Tech OnTap

Jay White
Technical Marketing Engineer
NetApp

Carlos Alvarez
Senior Technical Marketing Engineer
NetApp

This article is the third installment of Back to Basics, a series of articles that discusses the fundamentals of popular NetApp® technologies.

Explore

More Back to Basics

Learn the fundamentals of NetApp core technologies. Installments in this series so far:

- Chapter 1: Thin Provisioning
- Chapter 2: Deduplication
- Chapter 3: FlexClone
- Chapter 4: Volume SnapMirror
- Chapter 5: RAID-DP

With disk drives continuing to increase in size, providing the resiliency to protect critical data becomes more challenging. While disks have gotten larger, their overall reliability has remained about the same. Larger disk size means that the time needed to reconstruct a failed disk using Redundant Array of Independent Disks (RAID) parity information has become significantly longer, raising the possibility of a second disk failure or other error occurring before reconstruction can complete. The likelihood of bit and block errors also increases proportionally with the increased media size, making the chances of this type of event during reconstruction a distinct possibility and increasing the chances of a double failure that could disrupt business and cause data loss in single parity RAID implementations.

NetApp pioneered the development of its unique dual-parity RAID implementation, RAID-DP®, to address this resiliency problem. While other dual-parity RAID 6 implementations exist, RAID-DP is the only one that provides protection against double disk failures in the same RAID group with no significant decreases in performance.

RAID-DP performs so well that it is the default option for NetApp storage systems. Tests show a random write performance delta of only 2% versus the NetApp RAID 4 implementation. By comparison, another major storage vendor's RAID 6 random write performance decreases by 33% relative to RAID 5 on the same system. (RAID 4 and RAID 5 are both single-parity RAID implementations.) RAID 4 uses a designated parity disk. RAID 5 distributes parity information across all disks in a RAID group to avoid having a parity disk that becomes a hot spot. This is unnecessary with NetApp RAID 4 because of the way that Data ONTAP® writes data, as we'll see later.

RAID-DP offers significant advantages, including:

- **Maximum data protection.** With NetApp RAID-DP, the chance of data loss as a result of a double disk failure is hundreds of times less likely than in RAID 5 configurations. While RAID 1+0 offers better data protection than RAID 5, it still presents a risk of data loss in the event of double mirrored

disk failures. RAID-DP offers 100% double disk failure protection at half the cost of RAID 1+0.

- **Lowest cost.** RAID 5 implementations often limit RAID group size to 3+1 or 5+1 (which represents a 17% to 25% cost overhead). RAID 1+0 requires 1+1 (a 50% overhead). In contrast, NetApp supports RAID group sizes of up to 28 (26+2) disks for a 7% capacity overhead.

- **Uncompromising performance.** As described earlier, competing dual-parity technologies might incur a substantial write performance penalty and might be best suited for "read mostly" types of applications. NetApp RAID-DP incurs virtually zero performance penalty compared to single-parity RAID, is the NetApp default option, and is suitable for use with all workloads.

- **No software licensing fees.** The RAID-DP capability is standard on all NetApp systems. You incur no additional costs to use it save for the cost of adding parity disks, which can be offset by using larger RAID groups.

This chapter of Back to Basics explores how NetApp RAID-DP technology is implemented, applicable use cases, best practices for implementing RAID-DP, and more.

How RAID-DP Is Implemented in Data ONTAP

Close Integration with NVRAM and WAFL

The RAID-DP implementation within Data ONTAP is closely tied to the NetApp NVRAM and NetApp WAFL® (Write Anywhere File Layout). This is the key to the exceptional performance achieved by RAID-DP versus other RAID 6 implementations.

Because writing to memory is much faster than writing to disk, storage system vendors commonly use battery-backed, nonvolatile RAM (NVRAM) to cache writes and accelerate write performance. NetApp provides NVRAM in all of its storage systems, but the NetApp Data ONTAP operating environment uses NVRAM in a much different manner than typical storage arrays.

The NVRAM is used as a journal of the write requests that Data ONTAP has received since the last consistency point. Every few seconds, Data ONTAP creates a special Snapshot™ copy called a consistency point, which is a completely consistent image of the on-disk file system. A consistency point remains unchanged, even as new blocks are being written to disk, because Data ONTAP never overwrites existing disk blocks. With this approach, if a failure occurs, Data ONTAP simply has to revert to the latest consistency point and then replay the journal of write requests from NVRAM.

