Technical Report

NetApp ONTAP FlexGroup volumes

Best practices and implementation guide

Justin Parisi and Richard Jernigan, NetApp

October 2021 | TR-4571

Abstract

This document provides a brief overview of NetApp® ONTAP® FlexGroup and a set of best practices and implementation tips to use with this feature. The FlexGroup feature is an evolution of scale-out NAS containers that blends nearly infinite capacity with predictable, low-latency performance in metadata-heavy workloads. For information about FlexGroup volumes that is not covered in this document, email flexgroups-info@netapp.com, and we will add information to this technical report, as necessary.

TABLE OF CONTENTS

The evolution of NAS in NetApp ONTAP **10**
Flexible volumes: A tried-and-true solution 10
FlexGroup volumes: An evolution of NAS 10
Terminology **11**
What are large files? 12
NetApp ONTAP FlexGroup advantages **12**
Massive capacity and predictable low latency for high-metadata workloads 12
Efficient use of all cluster hardware 13
Simple, easy-to-manage architecture and balancing 13
Superior density for big data 13
Use cases **13**
Ideal use cases 13
Nonideal cases 13
FlexGroup volume use case examples 14
FlexGroup feature support and maximums **16**
Behavior of unsupported SMB features 19
Maximums and minimums 19
Deciding whether FlexGroup volumes are the right fit **20**
Scale-out performance 20
Feature compatibility limitations 21
Simplifying performance 22
Workloads and behaviors 41
Initial FlexGroup design considerations **47**
Cluster considerations 47
ONTAP version considerations 48
Failure domains 48
Aggregate layout considerations 48
Flash Cache and Flash Pool 51
Advanced Disk Partitioning 51
NetApp SyncMirror (mirrored aggregates) 51
NetApp MetroCluster 51
NetApp Cloud Volumes ONTAP 52
Capacity considerations 52

Potential issues when using 100TB member volumes 88
Networking considerations 89
Security and access control list style considerations 91

FlexGroup administration considerations 94
Initial deployment....................94
Viewing FlexGroup volumes 94
Viewing FlexGroup volume capacity....................99
Adding capacity to a FlexGroup volume 100
Nondisruptive volume move considerations 106
Considerations when deleting FlexGroup volumes....................107
Volume rename considerations 108

Qtrees.................... 108
Qtrees and file moves....................109
Managing quotas with FlexGroup 110

General NAS and high-file-count considerations.................... 119
High file count considerations....................119
64-bit file identifiers....................123
Effects of file system ID changes in ONTAP 128
Directory size considerations: maxdirsize....................128
File system analytics 132
Special character considerations....................134
File deletions/FlexGroup member volume balancing....................136
Listing files when a member volume is out of space....................138
File rename considerations 139
Symlink considerations....................139
NFS version considerations....................139
CIFS/SMB considerations 147
Virtualization workload considerations....................149
Databases on FlexGroup volumes 152
FlexCache volume considerations....................152
FlexClone 153

Encryption at-rest 154
Re-keying a FlexGroup volume or encrypting existing FlexGroup volumes 155
Drive-level encryption (NSE and SED) 155

FlexGroup sample designs **155**

General troubleshooting and remediation **163**

Failure scenarios 163

Capacity monitoring and alerting **165**

Monitoring FlexGroup performance 171

FlexGroup data protection best practices **175**

Migrating to NetApp ONTAP FlexGroup **175**

Migration using NDMP 176

FlexVol to FlexGroup volume conversion 177

Migrating from third-party storage to FlexGroup volumes 181

Migrating from NetApp Data ONTAP operating in 7-Mode 181

Migrating from SAN LUNs or Infinite Volume in ONTAP 182

NetApp XCP 182

Examples **188**

Thin provisioning example 188

Volume autosize example 189

Snapshot spill example 190

Capacity monitoring and alerting examples in Active IQ Unified Manager 192

Sample FlexVol to FlexGroup volume conversion 198

Converting FlexVol volumes in existing SnapMirror relationships: Example 202

Sample FlexVol to FlexGroup volume conversion — 500 million files 205

Event management system examples 208

Command examples **212**

FlexGroup capacity commands 212

Example of statistics show-periodic command for entire cluster 214

Real-time SVM-level statistics show-periodic for NFSv3 read and write operations 215

Real-time FlexGroup local and remote statistics 215

Example of creating a FlexGroup volume and specifying fewer member volumes than the default value 216

Sample REST API for creating a FlexGroup volume 216

Example of increasing a FlexGroup volume's size 218

Example of expanding a FlexGroup volume 219

Other command-line examples 220

Where to find additional information **221**

Technical reports 221

Miscellaneous content 221

Version history 222

Acknowledgments 222

LIST OF TABLES

Table 1) General ONTAP feature support. 16
Table 2) General NAS protocol version support. 18
Table 3) Unsupported SMB2.x and 3.x features. 18
Table 4) How unsupported SMB features behave with FlexGroup volumes. 19
Table 5) FlexGroup maximums. 20
Table 6) FlexGroup minimums. 20
Table 7) ONTAP volume family comparison. 21
Table 8) NetApp FlexGroup volumes versus competitor system — EDA benchmark. 36
Table 9) Best practices for aggregate layout with FlexGroup volumes. 49
Table 10) FlexGroup maximums. 53
Table 11) FlexGroup minimums. 53
Table 12) Theoretical maximums for FlexGroup based on allowed volume count in ONTAP. 54
Table 13) Situations in which you see out of space errors. 59
Table 14) Capacity Management Decision Matrix 64
Table 15) ONTAP storage efficiency support matrix: FlexVol and FlexGroup volumes. 71
Table 16) Storage efficiency comparisons: Deduplication. 78
Table 17) Autosize maximum size examples. 88
Table 18) Inode defaults and maximums according to FlexVol size. 119
Table 19) Inode defaults resulting from FlexGroup member sizes and member volume counts. 121
Table 20) High-file-count/small capacity footprint examples—increasing member volume counts. 122
Table 21) Async-delete performance. 123
Table 22) nconnect performance results. 141
Table 23) Other Potential SMB Issues. 149
Table 24) FlexGroup versus FlexVol VAAI results 151
Table 25) `flexgroup show` output column definitions. 174

LIST OF FIGURES

Figure 1) FlexVol design with junctioned architecture for >100TB capacity. 10
Figure 2) What is a large file? 12
Figure 3) SQL Server backup environment. 15
Figure 4) Throughput and total operations during test runs. 15
Figure 5) CPOC scale-out throughput results. 16

Figure 6) FlexGroup volume. 21
Figure 7) Overall response time, SPEC SFS 2014_swbuild submissions. 23
Figure 8) Throughput, SPEC SFS 2014_swbuild submissions. 23
Figure 9) IOPS, SPEC SFS 2014_swbuild submissions. 24
Figure 10) Latency versus number of builds, SPEC SFS 2014_swbuild submissions. 24
Figure 11) SPEC SFS 2014_swbuild—concurrent builds. 25
Figure 12) SPEC SFS 2014_swbuild—latency versus build ops/sec. 25
Figure 13) SPEC SFS 2014_swbuild—megabyte per second comparison. 26
Figure 14) FlexVol volume versus FlexGroup volume—maximum throughput trends under increasing workload. 27
Figure 15) FlexVol volume versus FlexGroup volume—maximum throughput trends under increasing workload, detailed. 27
Figure 16) FlexVol volume versus FlexGroup volume—maximum average total IOPS. 28
Figure 17) Standard NAS benchmark (EDA)—ONTAP 9.5: one node versus two nodes (operations/sec). 29
Figure 18) Standard NAS benchmark (EDA)—ONTAP 9.4 versus ONTAP 9.5 (operations/sec). 29
Figure 19) Standard NAS benchmark (EDA)—ONTAP 9.5: one node versus two nodes (MBps). 30
Figure 20) Standard NAS benchmark (EDA)—ONTAP 9.4 versus ONTAP 9.5 (MBps). 30
Figure 21) Standard NAS benchmark (software builds)—ONTAP 9.5 (operations/sec). 31
Figure 22) Standard NAS benchmark (software builds)—ONTAP 9.4 versus ONTAP 9.5 (operations/sec). 31
Figure 23) Standard NAS benchmark (software builds)—ONTAP 9.5 (MBps). 32
Figure 24) Standard NAS benchmark (software builds)—ONTAP 9.4 versus ONTAP 9.5 (MBps). 32
Figure 25) TeraSort benchmark statistics summary on a FlexGroup volume. 33
Figure 26) Kernel extract — Competitor vs. NetApp FlexVol volumes. 34
Figure 27) Kernel extract — Competitor vs. NetApp FlexVol volumes and NetApp FlexGroup volumes. 34
Figure 28) Kernel extract — Competitor vs. NetApp FlexGroup volumes: scale-out. 35
Figure 29) EDA workload — cell builder. 35
Figure 30) EDA workload — memory simulation and validation. 36
Figure 31) Average latency versus achieved ops — EDA benchmark, customer test. 37
Figure 32) Throughput (GBps) — EDA benchmark, customer test. 37
Figure 33) Storage QoS on FlexGroup volumes—single-node connection. 39
Figure 34) Storage QoS on FlexGroup volumes— multinode connection. 39
Figure 35) Qtree QoS use cases. 40
Figure 36) Capacity imbalance and likelihood of remote placement. 41
Figure 37) FlexGroup volume with a few large files; why usage can be suboptimal. 44
Figure 38) Potential worst case scenario for large files; all land in the same member volume. 44
Figure 39) Capacity imbalance example. 45
Figure 40) How FlexVol capacity can affect FlexGroup load distribution. 50
Figure 41) FlexVol and FlexGroup architecture comparison. 54
Figure 42) ONTAP System Manager FlexGroup volume creation. 57
Figure 43) Error when creating a FlexGroup volume beyond the allowed maximum in System Manager. 61
Figure 44) How capacity is divided among member volumes. 62

Figure 45) Effect of larger files in a FlexGroup member volume. 63
Figure 46) Fewer, larger member volumes....64
Figure 47) FlexGroup volumes—member sizes versus FlexGroup volume capacity. 66
Figure 48) Member volume size allocation after a `volume autosize` operation....70
Figure 49) Storage efficiency domains—when will deduplication be effective?....72
Figure 50) Storage efficiency domains with FlexGroup volumes....74
Figure 51) Storage efficiency savings — ONTAP System Manager....74
Figure 52) File write behavior before elastic sizing....80
Figure 53) File write behavior after elastic sizing....80
Figure 54) Latency impact of elastic sizing....82
Figure 55) Constituent volume latency increase from elastic sizing. 83
Figure 56) Initial FlexGroup data balance –— proactive resize, autosize disabled. 84
Figure 57) FlexGroup data balance, ~68% used — proactive resize, autosize disabled....84
Figure 58) FlexGroup data balance, job complete — proactive resize, autosize disabled. 84
Figure 59) FlexGroup data balance, new large file — proactive resize, autosize disabled....85
Figure 60) FlexGroup data balance, 80GB file — proactive resize, autosize disabled....85
Figure 61) FlexGroup data balance, out of space — proactive resize, autosize disabled. 85
Figure 62) Initial FlexGroup data balance — proactive resize, autosize enabled....86
Figure 63) FlexGroup data balance, ~68% used — proactive resize, autosize enabled. 86
Figure 64) FlexGroup data balance, job complete — proactive resize, autosize enabled....87
Figure 65) FlexGroup data balance, second test run — proactive resize, autosize enabled....87
Figure 66) FlexGroup data balance, autosize limit — proactive resize, autosize enabled....88
Figure 67) Modifying FlexGroup volume security styles in ONTAP System Manager. 92
Figure 68) ONTAP System Manager FlexGroup volume view. 95
Figure 69) Active IQ Unified Manager; FlexGroup capacity view. 96
Figure 70) Active IQ Unified Manager Capacity Trend. 96
Figure 71) Active IQ Performance Manager FlexGroup volume view....97
Figure 72) Member volume performance chart. 97
Figure 73) Member volume graphs. 98
Figure 74) Capacity effect when thin-provisioned FlexGroup volumes exist with space-guaranteed FlexVol volumes. 100
Figure 75) Adding aggregates with FlexGroup volumes....104
Figure 76) Adding nodes and expanding the FlexGroup volume....105
Figure 77) Removing nodes that contain FlexGroup member volumes....106
Figure 78) Quota reports – ONTAP System Manager....111
Figure 79) Quota volume status – ONTAP System Manager....111
Figure 80) Quota rules – ONTAP System Manager. 112
Figure 81) ONTAP 9.5 performance (operations/sec)—quotas on and off....115
Figure 82) ONTAP 9.5 performance (MBps)—quotas on and off....116
Figure 83) 64-bit file IDs in ONTAP System Manager 9.8 127

Figure 84) File System Analytics – enable. 133
Figure 85) File System Analytics – directory and file information. 133
Figure 86) File System Analytics – inactive and active data. 133
Figure 87) Capacity imbalance after deletion of larger files. 137
Figure 88) pNFS diagram. 144
Figure 89) pNFS operations diagram. 145
Figure 90) pNFS operations diagram—FlexGroup volumes. 145
Figure 91) ONTAP tools for VMware vSphere Support – FlexGroup Datastores. 150
Figure 92) FlexGroup volume, entire cluster (24 nodes). 157
Figure 93) Multiple nodes, partial cluster. 157
Figure 94) Git clone completion times comparison. 159
Figure 95) Average and maximum throughput comparison. 159
Figure 96) Maximum read throughput comparison. 160
Figure 97) Maximum write throughput comparison. 160
Figure 98) Total average IOPS comparison. 161
Figure 99) Average CPU utilization, throughput, and IOPS for a FlexGroup volume—AFF A700 HA pair, 128 threads. 161
Figure 100) Average CPU utilization, throughput, and IOPS for a FlexGroup volume—single-node AFF A700, 128 threads. 162
Figure 101) FlexGroup volume, single node. 162
Figure 102) FlexGroup volume mounted to FlexGroup volume. 163
Figure 103) Google Sheet — FlexGroup capacity view. 169
Figure 104) Google Sheet — FlexGroup member capacity view. 169
Figure 105) Google Sheet – Average inode size. 170
Figure 106) How logical space accounting works. 171
Figure 107) Active IQ Performance Manager graphs. 175
Figure 108) Converting a FlexVol volume that is nearly full and at maximum capacity. 177
Figure 109) Converting a FlexVol volume to a FlexGroup and adding member volumes. 179
Figure 110) XCP reporting graphs. 183
Figure 111) XCP report. 185
Figure 112) FlexGroup capacity breakdown—Active IQ Unified Manager. 188
Figure 113) Capacity view in Active IQ Unified Manager 192
Figure 114) Capacity related events in Active IQ Unified Manager 193
Figure 115) Event details, capacity — Active IQ Unified Manager 193
Figure 116) Editing volume thresholds. 196
Figure 117) Capacity: Global Level default settings. 196
Figure 118) Capacity: Object Level. 196
Figure 119) Add Alert: Inode monitoring 197
Figure 120) Edit Thresholds: Inodes. 197
Figure 121) Active IQ Unified Manager — fix out of inodes. 198

Figure 122) Sample statistics from conversion process. 206
Figure 123) Sample statistics during conversion process — Adding member volumes. 207
Figure 124) Sample statistics of conversion process — two times performance...208

The evolution of NAS in NetApp ONTAP

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 or even terabytes are over. Storage administrators face increasing demands from application owners for large buckets of capacity with enterprise-level performance.

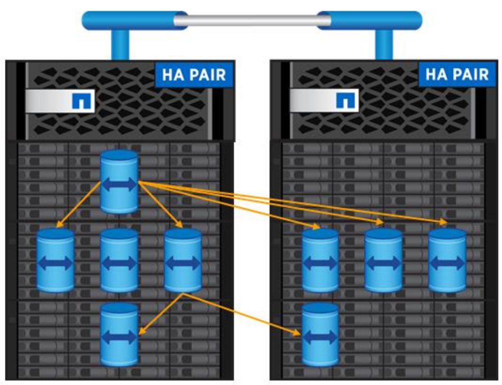
Machine learning and artificial intelligence (AI) workloads involve storage needs for a single namespace that can extend into the petabyte range (with billions of files). With the rise in these technologies, along with the advent of big data frameworks such as Hadoop, the evolution of NAS file systems is overdue. NetApp® ONTAP® FlexGroup is the ideal solution for these types of architectures.

Flexible volumes: A tried-and-true solution

The flexible volume, NetApp FlexVol® software, was introduced in NetApp Data ONTAP software in 2005 as part of the Data ONTAP 7.0 (Data ONTAP operating in 7-Mode) release. The concept took a storage file system and virtualized 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.

Figure 1) FlexVol design with junctioned architecture for >100TB capacity.



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 with.

Before FlexGroup, ONTAP administrators could create junction paths to attach FlexVol volumes to one another. In this way, they created a file system on the cluster that could act as a single namespace. Figure 1 shows an example of what a FlexVol volume junction design for a large namespace looks like.

Although this architecture worked for many environments, it was awkward to manage and did not give a “single-bucket” approach to the namespace, where the FlexVol volume’s capacity and file count constraints are limiting factors.

