Best Practice Guidelines for Volume and RAID Group Configuration on NearStore® R200

Chris Lueth, Network Appliance, Inc.

March 2006, TR-3295

Table of Contents

1. Introduction

This document provides an architectural overview of NearStore R200 and provides recommendations on configuring aggregates or traditional volumes and RAID groups to maximize both performance and fault tolerance. Configuration options for scaling across the capacity footprints on R200 are addressed, from the minimum of 8TB to the maximum 96TB footprint. Additional coverage is given to RAID Double Parity, or RAID-DP™, and how it significantly increases fault tolerance.

For the remainder of this paper the term volume, when used alone, means both traditional volumes and aggregates. Volumes have been the traditional unit of storage on NetApp appliances, but starting with Data ONTAP® 7G, volumes have two distinct versions. Although traditional volumes will continue to operate as they have for over 10 years, a new type of volume known as FlexVol™ is available starting in Data ONTAP 7G. As the name implies, FlexVol volumes offer extremely flexible and unparalleled functionality. Complete coverage of them is beyond the scope of this paper. FlexVol volumes are housed on a new construct known as an aggregate. At the RAID layer, both RAID-DP and RAID 4 operate at the traditional volume and aggregate level. For more information on aggregates, FlexVol, FlexClone™, and other functionality released in Data ONTAP 7G, please see the Network Appliance Portal and NOW™ (NetApp On the Web) Web sites.

