in a RAID 4 group, the additional RAID-DP parity disk stores diagonal parity across the disks in a RAID-DP group. With these two parity stripes in RAID-DP, one the traditional horizontal, and the other diagonal, data protection is obtained even in the event of two disk failure events occurring in the same RAID group.

2.2 AN EXAMPLE OF HOW RAID-DP WORKS

With RAID-DP, the traditional RAID 4 horizontal parity structure is still employed and becomes a subset of the RAID-DP construct. In other words, how RAID 4 works on NetApp storage hasn’t been modified with RAID-DP. The same process, in which data is written out in horizontal rows with parity calculated for each row, still holds in RAID-DP and is considered the row component of double parity. In fact, if a single disk fails or a read error from a bad block or bit error occurs, then the row parity approach of RAID 4 is the sole vehicle used to recreate the data without ever engaging RAID-DP. In this case, the diagonal parity component of RAID-DP is simply a protective envelope around the row parity component.

RAID 4 HORIZONTAL ROW PARITY

Figure 1 illustrates the horizontal row parity approach used in the traditional NetApp RAID 4 solution and is the first step in establishing an understanding of RAID-DP and double parity.

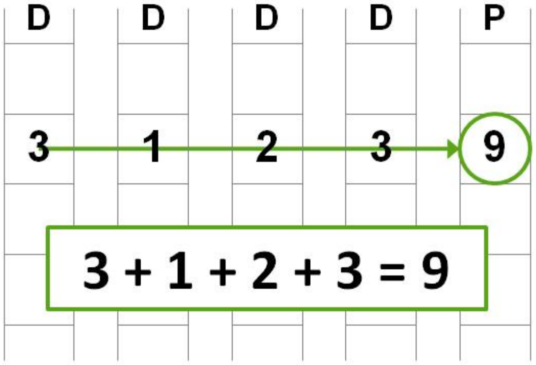


Figure 1) Simple RAID 4 parity.

The diagram represents a traditional RAID 4 group using row parity that consists of four data disks (the first four columns, labeled "D") and the single-row parity disk (the last column, labeled "P"). The rows in Figure 1 represent the standard 4KB blocks used by the traditional NetApp RAID 4 implementation. The second row in Figure 1 has been populated with some sample data in each 4KB block, and parity calculated for data in the row is then stored in the corresponding block on the parity disk. In this case, the way parity was calculated was to add the values in each of the horizontal blocks, then store the sum as the parity value (3 + 1 + 2 + 3 = 9). In practice, parity is calculated by an exclusive OR (XOR) process, but addition is fairly similar and works as well for the purposes of this example. If the need arose to reconstruct data from a single failure, the process used to generate parity would simply be reversed. For example, if the first disk were to fail, when RAID 4 recreated the data value 3 in the first column in Figure 1, it would subtract the values on remaining disks from what is stored in parity (9 – 3 – 2 – 1 = 3). This example of reconstruction with single-parity RAID should further assist with the conceptual understanding of why data is protected up to but not beyond one disk failure event.

ADDING RAID-DP DOUBLE-PARITY STRIPES

Figure 2 adds one diagonal parity stripe, denoted by the blue-shaded blocks, and a second parity disk, denoted with a "DP" in the sixth column, to the existing RAID 4 group from the previous section and shows the RAID-DP construct that is a superset of the underlying RAID 4 horizontal row parity solution.