FlexGroup volumes: An evolution of NAS

ONTAP 9.1 brought innovation to scale-out NAS file systems: the NetApp ONTAP FlexGroup volume.

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.

Terminology

Many of the usual NetApp ONTAP terms (such as SVM, LIF, and FlexVol) are covered in the NetApp Product Documentation and NetApp Knowledge Center. Terminology specific to NetApp ONTAP FlexGroup is covered in the following list.

- **Constituent/member volumes.** In a FlexGroup context, "constituent volume" and "member volume" are interchangeable terms. They refer to the underlying NetApp FlexVol volumes that make up a FlexGroup volume and provide the capacity and performance gains that are achieved only with a FlexGroup volume.
- **FlexGroup volume.** A FlexGroup volume is a single namespace that is made up of multiple constituent/member volumes. It is managed by storage administrators, and it acts like a NetApp FlexVol volume. 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 is an ONTAP subsystem that repairs FlexGroup inconsistencies dynamically, with no outage or administrator intervention required.

- **Ingest.** Ingest is the consumption of data by way of file or folder creations.
- **Junction paths.** Junction paths were used to provide capacity beyond a FlexVol volume's 100TB limit prior to the simplicity and scale-out of FlexGroup. Junction paths join multiple FlexVol volumes together to scale out across a cluster and provide multiple volume affinities. The use of a junction path in ONTAP is known as "mounting" the volume within the ONTAP namespace.
- **Large files.** See the next section, "What are large files?"
- **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 large sizes to avoid the need to grow them later, but it does present the management overhead of needing to monitor available space closely.

  In overprovisioned volumes, the available space reflects the actual physical available space in the aggregate. Therefore, the usage percentage and capacity available values might seem off a bit. However, they simply reflect a calculation of the actual space that is available when 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® system that allows a FlexGroup volume to balance ingest workloads across multiple FlexGroup constituents or members. It is also used with NetApp FlexCache® volumes.
- **Remote hard links.** Remote hard links are the building blocks of 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.

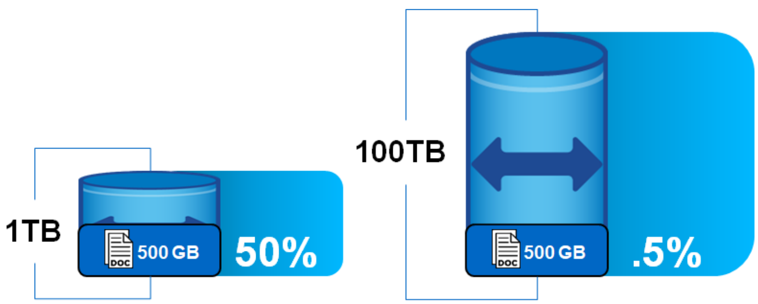
What are large files?

This document uses the term “large file” liberally. Therefore, it’s important to define exactly what a “large file” is in the context of FlexGroup volumes.

A FlexGroup volume operates optimally when a workload is ingesting numerous small files, because FlexGroup volumes maximize the system resources to address those specific workloads that might bottleneck because of serial processing in a FlexVol volume. FlexGroup volumes also work well with various other workloads (as defined in section 4, Use cases). One type of workload that historically created problems in a FlexGroup volume, however, is a workload with larger files or files that grow over time, such as database files.

In a 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 only take up 0.5% of the volume capacity.

Figure 2) What is a large file?



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. ONTAP has no visibility into how large a file will be when it is created, so it doesn’t know how to prioritize larger files. Instead, it works reactively to balance data to other member volumes with free space discrepancies.

Each ONTAP release is constantly improving the approach to large files in FlexGroup volumes.

- In ONTAP 9.6, elastic sizing helped mitigate concerns with larger files. ONTAP borrows space from other member volumes to allow large files to complete their writes.
- ONTAP 9.7 introduced ingest algorithm changes to help balance large files and/or datasets with mixed file sizes.
- ONTAP 9.8 brings capacity management simplicity by way of Proactive resizing.

NetApp ONTAP FlexGroup advantages

NetApp ONTAP FlexGroup provides various advantages for different workloads. The advantages are described in the following sections.

Massive capacity and predictable low latency for high-metadata workloads

FlexGroup volumes offer a way for storage administrators to easily provision massive amounts of capacity with the ability to nondisruptively scale out that capacity. FlexGroup also enables parallel performance for high metadata workloads that can increase throughput and total operations while still providing low latency for mission-critical workloads.

Efficient use of all cluster hardware

FlexGroup volumes allow storage administrators to easily span multiple physical aggregates and nodes with member FlexVol volumes, while maintaining a true single namespace for applications and users to dump data into. Although clients and users see the space as monolithic, ONTAP is working behind the scenes to distribute the incoming file creations evenly across the FlexGroup volume to provide efficient CPU and disk utilization.

Simple, easy-to-manage architecture and balancing

To make massive capacity easy to deploy, NetApp allows you to manage FlexGroup volumes like NetApp FlexVol volumes. ONTAP handles the underlying member volume creation and balance across the cluster nodes and provides a single access point for NAS shares.

Superior density for big data

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

- Thin provisioning
- Data compaction
- Data compression
- Deduplication

In addition, ONTAP supports large SSDs, which can deliver massive amounts of raw capacity in a single 24-drive shelf enclosure. It is possible to get petabytes of raw capacity in just 10U of rack space, which cuts costs on cooling, power consumption, and rack rental space and offers excellent density in the storage environment. These features, combined with a FlexGroup volume's ability to efficiently use that capacity and balance performance across a cluster, give you a solution that was made for big data.

Use cases

The NetApp ONTAP FlexGroup design is most beneficial in specific use cases (electronic design and automation, software development, and so on).

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 (EDA)
- 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 (such as video surveillance)
- Unstructured NAS data (such as home directories)
- Big data and data lakes (Hadoop with the NetApp NFS connector)
- Virtualized workloads (ONTAP 9.8 and later)

Nonideal cases

Some workloads are currently not recommended for FlexGroup volumes.

These workloads include:

- Workloads that require file 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 with a large amount of file renames
- Workloads with millions of files in a single directory that require frequent scans of all of the files
- Workloads with thousands of symlinks
- Workloads that require specific features and functionalities that are not currently available with FlexGroup volumes

If you have questions, feel free to email ng-flexgroups-info@netapp.com.

FlexGroup volume use case examples

The following sections describe two examples of real-world use cases.

FlexGroup use case example #1: NetApp Active IQ infrastructure

The NetApp on NetApp program involves NetApp's backing infrastructure teams and their use of NetApp products. This program serves not only to provide NetApp IT and other groups with the best solutions for their problems, but also to show that NetApp has confidence in their own offerings.

In Episode 182 of the Tech ONTAP Podcast, representatives from NetApp on NetApp describe how FlexGroup volumes are being used to power the NetApp Active IQ® infrastructure. For details about the solution, including statistics, read the blog post ONTAP FlexGroup Technology Powers NetApp's Massive Active IQ Data Lake.

FlexGroup use case example #2: Back up repository for SQL Server

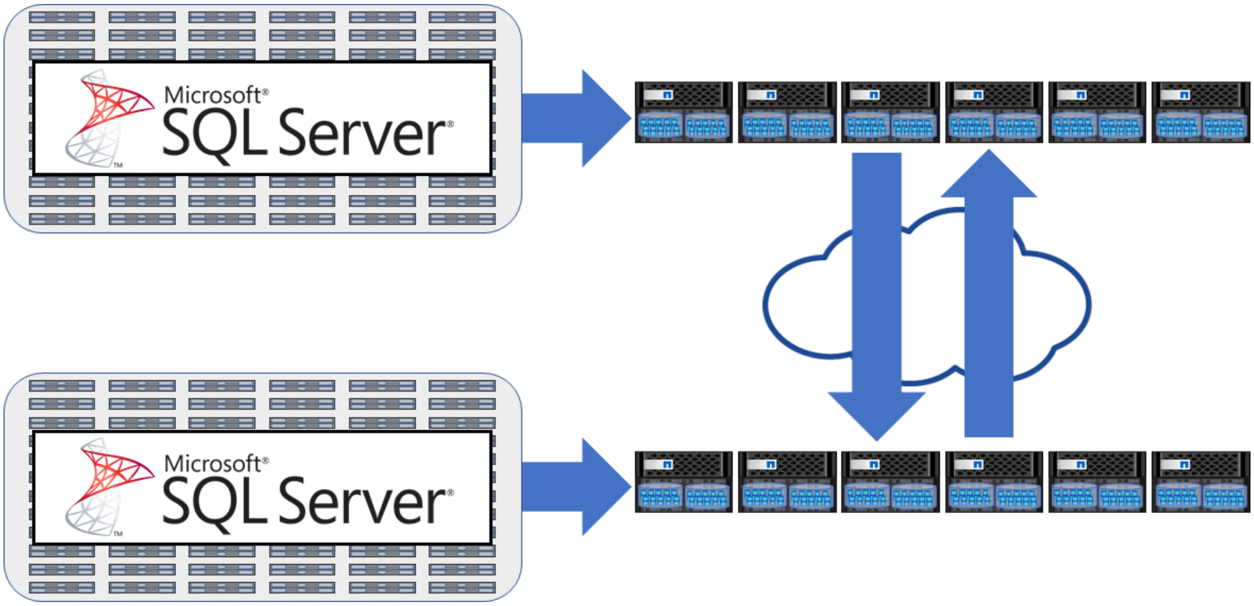
In this environment, the customer wanted to perform compressed backups of 5,000 Microsoft SQL Servers over SMB. This test was done with approximately 200 servers to vet out the solution, with a slow ramp up over the course of a few months.

But this database isn't only a backup target—it will also be replicated to a disaster recovery site by using NetApp SnapMirror® for extra data protection.

Each site has a six-node FAS8200 cluster running ONTAP 9.4 using 6TB near-line SAS (NL-SAS) encrypted drives. Each cluster holds 3PB of usable capacity. The clusters use 30 FlexGroup volumes and qtrees within the volumes for data organization.

The FlexGroup volumes are 64TB each and the member volumes are 2.6TB each, with four members per node across six nodes (24 total members per FlexGroup volume).

Figure 3) SQL Server backup environment.



The results

This customer needed a single namespace that could collect ~150TB worth of MSSQL backup data over a 12-hour period. That's ~12TB per hour at ~3.5GB per second.

During testing, we used 222 servers at site A and 171 servers at site B. During the test, each cluster's CPU was at 95% utilization and the backup jobs (sequential writes) were able to accomplish 8.4GB per second, which is ~2.4 times the amount of throughput the job needed. At this rate, the backups could complete in approximately 5 hours, rather than the 12-hour window. Also, this SMB workload performed approximately 120,000 IOPS. When more clients are added to this workload, we expect the throughput to max out at around 9GBps.

Figure 4) Throughput and total operations during test runs.

cpu avg	cpu busy	total ops	nfs-ops	cifs-ops	fcache ops	spin-ops	total recv	total sent	data busy	data recv	data sent	cluster busy	cluster recv	cluster sent	disk read	disk write	pkts recv	pkts sent
56%	81%	54530	0	54530	0	54420	6.16GB	2.65GB	44%	3.34GB	28.3MB	22%	2.82GB	2.62GB	128MB	3.31GB	968237	898917
65%	78%	70482	0	70482	0	70407	8.03GB	3.44GB	47%	4.33GB	30.9MB	24%	3.70GB	3.41GB	114MB	4.79GB	1178768	1102912
74%	87%	88725	0	88725	0	88105	10.2GB	4.30GB	49%	5.44GB	37.1MB	36%	4.78GB	4.26GB	157MB	5.56GB	1389743	1324559
86%	92%	111577	0	111577	0	110569	12.8GB	5.88GB	53%	6.84GB	41.9MB	31%	6.00GB	5.84GB	153MB	6.77GB	1726469	1679506
88%	92%	115036	0	115036	0	113509	13.2GB	6.44GB	51%	7.06GB	45.9MB	49%	6.14GB	6.40GB	142MB	7.65GB	1845760	1814549
92%	95%	118148	0	118148	0	117104	13.6GB	6.11GB	45%	7.26GB	49.9MB	42%	6.34GB	6.07GB	149MB	8.11GB	1802929	1769902
95%	98%	122953	0	122953	0	122123	14.3GB	7.10GB	47%	7.54GB	45.9MB	43%	6.75GB	7.06GB	135MB	8.29GB	1978205	1952416
96%	99%	126241	0	126241	0	125104	14.6GB	6.43GB	53%	7.75GB	54.3MB	44%	6.80GB	6.37GB	135MB	8.28GB	1865375	1849777
95%	97%	121948	0	121948	0	120719	13.9GB	7.25GB	44%	7.47GB	47.3MB	40%	6.41GB	7.20GB	108MB	8.30GB	1995908	1967271
95%	98%	123079	0	123079	0	121113	13.9GB	5.71GB	41%	7.56GB	49.0MB	38%	6.37GB	5.66GB	129MB	8.40GB	1761097	1712061
95%	97%	120567	0	120567	0	120493	13.7GB	7.01GB	42%	7.41GB	47.6MB	36%	6.34GB	6.96GB	116MB	8.48GB	1888934	1882711
95%	98%	119573	0	119573	0	119458	13.6GB	5.74GB	37%	7.33GB	44.4MB	35%	6.28GB	5.69GB	111MB	8.19GB	1702969	1671363
95%	97%	119538	0	119538	0	119829	13.5GB	6.98GB	41%	7.34GB	46.2MB	35%	6.17GB	6.93GB	120MB	8.44GB	1880298	1873821
95%	98%	118119	0	118119	0	118373	13.4GB	5.56GB	37%	7.25GB	45.4MB	37%	6.17GB	5.52GB	118MB	8.42GB	1666066	1630785
95%	98%	118862	0	118862	0	118327	13.6GB	6.29GB	39%	7.29GB	47.1MB	33%	6.30GB	6.24GB	114MB	8.31GB	1784134	1759266
96%	99%	121039	0	121039	0	121136	13.7GB	6.67GB	38%	7.44GB	44.5MB	34%	6.21GB	6.63GB	120MB	8.35GB	1832520	1827158
96%	99%	120852	0	120852	0	120920	13.7GB	5.77GB	39%	7.42GB	47.8MB	33%	6.24GB	5.72GB	111MB	8.51GB	1706939	1678778
94%	97%	119819	0	119819	0	120129	13.7GB	7.05GB	41%	7.36GB	42.6MB	35%	6.29GB	7.01GB	118MB	8.49GB	1882656	1877381

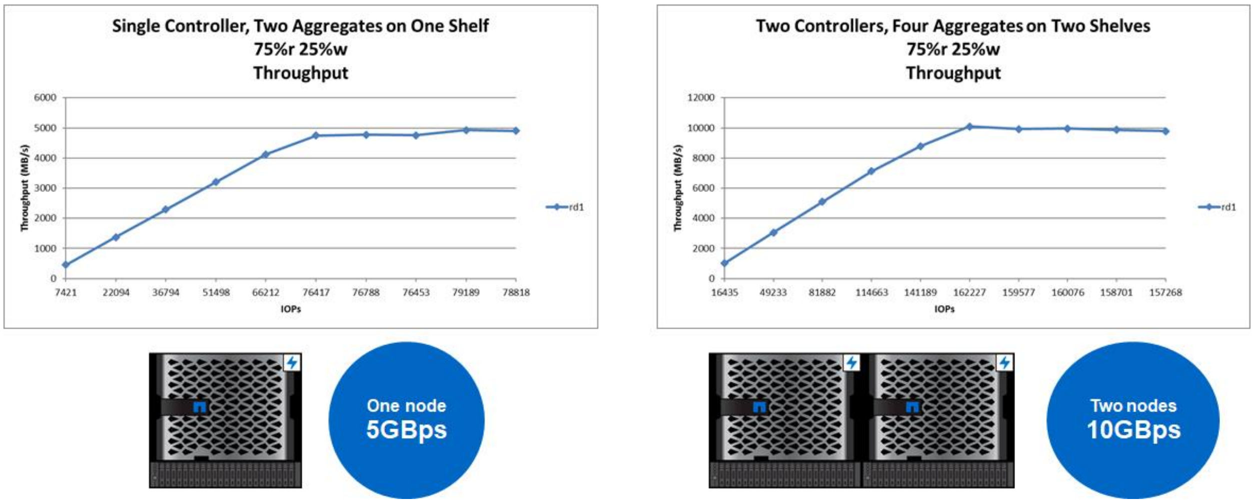
Data protection

In addition to the performance seen on the FlexGroup volume for the production workload, this customer was also able to achieve a high rate of transfer for the SnapMirror relationships between sites—8.4GB per second for the SnapMirror transfer. This rate means that the replication window for a 150TB dataset would be about 5.5 hours for the initial transfer. After that, the deltas should be able to complete well within the required transfer window, providing a solid disaster recovery plan for these MSSQL backups.

Scale-out performance

This six-node cluster was able to push over 8.4GB per second to a FlexGroup volume. In NetApp Customer Proof of Concept (CPOC) labs, we see near-linear performance gains by adding nodes to a cluster. The following graphs show throughput results for a single-node NetApp AFF A700 all-flash storage system and a two-node AFF A700.

