Abstract

This document is an overview of the new NetApp® ONTAP® feature NetApp FlexGroup volumes. NetApp FlexGroup is an evolution of scale-out NAS containers that blends near-infinite capacity with predictable, low-latency performance in metadata-heavy workloads. For information about NetApp FlexGroup not covered in this document, e-mail flexgroups-info@netapp.com and that information will be added as necessary. For best practices, see TR-4571: NetApp FlexGroup Best Practices and Implementation Guide.
Version History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Document Version History</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version 1.0</td>
<td>October 2016</td>
<td>Initial release</td>
</tr>
<tr>
<td>Version 1.1</td>
<td>December 2016</td>
<td>ONTAP 9.1RC2</td>
</tr>
<tr>
<td>Version 1.2</td>
<td>February 2017</td>
<td>ONTAP 9.1 GA</td>
</tr>
<tr>
<td>Version 1.3</td>
<td>June 2017</td>
<td>ONTAP 9.2RC1</td>
</tr>
<tr>
<td>Version 1.4</td>
<td>November 2017</td>
<td>ONTAP 9.3RC1</td>
</tr>
<tr>
<td>Version 1.5</td>
<td>May 2018</td>
<td>ONTAP 9.4RC1</td>
</tr>
<tr>
<td>Version 1.6</td>
<td>December 2018</td>
<td>ONTAP 9.5</td>
</tr>
</tbody>
</table>

TABLE OF CONTENTS

Version History ............................................................................................................................................. 2

1 Data Is Growing ........................................................................................................................................ 5
   1.1 Flexible Volumes: A Tried-and-True Solution .................................................................................. 5
   1.2 Infinite Volumes: Massive Capacity with Limitations ...................................................................... 6
   1.3 NetApp FlexGroup: An Evolution of NAS ......................................................................................... 6

2 Terminology .............................................................................................................................................. 7

3 Advantages of NetApp ONTAP FlexGroup ............................................................................................... 8
   3.1 Massive Capacity and Predictable Low Latency for High-Metadata Workloads .............................. 8
   3.2 Efficient Use of All Cluster Hardware ............................................................................................. 9
   3.3 Simple, Easy-to-Manage Architecture and Balancing ....................................................................... 9
   3.4 Superior Density for Big Data ....................................................................................................... 9
   3.5 Superior Resiliency ......................................................................................................................... 9

4 Use Cases ................................................................................................................................................ 9
   4.1 Supported Features with NetApp FlexGroup ................................................................................... 10
   4.2 Ideal Use Cases ............................................................................................................................. 12
   4.3 Non-Ideal Cases ............................................................................................................................ 12

5 Performance ............................................................................................................................................. 12
   5.1 FlexVol Versus FlexGroup: Software Build .................................................................................... 12
   5.2 FlexGroup Versus Scale-Out NAS Competitor: Do More with Less ............................................. 13
   5.3 SPEC SFS 2014_swbuild Submission – FlexGroup Volume, ONTAP 9.2 ...................................... 14
   5.4 AFF A700 testing ............................................................................................................................. 17

6 NetApp FlexGroup Technical Overview .................................................................................................. 25
LIST OF FIGURES

Figure 1) FlexVol design with junctioned architecture for >100TB capacity. .................................................................5
Figure 2) Evolution of NAS file systems in ONTAP. ..............................................................................................................6
Figure 3) What is a large file? ................................................................................................................................................8
Figure 4) Git benchmark—Linux compile in FlexGroup versus FlexVol. .............................................................................13
Figure 5) Git benchmark—gcc compile in FlexGroup versus FlexVol ..............................................................................13
Figure 6) NetApp FlexGroup (2-node cluster) versus competitor (14-node cluster): standard NAS workload ..........14
Figure 7) Overall response time, SPEC SFS 2014_swbuild submissions. ............................................................15
Figure 8) Throughput, SPEC SFS 2014_swbuild submissions. ..................................................................................15
Figure 9) IOPS, SPEC SFS 2014_swbuild submissions. ............................................................................................16
Figure 10) Latency vs. number of builds, SPEC SFS 2014_swbuild submissions. .....................................................16
Figure 11) Standard NAS benchmark (EDA) – ONTAP 9.5: one node versus two nodes (operations/sec). .............19
Figure 12) Standard NAS benchmark (EDA) – ONTAP 9.4 versus ONTAP 9.5 (operations/sec). ........................20
Figure 13) Standard NAS benchmark (EDA) – ONTAP 9.5: one node versus two nodes (MBps). ...........................20
Figure 14) Standard NAS benchmark (EDA) – ONTAP 9.4 versus ONTAP 9.5 (MBps). ............................................21
Figure 15) Standard NAS benchmark (software builds) – ONTAP 9.5 (operations/sec). ......................................21
Figure 16) Standard NAS benchmark (software builds) – ONTAP 9.4 versus ONTAP 9.5 (operations/sec) ......22
Figure 17) Standard NAS benchmark (software builds) – ONTAP 9.5 (MBps). ..........................................................22
Figure 18) Standard NAS benchmark (software builds) – ONTAP 9.4 versus ONTAP 9.5 (MBps). .......................23
1 Data Is Growing

As hard-drive costs are driven down and flash hard-drive capacity grows exponentially, file systems are following suit. The days of file systems that number in the tens of gigabytes are over. Storage administrators face increasing demands from application owners for large buckets of capacity with enterprise-level performance.

With the advent of big data frameworks such as Hadoop, in which storage needs for a single namespace can extend into the petabyte range (with billions of files), an evolution of NAS file systems is overdue. NetApp FlexGroup is the ideal solution for these architectures.

1.1 Flexible Volumes: A Tried-and-True Solution

The flexible volume, NetApp FlexVol® software, was introduced in NetApp Data ONTAP® technology in 2005 as part of the Data ONTAP 7.0 (7-Mode) release. The concept was to take a storage file system and virtualize it across a hardware construct to provide flexible storage administration in an ever-changing data center.

FlexVol volumes could be grown or shrunk nondisruptively and be allocated to the storage operating system as thin-provisioned containers to enable overprovisioning of storage systems. Doing so allowed storage administrators the freedom to allocate space as consumers demanded it.

However, as data grew, file systems needed to grow. FlexVol can handle most storage needs with its 100TB capacity, and Data ONTAP provided a clustered architecture that those volumes could work in conjunction with. But the use case for very large buckets of storage in a single namespace required petabytes of storage. Figure 1 shows an example of what a FlexVol volume junction design for a large namespace would look like.

Figure 1) FlexVol design with junctioned architecture for >100TB capacity.
1.2 Infinite Volumes: Massive Capacity with Limitations

In NetApp Data ONTAP 8.1.1, the Infinite Volume solution was presented as a potential solution to enterprises with massively large storage needs. With a 20PB maximum and the capability to grow a single namespace nondisruptively, the Infinite Volume solution provided a more than capable method of storing large amounts of data.

Single Namespace Metadata Volume Limitations: Infinite Volume

Because the Infinite Volume solution used a single namespace volume for all metadata operations, several limitations applied:

- Less than stellar performance, with large amounts of metadata because volume affinity limits and serial operations created CPU efficiencies
- A 2-billion-file maximum due to the single FlexVol volume limit that was imposed by the metadata volume
- The inability to share SVMs with FlexVol volumes
- No SMB 2.x and 3.x support

Therefore, although Infinite Volume provided an excellent method to store archival data, it did not offer a way to cover multiple use cases in big data environments with predictable low latency.

Note: In ONTAP 9.5 and later, the Infinite Volume feature is no longer supported. See TR-4571 for details.

1.3 NetApp FlexGroup: An Evolution of NAS

ONTAP 9.1 brought innovation in scale-out NAS file systems: NetApp FlexGroup.

With FlexGroup volumes, a storage administrator can easily provision a massive single namespace in a matter of seconds. FlexGroup volumes have virtually no capacity or file count constraints outside of the physical limits of hardware or the total volume limits of ONTAP. Limits are determined by the overall number of constituent member volumes that work in collaboration to dynamically balance load and space allocation evenly across all members. There is no required maintenance or management overhead with a FlexGroup volume. You simply create the FlexGroup volume and share it with your NAS clients. ONTAP does the rest.

Figure 2) Evolution of NAS file systems in ONTAP.
2 Terminology

Many of the usual NetApp ONTAP terms (such as storage virtual machine, logical interface [LIF], FlexVol, and so on) are covered in TR-3982: NetApp Clustered Data ONTAP 8.3.x and 8.2.x. FlexGroup specific terminology is covered in the following sections.

FlexGroup

A FlexGroup volume is a single namespace that is made up of multiple constituent member volumes and that is managed and acts like a NetApp FlexVol volume to storage administrators. Files in a FlexGroup volume are allocated to individual member volumes and are not striped across volumes or nodes.

Affinity

Affinity describes the tying of a specific operation to a single thread.

Automated Incremental Recovery (AIR)

Automated Incremental Recovery (AIR) is an ONTAP subsystem that repairs FlexGroup inconsistencies dynamically, with no outage or administrator intervention required.

Constituent and Member

Constituent and member are interchangeable terms when referring to FlexGroup. These items are the underlying FlexVol volumes that make up a FlexGroup volume and provide the capacity and performance gains that are seen only with a FlexGroup volume.

Ingest

Ingest is the consumption of data by way of file or folder creations.

Junction Paths

Junction paths are the process that joins multiple FlexVol volumes together to scale out across a cluster, provide multiple volume affinities, and provide capacity beyond the 100TB limit of a FlexVol volume. The use of a junction path in ONTAP is known as “mounting” the volume within the ONTAP namespace.

Large Files

See What Are Large Files?, later in this section.

Overprovisioning and Thin Provisioning

Overprovisioning (or thin provisioning) storage is the practice of disabling a volume’s space guarantee (guarantee = none). This practice allows the virtual space allocation of the FlexVol volume to exceed the physical limits of the aggregate that it resides on. For example, with overprovisioning, a FlexVol volume can be 100TB on an aggregate that has a physical size of only 10TB. Overprovisioning allows storage administrators to grow volumes to very large sizes to avoid the need to grow them later, but it does present the management overhead of needing to monitor available space very closely.

In the case of overprovisioned volumes, the available space reflects the actual physical available space in the aggregate. As such, the usage percentage and capacity available values might seem a bit off. However, they simply reflect a calculation of the actual space that is available compared with the virtual space that is available in the FlexVol volume. For a more accurate portrayal of space allocation when using overprovisioning, use the aggregate show-space command.

Remote Access Layer (RAL)

The remote access layer (RAL) is a feature in the NetApp WAFL® (Write Anywhere File Layout) system that allows a FlexGroup volume to balance ingest workloads across multiple FlexGroup constituents or members.
Remote Hard Links

*Remote hard links* are the building blocks of NetApp FlexGroup. These links act as normal hard links but are unique to ONTAP. The links allow a FlexGroup volume to balance workloads across multiple remote members or constituents. In this case, “remote” simply means “not in the parent volume.” A remote hard link can be another FlexVol member on the same aggregate or node.

What Are Large Files?

This document uses the term *large file* liberally. Therefore it’s important to define up front exactly what a “large file” is with regard to NetApp FlexGroup volumes.

A FlexGroup volume operates best when a workload contains numerous small files or metadata operations, because FlexGroup can maximize the system resources to address those specific workloads. FlexGroup volumes also work well with various other workloads (as defined in section 4, Use Cases). One type of workload that can create problems, however, is a workload with large files or files that grow over time, such as database files.

In a NetApp FlexGroup volume, a large file is a product of the percentage of allocated space, not of any specific file size. Thus, in some FlexGroup configurations—for example, in which the member volume size is only 1TB—a “large file” might be 500GB (50% of the member volume size). In other configurations—for example, in which the member volume size is 100TB—that same 500GB file size would take up only 0.5% of the volume capacity. This type of file could be large enough to throw off the ingest heuristics in the FlexGroup volume, or it could potentially create problems later when the member volume gets closer to full.