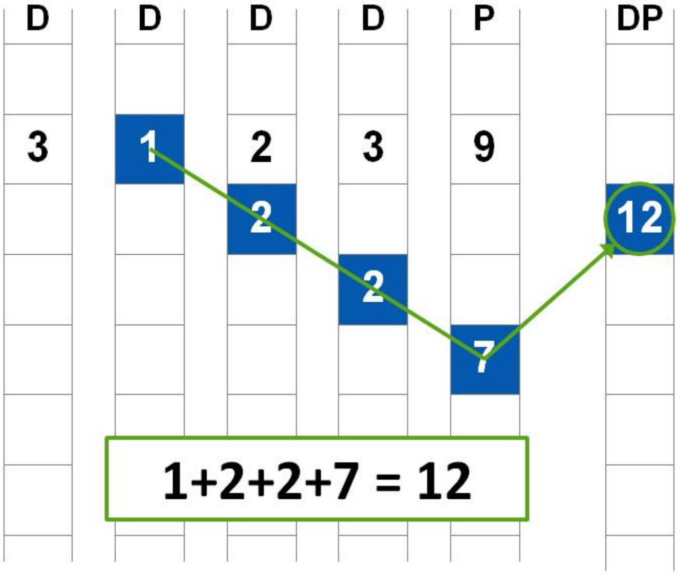


Figure 2) Add diagonal parity.

The diagonal parity stripe has been calculated using the addition approach for this example rather than the XOR used in practice as discussed earlier and stored on the second parity disk (1 + 2 + 2 + 7 = 12). One of the most important items to note at this time is that the diagonal parity stripe includes an element from row parity as part of its diagonal parity sum. RAID-DP treats all disks in the original RAID 4 construct, including both data and row parity disks, as the same. Figure 3 adds in the rest of the data for each block and creates corresponding row and diagonal parity stripes.

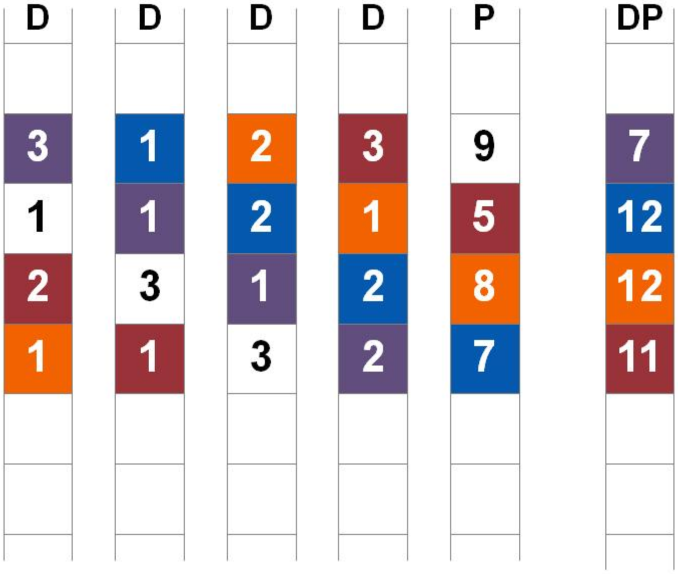


Figure 3) Numerical example of diagonal parity.

One RAID-DP condition that is apparent from Figure 3 is that the diagonal stripes wrap at the edges of the row parity construct. Two important conditions for the ability of RAID-DP to recover from double disk failures might not be readily apparent in this example. The first condition is that each diagonal parity stripe misses one and only one disk, but each diagonal misses a different disk. This results in the second condition, that there is one diagonal stripe that doesn't get parity generated on it or get stored on the second diagonal parity disk. In this example the omitted diagonal stripe is the white noncolored blocks. In the reconstruction example that follows it will be apparent that omitting the one diagonal stripe doesn't affect the ability of RAID-DP to recover all data in a double disk failure.

It is important to note that the same RAID-DP diagonal parity conditions covered in this example hold in real storage deployments that involve dozens of disks in a RAID group and millions of rows of data written horizontally across the RAID 4 group. And, while it is easier to illustrate RAID-DP with the smaller example above, recovery of larger size RAID groups works exactly the same regardless of the number of disks in the RAID group.

Proving that RAID-DP really does recover all data in the event of a double disk failure can be done in two manners. One is using mathematical theorems and proofs, and the other is to simply go through a double disk failure and subsequent recovery process. This document will use the latter approach to prove the concept of RAID-DP double-parity protection. For more extensive coverage on the mathematical theorems and proofs that RAID-DP is built upon, please review Row-Diagonal Parity for Double Disk Failure Correction available at the USENIX Organization Web site.