Figure 5) CPOC scale-out throughput results.



Note: If you want to add more performance to your backup workload, you can add more nodes.

Conclusion

Not only is a FlexGroup volume great for small or high-file-count workloads such as EDA and software builds, it also can handle high throughput requirements for larger streaming files. It also reduces backup windows by scaling out storage across multiple nodes and applies all your cluster resources while maintaining performance even with spinning drives.

FlexGroup feature support and maximums

This section shows which NetApp ONTAP features are supported for use with FlexGroup volumes; it also notes the ONTAP version in which feature support was added. If a feature is not listed in this section, email flexgroups-info@netapp.com for information. For features specific to data protection, see TR-4678: Data Protection Best Practices with FlexGroup Volumes.

Table 1) General ONTAP feature support.

Supported feature	Version of ONTAP first supported
NetApp Snapshot™ copy technology	ONTAP 9.0
NetApp SnapRestore® software (FlexGroup level)	ONTAP 9.0
Hybrid aggregates	ONTAP 9.0
Constituent or member volume move	ONTAP 9.0
Postprocess deduplication	ONTAP 9.0
NetApp RAID-TEC™ technology	ONTAP 9.0
Per-aggregate consistency point	ONTAP 9.0
Sharing FlexGroup with FlexVol in the same SVM	ONTAP 9.0

Supported feature	Version of ONTAP first supported
Active IQ Unified Manager	ONTAP 9.1
Inline adaptive compression	ONTAP 9.1
Inline deduplication	ONTAP 9.1
Inline data compaction	ONTAP 9.1
Thin provisioning	ONTAP 9.1
NetApp AFF	ONTAP 9.1
Quota reporting	ONTAP 9.1
SnapMirror technology	ONTAP 9.1
User and group quota reporting (no enforcement)	ONTAP 9.1
Access Based Enumeration (ABE) for SMB shares	ONTAP 9.1
Aggregate inline deduplication (cross-volume deduplication)	ONTAP 9.2
NetApp Volume Encryption (NVE)	ONTAP 9.2
NetApp SnapVault® technology	ONTAP 9.3
Qtrees	ONTAP 9.3
Automated deduplication schedules	ONTAP 9.3
Version-independent SnapMirror and unified replication	ONTAP 9.3
Antivirus scanning for SMB	ONTAP 9.3
Volume autogrow	ONTAP 9.3
QoS maximums/ceilings	ONTAP 9.3
FlexGroup expansion without SnapMirror return to baseline	ONTAP 9.3
Improved ingest heuristics	ONTAP 9.3
SMB change/notify	ONTAP 9.3
File audit	ONTAP 9.4
NetApp FPolicy™	ONTAP 9.4
Adaptive QoS	ONTAP 9.4
QoS minimums (AFF only)	ONTAP 9.4
Relaxed SnapMirror limits	ONTAP 9.4
SMB 3.x Multichannel	ONTAP 9.4
FabricPool	ONTAP 9.5
Quota enforcement	ONTAP 9.5
Qtree statistics	ONTAP 9.5
Inherited SMB watches and change notifications	ONTAP 9.5
SMB copy offload (offloaded data transfer)	ONTAP 9.5
Storage-Level Access Guard	ONTAP 9.5
NetApp FlexCache (cache only; FlexGroup as origin supported in ONTAP 9.7)	ONTAP 9.5
Volume rename	ONTAP 9.6
Volume shrink	ONTAP 9.6
NetApp MetroCluster™	ONTAP 9.6
Elastic sizing	ONTAP 9.6
Continuously Available Shares (SMB)*	ONTAP 9.6

Supported feature	Version of ONTAP first supported
* SQL Server and Hyper-V workloads only	
NetApp Aggregate Encryption (NAE)	ONTAP 9.6
NetApp Cloud Volumes ONTAP	ONTAP 9.6
NetApp FlexClone®	ONTAP 9.7
In-place conversion of FlexVol to FlexGroup volumes (see "Deploying a FlexGroup volume on aggregates with existing FlexVol volumes")	ONTAP 9.7
vStorage APIs for Array Integration (VAAI)	ONTAP 9.7
NDMP	ONTAP 9.7
NFSv4.0 and NFSv4.1 (including parallel NFS, or pNFS)	ONTAP 9.7
FlexGroup volumes as FlexCache origin volumes	ONTAP 9.7
File cloning	ONTAP 9.8
Proactive resizing	ONTAP 9.8
NFSv4.2 (base protocol support)	ONTAP 9.8
NDMP enhancements: EXCLUDE, RBE (Restartable Backup Extension), MULTI_SUBTREE_NAMES	ONTAP 9.8
1,023 Snapshots	ONTAP 9.8
Qtree QoS	ONTAP 9.8
Logical space reporting and enforcement	ONTAP 9.9.1
Storage VM disaster recovery (SVM-DR)	ONTAP 9.9.1
Fan-out SnapMirror	ONTAP 9.9.1
Cascading SnapMirror	ONTAP 9.9.1

Table 2) General NAS protocol version support.

Supported NAS Protocol Version	Version of ONTAP First Supported
NFSv3	ONTAP 9.0
SMB2.1, SMB3.x	ONTAP 9.1 RC2
NFSv4.0, NFSv4.1, pNFS	ONTAP 9.7
NFSv4.2 (base protocol support)	ONTAP 9.8 Labeled NFS support – ONTAP 9.9.1

Table 3) Unsupported SMB2.x and 3.x features.

Unsupported SMB2.x features	Unsupported SMB 3.x features
• SMB Remote Volume Shadow Copy Service (VSS)	• VSS for SMB file shares. • SMB directory leasing • SMB direct or remote direct memory access (RDMA) **Note:** SMB 3.0 encryption is supported.

Note: Remote VSS is not the same as the SMB Previous Versions tab. Remote VSS is application-aware Snapshot functionality and is most commonly used with Hyper-V workloads. FlexGroup volumes have supported the SMB Previous Versions tab since it was introduced.

Behavior of unsupported SMB features

Usually, if an SMB feature is unsupported in ONTAP, it simply does not work. With NetApp ONTAP FlexGroup, there are some considerations regarding unsupported SMB features and functionality.

Table 4) How unsupported SMB features behave with FlexGroup volumes.

Feature	Behavior with FlexGroup Volumes
SMBv1.0	Access fails or is denied for any shares accessing with SMB 1.0. This can affect Windows XP, Windows 2003, and office equipment such as scanners or copiers that attempt to connect to the NAS with SMB.
Change notification/SMB watches	Change notifications are supported as of ONTAP 9.2. For details on change notifications, see "Use of change notifications with SMB." Before ONTAP 9.2, the behavior is as follows: • There are no warning or failures. Change notification simply does not take place. For more information about change notification with SMB, see this MSDN article. • Lack of change notifications in SMB can affect applications that depend on the immediate appearance of newly created files in Windows folders. **Note:** ONTAP 9.5 and later releases provide support for inherited watches, which applies change notifications at a parent directory level and performs better than previous releases. **Note:** ONTAP 9.5P7 and 9.6P3 and later releases increase the cache size for change notifications to reduce cache churn impact on performance (bug 1232663). **Note:** ONTAP 9.7 and later releases offer parallel processing of change notification operations for a better overall experience. **Note:** ONTAP 9.10.1 moves the processing of change notifications in ONTAP to a process domain better equipped to handle large numbers of notifications (bug 1304356) and ensures that removing the `changenotify` share property takes place for existing connections to the share (bug 1413224). If you wish to use change notifications, use ONTAP 9.10.1 and later releases for the best results.
Offloaded data transfer (ODX)	In versions earlier than ONTAP 9.5, failovers occur to the traditional client-side copy. The effect is low; failovers are not as fast. In ONTAP 9.5 and later, ODX works as expected. For more information about ODX, see this TechNet article.
Remote Volume Shadow Copy Service (VSS)	There is no warning; Remote VSS just does not work. The effect should be low because the primary use case for Remote VSS is with Hyper-V, which is not a recommended workload for FlexGroup volumes. For more information about Remote VSS, see this TechNet article.
Continuously available shares	Continuously available shares are not allowed on FlexGroup volumes before ONTAP 9.6. Attempting to set the share property fails. For more information about continuously available shares, see this TechNet article.

Maximums and minimums

This section covers the maximums and minimums that are specific to NetApp ONTAP FlexGroup volumes. Table 5 lists the maximum values and shows whether the maximum is hard-coded/enforced or a recommended or tested value.

Table 5) FlexGroup maximums.

	Value	Value Type
FlexGroup volume size	20PB	Tested/recommended*
FlexGroup total file count	400 billion	Tested/recommended*
Cluster node count	24 (12 HA pairs)	Hard-coded/enforced
NetApp FlexVol member volume size	100TB	Hard-coded/enforced
FlexVol member volume file count	2 billion	Hard-coded/enforced
NetApp SnapMirror volume count (member per FlexGroup)	32 (ONTAP 9.4 and earlier) 200 (ONTAP 9.5 and later)	Hard-coded/enforced
SnapMirror volume count (FlexGroup total per cluster)	100 (ONTAP 9.4 and earlier) 6,000 (ONTAP 9.5 and later)	Hard-coded/enforced
File size	16TB	Hard-coded/enforced
FlexVol member constituent count	200	Tested/recommended*
Aggregate size/count	Same as NetApp ONTAP limits	Hard-coded/enforced

Table 6) FlexGroup minimums.

	Value	Value Type
FlexVol member size	100GB	Tested/recommended*
Data aggregate count	1	Hard-coded/enforced
SnapMirror schedule	30 minutes	Tested/recommended*
NetApp Snapshot copy schedule	30 minutes	Tested/recommended*

Note: *Limits described as tested/recommended are tested limits based on a 10-node cluster. If allowed by the platform, actual limits are not hard-coded and can extend beyond these limits up to 24 nodes. For more information, see the section “Theoretical or absolute maximums.” However, official support for the number of member volumes is 200. If you need to exceed this limit, contact your NetApp sales representative to start the qualification process for more member volumes.

Deciding whether FlexGroup volumes are the right fit

NetApp ONTAP FlexGroup volumes are an ideal fit for many use cases—particularly the ones that are listed in the section “Ideal use cases.”

However, not all use cases make sense for FlexGroup volumes. This section provides information to help you decide whether FlexGroup volumes are the right fit for your workloads.

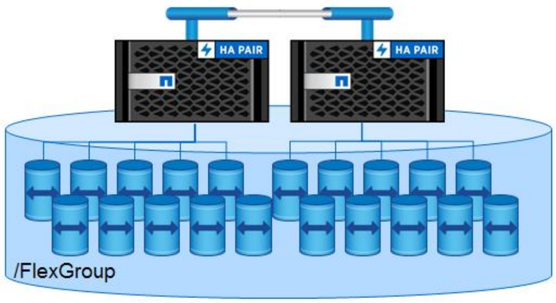
Scale-out performance

FlexGroup volumes distribute their data and load among the multiple constituents that make up the collective FlexGroup volume. This model allows a FlexGroup volume to use more of the resources within each node (CPU, network adapters, disks, and so on) and to use more nodes within a cluster to address a workload.

In addition, the concept ties in nicely with the NetApp ONTAP clustered architecture, which allows the nondisruptive addition of nodes and disks to increase performance without negatively affecting applications. With a FlexGroup volume, you can simply expand the FlexGroup to add more members or use nondisruptive volume move technology to redistribute the member volumes across the new nodes.

A single FlexGroup volume internally comprises multiple separate NetApp FlexVol volumes, which in turn can be stored on any aggregates and can span multiple nodes in your cluster.

Figure 6) FlexGroup volume.



In addition, the concept ties in nicely with the NetApp ONTAP clustered architecture, which allows the nondisruptive addition of nodes and disks to increase performance without negatively affecting applications. With a FlexGroup volume, you can simply expand the FlexGroup to add more members or use nondisruptive volume move technology to redistribute the member volumes across the new nodes.

A single FlexGroup volume internally comprises multiple separate NetApp FlexVol volumes, which in turn can be stored on any aggregates and can span multiple nodes in your cluster.

When clients add files and subdirectories to the FlexGroup volume, ONTAP automatically determines the best FlexVol member to use for storing each new file and subdirectory. The FlexGroup volume attempts to organize your data, both for best performance and for data and load distribution.

Because of this workload distribution, FlexGroup volumes can handle much more metadata traffic than a FlexVol volume. Thus, FlexGroup volumes can be useful for a variety of workloads that are metadata-intensive or that require a large amount of throughput.

Feature compatibility limitations

FlexGroup volumes in ONTAP 9.1 and later support some common NAS protocols, such as NFSv3, NFSv4.0, v4.1 and v4.2, SMB2.x, and SMB3. For details about the support of those protocols and which ONTAP release they were supported, see Table 1, Table 2, and Table 3 in "FlexGroup feature support and maximums."

Additionally, FlexGroup volumes are what makes up FlexCache volumes and ONTAP S3 object buckets. However, only FlexCache volumes are able to leverage NAS protocol interaction. S3 buckets are only accessible via the S3 protocol. ONTAP S3 is available for general use in ONTAP 9.8.

SMB 1.0 is not supported for use with FlexGroup volumes. FlexGroup volumes do not support block protocol/SAN access (iSCSI, FCP, NVMe).

Table 7 provides information for deciding whether FlexGroup volumes are the right fit for an environment by comparing the currently available container types in ONTAP.

Table 7) ONTAP volume family comparison.

	FlexVol Volumes	FlexGroup Volumes
Client access protocols (current support)	SAN (FCP, iSCSI, NVMe) NAS • SMB1.0, 2.1, 3.x • NFSv3 • NFSv4.0, NFSv4.1, NFSv4.2	S3 NAS • SMB2.x, 3.x • NFSv3 • NFSv4.0, NFSv4.1, NFSv4.2
Capacity scaling	• Single FlexVol volume	• Can be mounted to FlexGroup or FlexVol volumes in the namespace

	FlexVol Volumes	FlexGroup Volumes
	• Can be mounted to FlexVol or FlexGroup volumes in the namespace • 100TB, 2 billion file limit	• 20PB* • 400 billion files* • Nondisruptive capacity increases *Current tested limits on 10-node cluster; can extend beyond these values
Metadata scaling	FlexVol volumes are limited to a single node for metadata processing and serial processing of metadata, which does not take full advantage of the node's CPU threads.	FlexGroup volumes can use multiple nodes (and their resources) and multiple aggregates. In addition, FlexGroup can use multiple volume affinities to maximize CPU thread utilization potential.
ONTAP feature compatibility	Compatible with all ONTAP features	Supports most ONTAP features. • For information, see the "Supported Features" section of this document.
Throughput scaling	Limited to: • One node (set of CPU, RAM, network ports, connection limits, and so on) • One aggregate	FlexGroup volumes can use the resources of an entire cluster in service of I/O, providing much higher throughput than a single FlexVol volume can, with linear scale of performance as you add nodes to the FlexGroup.
Cloud support	All NetApp cloud integration, such as: • Cloud Volumes ONTAP • Cloud Volumes Services • SnapMirror Cloud	Cloud Volumes ONTAP (CVO) – CLI only • Capacity limitations of CVO apply
ONTAP upgrades and reverts	Data stored in any volume family is safely retained during ONTAP version changes, with one exception: If reverting to a release earlier than ONTAP 9.1, FlexGroup volumes cannot be retained.	
GUI compatibility	• ONTAP System Manager • Active IQ Performance Manager • Active IQ Unified Manager • Cloud Insights	

Simplifying performance

A single FlexGroup volume can consist of multiple FlexVol member volumes, which in turn can reside on any aggregate and on any node in your cluster. As clients drive traffic against that FlexGroup volume, ONTAP automatically breaks that traffic into tasks for different constituent FlexVol volumes to perform. This approach provides for a concurrency of operations that a single FlexVol volume is incapable of handling.

The benefit of this scale-out behavior is a dramatic increase in processing power that can scale linearly as you add nodes to your ONTAP cluster. A single FlexGroup volume can service much heavier workloads than a single FlexVol volume can at more predictable latencies.

SPEC SFS 2014_swbuild submission: FlexGroup volume, ONTAP 9.2

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

See the results for competitor systems here.