Figure 3) What is a large file?

3 Advantages of NetApp ONTAP FlexGroup

3.1 Massive Capacity and Predictable Low Latency for High-Metadata Workloads

Previously, NetApp ONTAP technology did not have a solution for this challenge. Earlier versions were constrained by architectural limitations and the notion of volume affinity: the tendency of ONTAP operations, particularly metadata operations, to operate in a single serial CPU thread.

The FlexGroup feature solves this problem by automatically balancing ingest workloads across multiple constituent NetApp FlexVol members to provide multiple affinities to handle high-metadata workloads.
3.2 Efficient Use of All Cluster Hardware

Previously, file systems in ONTAP were tied to a single FlexVol container. Although it was possible to scale volumes across multiple nodes in a cluster, the management overhead was cumbersome, and the process did nothing to increase the total capacity of a single namespace. To achieve this type of scale, volumes could be junctioned to one another. For an example, see Figure 1.

3.3 Simple, Easy-to-Manage Architecture and Balancing

With FlexVol volumes, to achieve scale beyond the single node or aggregate that owns the FlexVol volume, several volumes had to be junctioned to one another. This concept required design, architecture, and management overhead that took valuable time away from storage administrators’ day-to-day operations. A FlexGroup volume can provision storage across every node and aggregate in a cluster in less than a minute through the FlexGroup tab in NetApp OnCommand® System Manager.

3.4 Superior Density for Big Data

A FlexGroup volume provides the ability to condense large amounts of data into smaller data center footprints by way of the superb storage efficiency features of ONTAP, including:

- Thin provisioning
- Data compaction
- Inline data compression
- Inline deduplication

In addition, ONTAP supports 30.2TB solid-state drives (SSDs), which can deliver ~724TB of raw capacity in a single 24-drive enclosure. It is possible to get over a petabyte of raw capacity in just 10U of rack space with NetApp FlexGroup, which cuts costs on cooling, power consumption, and rack rental space and offers excellent density in the storage environment.

3.5 Superior Resiliency

Another benefit of NetApp FlexGroup is the recovery infrastructure known as Automated Incremental Recovery (AIR). This feature addresses metadata inconsistencies that are found on data access from clients in real time, with little to no disruption of file system access during the recovery. AIR repairs inconsistencies without administrative intervention and without the FlexGroup volume taking itself offline.

4 Use Cases

The NetApp FlexGroup design is most beneficial in specific use cases, but those use cases are listed as “ideal.” Other use cases for a FlexGroup volume are possible, but they generally depend on feature support. In most instances, the use case is limited to the supported feature set. For example, virtualization workloads can work on FlexGroup volumes, but they currently lack support for SIS clone, FlexClone®, NFSV4.1, and VAAI/copy offload.
### 4.1 Supported Features with NetApp FlexGroup

The following table shows the current list of supported ONTAP features for NetApp FlexGroup. For questions about supported features not listed, e-mail flexgroups-info@netapp.com.

Table 1) General ONTAP feature support.

<table>
<thead>
<tr>
<th>Supported Feature</th>
<th>Version of ONTAP First Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>NetApp Snapshot™ technology</td>
<td>ONTAP 9.0</td>
</tr>
<tr>
<td>NetApp SnapRestore® software (FlexGroup level)</td>
<td>ONTAP 9.0</td>
</tr>
<tr>
<td>Hybrid aggregates</td>
<td>ONTAP 9.0</td>
</tr>
<tr>
<td>Constituent or member volume move</td>
<td>ONTAP 9.0</td>
</tr>
<tr>
<td>Postprocess deduplication</td>
<td>ONTAP 9.0</td>
</tr>
<tr>
<td>NetApp RAID-TEC™ technology</td>
<td>ONTAP 9.0</td>
</tr>
<tr>
<td>Per-aggregate consistency point</td>
<td>ONTAP 9.0</td>
</tr>
<tr>
<td>Sharing NetApp FlexGroup with FlexVol in the same SVM</td>
<td>ONTAP 9.0</td>
</tr>
<tr>
<td>NetApp OnCommand support</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>Inline adaptive compression</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>Inline deduplication</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>Inline data compaction</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>Thin provisioning</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>NetApp AFF</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>Quota reporting</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>NetApp SnapMirror® technology</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>User and group quota reporting (no enforcement)</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>Aggregate inline deduplication (cross-volume deduplication)</td>
<td>ONTAP 9.2</td>
</tr>
<tr>
<td>NetApp Volume Encryption (NVE)</td>
<td>ONTAP 9.2</td>
</tr>
<tr>
<td>NetApp SnapVault® technology</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>Qtrees</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>Automated deduplication schedules</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>Version independent SnapMirror/unified replication</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>Antivirus scanning for SMB</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>Volume autogrow</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>QoS maximu/ceilings</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>FlexGroup expansion without SnapMirror rebaseline</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>Improved ingest heuristics</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>Supported Feature</td>
<td>Version of ONTAP First Supported</td>
</tr>
<tr>
<td>-------------------------------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>SMB change/notify</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>File audit</td>
<td>ONTAP 9.4</td>
</tr>
<tr>
<td>NetApp FPolicy™</td>
<td>ONTAP 9.4</td>
</tr>
<tr>
<td>Adaptive QoS</td>
<td>ONTAP 9.4</td>
</tr>
<tr>
<td>QoS minimums (AFF only)</td>
<td>ONTAP 9.4</td>
</tr>
<tr>
<td>Relaxed SnapMirror limits</td>
<td>ONTAP 9.4</td>
</tr>
<tr>
<td>SMB 3.x Multichannel</td>
<td>ONTAP 9.4</td>
</tr>
<tr>
<td>FabricPool</td>
<td>ONTAP 9.5</td>
</tr>
<tr>
<td>Quota enforcement</td>
<td>ONTAP 9.5</td>
</tr>
<tr>
<td>Qtree statistics</td>
<td>ONTAP 9.5</td>
</tr>
<tr>
<td>Inherited SMB watches and change notifications</td>
<td>ONTAP 9.5</td>
</tr>
<tr>
<td>SMB copy offload (ODX)</td>
<td>ONTAP 9.5</td>
</tr>
<tr>
<td>Storage-Level Access Guard</td>
<td>ONTAP 9.5</td>
</tr>
<tr>
<td>NetApp FlexCache® (cache only; FlexGroup as origin not supported yet)</td>
<td>ONTAP 9.5</td>
</tr>
</tbody>
</table>

Table 2) General NAS protocol version support.

<table>
<thead>
<tr>
<th>Supported NAS Protocol Version</th>
<th>Version of ONTAP First Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFSv3</td>
<td>ONTAP 9.0</td>
</tr>
<tr>
<td>SMB 2.1, SMB 3.x</td>
<td>ONTAP 9.1 RC2</td>
</tr>
</tbody>
</table>

Table 3) Unsupported SMB 2.x and 3.x features.

<table>
<thead>
<tr>
<th>Unsupported SMB 2.x Features</th>
<th>Unsupported SMB 3.x Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>• SMB Remote Volume Shadow Copy Service (VSS)</td>
<td>• Continuously available shares and SMB transparent failover</td>
</tr>
<tr>
<td></td>
<td>• SMB scale-out</td>
</tr>
<tr>
<td></td>
<td>• VSS for SMB file shares</td>
</tr>
<tr>
<td></td>
<td>• SMB directory leasing</td>
</tr>
<tr>
<td></td>
<td>• SMB direct or remote direct memory access (RDMA)</td>
</tr>
</tbody>
</table>

Note: SMB 3.0 encryption is supported with FlexGroup volumes.
4.2 Ideal Use Cases

A FlexGroup volume works best with workloads that are heavy on ingest (a high level of new data creation), heavily concurrent, and evenly distributed among subdirectories:

- Electronic Design Automation
- Artificial intelligence and machine learning
- Log file repositories
- Software build/test environments (such as GIT)
- Seismic/oil and gas
- Media asset or HIPAA archives
- File streaming workflows
- Unstructured NAS data (such as home directories)
- Big data and data lakes ([Hadoop with the NetApp NFS connector](#))

4.3 Non-Ideal Cases

The general guidance for workloads that currently should not be used on a FlexGroup volume include:

- Very large files or files that grow over time
- Virtualized workloads
- Workloads that require striping (large files spanning multiple nodes or volumes)
- Workloads that require specific control over the layout of the relationships of data to NetApp FlexVol volumes
- Workloads that require specific features and functionality that are not currently available with FlexGroup volumes
- Workloads that leverage ONTAP Cloud or NetApp Cloud Volumes (currently unsupported)

5 Performance

Although NetApp FlexGroup is positioned as a capacity play, it's a performance play as well. With a FlexGroup volume, there are no trade-offs; you can have massive capacity and predictable low-latency and high-throughput performance with the same storage container. A FlexGroup volume can accomplish this goal by adding concurrency to workloads and presenting multiple volume affinities to a single storage container without the clients or storage administrators needing to manage anything. In metadata-intensive workloads with high file counts, being able to present multiple volumes and cluster nodes quickly and easily enables ONTAP to use multiple hardware assets and CPU cores to perform at a higher performance threshold.

5.1 FlexVol Versus FlexGroup: Software Build

In a simple workload benchmark using a software build tool (Git), a Linux kernel was compiled on a single FAS8080 node running ONTAP 9.1 with two aggregates of SAS drives and eight FlexVol member constituents in a FlexGroup, versus a single FlexVol on the same hardware. The metric being measured was a simple “time to completion” test. In this benchmark, the FlexGroup volume outperformed the FlexVol volume by 2 to 6 times across multiple Git operations. In addition, the same Git test was run with a gcc compile on All Flash FAS.

**Note:** The gcc compile works with a higher file count, thus the differences in completion times.
5.2 FlexGroup Versus Scale-Out NAS Competitor: Do More with Less

In another benchmark, we compared a FlexGroup volume on a 2-node FAS8080 cluster running ONTAP 9.1 using SAS drives against a competitor system leveraging 14 nodes. The competitor system also leveraged some SSDs for metadata caching. This test used a standard NAS workload generation tool to simulate workloads.

In the test, we saw that a single FlexGroup volume with 8 member constituents was able to ingest nearly the same amount of ops/second at essentially the same latency curve as the competitor’s 14-node cluster.
5.3 SPEC SFS 2014_swbuild Submission – FlexGroup Volume, ONTAP 9.2

NetApp also submitted results from the official SPEC SFS 2014_swbuild benchmark test, which allow storage vendors to test their systems against a standardized test that is approved by an independent benchmarking consortium. For the NetApp results of this test, see:


For results of competitor systems, see:


NetApp Results

The benchmark includes a metric known as overall response time (ORT), defined here:

The overall response time is a measure of how the system will respond under an average load. Mathematically, the value is derived by calculating the area under the curve divided by the peak throughput.
In that test, NetApp FlexGroup volumes achieved the lowest ORT ever recorded for a storage system.

Figure 7) Overall response time, SPEC SFS 2014_swbuild submissions.

![SPEC SFS®2014_swbuild](image)

FlexGroup volumes also outperformed other submissions in throughput. In the benchmark, FlexGroup volumes achieved over 4GBps.

Figure 8) Throughput, SPEC SFS 2014_swbuild submissions.

![SPEC SFS®2014_swbuild](image)

The results of this performance benchmark were achieved with more than 500 concurrent jobs, providing 260,000 IOPS.
If latency is important to your business, FlexGroup volumes also saw the most predictable low latency of all the submissions.