2. NearStore R200 System Architecture

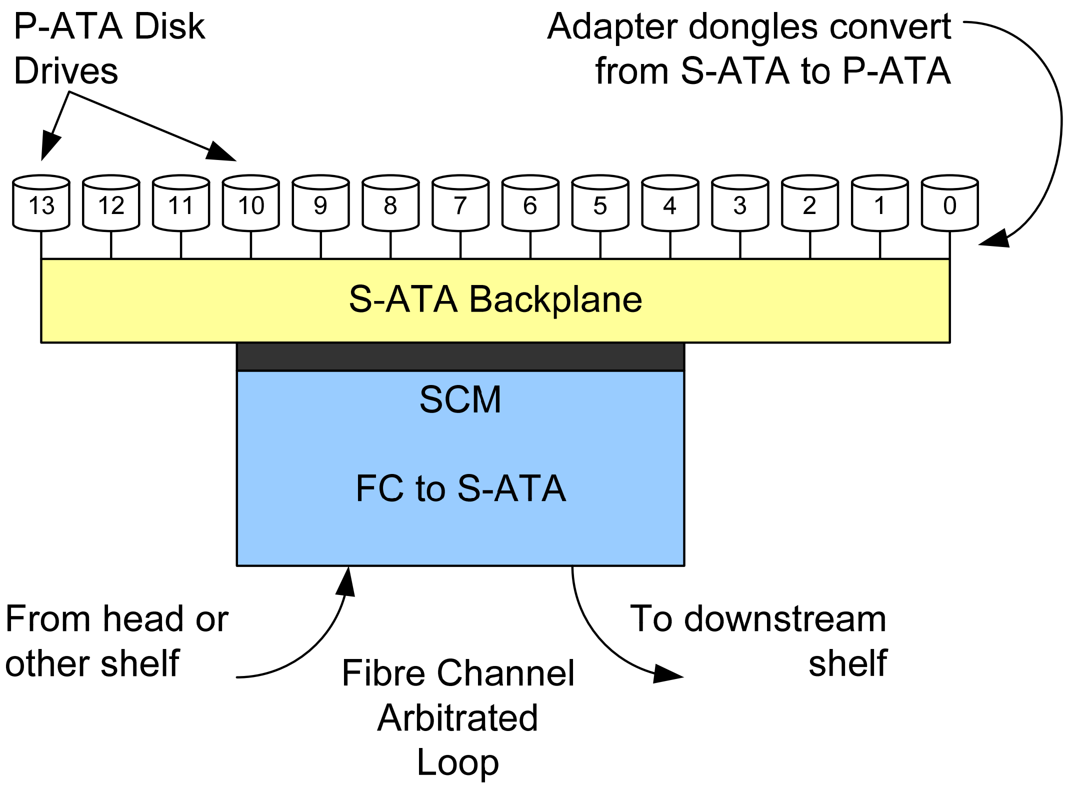
2.1 Overview

NearStore R200 consists of the NetApp Data ONTAP operating system installed on a hardware platform that includes a traditional filer head containing CPU, memory, NVRAM, etc., and shelves containing disk drives and associated control hardware. The shelves are connected to the head with up to four 2Gbps Fibre Channel Arbitrated Loops (FCALs), and each of the four loops is capable of supporting up to six shelves. Each shelf contains fourteen 274GB ATA hard drives, and the R200 capacity scales two shelves at a time for a raw capacity gain of 8TB. The minimum capacity option is with two shelves for 8TB and the maximum capacity of 24 shelves for 96TB. All components in an R200 are normally packaged in a standard data center cabinet, but they can be ordered without the cabinet for mounting in typical Telco racks. The R200 uses the DS14-Mark II-AT shelf, which is very similar to regular DS14 Mark II shelves used on NetApp filers, with a few notable differences. The R200 shelf contains a Serial-ATA Controller Module (SCM), which bridges from FC (Fibre Channel) to Serial-ATA (S-ATA). The SCM is what the FC loop connects to on the rear of the shelf, and its equivalent on filer shelves is the Loop Resiliency Circuit (LRC) or Electronically Switched Hub (ESH) module. To prevent possible FC and AT shelf mismatches, the SCM and LRC or ESH are keyed so they can only be inserted into the appropriate DS14-Mark II. The SCM connects into the rear side of the back plane in the shelf. The front side of the back plane provides 14 individual S-ATA connections for each disk. Completing the conversion from S-ATA to Parallel-ATA (P-ATA) disks, an adapter (known as a dongle) is included on each disk enclosure. Again, to highlight I/O keying to prevent possible component mismatches, R200 disk enclosures and the back plane are keyed in such a way that FC disks cannot be inserted into the shelf. Visual determination of whether a shelf belongs to an R200 is easy: simply look for AT on the top front right corner or on the SCM module in the rear.

The SCM and dongle adapter translate from FC to S-ATA and finally to P-ATA for communication with the individual drives. This translation series allows Data ONTAP to treat the P-ATA disks as standard FC disk devices. The SCM and dongle adapter also allow all of the P-ATA disks to be factory configured as "master" devices, eliminating the need to adjust jumpers or dip switches when moving drives between shelves or slots.

The following figure illustrates the shelf layout for NearStore R200. It should be noted that there are redundant power supplies on each side of the SCM module. FCAL connectivity from the R200 head to the first shelf in a loop uses optical FC. Connectivity to all downstream shelves in the loop uses copper FC connections. In either case, the data rate is still 2Gbps on the entire loop, whether over the first optical connection or fiber ones after that.

2.2 NearStore R200 Disk and Shelf Architecture



2.3 NearStore R200 Capacity Architecture

The R200 supports up to four FCAL loops, with each loop connecting to up to six shelves, for a maximum of 24 shelves. Capacity on the R200 scales in minimum increments of two shelves, for an increase of 8TB in raw capacity. Scaling past 48TB of capacity requires a second cabinet or Telco rack for mounting the shelves.

On the R200 with the minimum configuration option of two shelves with 8TB of capacity, each of the shelves is on a different loop. Since by default the FC-AL card is in slot 2 inside the head, the first two loops are 2a and 2b. Once the R200 capacity goes past 48TB, the second FCAL card goes in slot 3, and the remaining two loops are 3a and 3b. As capacity is added to the R200 in two-shelf increments, one shelf should go on loop 2a and the second on 2b. When scaling past 48TB, again, put one of the two shelf upgrades on 3a and the second on 3b. The following diagram outlines how capacity should be configured on the R200.