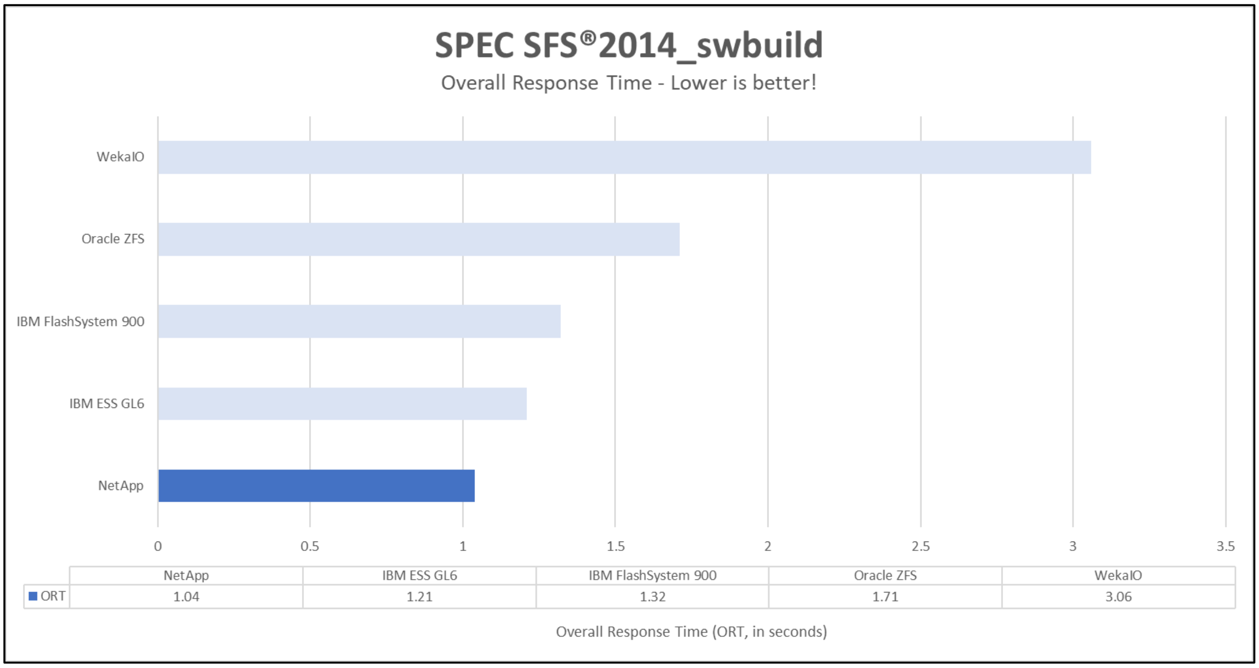
NetApp results

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."

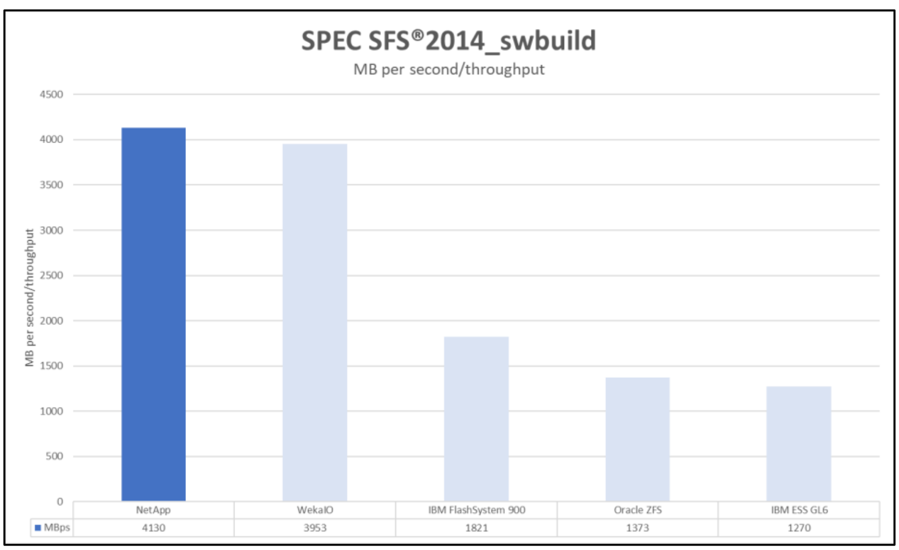
In this test, FlexGroup volumes achieved the lowest ORT ever recorded for a storage system at the time.

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



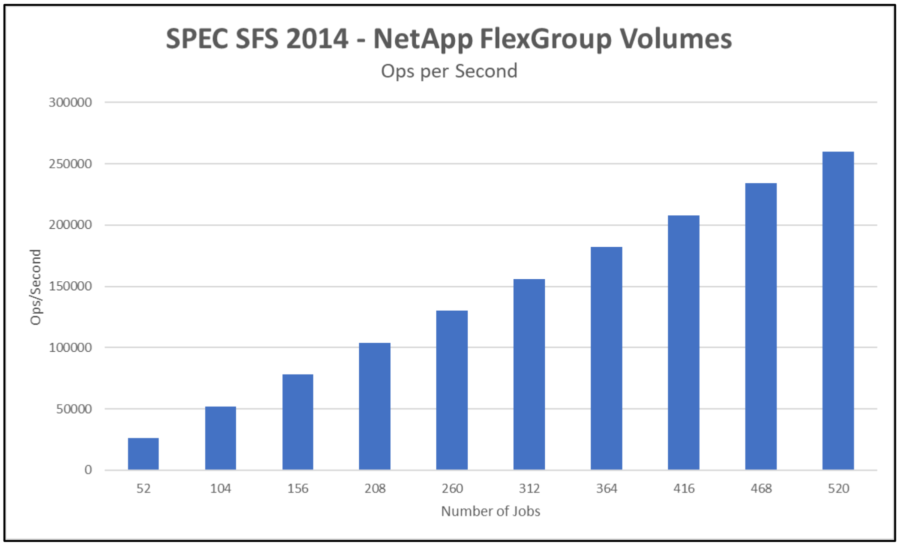
FlexGroup volumes also outperformed other submissions in throughput. In the benchmark, FlexGroup volumes achieved over 4GBps.

Figure 8) Throughput, SPEC SFS 2014_swbuild submissions.



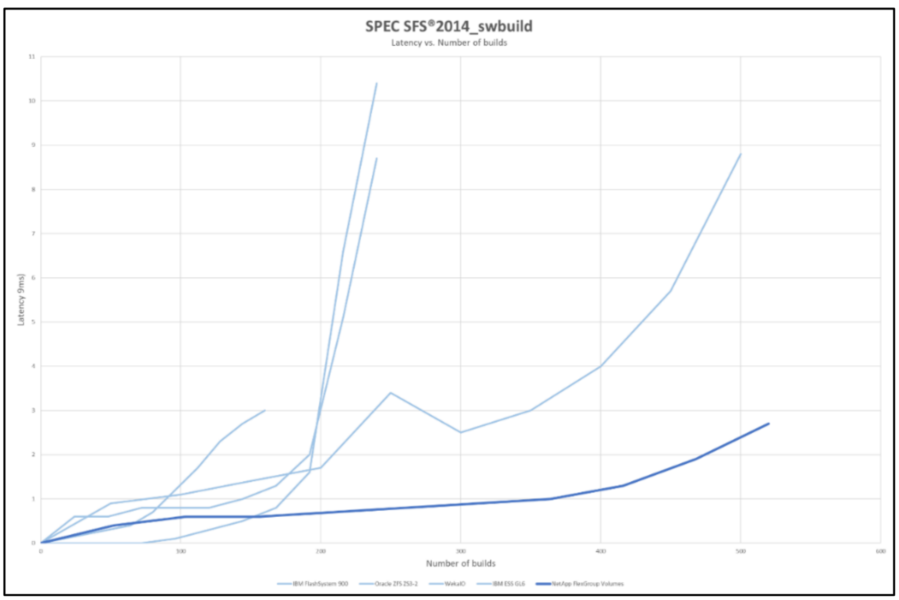
The results of this performance benchmark were achieved with more than 500 concurrent jobs providing 260,000 IOPS.

Figure 9) IOPS, SPEC SFS 2014_swbuild submissions.



If latency is important to your applications, FlexGroup volumes also saw the most predictable low latency of all the submissions.

Figure 10) Latency versus number of builds, SPEC SFS 2014_swbuild submissions.



SPEC SFS 2014_swbuild submission: FlexGroup volume, ONTAP 9.5

In November of 2018, NetApp submitted a second round of SPEC SFS 2014_swbuild results for publication, this time on an AFF system. NetApp ONTAP using FlexGroup volumes achieved the highest throughput with the most concurrent builds of all other systems by a factor of 2.5 times more than the next highest system. See the following links for the results:

- NetApp four-node AFF A800 with FlexGroup volumes

- [NetApp eight-node AFF A800 with FlexGroup volumes]
- [NetApp, twelve-node AFF A800 with FlexGroup volumes]

Figure 11, Figure 12, and Figure 13 show the testing results and comparisons with other submissions.

Figure 11) SPEC SFS 2014_swbuild—concurrent builds.

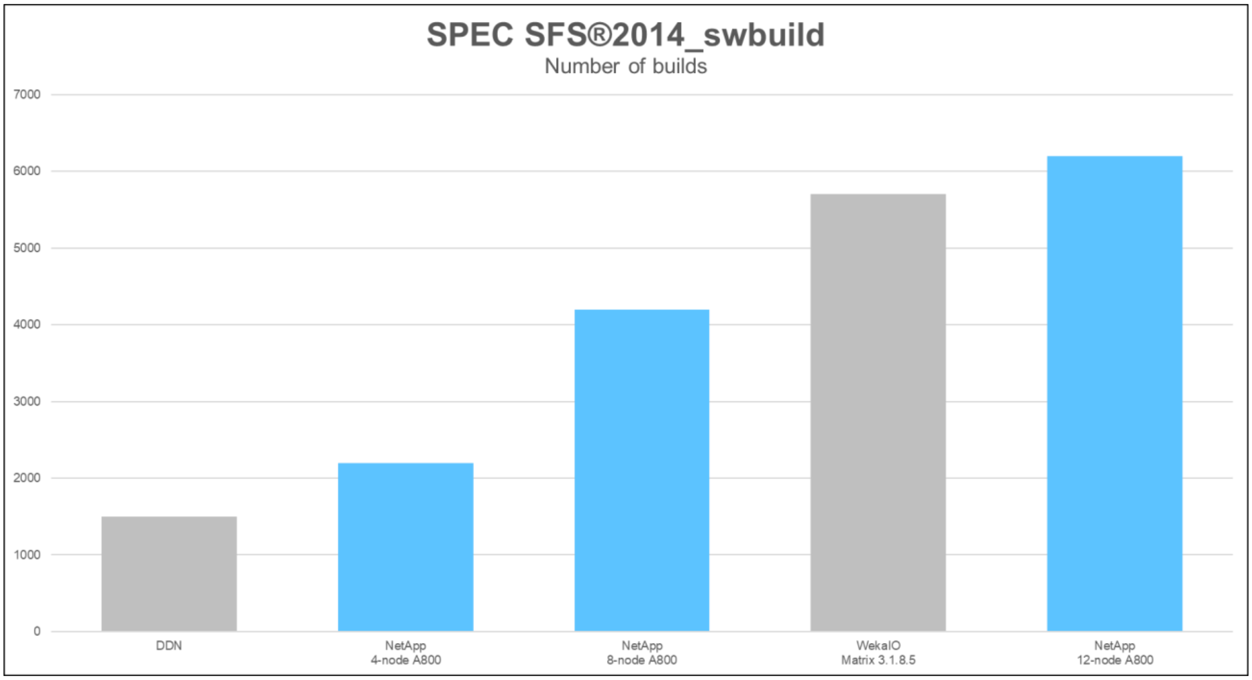


Figure 12) SPEC SFS 2014_swbuild—latency versus build ops/sec.

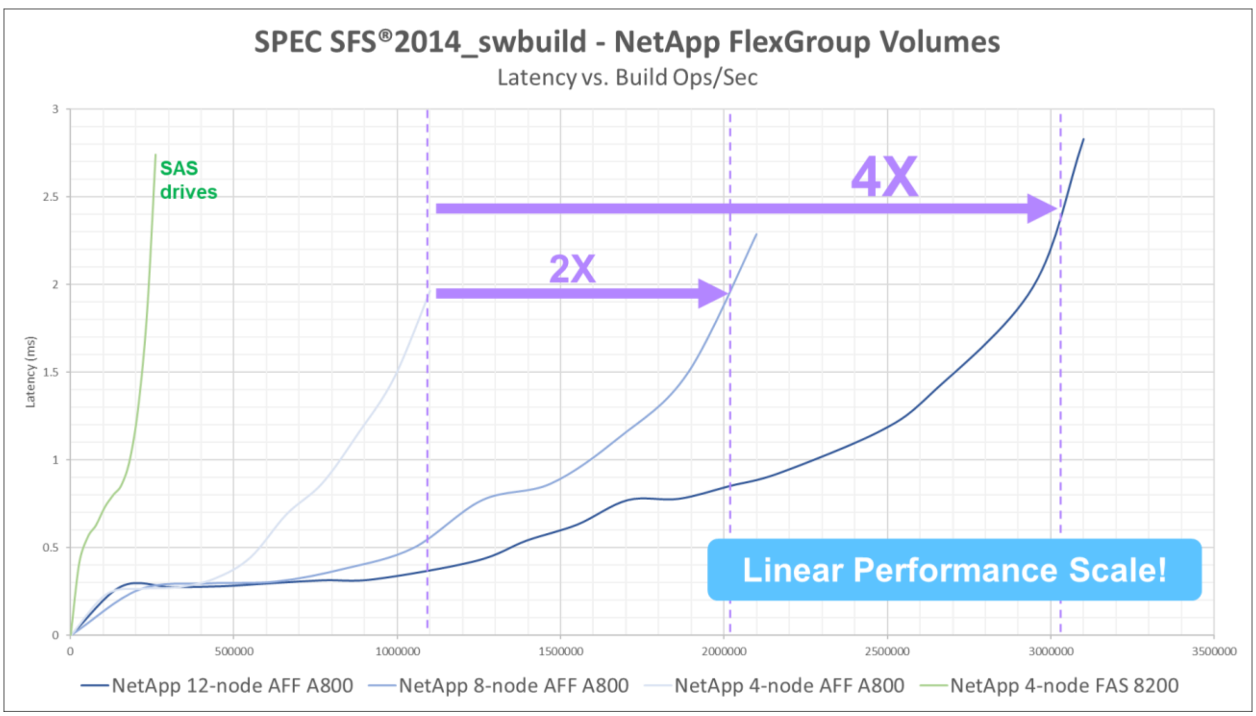
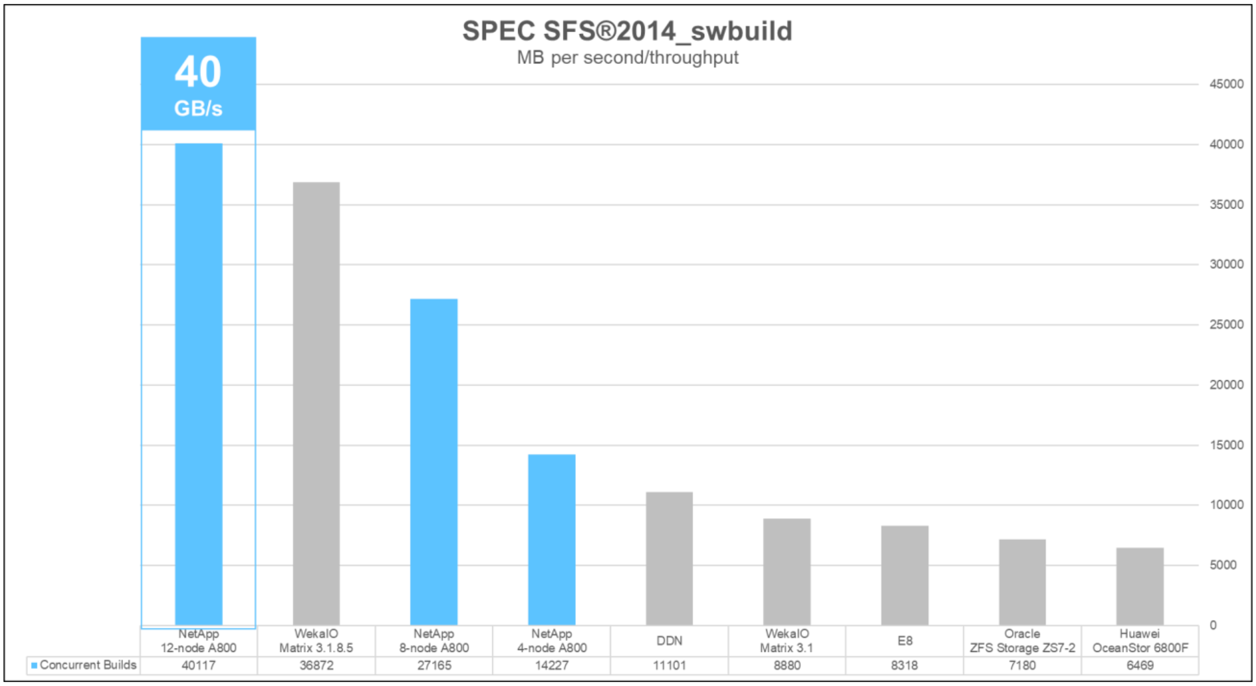


Figure 13) SPEC SFS 2014_swbuild—megabyte per second comparison.



AFF A700 testing

In a simple workload benchmark using a software build tool (Git), a Linux kernel was compiled on a two-node A700 cluster.

The following configuration was used:

- Two-node AFF A700 cluster
- A single aggregate of 800GB SSDs per node
- FlexVol volume: single node, 100% local
- FlexGroup volume: spans high-availability (HA) pair, eight members per node (16 members total)

The workload was as follows:

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

Figure 14 compares the maximum achieved throughput (read + write) on Git clone operations on a single FlexVol volume versus a single FlexGroup volume that spanned two nodes.