This is a much different use of NVRAM than that of traditional storage arrays, which cache write requests at the disk driver layer, and it offers several advantages, including reducing the amount of NVRAM required, improving response times to the writer, and allowing writes to disk to be optimized.

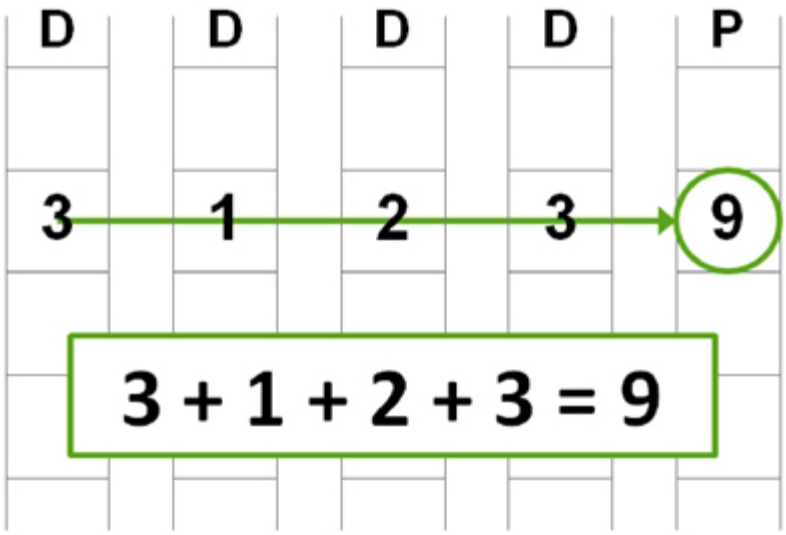
Optimizing Writes: RAID 4

This approach to caching writes is closely integrated with NetApp RAID implementations and allows NetApp to schedule writes such that disk write performance is optimized for the underlying RAID array. I'll begin by explaining how NetApp optimizes this process for its RAID 4 implementation before tackling RAID-DP.

RAID arrays manage data in stripes, where a stripe is composed of one block from each disk in a RAID group. For each stripe, one block is designated as the parity block. Figure 1 shows a traditional RAID 4

group using horizontal parity that consists of four data disks (the first four columns, labeled “D”) and a single parity disk (the last column, labeled “P”).

Figure 1) Example of RAID 4 parity.



In this example parity has been calculated by adding the values in each horizontal stripe and then storing the sum as the parity value (3 + 1 + 2 + 3 = 9) for demonstration purposes. In practice, parity is calculated using an exclusive OR (XOR) operation.

If the need arises 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, RAID 4 would recalculate the data in each block of disk 1 from the remaining data; in our example, this is achieved by simply subtracting the values from the remaining disks from the value stored in parity (9 – 3 – 2 – 1 = 3). This also illustrates why single parity RAID only protects against a single disk failure. You can see that if two values are missing, there is not enough information to recalculate the missing values.

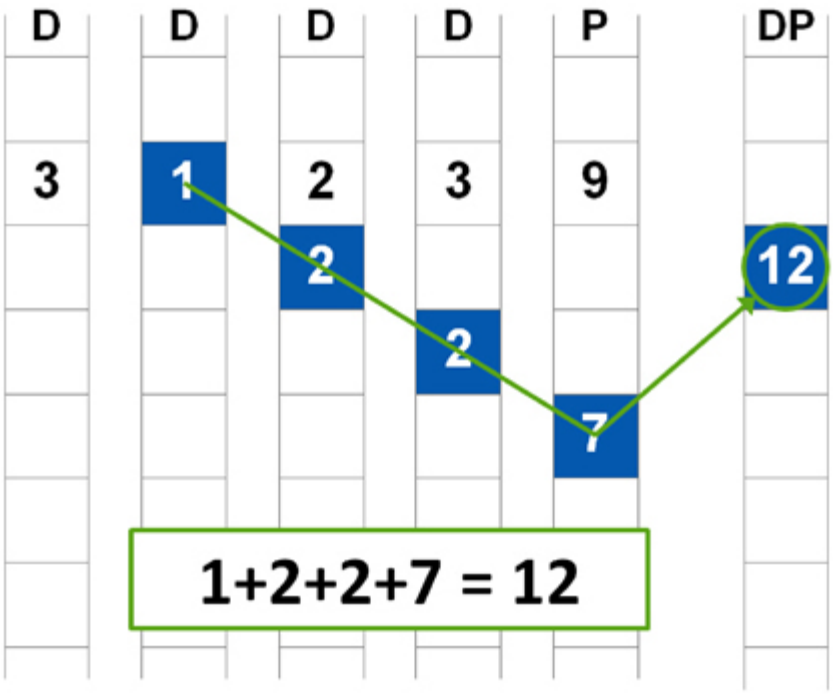
In typical RAID implementations, in order to write new data into a stripe that already contains data (and parity), you have to read the parity block and calculate a new parity value for the stripe before writing the data block and the new parity block. That’s a significant amount of overhead required for each block to be written.