Figure 10) Latency vs. number of builds, SPEC SFS 2014_swbuild submissions.
5.4 AFF A700 testing

In addition to the 4-node AFF8080 tests, the same Git workload was also run on an AFF A700 cluster. The following configuration was used:

- 2-node AFF A700 cluster
- Single aggregate of 800GB SSDs per node
- FlexVol: single node, 100% local
- FlexGroup: spans HA pair, 8 members per node (16 members total)

The workload was as follows:

- GCC library compile
- Clone operations only (these showed highest maximum throughput for both FlexVol and FlexGroup)
- 4 physical servers
- User workload/threads on the clients ranging from 4 to 224

Figure 10 compares the maximum achieved throughput (read + write) on Git clone operations on a single FlexVol volume versus a single FlexGroup spanning 2 nodes. Note in the graph how the maximum throughput reaches nearly 5x the amount of the FlexVol without seeing the same degradation the FlexVol sees as the workload reaches 64 threads.

Figure 10) FlexVol versus FlexGroup: Maximum throughput trends under increasing workload.

Figure 11 compares a FlexVol and FlexGroup in the same configurations. This time, we break down the maximum read and write throughput individually, as well as comparing that against the average throughput for the FlexVol and FlexGroup.
Figure 11) FlexVol versus FlexGroup: Maximum throughput trends under increasing workload – detailed.

Figure 12 shows the maximum total average IOPs for a FlexGroup versus a FlexVol on the AFF A700. Again, note the drastic increase of IOPs for the FlexGroup versus the degradation of IOPs at 64 threads for the FlexVol.

Figure 12) FlexVol versus FlexGroup: Maximum average total IOPs.
ONTAP 9.4 and 9.5 Performance Testing

For ONTAP version 9.4 and 9.5, we ran a new set of performance tests using standard NAS benchmark suites that simulate both EDA and software build workloads. The goal was to show that ONTAP improves performance with each release.

As discussed earlier in “NetApp AFF8080 Testing – ONTAP 9.1,” we ran standard NAS benchmark tests for a four-node AFF8080 cluster running ONTAP 9.1 for software builds. Our tests showed that the total IOPS achieved before latency started to spike was around 150,000–200,000.

The ONTAP 9.4 and 9.5 tests featured the following configurations:

- An AFF A700s cluster
- A FlexGroup volume spanning a single node and two nodes
- 14 NFSv3 clients
- 32 10GB LIFs (16 LIFs per node)
- 32 mount points on each client

The following graphs show that performance in a FlexGroup volume can scale, and that each release provides fairly substantial performance improvements in ONTAP. This can be accomplished with a nondisruptive upgrade.

The first set of graphs is for an EDA workload.

Figure 11) Standard NAS benchmark (EDA) – ONTAP 9.5: one node versus two nodes (operations/sec).
Figure 12) Standard NAS benchmark (EDA) – ONTAP 9.4 versus ONTAP 9.5 (operations/sec).

Figure 13) Standard NAS benchmark (EDA) – ONTAP 9.5: one node versus two nodes (MBps).
This next set of graphs shows the performance for a standard NAS benchmark running a software build workload (such as Git or Perforce). Both types of workloads are ideal for FlexGroup volumes because of the high file ingest rates and need for parallel processing of write metadata.

Figure 14) Standard NAS benchmark (EDA) – ONTAP 9.4 versus ONTAP 9.5 (MBps).

Figure 15) Standard NAS benchmark (software builds) – ONTAP 9.5 (operations/sec).
Figure 16) Standard NAS benchmark (software builds) – ONTAP 9.4 versus ONTAP 9.5 (operations/sec).

Figure 17) Standard NAS benchmark (software builds) – ONTAP 9.5 (MBps).
Figure 18) Standard NAS benchmark (software builds) – ONTAP 9.4 versus ONTAP 9.5 (MBps).

For more performance information, see TR-4571: FlexGroup Volumes Best Practices and Implementation Guide.
Reducing Costs Through Density

The test just described leveraged 13 shelves of 600GB SAS drives. However, NetApp FlexGroup also supports configurations using All Flash FAS systems.

With the support of high-capacity SSDs, in addition to superior storage efficiency technologies such as data compaction and inline compression and deduplication, you can generate the same workload, with 2PB usable* in just 4U of rack space.

Being able to use large-capacity drives and NetApp FlexGroup volumes to reduce rack space usage in the data center can provide cost savings for the following:

- Cooling and power
- Rack space rental
- Person-hours for installation and maintenance
- Better drive lifetimes for SSDs mean fewer replacements

* Usable space includes 4:1 storage efficiency calculation.

Note: Power/cooling estimates vary based on region and provider. Publicly available power calculator used for estimates.
6 NetApp FlexGroup Technical Overview

NetApp FlexGroup has taken the concept of FlexVol in ONTAP and leveraged a WAFL subsystem known as the Remote Access Layer (RAL). This layer directs the ingest of new files and folders and keeps track of existing files and folders for fast redirection on reads.

6.1 Overview of a FlexGroup Volume

At a high level, a FlexGroup volume is simply a collection of FlexVol volumes acting as a single entity. NAS clients access the FlexGroup volume just as they would any normal FlexVol volume: from an export or a CIFS/SMB share.

Figure 14) A FlexGroup volume.

Although the underlying concept of a FlexGroup volume is a FlexVol volume, there are several benefits a FlexGroup volume offers that a normal FlexVol volume cannot. See section 3, Advantages of NetApp ONTAP FlexGroup, for details.

A FlexGroup volume creates files on a per FlexVol basis—there is no file striping. The throughput gains for a FlexGroup volume are seen by way of concurrency of operations across multiple FlexVol volumes, aggregates, and nodes. Now a series of operations can occur in parallel across all hardware on which the FlexGroup volume resides. **NetApp FlexGroup is the perfect complement** to the clustered ONTAP scale-out architecture.

6.2 File Creation and Automatic Load Balancing

When a file is created in a FlexGroup volume, that file is directed to the “best available” FlexVol member in the FlexGroup volume. “Best available” in this case means “most free space available” and the recent load on a FlexVol member. The goal of the FlexGroup volume is to ensure that member volumes are as evenly allocated as possible. The goal is also to have the ingest workload of a FlexGroup volume as evenly distributed across members as possible, with the fewest number of remote hard links possible. This concept is known as “automatic load balancing” and is transparent to clients and storage administrators. This concept adds to the overall simplicity of the NetApp FlexGroup story: Storage administrators provision the storage in seconds and don’t have to think about the design or layout.

**Note:** If a FlexGroup member fills prematurely, the entire FlexGroup volume issues an ENOSPC error to the client. Therefore, it is important to ensure that member constituents are properly sized and workloads are properly vetted.
Local Versus Remote Placement

Because a FlexGroup volume has multiple constituent volumes and FlexGroup is designed to place data evenly in all constituents, there is a notion of “remote placement” of files.

In addition, ONTAP can operate in up to 24-node clusters in NAS-only configurations, so there is also a notion of “remote traffic.”

However, these two concepts are not synonymous. *Remote traffic* is traffic over the cluster interconnect network and is present in all ONTAP configurations, not just in a FlexGroup volume. *Remote placement* is unique to NetApp FlexGroup and occurs when a file or folder is created on a FlexGroup member that does not own the parent directory. Remote placement can occur even on the node that owns the parent volume by way of a remote hard link allocated through the Remote Access Layer (RAL).

Remote placement is traffic over the cluster interconnect network and is present in all ONTAP configurations, not just in a FlexGroup volume. Remote placement is unique to NetApp FlexGroup and occurs when a file or folder is created on a FlexGroup member that does not own the parent directory. Remote placement can occur even on the node that owns the parent volume by way of a remote hard link allocated through the Remote Access Layer (RAL).

Local placement involves creating the new object so that it is stored in the same constituent as its parent directory. Local placement provides the highest metadata performance for that file or subdirectory in future operations. Remote placement, in contrast, causes that file or subdirectory to suffer a slight metadata access penalty. However, it allows the FlexGroup volume to place the new content on a different constituent from its parent directory to better distribute the collective dataset and workload. The penalty seen from remote placement of files or folders is more than offset by the performance and throughput gains of having multiple volume affinities for workloads.

To accomplish an evenly balanced workload, ONTAP monitors the status and member volume attributes every second for optimal placement of data. This process seeks to accomplish a balance of multiple goals:

- Space usage remains balanced; if a member fills prematurely, the FlexGroup volume reports being out of space even if additional space is available on other members.
- Avoidance of overutilized member constituents by analyzing data creation patterns.
- Local placement of files to ensure that latency is as low as possible.

*Figure 15*) Remote placement of files through remote hard links.
Example of File/Folder Ingest

On GitHub, we added a `dd` script to do parallel `dd` operations on a client to provide a better throughput test than a single-threaded `dd` command. This script was used to illustrate how ONTAP places files across a FlexGroup on ingest. We then used the `volume explore` command to find the location of the files based on their inodes and mapped out some of their locations in the following scripts.

The configuration:
- ONTAP 9.3
- Single-node FlexGroup with 8 members
- Single client running the `dd` script

This is the FlexGroup volume size output before running the script:

```
cluster::*> vol show -vserver DEMO -volume flexgroup_local* -fields used,percent-used,size -sort-by used
vserver volume size   used   percent-used
------- ----------------- ----------- -----------
DEMO   flexgroup_local_0003 2.50TB 57.28MB 5%
DEMO   flexgroup_local_0004 2.50TB 57.28MB 5%
DEMO   flexgroup_local_0005 2.50TB 57.28MB 5%
DEMO   flexgroup_local_0007 2.50TB 57.28MB 5%
DEMO   flexgroup_local_0002 2.50TB 57.28MB 5%
DEMO   flexgroup_local_0001 2.50TB 57.29MB 5%
DEMO   flexgroup_local_0006 2.50TB 57.29MB 5%
DEMO   flexgroup_local_0008 2.50TB 57.29MB 5%
DEMO   flexgroup_local       20TB  458.2MB 61%
9 entries were displayed.
```

The script creates folders at the top level specified in the script, and then populates each folder with the specified number of files at a specific size. In this test, we chose 8 folders and 8 files to mirror the FlexGroup member volume count. We chose a file size of 1GB.

After running the script, the FlexGroup members looked like this (deduplication and compression kept the file sizes below 1GB):

```
cluster::*> vol show -vserver DEMO -volume flexgroup_local* -fields used,percent-used,size,files,files-used -sort-by used
vserver volume size      used percent-used files     files-used
------- ----------------- ----------- ----------- ---------- ----------
DEMO   flexgroup_local_0004 2.50TB 86.88MB 5%      21251126  106
DEMO   flexgroup_local_0007 2.50TB 90.85MB 5%      21251126  106
DEMO   flexgroup_local_0005 2.50TB 91.20MB 5%      21251126  107
DEMO   flexgroup_local_0006 2.50TB 91.23MB 5%      21251126  108
DEMO   flexgroup_local_0008 2.50TB 91.58MB 5%      21251126  107
DEMO   flexgroup_local_0002 2.50TB 92.48MB 5%      21251126  107
DEMO   flexgroup_local_0001 2.50TB 92.93MB 5%      21251126  114
DEMO   flexgroup_local_0003 2.50TB 96.44MB 5%      21251126  108
DEMO   flexgroup_local     20TB  733.6MB 61%     170009008  863
9 entries were displayed.
```

The files were placed relatively evenly across all member volumes, as shown in the above output. A closer look at the layout (by using the `volume explore` command) shows that the folders were allocated evenly across the member volumes.


However, on a per-folder basis, the files were not allocated evenly. This is because the files were 1GB in size and were more likely to go remote from their parent folder due to their size.

Figure 16) Folder distribution in a FlexGroup volume, dd script.

Figure 17) File distribution in a FlexGroup volume, folder c4.
Figure 18) File distribution in a FlexGroup volume, folder c6.