Note: The maximum throughput of the FlexGroup volume reaches nearly five times the amount of the FlexVol volume without the same degradation of the FlexVol volume as the workload reaches 64 threads.

Figure 14) FlexVol volume versus FlexGroup volume—maximum throughput trends under increasing workload.

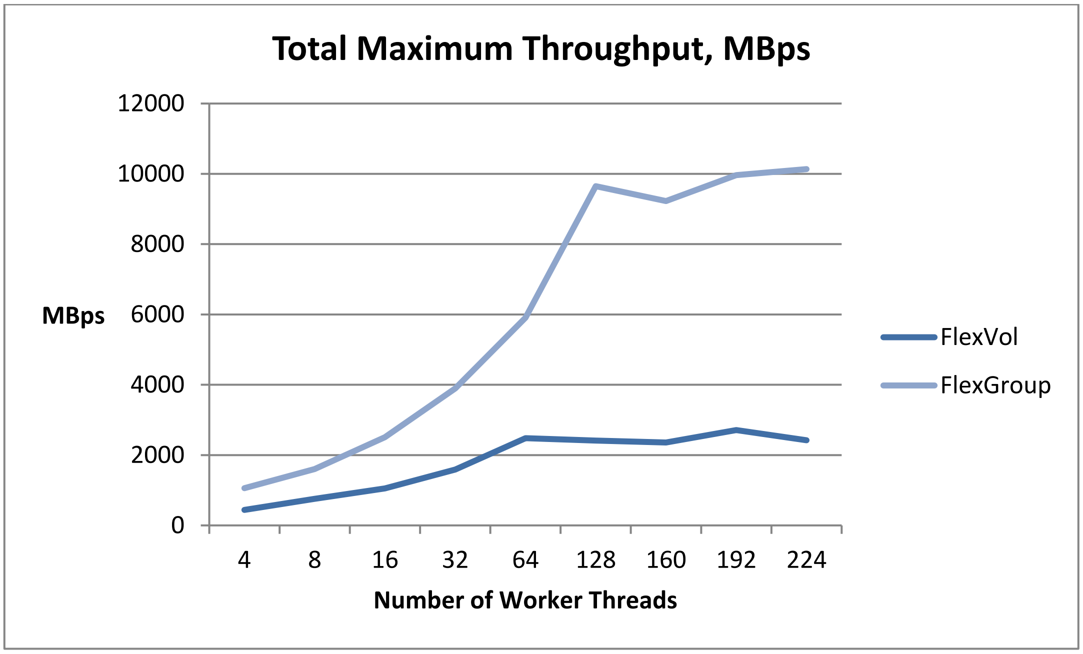


Figure 15 compares a FlexVol volume and a FlexGroup volume in the same configurations. This time, we break down the maximum read and write throughput individually, as well as comparing that against the average throughput for the FlexVol volume and the FlexGroup volume.

Figure 15) FlexVol volume versus FlexGroup volume—maximum throughput trends under increasing workload, detailed.

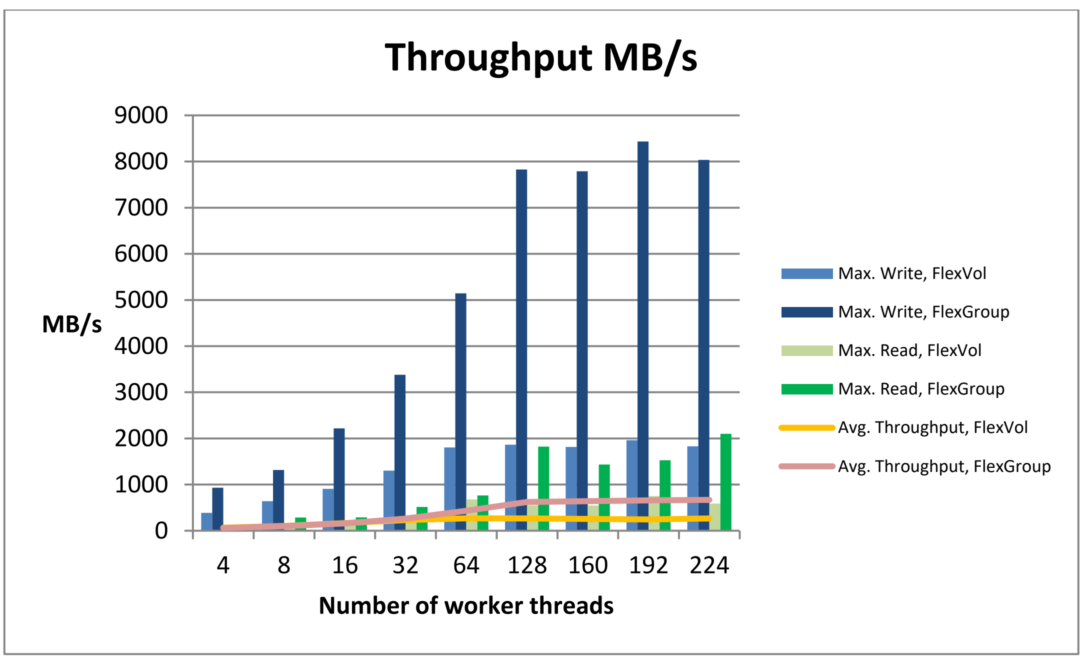
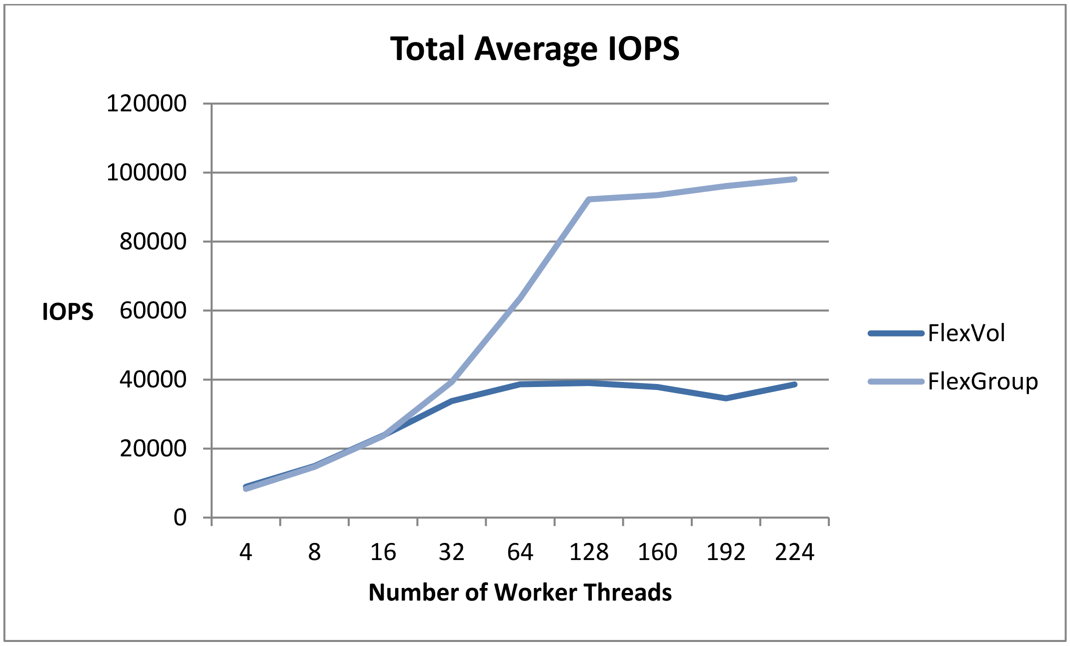


Figure 16 shows the maximum total average IOPS for a FlexGroup volume versus a FlexVol volume on the AFF A700. Again, note the dramatic increase of IOPS for the FlexGroup volume versus the degradation of IOPS at 64 threads for the FlexVol volume.

Figure 16) FlexVol volume versus FlexGroup volume—maximum average total IOPS.



ONTAP 9.4 and 9.5 performance testing

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

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 substantial performance improvements in ONTAP. These improvements can be accomplished with a nondisruptive upgrade.

Figure 17) Standard NAS benchmark (EDA)—ONTAP 9.5: one node versus two nodes (operations/sec).

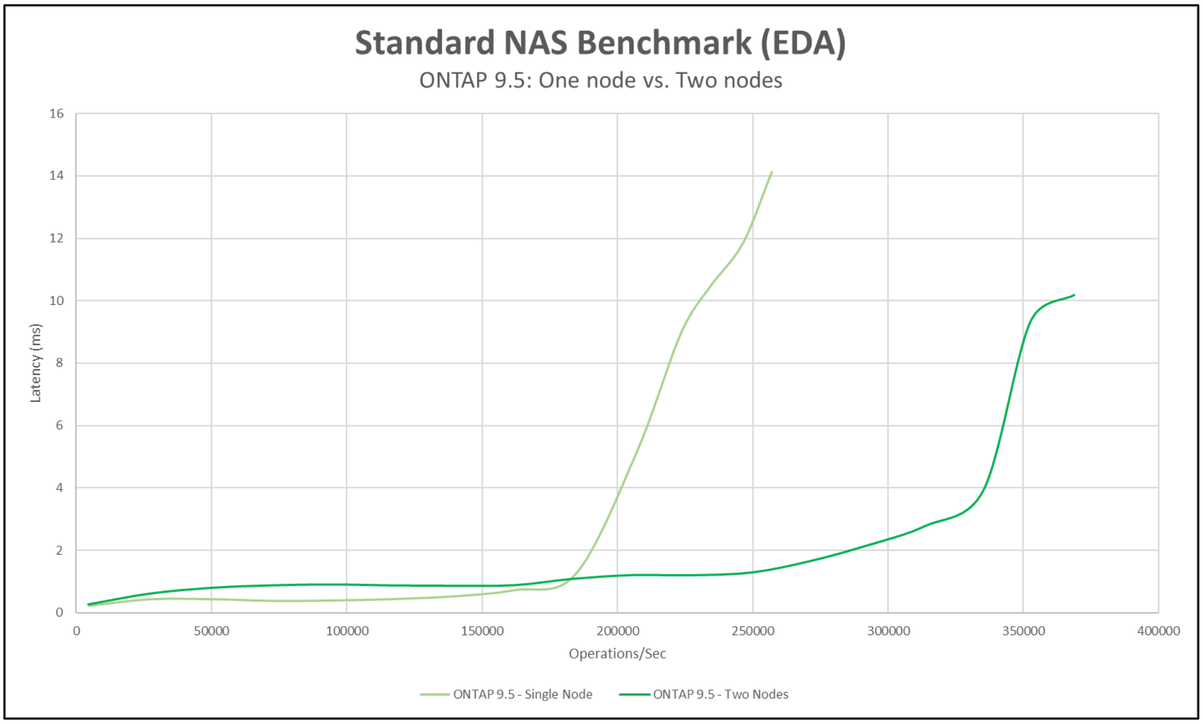


Figure 18) Standard NAS benchmark (EDA)—ONTAP 9.4 versus ONTAP 9.5 (operations/sec).

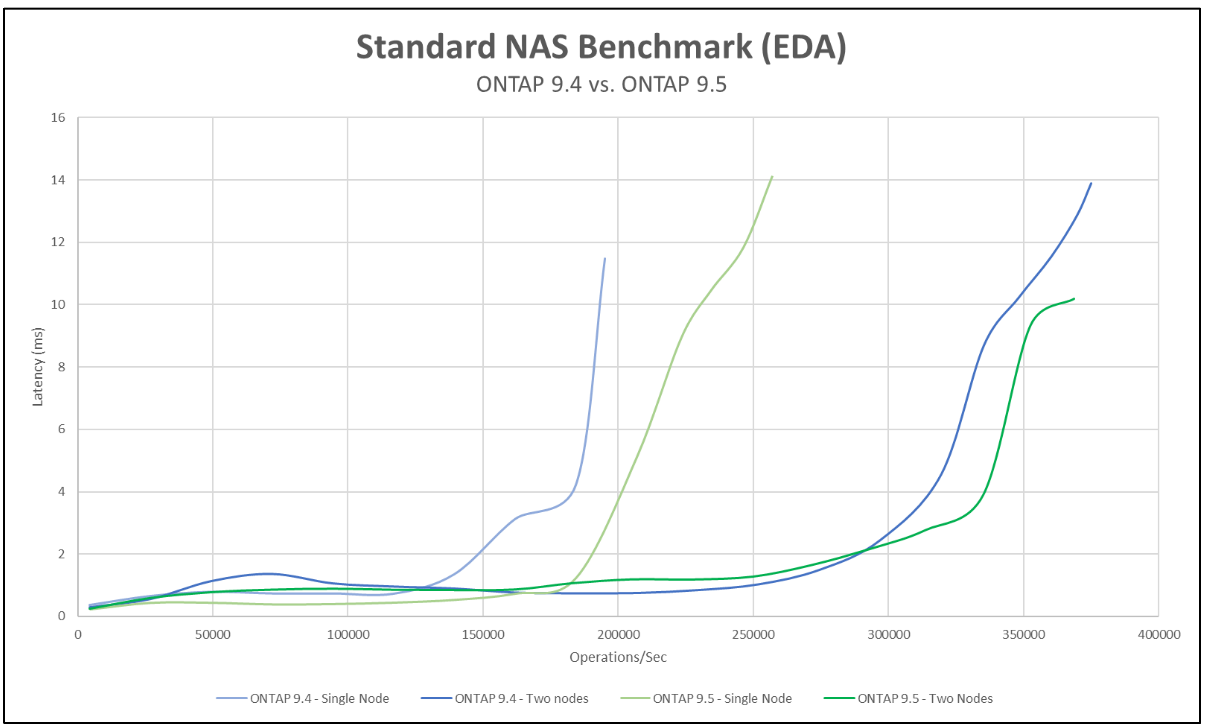


Figure 19) Standard NAS benchmark (EDA)—ONTAP 9.5: one node versus two nodes (MBps).

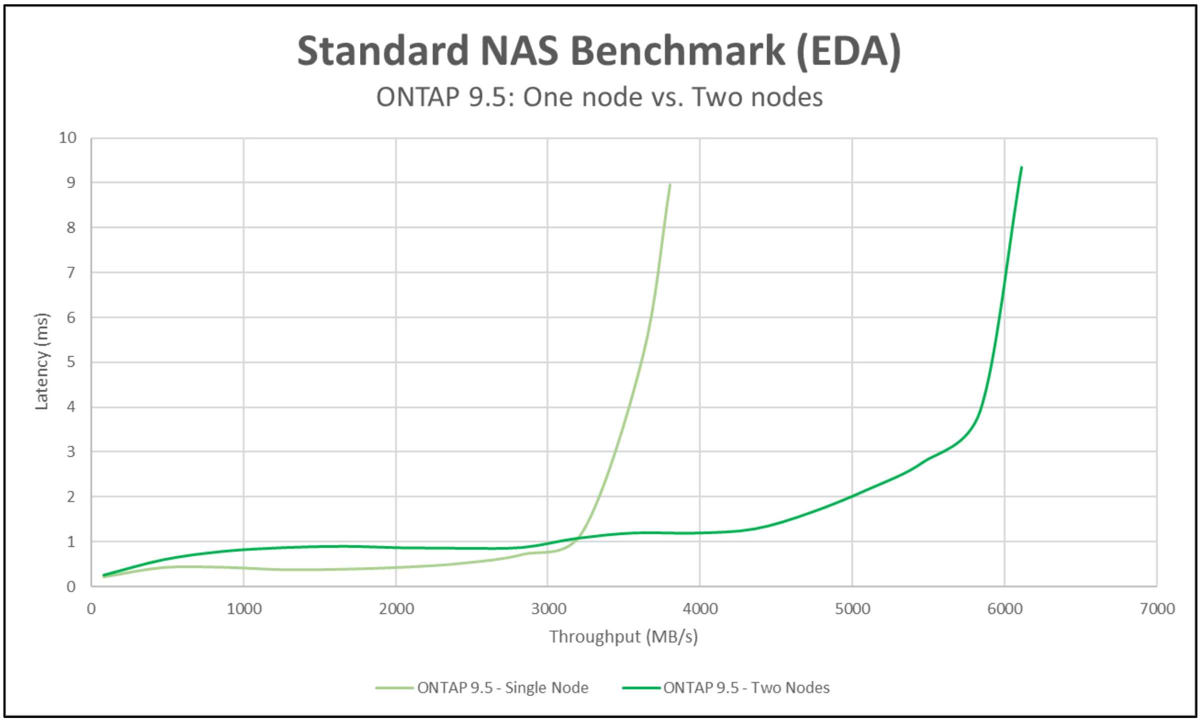


Figure 20) Standard NAS benchmark (EDA)—ONTAP 9.4 versus ONTAP 9.5 (MBps).