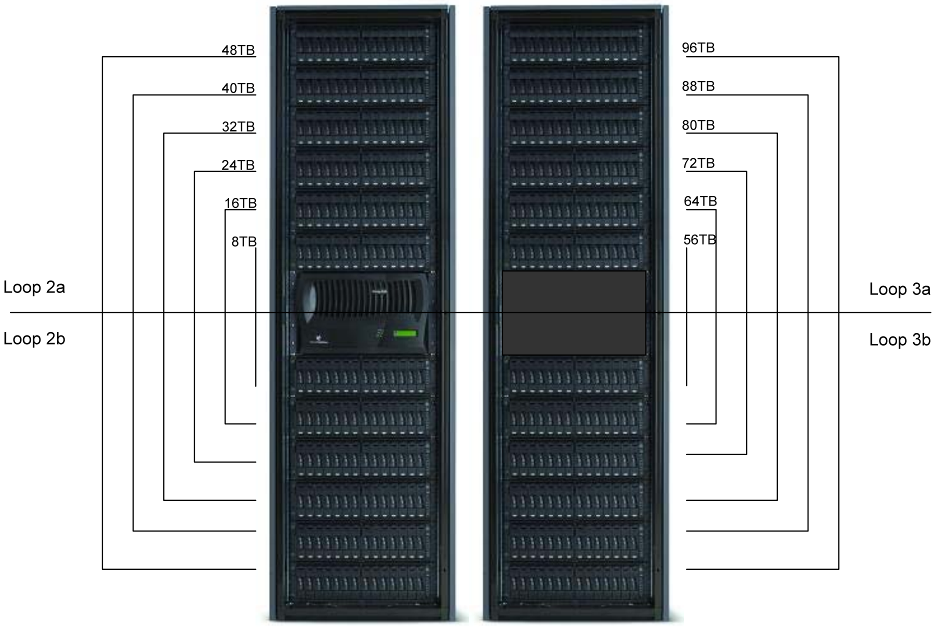


Figure 2) R200 in the 96TB configuration option

2.4 Benefits of Fibre Channel Connectivity to ATA Disks

The biggest benefit of FCAL connectivity to disk shelves is a strategic and tightly coupled integration of ATA-based disks into traditional NetApp architecture. This level of integration is readily apparent if you consider that the FCAL adapter in the FAS 960 filer is the same one used in the R200. Data ONTAP does not know that ATA-based disks are being used for storage on the R200. In addition, Data ONTAP now communicates with the ATA disks in the same tried-and-true manner it has used with FC disks for many years. This giant step toward comprehensive integration between product lines, like filers and NearStore in the case of NetApp, is something most storage vendors do not give much thought to, much less have hope of accomplishing.

3. NearStore R200 Volume and RAID Group Background

Traditional volumes and aggregates on all NetApp storage appliances are created using an underlying construct known as a RAID group. Depending on the size of the traditional volume or aggregate, it can have one or more RAID groups. In typical deployments, large traditional volumes and aggregates are created using multiple RAID groups, which are collections of disks, and the raidsize setting controls the number of disks in a RAID group.

In Data ONTAP 6.5 and later, there are two choices for the type of underlying RAID group that a traditional volume or aggregate uses. The standard RAID group type on NetApp appliances has been RAID4, and this RAID group type continues as an option going forward. The new type of RAID group available is RAID-DP (DP stands for Double Parity). Differences between the two types of RAID groups are covered in the following topic.

This paper assumes some familiarity with the NetApp Data ONTAP operating system and RAID4 implementation. For implementation details about both, please review Network Appliance - A Storage Networking Appliance. This paper discusses RAID-DP at a high level, but a complete description of RAID-DP is outside the scope of the paper. For more details on RAID-DP, please review the technical report Double Parity RAID for Enhanced Data Protection with RAID-DP.

3.1 RAID4 Groups Versus RAID-DP Groups

RAID4 is the traditional approach that NetApp has employed to provide fault tolerance against disk drive failure. All traditional RAID approaches, including the NetApp RAID4 solution, have one main limitation. If more than one disk drive in a RAID group fails, the result can be a loss of at least some, if not all, data contained in the aggregate or traditional volume. The operational impact of this scenario would be downtime until the volume could be restored from backup (either remote disk or tape). Using recommended best practice guidelines for limiting the maximum size of RAID4 groups, the chance of two drives failing in a RAID group and causing a loss of data is greatly reduced. The reason for this is that self-healing RAID reconstructs data contained on the failed disk to a hot standby, and a smaller RAID group reconstructs the data faster than a large RAID group. However, making RAID group sizes smaller carries its own costs, such as lost capacity to additional parity disks.