RAID-DP RECONSTRUCTION

Using the most recent diagram as the starting point for the double disk failure, assume that the RAID group is functioning normally when a double disk failure occurs. This is denoted by all data in the first two columns now missing in Figure 4.

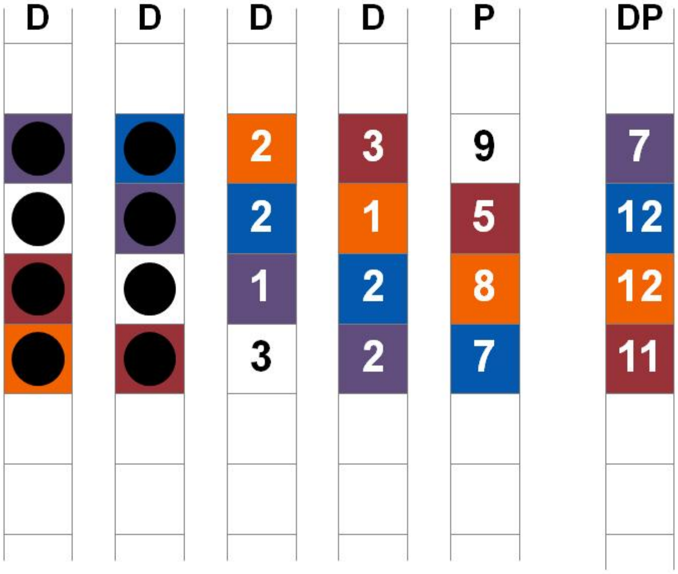


Figure 4) Two failed drives.

When engaged after a double disk failure, RAID-DP first begins looking for a chain on which to start reconstruction. In this case, let’s say the first diagonal parity stripe in the chain it finds is represented by the blue diagonal stripe. Remember when reconstructing data for a single disk failure under RAID 4 that this is possible if and only if no more than one element is missing. With this in mind, traverse the blue diagonal stripe in Figure 4 and notice that only one of the five blue blocks is missing. With four out of five elements available, RAID-DP has all of the information needed to reconstruct the data in the missing blue block. Figure 5 reflects this data having been recovered onto an available hot spare disk.

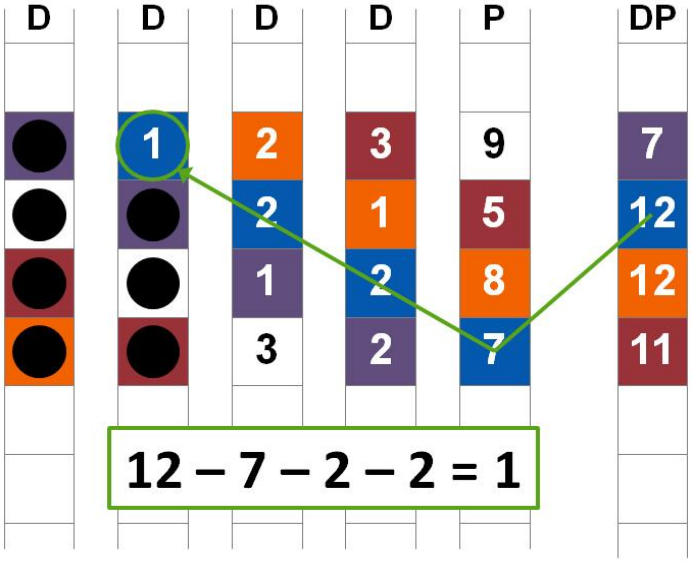


Figure 5) Recover by blue diagonal parity.

The data has been recreated from the missing blue diagonal block using the same arithmetic discussed earlier (12 – 7 – 2 - 2 = 1). Now that the missing blue diagonal information has been recreated, the recovery process switches from using diagonal parity to using horizontal row parity. Specifically, in the top row after the blue diagonal has recreated the missing diagonal block, there is now enough information available to reconstruct the single missing horizontal block from row parity (9 – 3 – 2 – 1 = 3). This occurs in Figure 6.