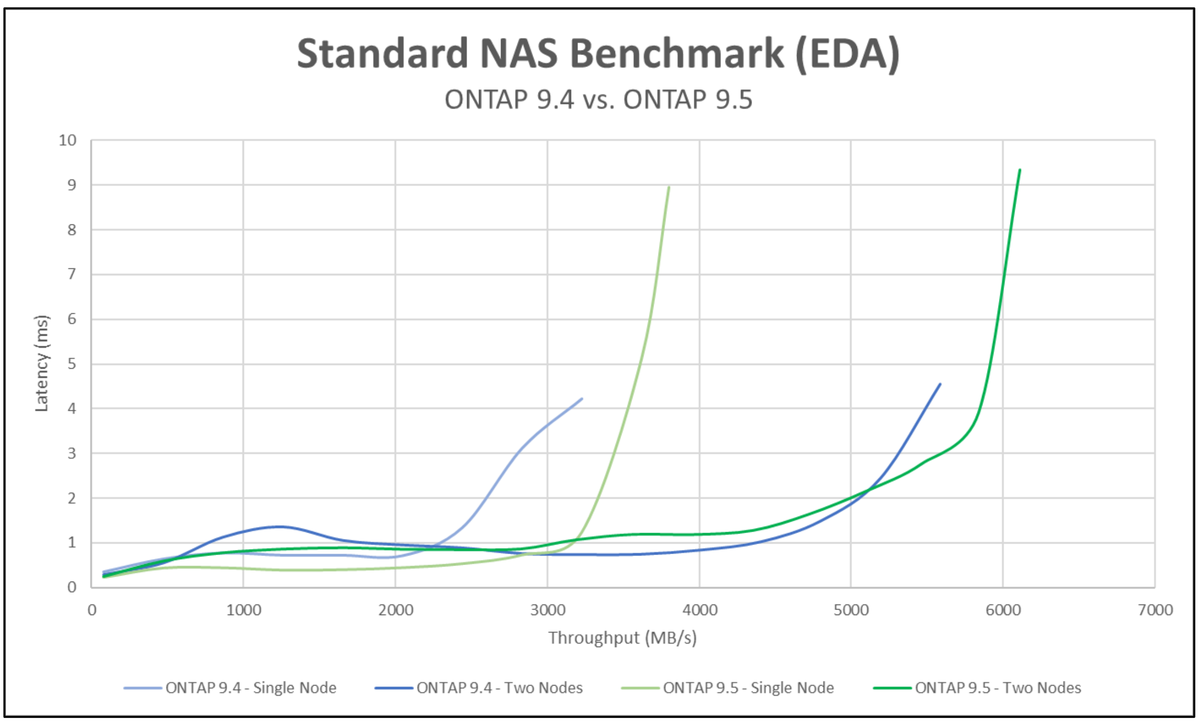


Figure 21 shows the performance for a standard NAS benchmark running a software build workload (such as Git or Perforce). Both types of workloads are ideal for FlexGroup volumes because of the high file ingest rates and need for parallel processing of write metadata.

Figure 21) Standard NAS benchmark (software builds)—ONTAP 9.5 (operations/sec).

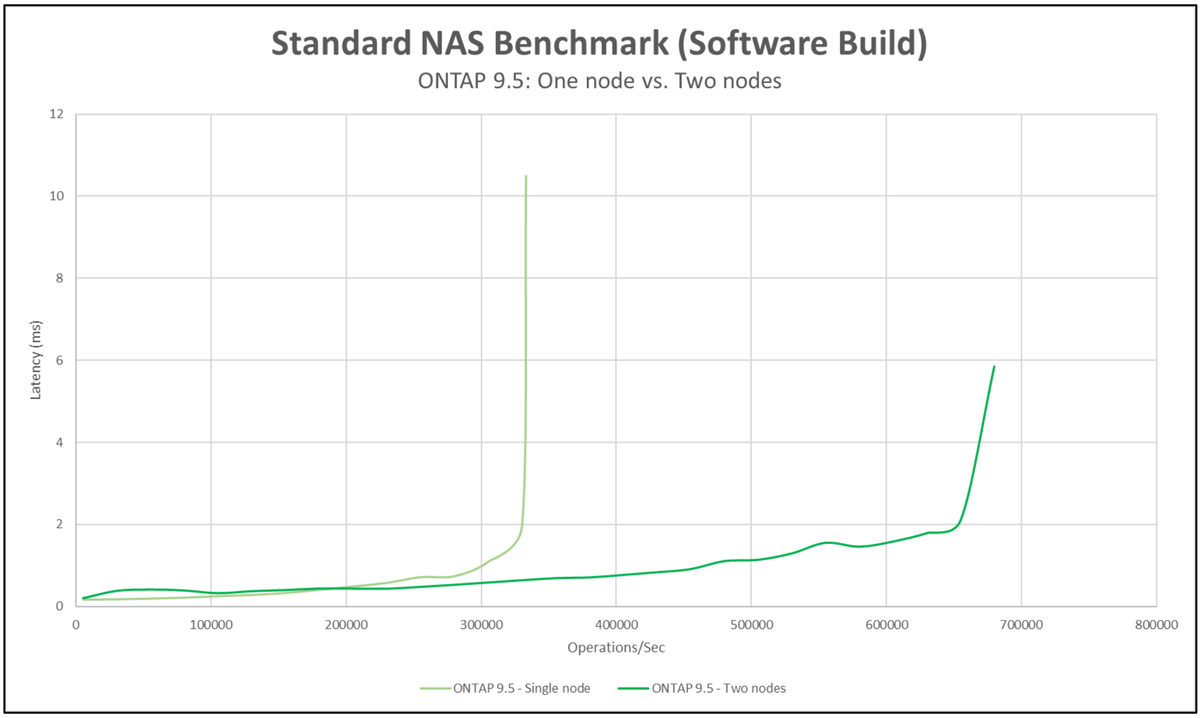


Figure 22) Standard NAS benchmark (software builds)—ONTAP 9.4 versus ONTAP 9.5 (operations/sec).

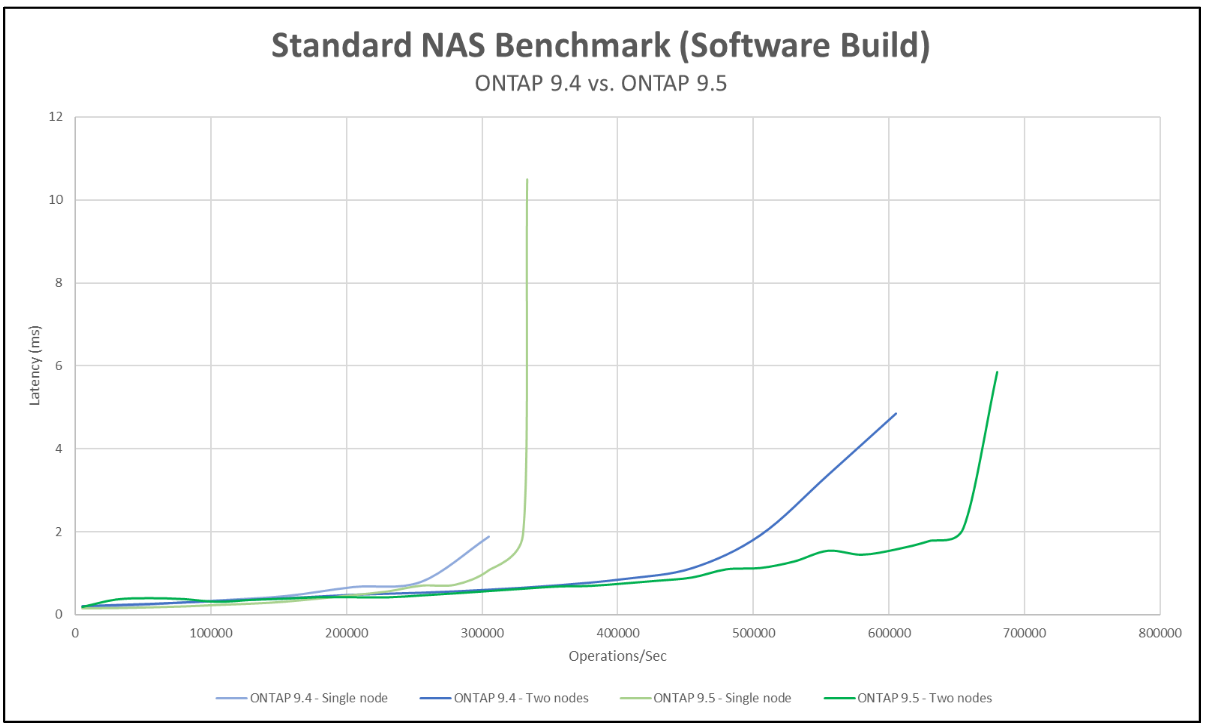


Figure 23) Standard NAS benchmark (software builds)—ONTAP 9.5 (MBps).

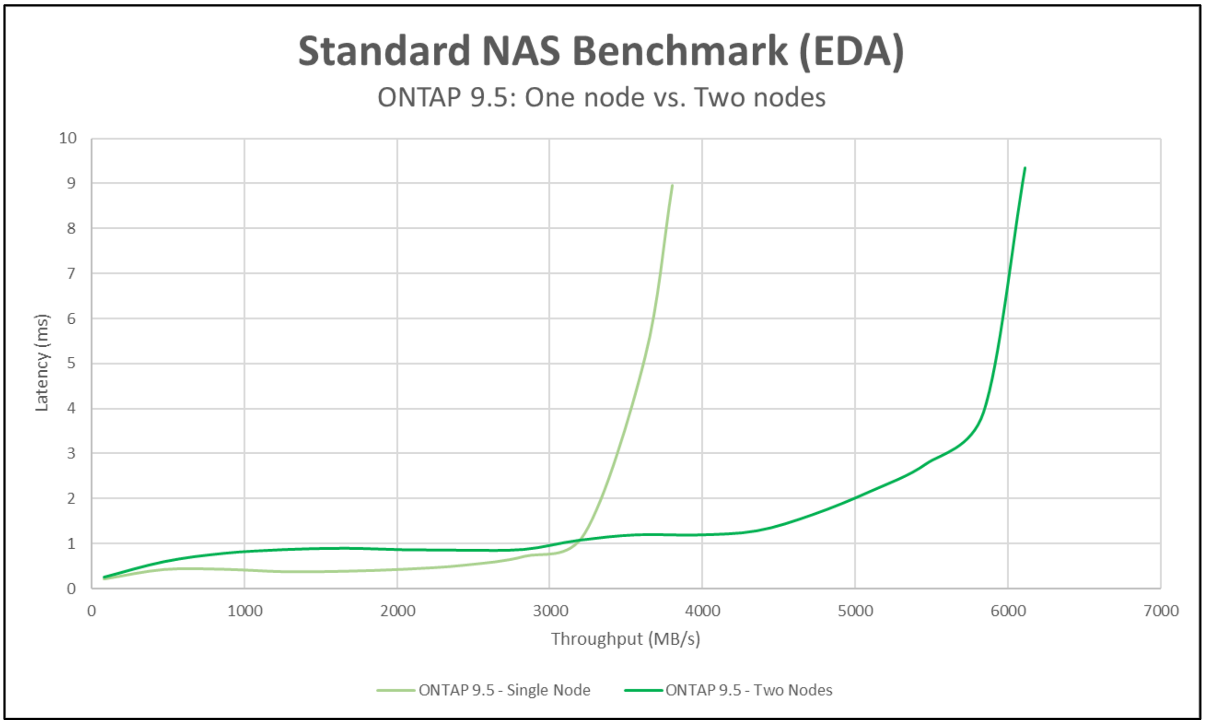
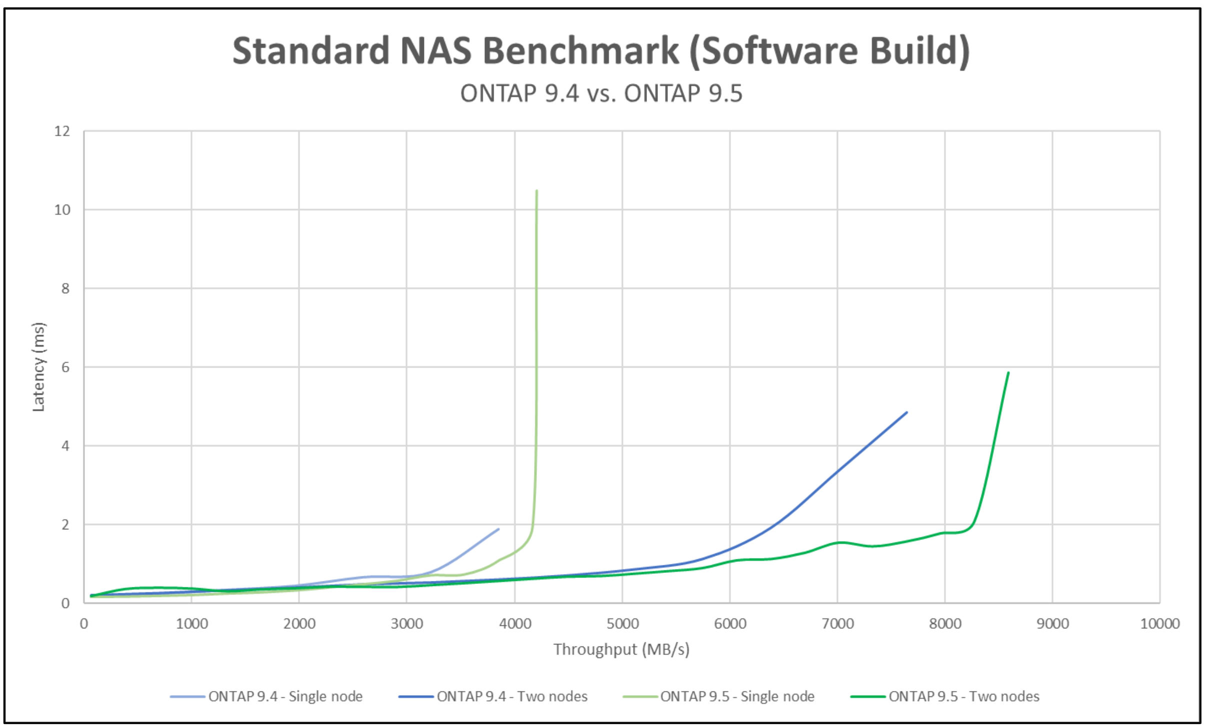


Figure 24) Standard NAS benchmark (software builds)—ONTAP 9.4 versus ONTAP 9.5 (MBps).



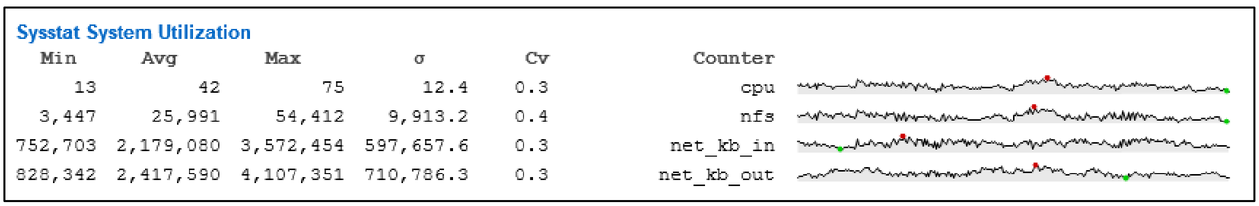
FlexGroup performance with big data workloads

Due to the FlexGroup volume's capacity and ability to scale a single namespace across multiple compute nodes in a cluster, it provides an interesting use case for big data workloads, such as Apache Hadoop, Splunk, and Apache Spark. These applications generally expect only one or two directories to dump large amounts of data and high file counts, requiring high throughput at a low latency. FlexVol volumes were able to accomplish this performance, but not without some tweaks to the application to make it aware of multiple volumes. TR-4570 describes an example of using Apache Spark on FlexGroup volumes.

Also, the NetApp Customer Proof-of-Concept (CPOC) lab conducted some performance testing using the TeraSort benchmark, which is used to test Apache big data workloads. In this testing, a two-node AFF A700 cluster running ONTAP 9.2 was used to push a maximum of 8GBps in and out of the cluster at an average read latency from ~3ms to 5ms and an average write latency from ~4ms to 8ms, while keeping the average CPU utilization around 55% on both nodes. Using a FlexGroup volume with big data workloads allows all available hardware to be used and provides a way to nondisruptively scale the capacity and performance by adding nodes for the workload as needed.

Figure 25) TeraSort benchmark statistics summary on a FlexGroup volume.