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. In the scenario in the previous paragraph, RAID-DP provides protection against up to two failed disks in the same RAID group. A more likely scenario is for a bit error to occur that the disk drive’s error-checking and correction firmware (ECC) cannot correct. When a RAID4 group is operating in normal mode, Data ONTAP detects this event and recalculates the data being read from parity. However, with traditional RAID, while in reconstruct mode after a failed disk drive, the ability to recreate data from parity is lost, because not enough information exists to do so. In short, customers and analysts demanded a better story about improving RAID reliability from storage vendors.

To meet this demand, NetApp has released a new type of RAID protection named RAID-DP, which drastically increases the fault tolerance from failed disk drives over traditional RAID. Fundamentally, RAID-DP adds a second parity disk to each RAID group in a volume. Whereas in RAID4, the parity disk stores row parity across the disks in a RAID4 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 horizontal and the other diagonal, data protection is obtained even in the event of two disk drives failing in the same RAID group.

3.2 Performance Overview

Improving performance on NearStore R200 parallels a key performance consideration on filers. That is, larger RAID groups enable larger aggregates and traditional volumes, which perform better than smaller ones would. As with filers, the first criterion to realize optimal performance on the R200 is to use the largest recommended RAID group size. This is seven disks per RAID4 group and 14 per RAID-DP group. After initial aggregate or traditional volume creation, increasing or adding to the space in a volume by an entire RAID group helps keep performance optimal. Conversely, adding one or two disks to increase the size of an aggregate or traditional volume, especially one that is almost full, is a certain way to severely degrade performance and should be avoided.

Better performance can also be obtained by laying out the RAID groups in an aggregate or traditional volume in such a way that parity disks for each RAID group are on a different shelf. The reason for this is that a parity disk in a RAID group on the R200, when used primarily for write-intensive data archiving, is the busiest disk in the group. Spreading the slightly increased load associated with a parity disk as evenly as possible across shelves ensures that no one pathway to data is overworked relative to other data pathways. On the R200, Data ONTAP automatically spreads parity disks for both RAID4 and RAID-DP across shelves, and no manual intervention is required to realize this performance optimization.

4. NearStore R200 Storage Configuration Strategies

4.1 Aggregate, Traditional Volume, and RAID Group Configuration

The best-practice recommendation for creating aggregates or traditional volumes on the R200 is to have Data ONTAP automatically select the disks rather than manually specifying which disks are included in which RAID group or volume. Data ONTAP automatically optimizes for performance and fault tolerance when selecting disks for creating or expanding an aggregate or traditional volume.

Highlights

Letting Data ONTAP automatically select disks is the quickest and easiest way to create aggregates or traditional volumes.

- No need to determine which disks are available, then manually specify them.
- Less chance for human error.

Data ONTAP automatic disk selection recommendation applies to both RAID4 and RAID-DP groups.

- RAID-DP by default, but RAID4 can be selected when creating the volume.

Automatically uses recommended maximum RAID4 group size of seven disk drives or RAID-DP group size of 14 disk drives.

- Using maximum RAID group size provides best capacity utilization.

By default, the root volume on R200 ships as RAID4.

- Recommend leaving as RAID4, because conversion to RAID-DP would require an additional disk for parity without significantly increasing fault tolerance of a two-disk traditional volume.

  In the event of a disk failure in the root traditional volume, a two-disk RAID4 group can be reconstructed rapidly enough to offset the risk of second disk failing before reconstruction is complete.

- When running Data ONTAP 7G, better capacity utilization can be realized by converting the root volume to a FlexVol volume contained in an aggregate. A 90GB FlexVol volume can house the root volume, thus freeing the existing two disks in the root RAID 4 volume for other uses. Guidelines and steps for migrating the root volume to a FlexVol volume can be found in the Storage Administrators Guide on the NOW Web site.

Possible Drawback

With RAID-DP, using the maximum recommended RAID group size of 14 results in an aggregate or traditional volume size of 2.5TB with a single RAID group.

- With Data ONTAP 7G, once a large aggregate is created, the FlexVol flexible volumes it contains can vary in size from 20MB to 16TB. In this scenario, although the underlying aggregate would be large, the size granularity of the flexible volumes it contains can be as small as 20MB, allowing substantial storage configuration options.

Configuration Steps

Simply create aggregates or traditional volumes and allow Data ONTAP to select drives automatically.