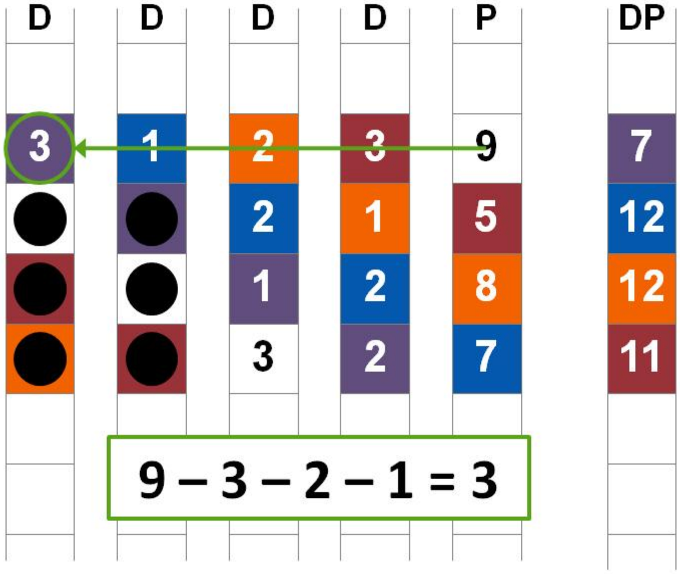


Figure 6) Recover by row parity.

RAID-DP next continues in the same chain to determine if other diagonal stripes can be recreated. With the top left block having been recreated from row parity, RAID-DP can now recreate the missing diagonal block in the gray diagonal stripe, as shown in Figure 7.

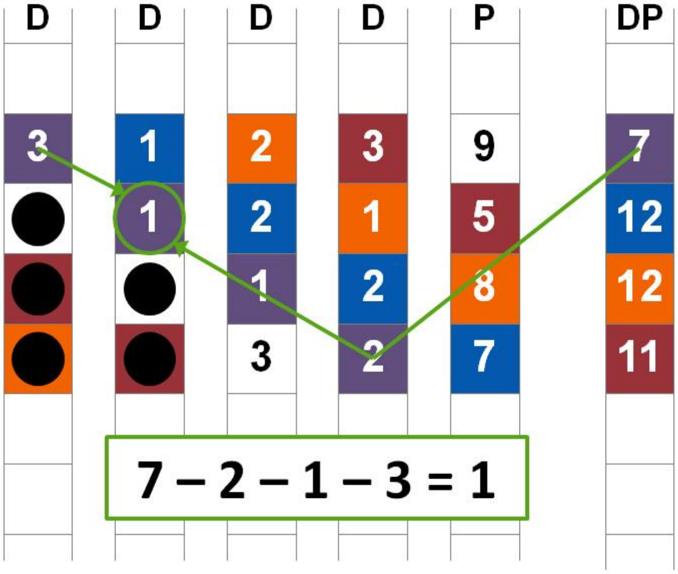


Figure 7) Recover by purple diagonal parity.

Once again, after RAID-DP has recovered a missing diagonal block in a diagonal stripe, enough information exists for row parity to recreate the one missing horizontal block in the first column, as illustrated in Figure 8.

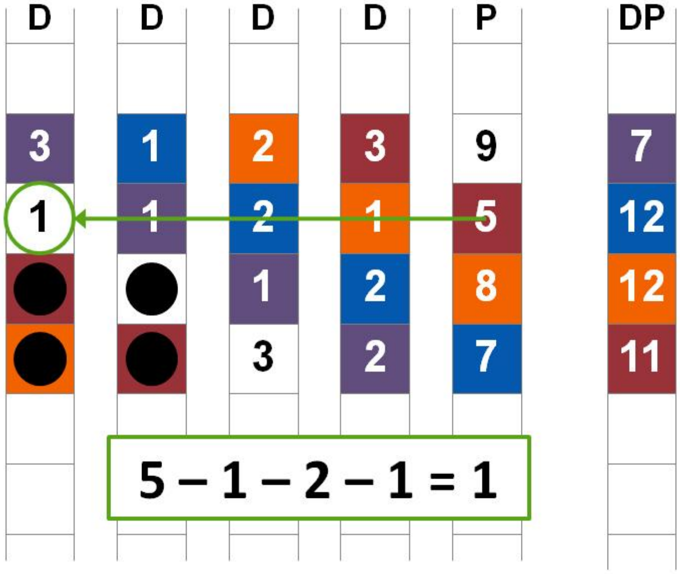


Figure 8) Recover by row parity.