Figure 19) File distribution in a FlexGroup volume, folder c7.
Although the files were not evenly distributed on a per-folder basis, the result was that the files across all folders were evenly allocated over time. This result is shown in the space distribution output:

```
cluster:*> vol show -vserver DEMO -volume flexgroup_local* -fields used,percent-used,size -sort-by used

vserver volume size used percent-used
----- --------------------- ---- ------ ------------
DEMO flexgroup_local__0002 2.50TB 85.77MB 5%
DEMO flexgroup_local__0008 2.50TB 85.92MB 5%
DEMO flexgroup_local__0003 2.50TB 86.04MB 5%
DEMO flexgroup_local__0006 2.50TB 89.72MB 5%
DEMO flexgroup_local__0004 2.50TB 90.34MB 5%
DEMO flexgroup_local__0007 2.50TB 94.10MB 5%
DEMO flexgroup_local__0005 2.50TB 94.18MB 5%
DEMO flexgroup_local__0001 2.50TB 94.23MB 5%
DEMO flexgroup_local     20TB  720.3MB 61%
9 entries were displayed.
```

**Figure 20** Space distribution across member volumes – `dd` script.

### Remote Access Layer (RAL)

The Remote Access Layer (RAL) is a new mechanism in ONTAP provided with the NetApp FlexGroup feature. RAL allows a single WAFL message that runs against one member volume to manipulate inodes in that member volume and inodes on other member volumes.

### Ingest Heuristics

Ingest heuristics allow ONTAP to make intelligent decisions for file ingest based on a series of decision points. Member volumes participating in a FlexGroup volume refresh every second to ensure an up-to-date view of the current state of the volumes.

### Remote Allocation

A FlexGroup volume allocates workloads based on the following:

- How much data a member constituent holds (% used)
- Amount of available free space in members
- Last N seconds’ worth of new content allocation requests
- Last N seconds’ worth of inode allocations and where they were drawn from
- Amount of free inodes/files in a member volume (ONTAP 9.3 and later)
A FlexGroup volume generally favors local placement when possible, but in some cases, remote allocation of data is more likely. These scenarios include:

- Creating subdirectories near the top-level junction of the FlexGroup volume
- Creating new files or folders in directories that already have many files or folders
- When member constituent space allocation has a high discrepancy in capacity utilization
- When member constituent volumes approach 90% capacity or inode utilization
- If there is an unbalanced load (one member is getting more traffic than others)

In some cases, the member will be a different FlexVol volume than the parent directory, but the FlexGroup ingest heuristics tend to favor local traffic for files over remote traffic. For directories, a FlexGroup volume tends to favor remote creation over local creation in FlexVol members. As traffic normalizes on a FlexGroup volume and the FlexGroup volume begins to allocate files, the allocation favors local traffic more and remote traffic hovers around 5 to 10%.

**Urgency**

_Urgency_ in a FlexGroup volume is how full a member is (or, as of ONTAP 9.3, how close to the max file count the volume is) versus how likely it is to be used for ingest of data. Each node maintains two global variables used in determining the probability of how likely new data will hit a member volume in a FlexGroup volume:

**Free-warning**

This variable is set to 50GB by default. This number represents the amount of free space available in a FlexGroup member and is used to calculate the probability of how urgent it is to place content in remote members.

**Free-danger**

This variable is set to 10GB by default. This number represents the threshold at which the member volume’s urgency will be set to 100%. All ingest traffic will avoid the member volume until sufficient free space is added to the member or data is deleted.

**Tolerance**

_Tolerance_ in a FlexGroup volume is a measure of usage disparity between members or the percentage of disparity of used space between members that a FlexGroup volume can tolerate before generating more remote allocation decisions.

Tolerance is controlled through three node-level global variables:

**Max-tolerance**

This value is set to a default of 10%. This means that a member volume can tolerate up to 10% of the working-set value (100GB) of used space before it has to send traffic remote a higher percentage of the time. For example, if one member is >10% more full than another member, traffic will be diverted elsewhere. Empty member volumes always use the max-tolerance value.

**Min-tolerance**

This value is set to a default of 0%. When a member volume is full, the min-tolerance value is enforced and traffic is sent remote 100% of the time in an attempt to even up the space distribution.

**Working-set**

This variable defines the free space level that the max- and min-tolerance percentages use for their calculations. The default of this value is 100GB.
Caveats

In cases in which a member volume starts to become “more full” than other member volumes, performance can deteriorate on the FlexGroup volume because the workload is creating more remote hard links.

In cases in which member volumes fill up completely in a FlexGroup volume, the entire FlexGroup volume reports ENOSPC errors to the client. Remediation steps must be taken to correct the issue (such as growing the FlexGroup members or deleting data). This also applies to member volumes running out of inodes. However, ONTAP 9.3 improves the ingest calculations to take member volume inode counts into consideration when allocating files.

Local Versus Remote Test

A simple file and directory creation test was performed to measure the local versus the remote placement for files and directories. The following setup was used:

- Two FAS8040 nodes
- Two SSD aggregates (non-AFF personality)
- Four FlexVol member constituents per aggregate; eight total members
- 100,000 directories
- 10,000,000,000 files (100,000 per directory x 100,000 directories)
- Red Hat 7.x client
- Simple `mkdir` and `truncate` commands for loops:

```bash
for x in `seq 1 100000`; do mkdir dir$x; done
for x in dir{1..100000}; do truncate -s 1k /mnt/$x/file{1..100000}; done
```

In the above scenario, the remote allocation of directories was at 90%:

```
remote_dirs 90
```

The remote allocation of files was only 10%:

```
remote_files 10
```

The statistics above were pulled from the command `statistics show -object flexgroup in advanced privilege`. See the appendix for information about how to collect and view FlexGroup statistics.

6.3 Volume Autogrow

In ONTAP 9.3, support for volume autogrow was added for NetApp FlexGroup volumes. This support enables a storage administrator to set an autogrow policy for the FlexGroup volume that allows ONTAP to increase the FlexVol size to a predefined threshold when a volume approaches capacity. This ability is especially useful in a FlexGroup volume, because volume autogrow can help prevent member volumes from filling prematurely and causing premature “out of space” scenarios in the entire FlexGroup. Applying volume autogrow to a FlexGroup is done in the same way as with a FlexVol volume. See the appendix for an example of how to apply volume autogrow.

Note: Autoshrink is also supported with autosize.

6.4 64-Bit File Identifiers

Today, NFSv3 in ONTAP uses 32-bit file IDs by default. Doing so provides 2,147,483,647 maximum unsigned integers. With the 2 billion inode limit in FlexVol, this value fit nicely into the architecture.
However, because NetApp FlexGroup can support up to 400 billion files in a single container, the implementation of 64-bit file IDs was needed, because they support up to 9,223,720,368,547,758,775,807 maximum unsigned integers.

NetApp highly recommends enabling the NFS server option `-v3-64bit-identifiers` at advanced privilege level before creating a FlexGroup volume.

The 64-bit file identifier option is off by default, to make certain that legacy applications that require 32-bit file identifiers are not unexpectedly affected by ONTAP changes before administrators can properly evaluate their environments.

This option can be found at the advanced privilege level.

```
[-v3-64bit-identifiers {enabled|disabled}] - Use 64 Bits for NFSv3 FSIDs and File IDs (privilege: advanced)
```

This optional parameter specifies whether Data ONTAP uses 64 bits (instead of 32 bits) for file system identifiers (FSIDs) and file identifiers (file IDs) that are returned to NFSv3 clients. If you change the value of this parameter, clients must remount any paths over which they are using NFSv3. When `-v3-fsid-change` is disabled, enable this parameter to avoid file ID collisions.

To enable:

```
cluster::> set advanced
cluster::*> nfs server modify -vserver SVM -v3-64bit-identifiers enabled
```

After enabling or disabling this option, you must remount all clients. Otherwise, because the file system IDs change, the clients might receive stale file handle messages when attempting NFS operations.

If a FlexGroup volume will not exceed 2 billion files, you can leave this value unchanged. However, to prevent any file ID conflicts, the inode maximum on the FlexGroup volume should also be increased to no more than 2,147,483,647.

**System Manager Support**

Starting in ONTAP 9.2, it is also possible to enable/disable the NFS server option from OnCommand System Manager. To do this, navigate to the SVM and select SVM Settings from the menu tabs.

**Figure 21** SVM Settings tab.

![SVM Settings tab](image)

Click the NFS link on the left, under Protocols and then click Edit. The SVM Settings dialog box that opens contains a checkbox that allows you to enable or disable 64-bit file identifiers.
After enabling or disabling this option, all clients must be remounted, because the file system IDs change and the clients might receive stale file handle messages when attempting NFS operations.

Impact of File ID Collision

If 64-bit file IDs are not enabled, the risk of file ID collisions increases. When a file ID collision occurs, the impact can range from a “stale file handle” error on the client to directory and file listings failing, to an application failing entirely. In most cases, it is imperative to enable the 64-bit file ID option when using NetApp FlexGroup volumes.

Effects of File System ID (FSID) Changes in ONTAP

NFS makes use of a file system ID (FSID) when interacting between client and server. This FSID lets the NFS client know where data lives in the NFS server’s file system. Because ONTAP can span multiple file systems across multiple nodes by way of junction paths, this FSID can change depending on where data lives. Some older Linux clients can have problems differentiating these FSID changes, resulting in failures during basic attribute operations, such as chown, chmod, and so on.

An example of this issue can be found in bug 671319. If disabling the FSID change with NFSv3, be sure to enable the new -v3-64bit-identifiers option in ONTAP 9, but keep in mind that this option could affect older legacy applications that require 32-bit file IDs. NetApp recommends leaving the FSID change option enabled with NetApp FlexGroup volumes to help prevent file ID collisions.

How FSIDs Operate with Snapshot Copies

When a Snapshot copy of a volume is taken, a copy of a file’s inodes is preserved in the file system for access later. The file theoretically exists in two locations.

With NFSv3, even though there are two copies of essentially the same file, the FSIDs of those files are not identical. FSIDs of files are formulated using a combination of NetApp WAFL inode numbers, volume identifiers, and Snapshot IDs. Because every Snapshot copy has a different ID, every Snapshot copy of a file has a different FSID in NFSv3, regardless of the setting of the option -v3-fsid-change. The NFS RFC spec does not require FSIDs for a file to be identical across file versions.

Note: The -v4-fsid-change option does not apply to NetApp FlexGroup volumes, because NFSv4 is currently unsupported with FlexGroup volumes.

Directory Size Considerations

In ONTAP, there are limitations to the maximum directory size on disk. This limit is known as maxdirsize. The maxdirsize for a volume is capped at 320MB, regardless of platform. That means the memory allocation for the directory size can reach a maximum
of only 320MB before a directory can no longer grow larger. How many files can fit into a single directory with the default maxdirsize?

This KB article (internal and partner only) does a great job of breaking down the math to figure out whether a directory’s index will cause maxdirsize issues. If you don’t have access to that KB, here’s an example of how many files can fit into a single directory in a larger memory system by default.

CIFS/SMB files use up more space per file than NFS due to the namelength size and Unicode encoding. Therefore, volumes that are accessed via SMB (either dedicated SMB or multiprotocol NAS) can hold fewer files per folder than NFS-only volumes.

The formulas to figure out the number of files are:

- Memory in KB * 53 * 25% for SMB/multiprotocol
- Memory in KB * 128 * 25% for NFS

Since the maxdirsize is set to 320MB by default on larger systems, the maximum number of files would be:

- 4,341,760 for SMB/multiprotocol
- 10,485,760 for NFS

What EMS messages are sent when maxdirsize is exceeded?