NetApp reduces this penalty by buffering writes in memory (protected by the journal in NVRAM) and then writing full RAID stripes plus parity whenever possible. This makes reading parity data before writing unnecessary and allows WAFL to perform a single parity calculation for a full stripe of data blocks. (The exact number of blocks depends on the size of the RAID group.) This is possible because WAFL never overwrites existing blocks when they are modified and because it can write data and metadata (the accounting information that describes how data is organized) to any location. In other data layouts, modified data blocks are usually overwritten, and metadata is often required to be at fixed locations.

Adding Diagonal Parity: RAID-DP

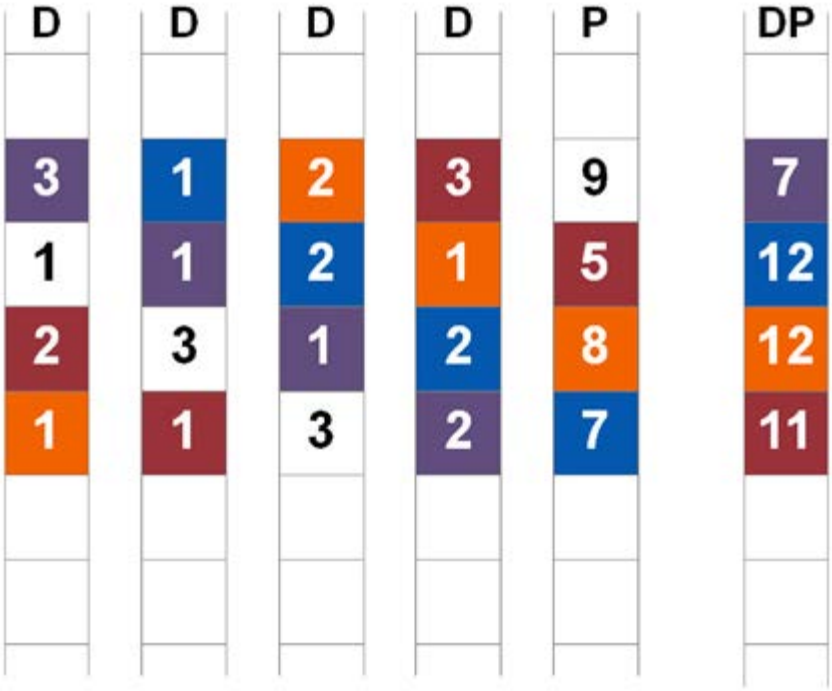
NetApp RAID-DP uses two parity disks per RAID group. One parity disk stores parity calculated for horizontal stripes, as described earlier. The second parity disk stores parity calculated from diagonal stripes. Figure 2 adds one diagonal parity stripe, denoted by the blue-shaded blocks, and a second parity disk, denoted as “DP,” to the horizontal parity illustration from Figure 1.

Figure 2) The addition of diagonal parity.



The diagonal parity stripe includes a block from the horizontal parity disk as part of its calculation. RAID-DP treats all disks in the original RAID 4 construct—including both data and parity disks—the same. Note that one disk is omitted from the diagonal parity stripe. Figure 3 shows additional horizontal and diagonal parity stripes.

Figure 3) Multiple stripes showing both horizontal and diagonal parity.