As we noted earlier, the white diagonal stripe is not stored, and no additional diagonal blocks can be recreated on the existing chain. RAID-DP will start to look for a new chain to start recreating diagonal blocks on and, for the purposes of this example, determines it can recreate missing data in the orange diagonal stripe, as Figure 9 shows.

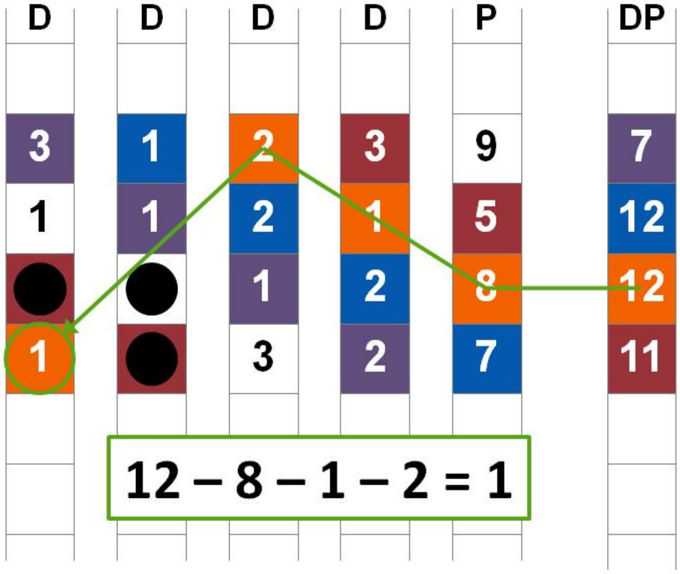


Figure 9) Recover by orange diagonal parity.

After RAID-DP has recreated a missing diagonal block, the process again switches to recreating a missing horizontal block from row parity. When the missing diagonal block in the gold diagonal has been recreated, enough information is available to recreate the missing horizontal block from row parity, as evident in Figure 10.

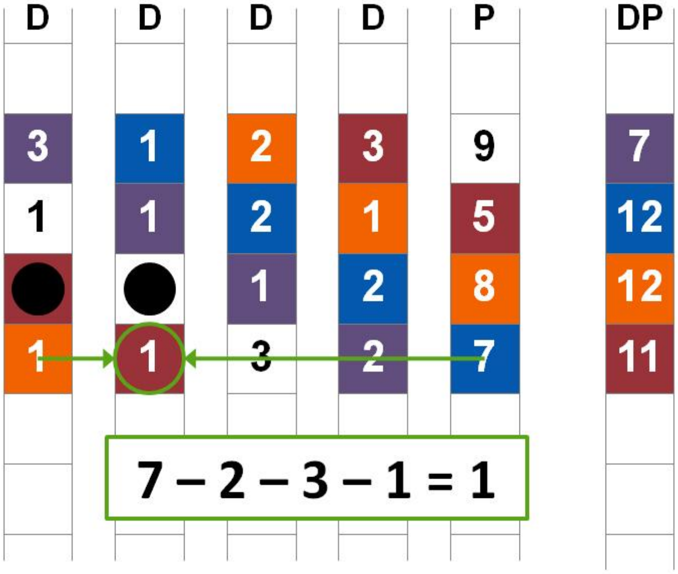


Figure 10) Recover by row parity.

After the missing block in the horizontal row has been recreated, reconstruction switches back to diagonal parity to recreate a missing diagonal block. RAID-DP can continue in the current chain on the red diagonal stripe, as shown in Figure 11.