The following EMS messages are triggered when maxdirsize is either exceeded or close to being exceeded. Warnings are sent at 90% of the maxdirsize value and can be viewed with the event log show command or via the OnCommand System Manager event section (example below). OnCommand Unified Manager can be used to monitor maxdirsize, trigger alarms, and notify before the 90% threshold.

<table>
<thead>
<tr>
<th>Message Name: wafl.dir.size.max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severity: ERROR</td>
</tr>
<tr>
<td>Corrective Action: Use the &quot;volume file show_inode&quot; command with the file ID and volume name information to find the file path. Reduce the number of files in the directory. If not possible, use the (privilege:advanced) option &quot;volume modify -volume vol_name -maxdir-size new_value&quot; to increase the maximum number of files per directory. However, doing so could impact system performance. If you need to increase the maximum directory size, work with technical support.</td>
</tr>
<tr>
<td>Description: This message occurs after a directory has reached its maximum directory size (maxdirsize) limit.</td>
</tr>
<tr>
<td>Supports SNMP trap: true</td>
</tr>
<tr>
<td>Destinations: -</td>
</tr>
<tr>
<td>Number of Drops Between Transmissions: 0</td>
</tr>
<tr>
<td>Dropping Interval (Seconds) Between Transmissions: 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Message Name: wafl.dir.size.max.warning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severity: ERROR</td>
</tr>
<tr>
<td>Corrective Action: Use the &quot;volume file show_inode&quot; command with the file ID and volume name information to find the file path. Reduce the number of files in the directory. If not possible, use the (privilege:advanced) option &quot;volume modify -volume vol_name -maxdir-size new_value&quot; to increase the maximum number of files per directory. However, doing so could impact system performance. If you need to increase the maximum directory size, work with technical support.</td>
</tr>
<tr>
<td>Description: This message occurs when a directory has reached or surpassed 90% of its current maximum directory size (maxdirsize) limit, and the current maxdirsize is less than the default maxdirsize, which is 1% of total system memory.</td>
</tr>
<tr>
<td>Supports SNMP trap: true</td>
</tr>
<tr>
<td>Destinations: -</td>
</tr>
<tr>
<td>Number of Drops Between Transmissions: 0</td>
</tr>
<tr>
<td>Dropping Interval (Seconds) Between Transmissions: 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Message Name: wafl.dir.size.warning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severity: ERROR</td>
</tr>
</tbody>
</table>
| Corrective Action: Use the "volume file show_inode" command with the file ID and volume name information to find the file path. Reduce the number of files in the directory. If not possible, use the (privilege:advanced) option "volume modify -volume vol_name -maxdir-size new_value" to increase the maximum number of files per directory. However, doing so
could impact system performance. If you need to increase the maximum directory size, work with technical support.

Description: This message occurs when a directory surpasses 90% of its current maximum directory size (maxdirsize) limit.

Supports SNMP trap: true

Destinations: -

Dropping Interval (Seconds) Between Transmissions: 0

| Figure 23 | OnCommand System Manager Event screen with maxdirsize warning. |

Impact of Increasing maxdirsize

When a single directory contains a large number of files, the lookups (such as in a “find” operation) can consume large amounts of CPU and memory. In ONTAP 9.2, the introduction of “directory indexing” creates an index file for directory sizes exceeding 2MB to help offset the need to perform so many lookups and avoid cache misses. In most cases, this helps large directory performance. However, in the case of wildcard searches and readdir operations, indexing is not of much use.

Do FlexGroup volumes avoid maxdirsize limitations?

In NetApp FlexGroup volumes, each member volume has the same maxdirsize setting. Even though a directory could potentially span multiple FlexVol member volumes and nodes, the maxdirsize performance impact can still come into play, because directory size is the key component, not individual FlexVol volumes. Therefore, the size of a directory will still be an issue. Thus, NetApp FlexGroup volumes do not provide relief for environments facing maxdirsize limitations. Although newer platforms offer more memory and CPU, and the AFF systems provide performance benefits, the best way to reduce performance impact in directories with large numbers of files is to spread files across more directories in a file system.

Impact of Exceeding maxdirsize

When maxdirsize is exceeded in ONTAP, an “out of space” error (ENOCHD) is issued to the client and an EMS message is triggered. To remediate this, the storage administrator must increase the maxdirsize setting or move files out of the directory. For more information on remediation, see KB 000002080 on the NetApp Support site.
7 Features of NetApp FlexGroup

NetApp FlexGroup features are grouped into the following categories:

- Simplicity
- Data Protection
- Storage Efficiencies

In addition, the functionality and advantages of NetApp FlexGroup are in and of themselves a feature and can be reviewed in section 3, Advantages of NetApp ONTAP FlexGroup.

7.1 Simplicity

One of the key value-adds of a NetApp FlexGroup volume is the capability to create a massive container for capacity that delivers superior performance with the same ease as a normal FlexVol volume. NetApp FlexGroup offers support with OnCommand System Manager and Performance Manager and automated commands in the CLI, such as `volume create -auto-provision-as flexgroup` and `flexgroup deploy` to enable quick and easy deployment of a FlexGroup volume.

Command Line (CLI)

Although most people think of GUIs when they think of simplicity, the command line is also a place where tasks can be made easier. The NetApp FlexGroup CLI has some improvements to better the overall usability experience.

FlexGroup Deploy

If you are not a fan of using graphical interfaces to manage your storage and have found the ONTAP CLI to be a bit unwieldy, then the new `flexgroup deploy` command is for you.

This command automates the steps needed to deploy a FlexGroup volume, including:

- Member volume count and placement
- Space guarantee settings

Some things to consider with the `flexgroup deploy` command:

- Aggregates must be provisioned before using the command.
- The `-vserver` parameter is also mandatory, so an SVM must be in place in addition to the aggregates.
- The Snapshot policy will be set to default. To disable Snapshot copies at volume creation, use `volume create -snapshot-policy set to none` or use `volume modify` after the creation.
- The security style (UNIX, NTFS, mixed) will be set to the same security style as the vsroot volume. To control the FlexGroup security style at creation, use the `-security-style` option with `volume create` or use `volume modify` after the creation.
- The `flexgroup deploy` command does not set the `advanced privilege` NFS server option `-v3-64bit-identifiers` to enabled. NetApp highly recommends this option for the FlexGroup volume to avoid file ID collisions.
With a single command, you can have petabytes of storage created and deployed within minutes.

```
cluster::> flexgroup deploy ?
(volume flexgroup deploy)
   [-size] {<integer>[KB|MB|GB|TB|PB]}  Size of the FlexGroup
   [ -vserver] <vserver name>           Vserver Name
   [ -volume <volume name>]             Name of the FlexGroup to Create
   [ -type {RW|DP}]                     Volume Type (default: RW)
   [ -space-guarantee {none|volume}]    Space Guarantee Style (default: volume)
   [ -foreground {true|false}]         Foreground Process (default: true)
```

**Note:** Currently, the `flexgroup deploy` command is supported only on clusters with four or fewer nodes.

**Volume Create (with -auto-provision-as flexgroup)**

In ONTAP 9.2, a new volume option was introduced for provisioning FlexGroup volumes at `admin privilege`. When specified, this option defaults to 8 member FlexVol volumes per node. If no size is specified, the command creates member FlexVol volumes of 200MB each, so it’s important to specify a size with the command. Keep in mind that the formula is `(8 * aggregates specified/total specified size)`. This is important because a member volume must be at least 100GB and can be no larger than 100TB. Review TR-4571 for the best practices for sizing your FlexGroup volume capacity.

If no aggregates are specified, ONTAP attempts to select all aggregates available to the specified SVM. So although it’s possible to run a simplified command, it’s best to be as prescriptive as possible for the FlexGroup configuration.

**At a minimum, specify:**
- Autoprovision as FlexGroup (`-auto-provision-as`)
- Volume name (`-volume`)
- SVM (`-vserver`)
- Volume size (`-size`)
- Export/mount point (`-junction-path`)

**Optionally specify:**
- Volume security style (`-security-style {unix|ntfs|mixed}`)
- UNIX permissions (`-unix-permissions`, if security style is UNIX)
- Thin provisioning (`-space-guarantee none`)

**Volume Create (Advanced)**

If customization outside of best practices is needed (such as when fewer/more member volumes are needed), `volume create -auto-provision-as flexgroup` and `flexgroup deploy` might not be the right commands to use to create a FlexGroup volume. If a cluster has more than 4 four nodes, or if more granular control over the design and placement of the FlexGroup constituent members is desired, then the alternative is the command `volume create` without the `-auto-provision-as` option specified. Several new options were added that are specific to FlexGroup creation.
Table 4) New volume command options for use with FlexGroup.

<table>
<thead>
<tr>
<th>Volume Option</th>
<th>What It Does</th>
</tr>
</thead>
<tbody>
<tr>
<td>-aggr-list</td>
<td>Specifies an array of names of aggregates to be used for FlexGroup constituents. Each entry in the list creates a constituent on the specified aggregate. An aggregate can be specified multiple times to have multiple constituents created on it. This parameter applies only to FlexGroup.</td>
</tr>
<tr>
<td>-aggr-list-multiplier</td>
<td>Specifies the number of times to iterate over the aggregates listed with the -aggr-list parameter when creating a FlexGroup volume. The aggregate list will be repeated the specified number of times.</td>
</tr>
<tr>
<td>-max-constituent-size</td>
<td>Optionally specifies the maximum size of a FlexGroup constituent. The default value is determined by checking the maximum FlexVol size setting on all nodes used by the FlexGroup volume. The smallest value found is selected as the default for the max-constituent size for the FlexGroup. This parameter applies only to FlexGroup volumes.</td>
</tr>
</tbody>
</table>

Volume Modify

Once a FlexGroup is created, changes to the volume options or size must be carried out using the command-line command `volume modify` or the OnCommand System Manager GUI.

Volume Expand

Another new command added to ONTAP for management of a FlexGroup volume is `volume expand`. This command enables storage administrators to add members to an existing FlexGroup volume by using the `-aggr-list` and `-aggr-list-multiplier` options. Simply specify the aggregates to add members to and the number of desired members per aggregate. ONTAP does the rest.

Volume Expand on Volumes in SnapMirror Relationships

The `volume expand` command does not work natively with FlexGroups participating in SnapMirror relationships earlier than ONTAP 9.3, because those required a rebaseline of the SnapMirror relationship. ONTAP 9.3 introduced the enhancement to allow volume expansion on FlexGroups participating in a SnapMirror relationship without the need to rebaseline. As of ONTAP 9.3, ONTAP adjusts the FlexGroup member volume count on the next SnapMirror update.

**Note:** NetApp recommends upgrading to ONTAP 9.3 when using SnapMirror with FlexGroup volumes.

Expanding FlexGroup Volumes in SnapMirror Relationships Prior to ONTAP 9.3

To expand a volume (to add more members) in a SnapMirror relationship prior to ONTAP 9.3, perform the following steps:

1. `snapmirror delete` of the existing relationship on the destination
2. `snapmirror release` on the source
3. `volume delete` of the destination FlexGroup DP volume
4. `volume expand` of the source FlexGroup volume
5. `volume create` of a new destination FlexGroup DP volume with same size and constituent count as source FlexGroup volume
6. `snapmirror initialize` of the new relationship (re-baseline)

Growing the member volumes without needing to re-baseline the relationship is supported with SnapMirror and FlexGroup.
**OnCommand System Manager**

ONTAP 9.1 offered OnCommand System Manager support for NetApp FlexGroup right out of the gate. A new FlexGroup tab was added to the GUI under the Volumes page. On this page, storage administrators can manage an existing FlexGroup volume or create a new FlexGroup volume with two clicks. ONTAP 9.4 has raised the stakes with an even more robust GUI to support FlexGroup volumes, with the ability to perform virtually all of the same tasks as a FlexVol volume.