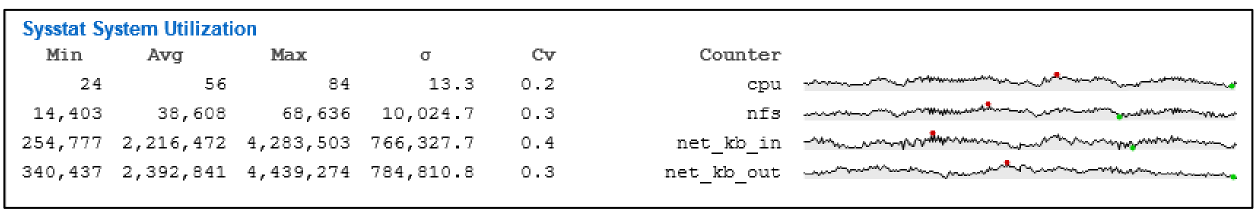
Node 1



Sysstat System Utilization

Min	Avg	Max	σ	Cv	Counter
13	42	75	12.4	0.3	cpu
3,447	25,991	54,412	9,913.2	0.4	nfs
752,703	2,179,080	3,572,454	597,657.6	0.3	net_kb_in
828,342	2,417,590	4,107,351	710,786.3	0.3	net_kb_out

Node 2



Sysstat System Utilization

Min	Avg	Max	σ	Cv	Counter
24	56	84	13.3	0.2	cpu
14,403	38,608	68,636	10,024.7	0.3	nfs
254,777	2,216,472	4,283,503	766,327.7	0.4	net_kb_in
340,437	2,392,841	4,439,274	784,810.8	0.3	net_kb_out

As a bonus, big data workloads running on ONTAP FlexGroup volumes have shown a space savings of nearly 50% with storage efficiency features such as inline aggregate deduplication, inline data compaction, and inline compression.

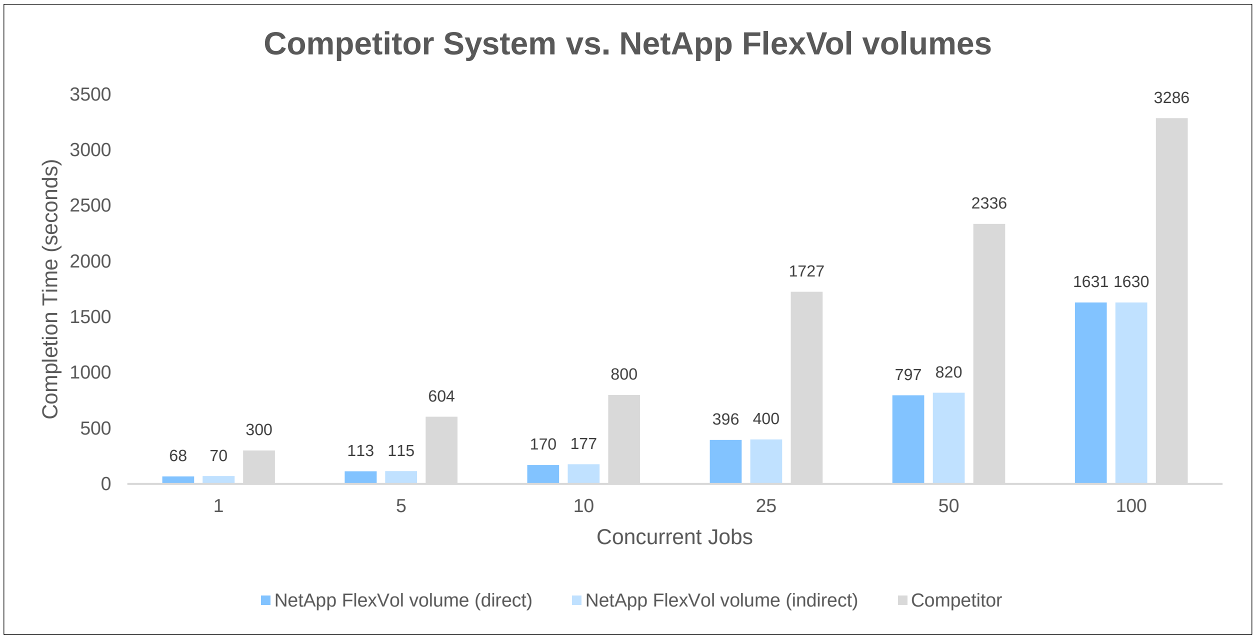
EDA workload comparison – ONTAP versus competitor

The best way to show that a system can handle your workload is by testing that workload. However, not everyone is able to run tests. In this section, we've collected real-world EDA testing data using specific tests from a chip-design manufacturer.

The first set of tests uses a kernel extract to compare a competitor system using twenty-one nodes with a NetApp AFF system using FlexVol volumes. In one test, the FlexVol volume used indirect access in which clients were attached to a network interface on a separate node than where the volume lived. In another test, the FlexVol used direct access for which the network connection was on the same node as the volume. Then, the number of concurrent jobs was increased incrementally.

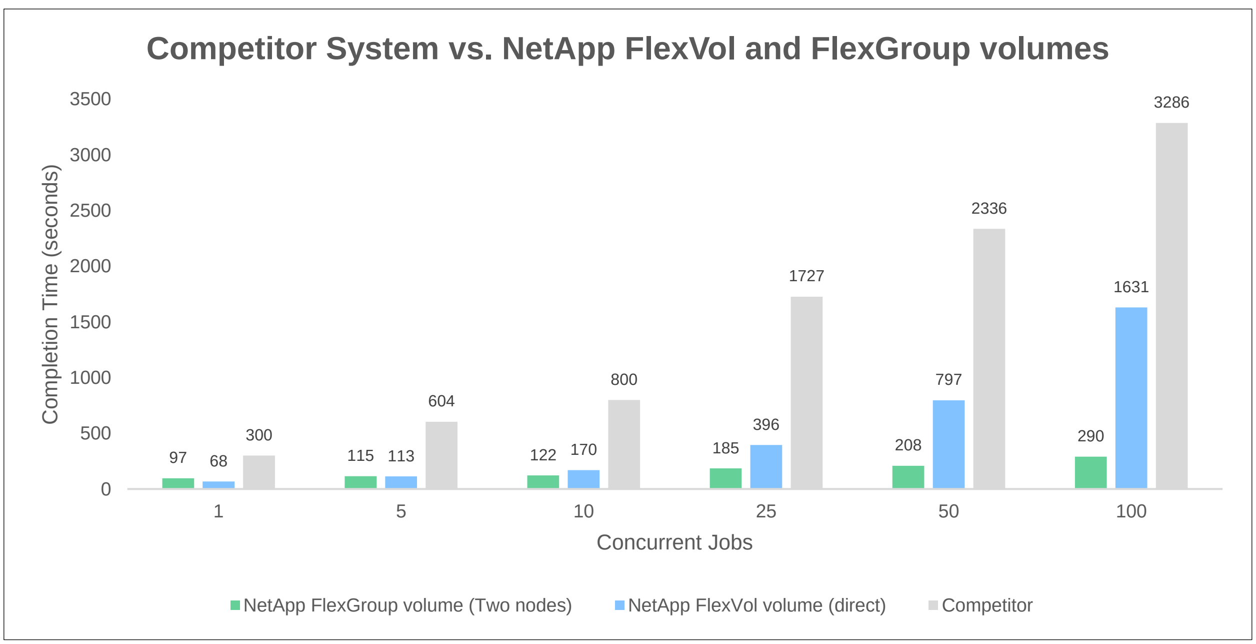
In all cases, the NetApp AFF system using FlexVol volumes exceeded the performance of the competitor system. In most of the tests, the local FlexVol volume performed slightly better than the indirectly accessed FlexVol volume, but not considerably so.

Figure 26) Kernel extract — Competitor vs. NetApp FlexVol volumes.



The same tests were used to compare the competitor system and FlexVol performance with NetApp FlexGroup volumes, which can use more hardware resources for these high metadata workloads. The FlexGroup volumes scaled across two of the nodes in the sixteen-node cluster. As we can see in Figure 27, the FlexGroup volume greatly outperformed even the locally accessed FlexVol volume on the same system – especially as the concurrent jobs increased.

Figure 27) Kernel extract — Competitor vs. NetApp FlexVol volumes and NetApp FlexGroup volumes.



The kernel extract test was also used to show the benefits of scaling a NetApp FlexGroup volume across more nodes in a single cluster. Figure 28 shows that the NetApp FlexGroup volume outperformed the competitor system's twenty-one node cluster with a fraction of the hardware needed, and it performed even better as the number of concurrent jobs scaled to the point where a single node’s resources would start to be exhausted.

Figure 28) Kernel extract — Competitor vs. NetApp FlexGroup volumes: scale-out.

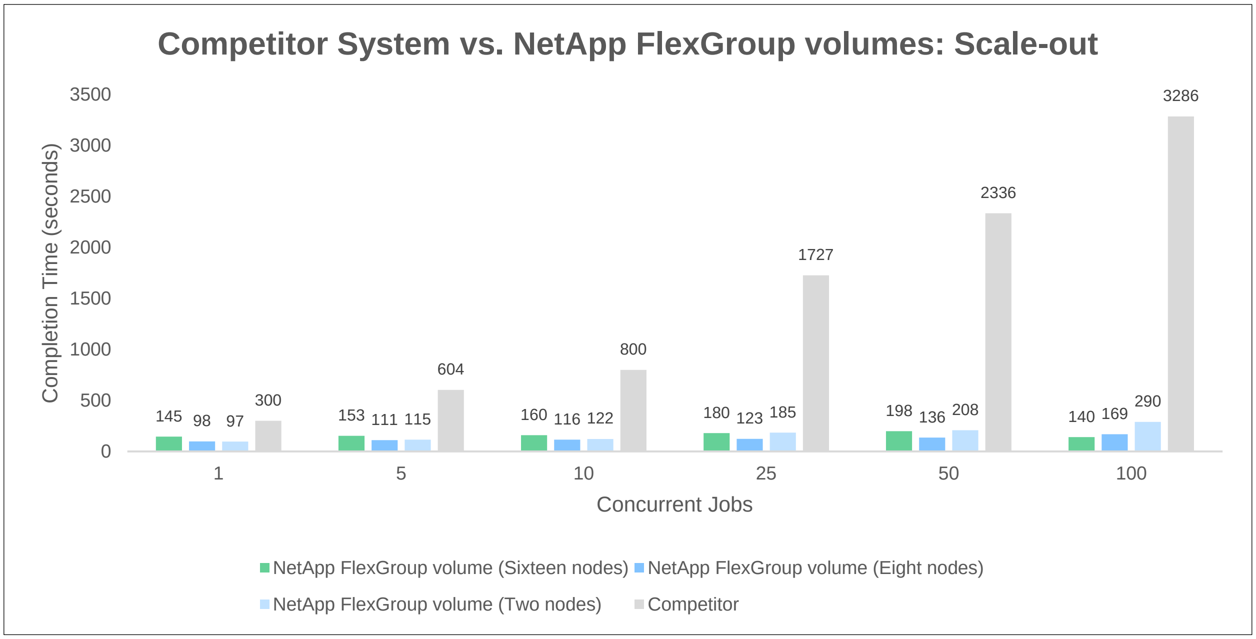
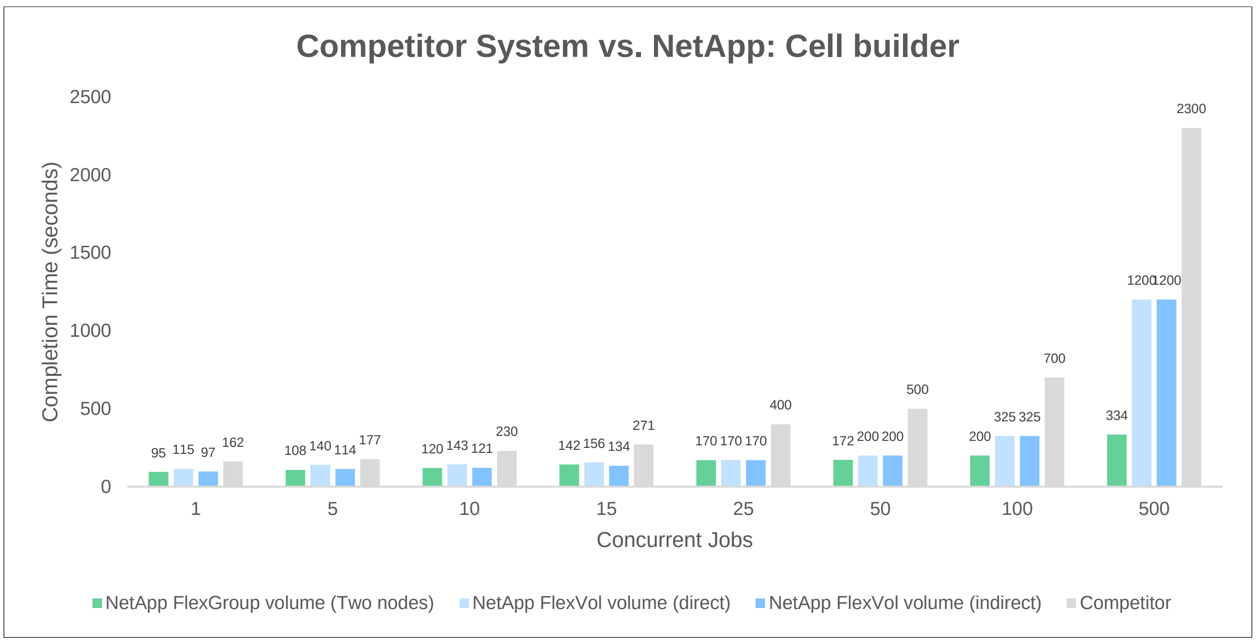


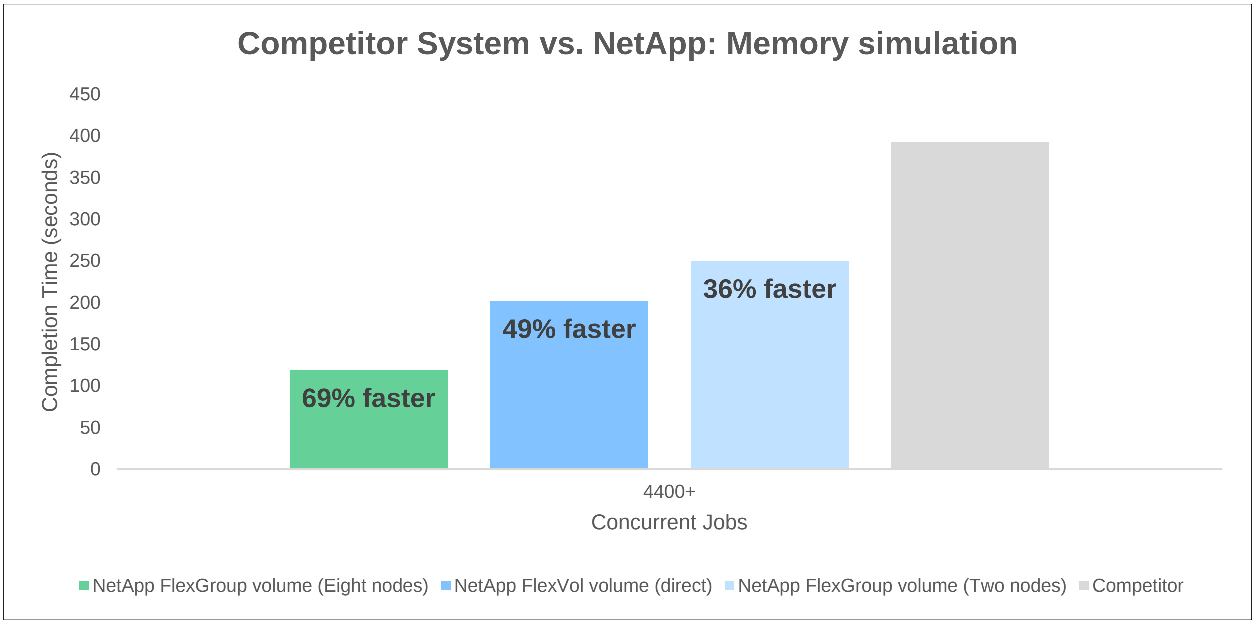
Figure 29 shows completion times for a cell builder workload, again comparing FlexVol, FlexGroup, and a competitor system. Note the lower completion times for the NetApp FlexGroup volume.

Figure 29) EDA workload — cell builder.



Memory simulation and validation tests are also a common EDA workload. Figure 30 shows the average run time for 4,400+ concurrent jobs. A FlexGroup volume across eight nodes was 69% faster than the competitor system across twenty-one nodes for this test.

Figure 30) EDA workload — memory simulation and validation.



Standard EDA benchmark — customer test

An additional set of tests using a standard EDA benchmark was run by the same customer comparing NetApp ONTAP 9.7 with FlexGroup volumes to the all-flash competitor system. The following table shows the configuration comparison and maximum results from the test.

Table 8) NetApp FlexGroup volumes versus competitor system — EDA benchmark.

Competitor test information	NetApp ONTAP test information
• Fourteen all-flash nodes • 624 concurrent jobs • ~6.8ms latency • 259,664 Achieved Ops • 4.36GBps	• Eight AFF A800 nodes • 2000 concurrent jobs • ~2.6ms latency • 897,241 Achieved Ops • 15.6GBps

In these tests, we observed lower overall latency, higher IOPS, and better overall throughput, which results in faster job completion times for EDA workloads.

Figure 31) Average latency versus achieved ops — EDA benchmark, customer test.