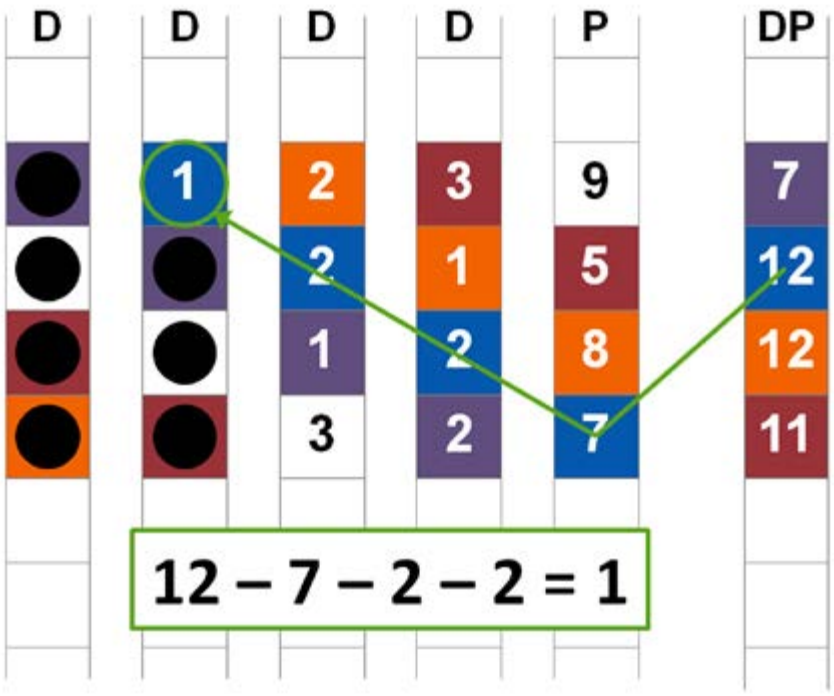
Each diagonal parity stripe misses one and only one disk, and each diagonal misses a different disk. There is also one diagonal stripe (white blocks in Figure 3) that doesn't have parity stored on the diagonal parity disk. This doesn't affect the ability to recover all data.

Recovering from Double Disk Failure

The combination of horizontal and diagonal parity makes it possible to recover from two disk failures within the same RAID group. If a single disk fails or a bad block or bit error occurs, then the horizontal parity is all that is needed to recreate the missing data.

After a double disk failure, RAID-DP first identifies a chain on which to start reconstruction, as shown in Figure 4. Remember that reconstructing data from parity is possible if and only if no more than one element is missing; this is why each diagonal parity stripe skips one of the data disks.

Figure 4) Start of RAID-DP recovery using diagonal parity.



The recovery of the first block using diagonal parity in turn makes it possible to recover a second block using horizontal parity (the first row in Figure 4). This in turn makes it possible to recover another missing block using diagonal parity. This chain of recovery continues until it terminates due to the stripe for which no diagonal parity exists. At that point, another entry point is found, and recovery on another chain of diagonal and horizontal stripes begins. Eventually, enough missing blocks are filled in that all values for the stripe without parity can be recalculated using horizontal parity alone. This process is illustrated more fully in TR-3298: RAID-DP.

I've simplified the examples here to make it easier to understand the basic concepts behind RAID-DP, but it's important to understand that the same processes apply in real storage deployments with dozens of disks in a RAID group and millions of rows of data. While the failure example shows the recovery of two adjacent data disks, the same process works whether the disks are adjacent or not and whether the failed disks are data disks or the parity disks themselves.

Optimizing Writes: RAID-DP

As you'll recall from the earlier RAID 4 discussion, WAFL always tries to buffer and write full stripes of blocks to disk whenever possible. RAID-DP buffers data blocks on memory to accomplish multiple horizontal and diagonal parity calculations on a single read operation. The 2% performance overhead of RAID-DP versus RAID 4 results from the additional overhead of calculating diagonal parity and writing the second parity block.

Use Cases

In terms of use cases for RAID-DP, the technology has become so widely used on NetApp storage that it's easier to talk about the few situations where you might choose not to use it than the situations where you should use it. Over 90% of NetApp customers already use RAID-DP, including for their most business-critical and performance-critical workloads. RAID-DP is the default for all new NetApp storage systems, we prescribe the use of RAID-DP in our best practices, and RAID-DP is used in our published performance benchmarks. All NetApp software is fully compatible with RAID-DP. No other vendor can make these claims about its RAID 6 implementation.

The only situations in which you might choose to use RAID 4 instead of RAID-DP are those few situations in which resiliency is not important, such as for scratch space, testing, and laboratory environments.

Using RAID-DP

RAID-DP was introduced in Data ONTAP version 6.5.

Creating RAID-DP Volumes

To create an aggregate (or traditional volume) with RAID-DP RAID groups, select the option when provisioning storage with NetApp graphical tools or add the –t raid_dp switch to the aggr create or vol create commands.