**Creating a FlexGroup Volume (ONTAP 9.1)**

To create a FlexGroup volume in System Manager in ONTAP 9.1, simply navigate to the storage virtual machine (SVM) being managed and click Volumes > FlexGroups. Then click the Create button. A splash page opens with four fields to choose from. The only required fields are Name and Size.

Figure 23) Creating a new FlexGroup volume in ONTAP 9.1.

**Creating a FlexGroup Volume (ONTAP 9.2)**

In ONTAP 9.2 and later, the FlexGroups tab was removed and replaced by a dropdown box on the Create button. Select either FlexVol or FlexGroup from this new dialog box in ONTAP 9.2 and later.

Figure 24) Creating a new FlexGroup volume in ONTAP 9.2 and later.
Creating a FlexGroup Volume (ONTAP 9.4)

In ONTAP 9.4, select Storage and Volumes to begin the FlexGroup creation process.

From the + Create menu icon, select Create FlexGroup.

Select the desired SVM to create the FlexGroup volume and then click Select.
Enter the FlexGroup volume name and desired capacity.

In this window, you can also configure a number of features and options, including:

- NetApp Volume Encryption
- Volume Protection

Aggregate selection is no longer available from this window, but it is available in the new Advanced Features page, which can be accessed by clicking the gear icon on the top right of this window.

This opens a page that allows configuration of:

- Space reservations
- Aggregate selection
- Security style
- UNIX permissions
- Atime updates
- Fractional reserve
- Volume autogrow
- Storage efficiencies and policies
- Quality of service (minimums, maximums and Adaptive QoS)

**Configuration Options for FlexGroup in OnCommand System Manager (Basic Options)**

**Name**: Name your FlexGroup volume.

**Protocols Enabled**: There is nothing to configure here. Protocols are fetched from the SVM’s enabled data protocols. If iSCSI or FCP shows up in this field, that does not mean that the FlexGroup volume can be used for LUNs; it is simply displaying what the SVM allows.

**Aggregates (moved to Advanced Features in ONTAP 9.4)**: Define the aggregates to be used with the FlexGroup volume.
Selecting Recommended per Best Practices creates eight member constituents per node. With All Flash FAS systems, the member constituents will reside on the same aggregate. In other configurations, four member constituents per aggregate per node are created. This option requires one aggregate per node for All Flash FAS and two aggregates per node for other configurations. If the requirements are not met, creating fails and the storage administrator must manually select aggregates.

In some cases, storage administrators might want to control the layout of the FlexGroup volume. The GUI gives the option of selecting the aggregates manually through the Select Aggregates button.

Figure 25) Manual selection of aggregates.

Space Reserve (moved to Advanced Features in ONTAP 9.4): This field allows a storage administrator to specify whether the FlexGroup volume is thin or thick provisioned. Thin provisioning disables the space guarantee of all member volumes and allows a FlexGroup volume to be overprovisioned in a cluster. Overprovisioning means being able to size a volume beyond the physical capacity of the cluster.

Size: This field specifies the total size of the FlexGroup volume. The size of member constituents depends on the number of nodes and aggregates in the cluster. Member constituents are automatically sized to equal amounts across the FlexGroup volume.

The available size allowed depends on the total number of aggregates available in the cluster. Remember that System Manager deploys four member volumes per aggregate. If only two aggregates are available in the cluster, then only eight members are created at a maximum of 100TB per member.

Table 5) FlexVol member to aggregate ratios when using System Manager.

<table>
<thead>
<tr>
<th>Aggregates per Cluster</th>
<th>Total Member Volumes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
</tr>
</tbody>
</table>

Additionally, in ONTAP 9.2, a new warning has been added if 64-bit file identifiers have not been enabled on the NFS server. If you left this disabled intentionally, you can ignore the warning. This option is found in the general SVM configuration section of OnCommand System Manager.

Figure 26) 64-bit file identifier warning in OnCommand System Manager.
**Volume Protection:** This section allows storage administrators to automatically set up data protection for FlexGroup volumes with just a few clicks. If a cluster/SVM is already peered to the source cluster/SVM, a SnapMirror or SnapVault relationship can be set up in the same window as the FlexGroup volume creation.

Figure 26) Protecting your FlexGroup volume.

Configuration Options for FlexGroup in OnCommand System Manager (Advanced Options)

When creating a FlexGroup volume, click the small gear icon at the top right of the page to view advanced options for configuring FlexGroup volumes. The following section covers the available advanced options for FlexGroup volumes in OnCommand System Manager.

Figure 27) Advanced FlexGroup Volume Option Icon.

The advanced options are divided into multiple sections.

**General Details – Advanced Options**

**Space Reserve (found in basic options earlier than ONTAP 9.4):** In this field, storage administrator can specify whether the FlexGroup volume is thin or thick provisioned. Thin provisioning disables the space guarantee of all member volumes and allows a FlexGroup volume to be overprovisioned in a cluster. (Overprovisioning means being able to size a volume beyond the physical capacity of the cluster.)

**Aggregates (found in basic options earlier than ONTAP 9.4):** Define the aggregates to be used with the FlexGroup volume. Selecting Recommended per Best Practices creates eight member constituents per node. With AFF systems, the member constituents reside on the same aggregate. In other configurations, four member constituents per aggregate per node are created. This option requires one aggregate per node for AFF and two aggregates per node for other configurations. If the requirements are not met, creating fails and the storage administrator must manually select aggregates.
**Volume security style**: Security style simply means “what style of permissions will be applied to the FlexGroup volume.” The available options are UNIX, NTFS, and mixed. UNIX security leverages UNIX mode bits (rwx). NTFS security style leverages Windows NT ACLs. Mixed security style toggles between UNIX and NTFS, depending on what type of client last set ACLs. For more information on volume security styles in FlexGroup volumes, see TR-4571.

**Note**: Mixed security style is not generally recommended. See TR-4571 for details.

**Note**: NFSv4.x ACLs are not currently supported with FlexGroup volumes.

**UNIX permissions**: Allows storage administrators to set rwx permissions on UNIX or mixed security style volumes at the time of creation.

**Update access time when a file is read**: This checkbox determines whether a file updates its access time (atime) when the file is read.

**Optimize Space – Advanced Options**

**Enable Fractional Reserve**: This option doesn’t really pertain to NAS volumes; it’s used to reserve a portion of a volume’s space when LUNs are present to help protect against space overruns. Since FlexGroup volumes do not currently support SAN, this option can be safely ignored.

**Volume Autogrow**: Starting in ONTAP 9.3, FlexGroup volumes added support for volume autogrow, which allows volumes to grow automatically once they approach a threshold of free space available. Although Grow or Shrink is an available option for FlexGroup volumes, only autogrow is currently supported.

**Storage Efficiency – Advanced Options**

For more information about storage efficiency options, see TR-4476.

**Background Deduplication**: Starting in ONTAP 9.4, FlexGroup volumes support automatic background deduplication schedules, which allow policies to be set to run deduplication tasks. The goal is to allow ONTAP to decide the best time to run deduplication jobs, based on data ingest rates. Alternatively, manual deduplication schedules can be used.

Background compression and inline deduplication/compression are also available storage efficiency options to enable/disable. For inline compaction and other storage efficiencies, use the CLI.

**Quality of Service (QoS) – Advanced Options**

For more information about storage QoS options, see TR-4211.

**QoS Policy**: This option allows you to enable or disable QoS for the FlexGroup volume, as well as to set or create new QoS Policy Groups.

**Managing a FlexGroup Volume**

Once the FlexGroup volume finishes creating, a dialog box opens with the option to create a CIFS share or to click Done to finish the process.
Figure 27) Create shares to a FlexGroup volume in OnCommand System Manager.

From there, the admin can manage the FlexGroup volume from the FlexGroup tab. OnCommand System Manager provides an overview of the FlexGroup volume, including:

- Volume overview
- Space allocation
- Data protection status (SnapMirror)
- Current performance statistics

Figure 28) FlexGroup overview in System Manager.
For more detailed information, you can click the hyperlinked volume name or Show More Details. This view gives information such as:

- Volume overview
- Snapshot copies
- Data protection details
- Storage efficiency details
- Performance details (real-time only)

In addition, the FlexGroup volume can be managed from the System Manager GUI via the Edit and More Actions buttons.

Figure 28) Managing an existing FlexGroup volume.

OnCommand Performance Manager (OPM)

In addition to OnCommand System Manager support, you can use OnCommand Performance Manager (OPM) to monitor a FlexGroup volume and its members at a granular level.

In OPM, a FlexGroup volume can be found with the other volumes in an SVM. When you click on the desired object, you see a screen with a member volume summary. You can also add these members to a graphical view over a specified range of time.

Figure 29) FlexGroup member volumes in OPM.
As member volumes are added to the OPM graphical view, they are assigned different graph line colors to differentiate them in the charts. In this way, any member that deviates from the expected performance output can be investigated and remediated.

**Figure 30)** Graphical representation of FlexGroup member volumes in OPM.

[Graphical representation of FlexGroup member volumes in OPM]

**Figure 31)** Graphical representation of FlexGroup member volumes in OPM—zoomed.

[Graphical representation of FlexGroup member volumes in OPM—zoomed]
Single Transparent Namespace

NetApp FlexGroup offers the advantage of massive capacity that exceeds that of a normal FlexVol volume without needing to implement a complicated architecture. The entire volume can be mounted as a single export or share and does not require additional changes on the application side, even when more storage is added. This benefit reduces the management complexity associated with managing numerous containers and numerous mount points or shares from the client side.

Qtrees

ONTAP 9.3 introduces support in FlexGroup volumes for logical directories called qtrees. Qtrees allow a storage administrator to create folders from the ONTAP GUI or CLI to provide logical separation of data within a large bucket. Qtrees are useful for home directory workloads, because folders can be named to reflect the usernames of users accessing data, and dynamic shares can be created to provide access based on a username. Qtrees are distributed across a FlexGroup volume in much the same way as a normal folder. Quota monitoring can be applied at the qtree level, and in ONTAP 9.5 and later, quota enforcement policies can be applied. Qtrees are created and managed the same way as a FlexVol qtree is managed. A maximum of 4,995 qtrees is supported per FlexGroup volume.

Quota Enforcement Example

When quota enforcement is enabled on a qtree or for a user, ONTAP disallows new file creations or writes after a quota is exceeded. In addition, an EMS message is logged at DEBUG severity level to notify storage administrators of the quota violation. You can configure these EMS messages so that the system forwards them as SNMP traps or as syslog messages.

In this example, a quota has been set with a hard limit of 1GB and 10 files.

```
classic::*> quota policy rule show -vserver DEMO
Vserver: DEMO    Policy: tree    Volume: flexgroup_local

<table>
<thead>
<tr>
<th>Type</th>
<th>Target</th>
<th>Qtree</th>
<th>User Mapping</th>
<th>Disk Limit</th>
<th>Soft Disk Limit</th>
<th>Soft Files Limit</th>
<th>Soft Files Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>tree</td>
<td>qtree</td>
<td>&quot;&quot;</td>
<td>-</td>
<td>1GB</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
```

When a user tries to copy a 1.2GB file to the qtree, ONTAP reports an “out of space” error:

```
[root@centos7 qtree]# cp /SANscreenServer-x64-7.3.1-444.msi /FGlocal/qtree/
cp: failed to close "/FGlocal/qtree/SANscreenServer-x64-7.3.1-444.msi”: No space left on device
```

The file is partially written, but it is unusable because it’s missing data:

```
# ls -alh
total 1.1G
drwxr-xr-x  2 root root  4.0K Jul 19 15:44 .
drwxr-xr-x 11 root root  4.0K Jun 28 15:10 ..
-rw-r--r--  1 root root  14M Dec 12 2017 First Draft TTDD Slide Deck on ONTAP 9.3 - Parisi.pptx
-rw-r--r--  1 root root  0 Dec 12 2017 newfile1
-rw-r--r--  1 root root  0 Dec 12 2017 newfile2
-rw-r--r--  1 root root 1021M Jul 19 2018 SANscreenServer-x64-7.3.1-444.msi
```

ONTAP then reports the quota as exceeded:

```
classic::*> quota report -vserver DEMO
Vserver: DEMO

<table>
<thead>
<tr>
<th>Volume</th>
<th>Tree</th>
<th>Type</th>
<th>ID</th>
<th>Used Limit</th>
<th>Used Limit</th>
<th>Specifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>flexgroup_local</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
The same behavior occurs for file count limits. In this example, the file count limit is 10 and the qtree already has 5 files. An additional 5 files meets our limit.