- Use recommended RAID group size when creating new aggregates or traditional volumes.
- If using aggregates, the next step is to create FlexVol flexible volumes on top of the underlying aggregate.

Hot spare disk drives.

- For capacity range from 8TB to 16TB, leave one hot spare disk.
- For capacity range from 24TB to 48TB, leave two hot spare disks.
- For capacity range from 48TB to 96TB, leave four hot spare disks.

Maintenance Steps

- Replace disk drive in the event of a failure.

4.2 Using Aggregates and FlexVol Volumes Versus Traditional Volumes

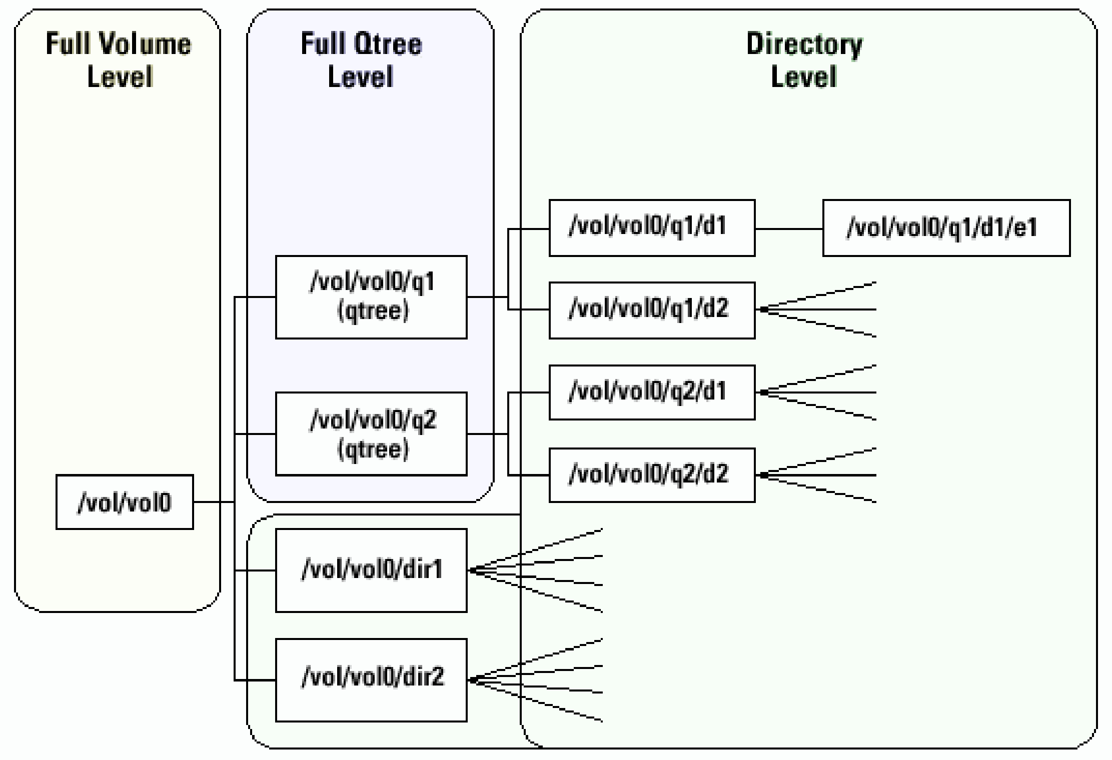
In virtually all storage use cases, there are substantial benefits to the new FlexVol technology available in Data ONTAP 7G versus traditional volumes. Where traditional volumes are tightly coupled to their underlying disk drives, a FlexVol volume simply occupies space within the aggregate, which in turn is tightly coupled to its underlying disks. An aggregate can host many flexible volumes, large and small, and each FlexVol volume can range in size from 20MB to 16TB. Being loosely coupled only to the aggregate and not to the disks themselves, a FlexVol volume can be not only dynamically increased in size, much like its traditional volume cousin, but also decreased in size as business needs change. Another benefit available with flexible volumes is that a small FlexVol will still share the I/O operations (IOPs) available through the large number of disks in the aggregate. The result of IOPs sharing with a large disk pool is that small flexible volumes can be created without incurring the performance penalty that occurs with smaller but busy traditional volumes. And since the parity disks are a part of the aggregate's underlying RAID group and not the FlexVol volume itself, there is no capacity penalty lost to additional parity disks when creating a smaller FlexVol volume.