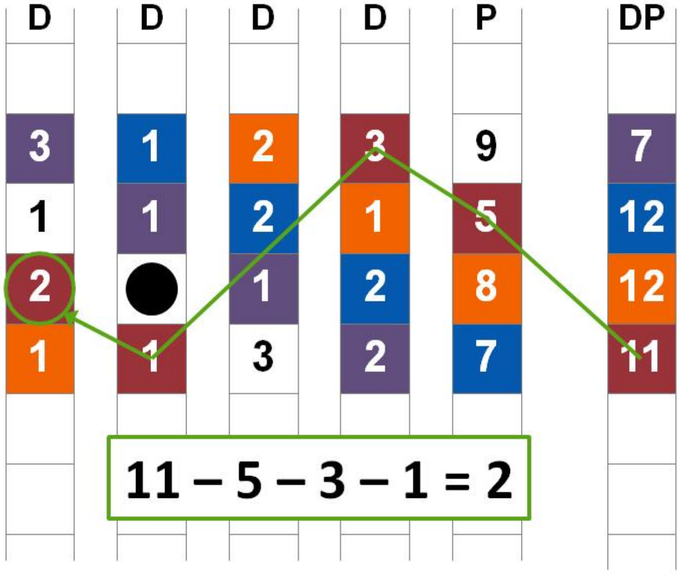


Figure 11) Recover by red diagonal parity.

Once again, after the recovery of a diagonal block the process switches back to row parity, as it has enough information to recreate data for the one horizontal block. The final diagram in the double disk failure scenario follows next, in Figure 12, with all data having been recreated with RAID-DP.

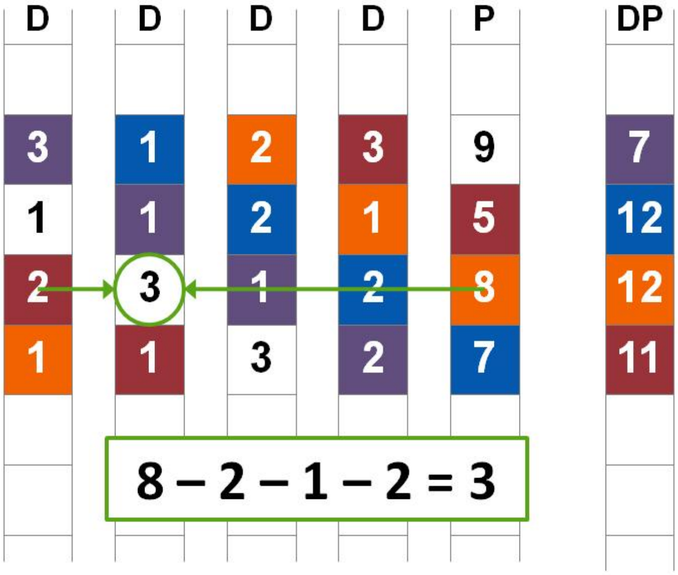


Figure 12) Recover by row parity.

RAID-DP OPERATION SUMMARY

The preceding recovery example goes a long way to give a pictorial description of RAID-DP in operation. But there are a few more areas about RAID-DP operations the example didn't make evident that need further discussion. If a double disk failure occurs, RAID-DP automatically raises the priority of the reconstruction process so the recovery completes faster. As a result, the time to reconstruct data from two failed disks is slightly less than the time to reconstruct data from a single disk failure. A second key feature of RAID-DP with double disk failure is that it is highly likely one disk failed some time before the second and at least some information has already been recreated with traditional row parity. RAID-DP automatically adjusts for this occurrence by starting recovery where two elements are missing from the second disk failure.

3 RAID-DP OVERVIEW

RAID-DP is available with no cost or special hardware requirements. The only requirement to start using RAID-DP is to upgrade to at least Data ONTAP version 6.5. In addition, there are some more technical details about RAID-DP usage that haven't been covered yet and will be addressed in this section.