```
[root@centos7 /]# su student1
sh-4.2$ cd ~
sh-4.2$ pwd
/home/student1
sh-4.2$ touch file1
sh-4.2$ touch file2
sh-4.2$ touch file3
sh-4.2$ touch file4
sh-4.2$ touch file5
```
touch: cannot touch 'file5': Disk quota exceeded

```
cluster::*> quota report -vserver DEMO
Vserver: DEMO

<table>
<thead>
<tr>
<th>Volume</th>
<th>Tree</th>
<th>Type</th>
<th>ID</th>
<th>Used</th>
<th>Limit</th>
<th>Used</th>
<th>Limit</th>
<th>Specifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>flexgroup_local qtree</td>
<td>tree</td>
<td>1</td>
<td>1.01GB</td>
<td>1GB</td>
<td>5</td>
<td>10</td>
<td>qtree</td>
<td></td>
</tr>
<tr>
<td>home</td>
<td>user</td>
<td>student1, NTAP@student1</td>
<td>4KB</td>
<td>1GB</td>
<td>10</td>
<td>10</td>
<td>student1</td>
<td></td>
</tr>
</tbody>
</table>
```

2 entries were displayed.

From the event logs, we can see the quota violations:

```
cluster::*> event log show -message-name quota.exceeded
Time     Node             Severity      Event
---------- ---------------------- --------------- -----------------  
7/19/2018 16:27:54 node02 DEBUG     quota.exceeded: ltype="hard", volname="home", app="", volident="@vserver:e3cc08e-d9b3-11e6-85e2-00a0986b1210", limit_item="file", limit_value="10", user="uid=1301", qtree="treeid=1", vfiler=""
7/19/2018 15:45:02 node01 DEBUG     quota.exceeded: ltype="hard", volname="flexgroup_local", app="", volident="@vserver:e3cc08e-d9b3-11e6-85e2-00a0986b1210", limit_item="disk", limit_value="104456", user="", qtree="treeid=1", vfiler=""
```

### 7.2 Integrated Data Protection

NetApp FlexGroup supports several methods of data protection, including NetApp RAID DP® software, RAID Triple Erasure Coding (NetApp RAID-TEC™ technology), Snapshot technology, SnapMirror replication technology, and NFS or CIFS mounted tape backup.

#### RAID DP and RAID Triple Erasure Coding (RAID-TEC)

RAID DP is known as “dual parity” RAID and can survive two simultaneous disk failures per RAID group. This means that in the event of a drive failure, data is still protected with another parity drive.

RAID Triple Erasure Coding (RAID-TEC) was new in ONTAP 9.0 and provides an extra parity drive for RAID groups using larger-capacity drives. This feature is intended to help protect against drive failures during longer rebuild times for larger-capacity drives. RAID-TEC also offers the ability to provide larger RAID groups in terms of drive numbers.

All RAID protection features are supported with NetApp FlexGroup.
NetApp Snapshot Technology

NetApp Snapshot copies are automatically scheduled point-in-time copies that take up no space and incur no performance overhead when created. Over time, Snapshot copies consume minimal storage space, because only changes to the active file system are written. Individual files and directories can be easily recovered from any Snapshot copy, and the entire volume can be restored back to any Snapshot state in seconds.

Snapshot copies are supported for use with NetApp FlexGroup. Each Snapshot copy is made as a consistency group of the FlexVol members in which all members are quiesced and prepared for a Snapshot copy to ensure a consistent point-in-time copy of all members in a FlexGroup volume.

If any member in a FlexGroup volume in unable to perform the Snapshot operation (out of space, offline, too busy to complete), then the entire FlexGroup Snapshot copy is considered partial and to have failed. ONTAP cleans up the remnants of the attempted Snapshot copy and issues an EMS event.

After a Snapshot copy is created, storage administrators can perform the following operations for restores:

- Use NetApp SnapRestore® technology to restore the entire FlexGroup volume.
- Navigate the Snapshot directories with the .snapshot (NFS) or ~snapshot (CIFS/SMB) folders to restore individual files and folders.

Note: Single File SnapRestore (SFSR) is currently not supported, nor is restoring a single FlexGroup member volume. Using SnapRestore for a NetApp FlexGroup volume restores the entire volume, not just single member volumes. Restores are only available at diag privilege currently.

Figure 32) FlexGroup Snapshot copy.
SnapMirror and SnapVault

SnapMirror provides asynchronous replication of volumes, independent of protocol for data protection and disaster recovery. SnapVault provides asynchronous snapshot retention for data protection and backup/archive use cases.

SnapMirror support with NetApp FlexGroup was provided starting in ONTAP 9.1. SnapVault support for FlexGroup volumes was added in ONTAP 9.3.

SnapMirror and SnapVault function in a similar manner as Snapshot copies. All member volumes need to have a successful Snapshot copy, and all member volumes are concurrently replicated to the DR site. If any component of that operation fails, mirroring with SnapMirror fails as well.

See the technical report TR-4678: FlexGroup Volume Data Protection Best Practices for details of best practices and limits for SnapMirror and SnapVault with FlexGroup volumes.

Tape Backup with CIFS/SMB or NFS

Tape backup for NetApp FlexGroup volumes can be performed by using external backup applications such as Commvault Simpana and Symantec NetBackup over CIFS or NFS mounts. NDMP is currently not supported.

7.3 Storage Efficiencies

NetApp FlexGroup also offers support for the following storage efficiency technologies:

- **Inline and postprocess deduplication** removes duplicate data blocks in primary and secondary storage, storing only unique blocks. This action results in storage space and cost savings. Deduplication runs on a customizable schedule.

- **Inline aggregate deduplication (or cross-volume deduplication)** provides inline storage efficiency at the aggregate level. This allows duplicate blocks to be reduced if they exist in multiple FlexVol volumes in the same aggregate. FlexGroup member volumes are an excellent use case for this new ONTAP 9.2 feature. ONTAP 9.3 introduces automated scheduling and scheduled background inline aggregate deduplication.

- **Inline adaptive compression** was introduced for primary workloads such as database and desktop virtualization with ONTAP 8.3.1. Inline compression is on by default in the All Flash FAS product family starting with 8.3.1.

- **Inline data compaction** was introduced in ONTAP 9.0 and further reduces the physical used space needed to store data. Data compaction is a significant addition to our storage efficiency portfolio and complements deduplication and compression technologies. Data compaction takes I/Os that normally consume a 4K block on physical storage and packs multiple such I/Os into one physical 4K block.

- **Thin provisioning** has been around for years and offers storage administrators the ability to overprovision virtual containers (FlexVol volumes) on physical storage (aggregates). With NetApp FlexGroup, thin provisioning can play an important role in how initial deployment of the FlexGroup volume is handled. Thin provisioning also allows member constituents to be much larger than their physical aggregate counterparts, which provides flexibility in the design of the container.

These features are applied by ONTAP at the member volume level individually, but configured by the storage administrator at the FlexGroup level for ease of management. In earlier releases of ONTAP, the features were applied at the FlexVol member level, so Table 5 gives guidance on what ONTAP versions support FlexGroup-level management of these features, as well as which ONTAP versions require more granular management of the efficiencies per member volume.

**Note:** NetApp highly recommends using ONTAP 9.3 or later for maximum storage efficiency with NetApp FlexGroup volumes.
Table 6) Storage efficiency guidance for FlexGroup in ONTAP versions.

<table>
<thead>
<tr>
<th></th>
<th>9.1RC1</th>
<th>9.1RC2 and later</th>
<th>9.2RC1 and later</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thin provisioning</td>
<td>FlexGroup level</td>
<td>FlexGroup level</td>
<td>FlexGroup level</td>
</tr>
<tr>
<td>Inline deduplication</td>
<td>FlexVol member</td>
<td>FlexGroup level</td>
<td>FlexGroup level</td>
</tr>
<tr>
<td>Post-process deduplication</td>
<td>FlexVol member</td>
<td>FlexGroup level</td>
<td>FlexGroup level</td>
</tr>
<tr>
<td>Inline data compaction</td>
<td>FlexVol member</td>
<td>FlexGroup level</td>
<td>FlexGroup level</td>
</tr>
<tr>
<td>Inline data compression</td>
<td>FlexVol member</td>
<td>FlexGroup level</td>
<td>FlexGroup level</td>
</tr>
<tr>
<td>Post-process data compression</td>
<td>FlexVol member</td>
<td>FlexGroup level</td>
<td>FlexGroup level</td>
</tr>
<tr>
<td>Aggregate inline deduplication</td>
<td>N/A</td>
<td>N/A</td>
<td>FlexGroup level</td>
</tr>
</tbody>
</table>

Note: For storage efficiency best practices and caveats, see TR-4476.

Applying Storage Efficiencies per FlexGroup Member Volume

If a FlexGroup volume does not currently have support to enable storage efficiencies at the FlexGroup level, use the following command to enable it on every FlexVol member. **This should be necessary only in ONTAP 9.1RC1.**

```
class::*> volume efficiency on -vserver SVM -volume flexgroup4*
```

```
cluster::*> volume efficiency modify -vserver SVM -volume flexgroup4* -compression true -data-compaction true -inline-compression true -inline-dedupe true
```

```
class::*> volume efficiency show -vserver SVM -volume flexgroup4* -fields data-compaction,compression,inline-compression,inline-dedupe
```

```
vserver volume   compression inline-compression inline-dedupe data-compaction
-----------------  ---------------  ------------------  ------------------  ------------------
SVM   flexgroup4TB__0001 true        true               true          true
SVM   flexgroup4TB__0002 true        true               true          true
SVM   flexgroup4TB__0003 true        true               true          true
SVM   flexgroup4TB__0004 true        true               true          true
SVM   flexgroup4TB__0005 true        true               true          true
SVM   flexgroup4TB__0006 true        true               true          true
SVM   flexgroup4TB__0007 true        true               true          true
SVM   flexgroup4TB__0008 true        true               true          true
```

Note: NetApp FlexClone® volumes and autoshrink in FlexGroup are not currently supported.

For more information about storage efficiencies in ONTAP, see TR-4476: NetApp Data Compression, Deduplication, and Data Compaction.

For more information about thin provisioning in ONTAP, see TR-3965: NetApp Thin Provisioning Deployment and Implementation Guide.

FabricPool

In ONTAP 9.2, the ability to automatically tier cold data blocks on SSD aggregates to the cloud or on-premises S3 object storage was added for FlexVol volumes. This functionality allowed storage administrators to preserve more costly SSD drives for active workloads, whereas cold or unused data was
moved to more cost-effective capacity tiers. This feature is known as FabricPool. You can learn more about the feature in TR-4598: FabricPool Best Practices.

ONTAP 9.5 introduces support for FabricPool for FlexGroup volumes. There are no special considerations to make for FlexGroup volumes; the same FlexVol considerations apply.

7.4 Maximum Storage Quality of Service (QoS) Policies

Starting in ONTAP 9.3, you can apply maximum storage QoS policies to a FlexGroup volume to help prevent a FlexGroup from acting as a bully workload in ONTAP. ONTAP 9.4 introduces support for QoS minimums and Adaptive QoS. Storage QoS can help you manage risks around meeting your performance objectives. You use storage QoS to limit the throughput to workloads and to monitor workload performance. You can reactively limit workloads to address performance problems, and you can proactively limit workloads to prevent performance problems. For more information on storage QoS, see TR-4211: Storage Performance Primer.

How Storage QoS Maximums Work with FlexGroup volumes

With a FlexGroup volume, storage QoS policies are applied to the entire FlexGroup. Because a FlexGroup volume contains multiple FlexVol member volumes and can span multiple nodes, the QoS policy gets shared evenly across nodes as clients connect to the storage system. Figure 33 shows how storage QoS gets applied to a FlexGroup volume spanning multiple nodes in a cluster.

Figure 33) Storage QoS on NetApp FlexGroup volumes – single-node connection.

Figure 34) Storage QoS on NetApp FlexGroup volumes – multi-node connection.
8 Appendix

The following sections cover FlexGroup information not covered in the previous sections of this document, including:

- Command-line examples of creating and managing FlexGroup
- Gathering FlexGroup statistics
- Viewing FlexGroup ingest usage distribution through the CLI
- Sample Python script for generating many files from a client

Command-Line Examples

This section shows command-line examples for various basic FlexGroup related operations.

Creating a FlexGroup Volume by Using `flexgroup deploy`