To find out if using FlexVol technology is a good fit in your business, or if you need assistance with migration from traditional to flexible volumes, please contact your NetApp sales team.

5. Backup Recommendations

All methods applicable to Network Appliance storage systems for backing up and restoring data to and from tape are also applicable to NearStore. However, these methods can exhibit very different performance characteristics on NearStore than on traditional Network Appliance storage systems. In addition, proper data protection strategies for NearStore depend on the characteristics of the data stored in it. This section discusses both of these points.

When performing a tape backup of a volume on a Network Appliance storage system or NearStore, the volume can be backed up at one of three levels, as shown in the following figure. The first is backing up the entire volume in one operation, also referred to as a full-volume backup. The second is backing up by individual qtree within the volume, known as a full-qtree backup. The third is backing up individual subdirectories, known as directory-level backup. Although a full-volume backup is usually the quickest way to back up an entire volume, it is sometimes possible to use multiple concurrent full-qtree operations to meet or exceed full-volume backup performance. Directory-level backups generally exhibit substantially slower performance and should be avoided.



Due to the inherent performance characteristics of NearStore volumes, recommendations for backup levels are more complicated than for other Network Appliance storage systems. When using tape drives with a native transfer rate less than 10MB per second, such as DLT 8000 and AIT-2, the best method of backing up a volume is using between two and four concurrent full-qtree backup operations. When using tape drives with native speeds over 10MB per second, such as LTO and AIT-3, multiple concurrent qtree backup operations from one volume will not transmit data quickly enough to the tape drives to keep them operating at their minimum required transfer rate, causing the drives to pause and drastically reducing overall performance. Therefore, with tape drives over 10MB per second, only full-volume backups are recommended.

Data on a NearStore system falls into one of two classifications. The first classification is data that is either copied or replicated to the device via SnapMirror®, SnapVault®, or some other data replication software. This data is logically organized on a NearStore system in a manner that makes it meaningful and easy to use. Incremental tape backup of data in this category is a valid option as long as the rate of change is low. However, daily incremental tape backups should be avoided in favor of online Snapshot™ backups.

The second classification is virtual tape images from third-party backup software. These images do not have a structure or organization comprehensible without significant knowledge of the internals of the software application that recorded them. They generally consist of several very large files. Incremental tape backups of this data are essentially full backups and therefore are not a valid option.

The best method for recording virtual tape images to tape is through the backup application itself via cloning or duplication. Cloning or duplication records the data onto tape in a format that is readable natively by the backup application, essentially transferring the virtual tape image into a real tape image. This may require some scripting or manual intervention, but completely eliminates the need to restore these images back onto the NearStore device before restoring the data back to the original source, thereby dramatically improving the performance and simplicity of recovering data from these images after they are recorded to tape.

With the notable exception described above, restoring data from tape to NearStore is no different from restoring data from tape to Network Appliance storage systems. It is generally appropriate when preparing for a large restore operation to allocate twice the amount of time required for the backup of the data. For this reason, whenever possible, restore operations should be made from online Snapshot copies. Restoring from a Snapshot copy is nearly instantaneous. Entire volumes can be restored in seconds, regardless of how large they are.

If NearStore is properly configured as outlined in this document, backup-and-restore performance will approach that of other Network Appliance storage systems. Special consideration should be taken when designing volumes initially. Volumes should consist of multiples of eight disks evenly distributed across the available SCSI buses and ATA-SCSI bridges. Backup operations to tape should generally be full backups at either the full-volume level or the qtree level if slower tape devices are in use. When recording virtual tape images to tape, use the cloning or duplication features of the backup software application whenever possible.

The recommendations made in this paper are targeted at providing best practices for the majority of environments. This paper does not and cannot provide the best practices for all environments. In general, following the guidelines presented here will result in one of the most flexible and scalable backup-and-recovery environments in existence today.

6. Additional Information

For more information about the NearStore platform and Network Appliance, see:

NearStore solutions and white papers: Network Appliance - Tech Library

NearStore product sheet: Network Appliance - NearStore

Network Appliance: Network Appliance

NetApp®

www.netapp.com