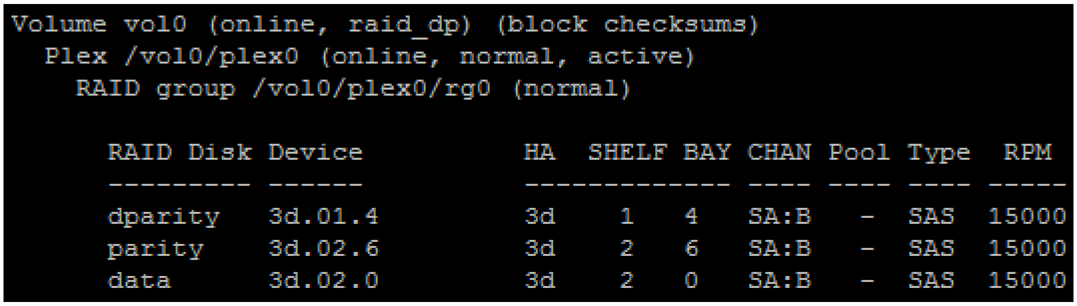
3.1 PROTECTION LEVELS WITH RAID-DP

At the lowest level, RAID-DP offers protection against either two failed disks within the same RAID group or from a single disk failure followed by a bad block or bit error before reconstruction has completed. A higher level of protection is available by using RAID-DP in conjunction with SyncMirror®. In this configuration, the protection level is up to five concurrent disk failures, four concurrent disk failures followed by a bad block or bit error before reconstruction is completed.

3.2 CREATING RAID-DP AGGREGATES AND TRADITIONAL VOLUMES

To create an aggregate or traditional volume with RAID groups based on RAID-DP, select the option in FilerView® when provisioning storage or add the `–t raid_dp` switch to the `aggr create` or `vol create` commands in the command line interface. The command line interface syntax would be `[vol | aggr] create` *`name`* `–t raid_dp X` (with X representing the number of disks the traditional volume or aggregate contains). If the type of RAID group is not specified, Data ONTAP will automatically use the default RAID group type. The default RAID group type used, either RAID-DP or RAID 4, depends on the platform, disk, and Data ONTAP version. To determine what the default RAID group type is for your storage system, select your Data ONTAP version from the Data ONTAP Information Library, then select the Storage Management Guide.

The following partial output from the `sysconfig -r` command shows a three-disk RAID-DP RAID group for a volume vol0 with the second parity disk for diagonal parity denoted as dparity.



```
Volume vol0 (online, raid_dp) (block checksums)
  Plex /vol0/plex0 (online, normal, active)
    RAID group /vol0/plex0/rg0 (normal)

      RAID Disk Device          HA  SHELF BAY CHAN Pool Type  RPM
      --------- ------          ------------- ---- ---- ---- -----
      dparity   3d.01.4         3d    1   4   SA:B   -  SAS  15000
      parity    3d.02.6         3d    2   6   SA:B   -  SAS  15000
      data      3d.02.0         3d    2   0   SA:B   -  SAS  15000
```

Figure 13) Double-parity RAID disk in command output.

3.3 CONVERTING EXISTING AGGREGATES AND TRADITIONAL VOLUMES TO RAID-DP

Once the appliance has been upgraded to at least Data ONTAP 6.5, existing aggregates and traditional volumes are easily converted to RAID-DP using the command `[aggr | vol] options` *`name`* `raidtype raid_dp`. When entered, the aggregate or traditional volume is instantly denoted as RAID-DP, but all diagonal parity stripes still need to be calculated and stored on the second parity disk. RAID-DP protection against double disk failure isn't available until all diagonal parity stripes have been calculated and stored on the diagonal parity disk. Calculating the diagonals as part of a conversion to RAID-DP takes time and affects performance slightly on the storage controller. The amount of time and performance effect for conversions to RAID-DP depend on what the storage controller is (NearStore® or FAS960, for instance) and how busy the storage controller is during the conversion. Generally, conversions to RAID-DP should be planned for off-peak hours to minimize potential performance effect to business or users. For conversions from RAID 4 to RAID-DP, certain conditions are required. Namely, conversions occur at the aggregate or traditional volume level, and there has to be an available disk for the second diagonal parity disk for each RAID 4 group in either class of storage. The size of the disks used for diagonal parity needs to be at least as big as the original RAID 4 row parity disks.