```
cluster::> flexgroup deploy -size 20PB -space-guarantee volume -vserver SVM -volume flexgroup
```

Using the ONTAP 9.2 New auto-provision-as Option

```
cluster::> vol create -auto-provision-as flexgroup -vserver SVM -volume flexgroup92 -junction-path /flexgroup92 -size 100t -space-guarantee none -security-style unix
```

Creating a FlexGroup Volume Across Multiple Nodes by Using `volume create`

```
cluster ::> volume create -vserver SVM -volume flexgroup -aggr-list aggr1_node1,aggr1_node2 -policy default -security-style unix -size 20PB -space-guarantee none -junction-path /flexgroup
```

Modifying the FlexGroup Snapshot Policy

```
cluster::> volume modify -vserver SVM -volume flexgroup -snapshot-policy [policyname|none]
```

Resizing the FlexGroup Volume

```
cluster::> volume size -vserver SVM -volume flexgroup -new-size 20PB
```

Adding Members to the FlexGroup Volume

```
cluster::> volume expand -vserver SVM -volume flexgroup -aggr-list aggr1_node1,aggr1_node2 -aggr-list-multiplier 2
```

Applying Storage QoS

```
cluster::> volume modify -vserver DEMO -volume flexgroup -qos-policy-group FlexGroupQoS
```

Applying Volume Autogrow

```
cluster::> volume autosize -vserver DEMO -volume Tech_ONTAP -mode grow -maximum-size 20t -grow-threshold-percent 80
```

```
cluster::> volume autosize -vserver DEMO -volume Tech_ONTAP
Volume autosize is currently ON for volume "DEMO:Tech_ONTAP".
The volume is set to grow to a maximum of 20t when the volume-used space is above 80%.
Volume autosize for volume 'DEMO:Tech_ONTAP' is currently in mode grow.
```
## FlexGroup Statistics

In ONTAP 9, a new statistic object called `flexgroup` was added. The object is available only at `diag` privilege. This object gathers the following counters:

<table>
<thead>
<tr>
<th>Counter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat1_tld_local</td>
<td>cat1_tld_remote</td>
</tr>
<tr>
<td>cat2_tld_local</td>
<td>cat2_tld_remote</td>
</tr>
<tr>
<td>cat3_dir_local</td>
<td>cat3_dir_remote</td>
</tr>
<tr>
<td>cat4_fil_local</td>
<td>cat4_fil_remote</td>
</tr>
<tr>
<td>dsidlist_factory_enomem</td>
<td>groupstate_analyze</td>
</tr>
<tr>
<td>groupstate_create</td>
<td>groupstate_delete</td>
</tr>
<tr>
<td>groupstate_enomem</td>
<td>groupstate_insert</td>
</tr>
<tr>
<td>groupstate_preupdate_fail</td>
<td>groupstate_update</td>
</tr>
<tr>
<td>indextable_factory_enomem</td>
<td>indextableload_factory_enomem</td>
</tr>
<tr>
<td>indextablesave_factory_enomem</td>
<td>instance_name</td>
</tr>
<tr>
<td>instance_uuid</td>
<td>memberstate_create</td>
</tr>
<tr>
<td>memberstate_delete</td>
<td>memberstate_enomem</td>
</tr>
<tr>
<td>memberstate_expired</td>
<td>memberstate_factory_enomem</td>
</tr>
<tr>
<td>memberstate_unhealthy</td>
<td>monitor_receive</td>
</tr>
<tr>
<td>monitor_respond</td>
<td>node_name</td>
</tr>
<tr>
<td>node_uuid</td>
<td>process_name</td>
</tr>
<tr>
<td>refresh_enomem</td>
<td>refreshclient_create</td>
</tr>
<tr>
<td>refreshclient_delete</td>
<td>refreshserver_create</td>
</tr>
<tr>
<td>refreshserver_delete</td>
<td>remote_dirs</td>
</tr>
<tr>
<td>remote_files</td>
<td>snapclient_create</td>
</tr>
<tr>
<td>snapclient_delete</td>
<td>snapcoord_create</td>
</tr>
<tr>
<td>snapcoord_delete</td>
<td>snapserver_create</td>
</tr>
<tr>
<td>snapserver_delete</td>
<td>snapserver_fail_fence_down</td>
</tr>
<tr>
<td>snapserver_fail_fence_raise</td>
<td>snapserver_fail_snapid</td>
</tr>
<tr>
<td>snapshot_create</td>
<td>snapshot_enomem</td>
</tr>
<tr>
<td>snapshot_restore</td>
<td>tally_enomem</td>
</tr>
<tr>
<td>vldb_enomem</td>
<td>vldb_enorecord</td>
</tr>
<tr>
<td>vldbclient_create</td>
<td>vldbcclient_delete</td>
</tr>
<tr>
<td>vldbcclient_factory_enomem</td>
<td>vldbcclient_delete</td>
</tr>
</tbody>
</table>

The counters are specific to the FlexGroup volume, measuring remote allocation percentages, number of local versus remote files and directories, refresh counters, and various other objects.

FlexGroup statistics can be captured in the same way that other statistics are captured. You must start a statistics collection with `statistics start`, which creates a sample_id file. Once this is done, the statistics can be viewed by using `statistics show`.

If you want to specify multiple objects or counters, use a pipe symbol ( | ).

**Example of statistics start for FlexGroup and NFSv3 statistics:**

```
cluster::> set diag
cluster::*> statistics start -object nfsv3|flexgroup
Statistics collection is being started for sample-id: sample_2144
```
Example of statistics show for FlexGroup counters:

```
cluster::*> statistics show -object flexgroup

Object: flexgroup
Instance: 0
Start-time: 8/9/2016 13:00:22
End-time: 8/9/2016 15:22:29
Elapsed-time: 8527s
Scope: node1
Counter                     Value
----------------------------------------------
cat4_fil_local                3623435
cat4_fil_remote               600298

For more information on statistics command, use the man statistics start command in the CLI.

Qtree Statistics

Starting in ONTAP 9.5, qtree statistics were made available for FlexGroup volumes. These statistics provide granular performance information about FlexGroup volumes and their qtrees. The following example shows a statistics capture for a FlexGroup volume running a large NFS workload.

```
cluster::> statistics qtree show -interval 5 -iterations 1 -max 25 -vserver DEMO -volume flexgroup_local

cluster : 11/7/2018 15:19:15

<table>
<thead>
<tr>
<th>Qtree</th>
<th>Vserver</th>
<th>Volume</th>
<th>NFS</th>
<th>CIFS</th>
<th>Internal</th>
<th>*Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEMO:flexgroup_local</td>
<td>DEMO flexgroup_local</td>
<td>22396</td>
<td>0</td>
<td>0</td>
<td>22396</td>
<td></td>
</tr>
<tr>
<td>DEMO:flexgroup_local/qtree</td>
<td>DEMO flexgroup_local</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
```

Viewing FlexGroup Ingest Distribution

Using the command line, it is possible to get a real-time view of FlexGroup data ingest during workloads to see how evenly allocated the member volumes are with the diag privilege node-level command flexgroup show. In addition, the command provides visibility into the urgency and tolerance percentages as well as calculated probabilities for remote versus local placement of files and folders. This command can be run in the cluster shell CLI across multiple nodes.

```
cluster::> set diag
cluster::*> node run * flexgroup show
```

The following graphic shows an “ideal” flexgroup show output in which traffic is evenly distributed.
Sample Python Script to Generate Files on a FlexGroup Volume

When testing a FlexGroup volume, it is possible to use normal load-generating utilities. In our lab testing, one of the benchmarks used was a basic GIT benchmark using a Linux source code compile. Although that type of test is available to everyone, it might be more involved and complicated than most storage administrators want to undertake.

Conversely, it is not ideal to use common file creation utilities such as dd or writing bash scripts to create files and folders. That is because these are single-threaded tests and do not fully use the benefit of the client’s or storage’s CPU and throughputs capabilities.

One simple way to create a lot of files and generate sufficient load on a FlexGroup volume to see its benefits is to use a Python script written by Chad Morgenstern, a senior performance systems engineer at NetApp.

The script leverages multiprocessor calls to create 1,000 directories, each with 1,000 subdirectories. Below those, the script writes 5 small text files, for a total of 5 million files. This script can be modified to change the number of files and folders being created.

This script is available on GitHub, but it is not officially supported by NetApp Support. This script is not intended to measure load generation or to max out a FlexGroup volume’s performance.
9 References and Additional Information

This section provides links to content that is directly or tangentially related to FlexGroup volumes.

Technical Reports

- NetApp Thin Provisioning Deployment and Implementation Guide
- TR-3982: NetApp Clustered Data ONTAP 8.3.x and 8.2.x
- TR-4037: Introduction to NetApp Infinite Volume
- NFS Best Practice and Implementation Guide
- TR-4379: Name Services Best Practices Guide (pre-ONTAP 9.3)
- TR-4668: Name Services Best Practices Guide (ONTAP 9.3 and later)
- NetApp Data Compression, Deduplication, and Data Compaction
- NetApp Storage Solutions for Apache Spark
- NetApp FlexGroup Best Practices and Implementation Guide
- NetApp FlexGroup Top Best Practices
- Electronic Design Automation Best Practices
- FabricPool Best Practices

Miscellaneous

- Tech ONTAP Podcast—Episode 46: FlexGroups
  https://soundcloud.com/techontap_podcast/episode-46-flexgroups-1
- What’s New For FlexGroup Volumes in ONTAP 9.3?
- FlexGroup Volumes: An Evolution of NAS
- 7 Myths about NetApp FlexGroup Volumes
  https://blog.netapp.com/blogs/seven-myths-about-netapp-ontap-flexgroup-volumes/
- Volume Affinities: How ONTAP and CPU Utilization Has Evolved
  https://blog.netapp.com/volume-affinities-how-ontap-and-cpu-utilization-has-evolved/
- FlexGroup lightboard video
  https://www.youtube.com/watch?v=Wp6jEd4VkgI&t=4s
Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

Copyright © 2018 NetApp, Inc. All rights reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

RESTRICTED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (c)(1)(ii) of the Rights in Technical Data and Computer Software clause at DFARS 252.277-7103 (October 1988) and FAR 52-227-19 (June 1987).

Trademark Information

NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners. TR-4557-1218