If the RAID type is not specified, Data ONTAP will automatically use the default RAID type, which is RAID-DP for all currently shipping versions of Data ONTAP. You can find out what the default for your system is by selecting your version of Data ONTAP from the Data ONTAP Information Library. (NetApp NOW® access is required.)

Data ONTAP 8.0.1 Default and Maximum RAID Group Size by Drive Type			
Drive Type	**RAID Type**	**Default RAID Group Size**	**Maximum RAID Group Size**
SSD	RAID-DP (default)	23 (21+2)	28 (26+2)
	RAID 4	8 (7+1)	14 (13+1)
SAS/FC	RAID-DP (default)	16 (14+2)	28 (26+2)
	RAID 4	8 (7+1)	14 (13+1)
SATA	RAID-DP (default)	14 (12+2)	20 (18+2)
	RAID 4	7 (6+1)	7 (6+1)

Existing RAID 4 RAID groups can be converted to RAID-DP. Conversions occur at the aggregate or

traditional-volume level, and there must be an available disk (that is at least as large as the largest disk in the RAID group) for the diagonal parity disk for each RAID group.

Choosing a RAID-DP RAID Group Size

The ability to use a larger RAID group size with RAID-DP can offset the effect on usable capacity of the additional disk required for parity. To reduce or even eliminate this effect, one option is to use the default RAID-DP group size for the disk drive type you are using. Create aggregates based on multiples of the default RAID-DP RAID group size.

For hard disk drives (SATA, FC, and SAS) the preferred sizing approach is to establish a RAID group size that is within the range of 12 (10+2) to 20 (18+2) and that achieves an even RAID group layout (all RAID groups containing the same number of drives). If multiple RAID group sizes achieve an even RAID group layout, it is recommended to use the higher RAID group size value within the range. If incomplete RAID groups are unavoidable (as is the case sometimes), it is recommended that the aggregate not be deficient more than a number of drives equal to one less than the number of RAID groups (otherwise you would just pick the next lowest RAID group size). Drive deficiencies leading to incomplete RAID groups should be distributed across RAID groups evenly so that no single RAID group is deficient more than a single drive.

RAID-DP Management

Little to no change is required to your operating procedures when using or switching to RAID-DP. A storage system can contain a mix of RAID 4 and RAID-DP aggregates and volumes, and the commands you use for management remain the same.

RAID-DP Reconstruction

If a double disk failure occurs, RAID-DP automatically raises the priority of the reconstruction process so the recovery completes more quickly. 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. With double disk failure it is highly likely that one disk failed some time before the second and at least some information has already been recreated using horizontal parity. RAID-DP automatically adjusts for this occurrence by starting recovery where two elements are missing from the second disk failure.

Data ONTAP includes options that allow a storage administrator to adjust the effect that a RAID reconstruction has on system performance.

The `raid.reconstruct.perf_impact` option is set to medium by default. The three possible values for this option are low, medium, and high. A setting of low for this option might increase the time it takes to complete a RAID reconstruction as system resources are prioritized to respond to foreground I/O. Setting this option to high will allow RAID recovery operations to compete with foreground I/O for an increased amount of systems resources (thereby reducing foreground I/O performance).

Some situations might warrant adjusting this option, but this should be a last resort. NetApp typically recommends keeping the default value.

For more complete information on managing RAID-DP, refer to relevant sections of TR-3437: Storage Subsystem Resiliency Guide.

Conclusion

NetApp RAID-DP technology is an important resiliency tool that can be used with virtually all common storage workloads. To learn more about NetApp RAID-DP, be sure to refer to NetApp TR-3298: Implementation of Double-Parity RAID for Data Protection and WP-7005: NetApp RAID-DP: Dual-Parity RAID 6 Protection Without Compromise.

By Carlos Alvarez, Senior Technical Marketing Engineer, and Jay White, Technical Marketing Engineer

Jay is the technical marketing engineer for system resiliency, storage subsystems (shelves, drives, and so on), and RAID. He has authored several technical reports and FAQs related to storage subsystem configuration and resiliency.

Carlos has been working for NetApp since 2008, specializing in storage efficiency with deep-dive expertise in deduplication, data compression, and thin provisioning. He regularly provides guidance for integrating the most effective and appropriate NetApp storage efficiency technologies into customer configurations.