Aggregates and traditional volumes may be converted back to RAID 4 with the command `[aggr | vol] options` *`name`* `raidtype raid4`. In this case, the conversion is instantaneous, since the old RAID 4 row parity construct is still in place as a subsystem in RAID-DP. If a RAID-DP group is converted to RAID 4, then each RAID group's second diagonal parity disk is released and put back into the spare disk pool. It is important to note that in order to revert to a previous version of Data ONTAP that doesn't support RAID-DP, volumes based on RAID-DP would first need to be converted to RAID 4 ones.

3.4 RAID-DP VOLUME MANAGEMENT

From a management and operational point of view, once created or converted to, RAID-DP aggregates and traditional volumes work exactly like their RAID 4 counterparts. The same practices and guidelines work for both RAID 4 and RAID-DP, so little to no changes are required for standard operational procedures used by NetApp storage administrators. While a storage controller might contain any mix of RAID 4 and RAID-DP aggregates or traditional volumes, the commands an administrator would use for management activities on the storage controller are the same.

For instance, if a RAID-DP aggregate or traditional volume requires additional capacity, the command `[aggr | vol] add` *`name`* `X` (where X is the number of disks to add) is run in exactly the same manner as an administrator would use for a RAID 4–based storage.

3.5 RAID-DP CAPACITY UTILIZATION

RAID-DP requires two parity disks per RAID group, and, although this could affect capacity utilization, there are ways to offset the effects of extra disks being used for parity. To reduce or even eliminate the possible effect on capacity utilization from the second parity disk in RAID-DP groups, the first step is to simply use the default RAID-DP group size for the storage platform and disk drive type. Afterward, create aggregates or traditional volumes in sizes that reflect full increments of the default RAID-DP RAID group size.

For instance, on the NearStore R200, the default RAID-DP group size is 14 disks per RAID group, and the maximum is 16 disks per RAID group. For either size RAID-DP group, two disks are used for parity and the rest to serve data. Conversely, the maximum RAID 4 group size allowed on the R200 is seven disks, resulting in one disk being used for parity and the rest to serve data. So, when using the default RAID group sizes and creating a traditional volume or aggregate that is a full increment of the underlying RAID group size, there is no capacity penalty for RAID-DP versus RAID 4. Using RAID-DP on the R200 with the maximum allowed RAID group size actually provides slightly better capacity utilization versus RAID 4.

3.6 RAID-DP PERFORMANCE

The performance of RAID-DP volumes is comparable to that of RAID 4. Read operation performance is exactly the same for each type of RAID group. Depending on the type of write, performance on RAID-DP can be about 2% to 3% slower than that of RAID 4. The reason for this small performance difference is that an extra write occurs to the second diagonal parity disk on RAID-DP volumes. There is no discernable effect to the CPU utilization from running RAID-DP versus RAID 4.

4 CONCLUSION

RAID-DP offers dramatic improvements in data protection that addresses the challenges to RAID implementation brought on by the rapid growth in size of modern disks. In addition, these improvements are provided by NetApp at no cost to the customer. For installations utilizing Data ONTAP 6.5 or greater, converting existing RAID 4 is a simple process. Unlike other approaches to double disk failures, RAID-DP performance is comparable to that of RAID 4 and does not require the additional storage resources that these other approaches require.

NetApp provides no representations or warranties regarding the accuracy, reliability, or serviceability of any information or recommendations provided in this publication, or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information in this document is distributed AS IS, and the use of this information or the implementation of any recommendations or techniques herein is a customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. This document and the information contained herein may be used solely in connection with the NetApp products discussed in this document.

NetApp™

www.netapp.com