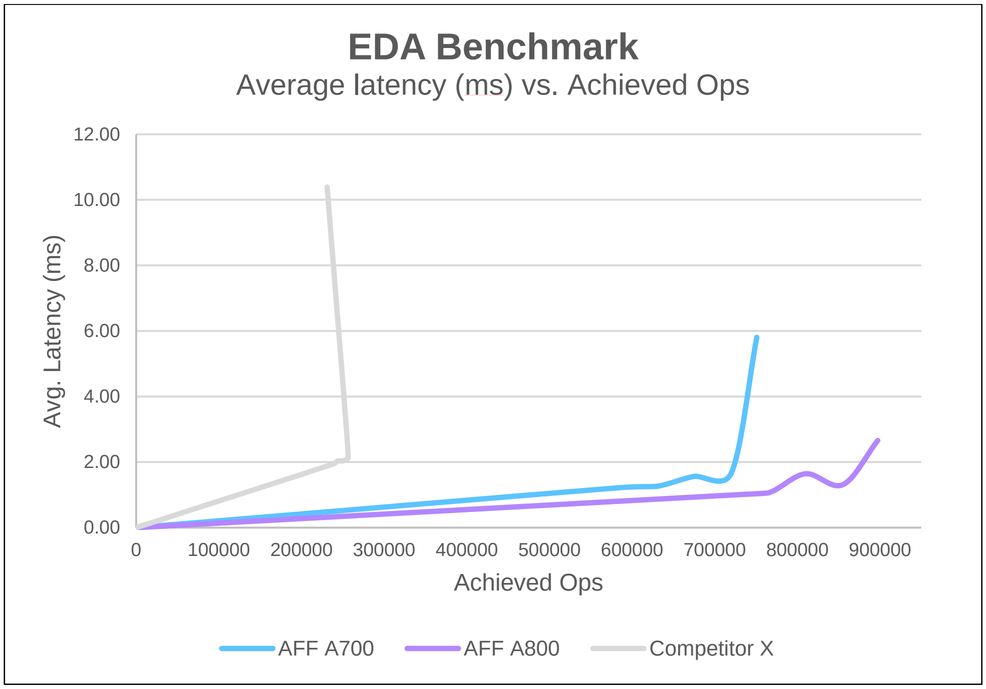
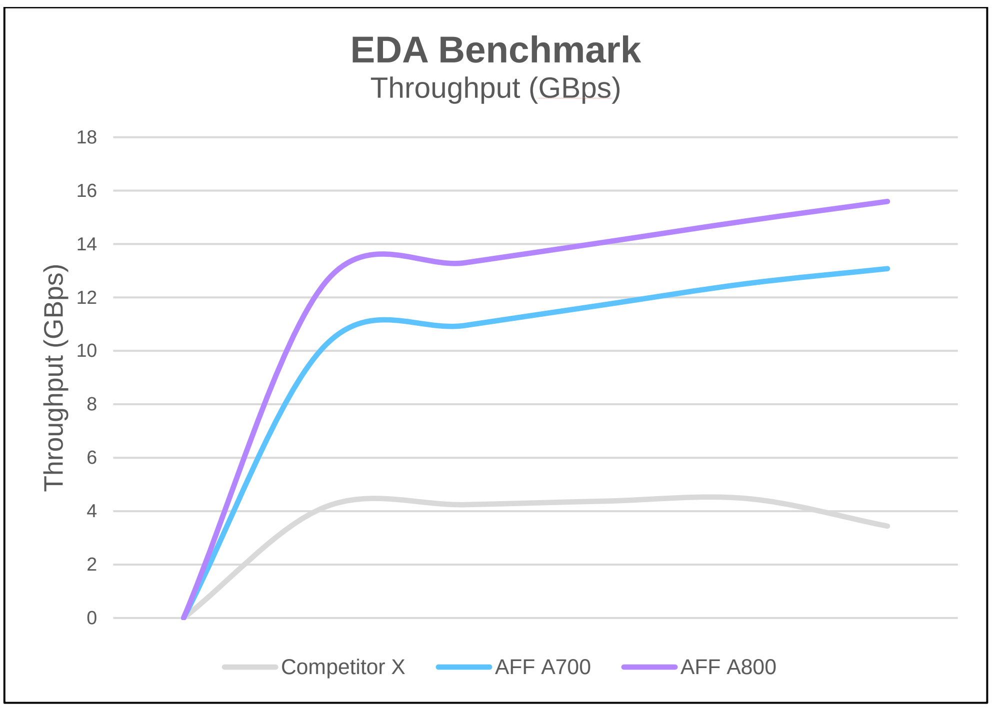


Figure 32) Throughput (GBps) — EDA benchmark, customer test.



Automatic workload adaptation

The FlexGroup volume continually adapts to the current conditions in the cluster, changing behavior constantly to keep usage evenly consumed and to keep dynamic load evenly balanced. Trade-offs are implicit in this continual balancing act. The cost of this automatic balancing is that a FlexGroup volume cannot attain the same theoretical maximum performance that a perfectly balanced and manually organized collection of FlexVol volumes could otherwise attain. However, the FlexGroup volume can get very close to that maximum, and it requires no foreknowledge of the workload to accomplish its work. In addition, a FlexGroup volume adds a simplicity aspect to large data layouts that a multiple FlexVol architecture cannot.

FlexGroup volumes perform better—balancing load and usage more smoothly—when faced with a broad variety of workloads and high data-creation rates. Thus, a single FlexGroup volume that performs many different roles can be a more effective use of your cluster’s resources than if you use different FlexGroup volumes for different workloads. You can, however, junction multiple FlexVol volumes and FlexGroup volumes together in the same ONTAP SVM if you require greater control and flexibility over your data.

If a workload is creating a high number of small files, then the FlexGroup volume places those files to balance them evenly across volumes while favoring folder locality to increase performance. If the workload is a smaller number of large files, then ONTAP recognizes that difference. Rather than favoring local folder placement (which could result in multiple large files ending up on the same member volume and creating an artificial imbalance of data), ONTAP instead places files in a more round-robin fashion to ensure even space allocation. This allows for a wider variety of workloads to perform optimally on FlexGroup volumes, preventing space imbalance scenarios and reducing the need for administrator intervention.

ONTAP 9.8 introduces a change to how capacity is managed called Proactive resizing. This change effectively maintains a free space buffer across all member volumes when a capacity threshold is reached to automatically to help protect against member volumes getting too full and guards against volumes having disparate free space. Additionally, before ONTAP 9.8, if a FlexGroup member volume reaches a 90% capacity threshold, performance of new file creations suffers, as ONTAP starts to create more remote hard links for new files. ONTAP 9.8 removes that 90% threshold since proactive resizing maintains enough free space to avoid the need to redirect traffic.

Ingest algorithm improvements

Every ONTAP release further improves the ingest algorithms for FlexGroup volumes that help ONTAP make better decisions about how new data is placed in FlexGroup volumes. The algorithms also improve the way FlexGroup volumes respond when member volumes approach “nearly full” status.

Best practice 1: Always run the latest ONTAP version

NetApp strongly recommends that you run the latest patched ONTAP version when using FlexGroup volumes for the best ingest results. You can download the latest release at NetApp Support for ONTAP 9.

Some of the changes to ingest that have taken place in various releases include:

- Inode counts factoring into placement in ONTAP 9.3
- SMB workload placement enhancements and NetApp FabricPool considerations in ONTAP 9.5
- Better handling of mixed workload types in ONTAP 9.7
- Proactive resizing and adjustment of remote placement triggers in ONTAP 9.8
- ONTAP 9.10.1 and later offer a way to adjust ingest of files and folders on a per-FlexGroup volume basis based on workload types. It is best to leave the default in place (ONTAP automatically adapts and adjusts ingest of data), but in cases in which you want more granular control over how a FlexGroup volume decides to place data, contact NetApp support.

Performance features

ONTAP provides various features to better control and monitor performance.

Quality of service (QoS)

Starting in ONTAP 9.3, you can apply maximum storage QoS policies to help prevent a FlexGroup volume's workload from overrunning other volume workloads. ONTAP storage QoS can help you manage risks around meeting your performance objectives.

ONTAP 9.4 added support to FlexGroup volumes for QoS minimums (also referred to as guarantees or floors), which provide a set threshold of performance that is allocated to a specified object.

You use storage QoS to limit the throughput to workloads, provide guaranteed performance to workloads, and to monitor workload performance. You can reactively limit workloads to address performance problems, and you can proactively manage workload performance to prevent problems.

How storage QoS policies work with FlexGroup

With FlexGroup, storage QoS policies are applied to the entire FlexGroup volume – not at the member volume level. 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 and Figure 34 show how storage QoS gets applied to a FlexGroup volume.

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

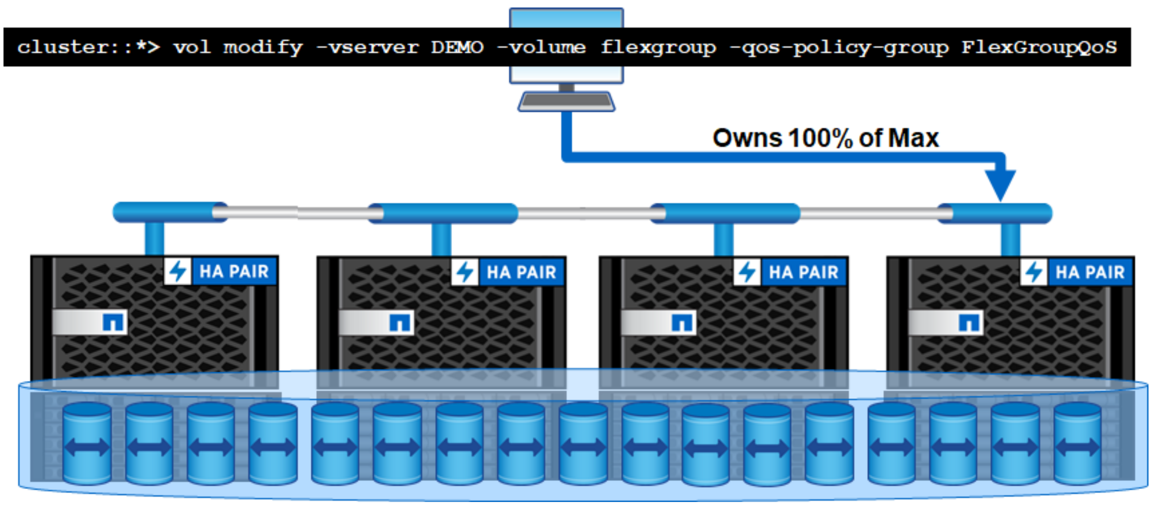
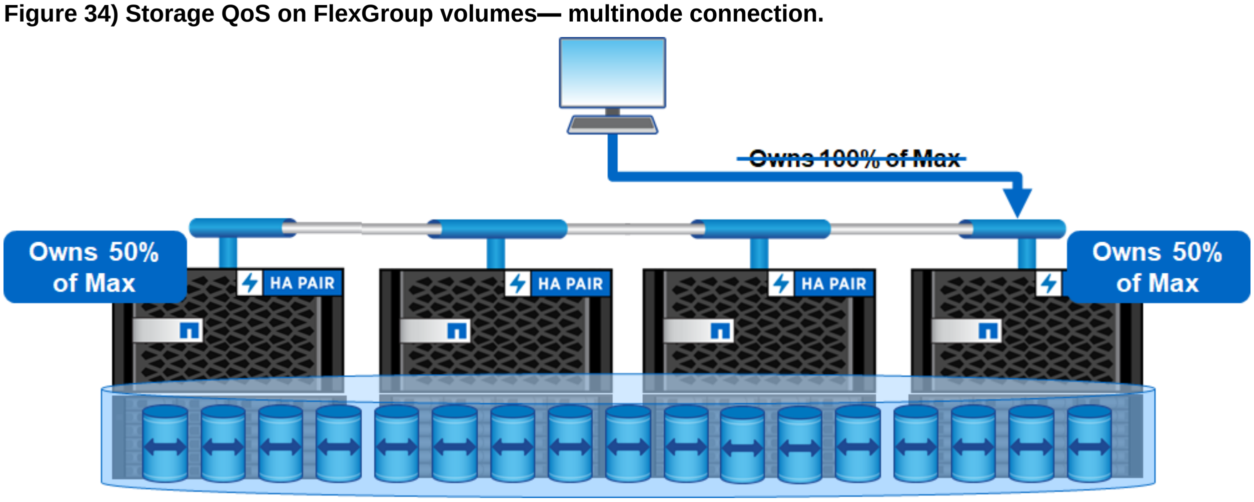


Figure 34) Storage QoS on FlexGroup volumes— multinode connection.

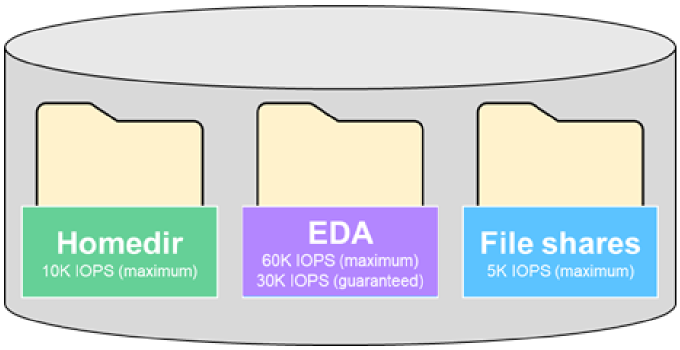


Note: Nested policies are currently not supported with FlexGroup volumes.

Qtree QoS

ONTAP 9.8 introduces the ability to apply QoS policies at the qtree level.

Figure 35) Qtree QoS use cases.



This means you can provision a FlexGroup volume and manage performance in that volume with qtrees, rather than creating multiple FlexVol or FlexGroup volumes to divide that workload up.

Qtree QoS also provides a more granular level of statistics for the qtree than previous Qtree Statistics offered.

Qtree QoS in ONTAP 9.8 can be used with FlexGroup volumes and FlexVol volumes, but it has the following limitations:

- NFS only
- CLI/REST API only; no current GUI support
- No adaptive QoS support

Qtree QoS also provides some enhanced statistics for performance monitoring, which aids in understanding specific workloads.

```
Policy Group                  IOPS       Throughput     Latency
-------------------- --------- --------------- ----------
qtree                          113       113.00MB/s      2.82ms
```

Adaptive QoS

Note: ONTAP 9.4 introduced adaptive QoS support for FlexGroup volumes, which allows ONTAP to adjust the IOPS and TB values of a QoS policy as the volume capacity is adjusted. Adaptive QoS is not supported with qtree QoS because qtrees are not objects you can grow or shrink.

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

                                                 NFS CIFS Internal *Total
                Qtree Vserver          Volume    Ops  Ops      Ops    Ops
--------------------- -------- --------------- ----- ---- -------- ------
```



```
DEMO:flexgroup_local/     DEMO flexgroup_local 22396     0          0  22396
DEMO:flexgroup_local/qtree
                          DEMO flexgroup_local     0     0          0
```

Workloads and behaviors

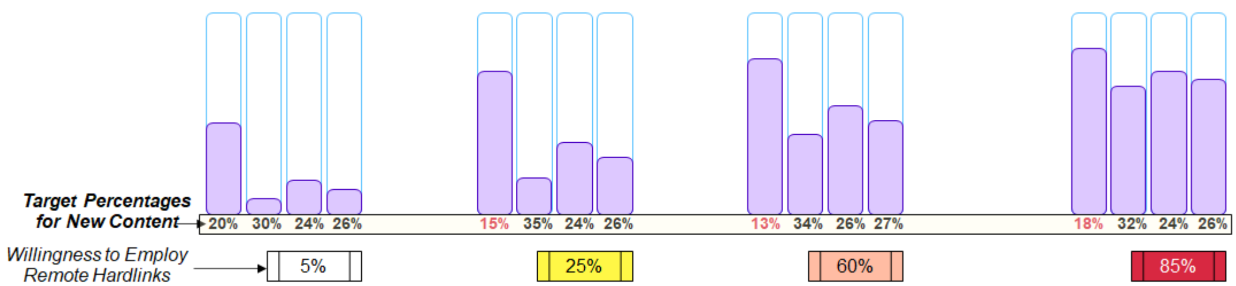
In an optimally balanced FlexGroup volume, all constituents have roughly the same amount of data and load, and the volume can maintain that state while using a high frequency of local placement for best performance. A workload with a good balance of folders and similarly sized files would be able to maintain local parent folder placement while also keeping a relatively even balance of capacity.

A less optimal FlexGroup volume might have some constituents that hold more or less data than their peers, or that are receiving much more or much less traffic. Workloads that have only a few folders with many files per folder, or workloads with highly variant file sizes can experience data usage imbalances in a FlexGroup.

Capacity balance, however, is not the most important function of a FlexGroup volume. Instead, a FlexGroup volume functions best when there is a mix of local placement for performance along with capacity and inode count balance. We don't want to sacrifice performance for the sake of perfectly balanced capacities across member volumes.

In the figures below, you can see several examples of capacity imbalances in a FlexGroup volume, and the varying degrees of remote placement based on the overall fullness of the FlexGroup volume. For example, a relatively empty FlexGroup with data imbalances is less likely to use remote hard links than a FlexGroup that is much closer to being full.

Figure 36) Capacity imbalance and likelihood of remote placement.



ONTAP constantly monitors the ongoing state of the member volumes and adjust placement decisions based on the current state of the FlexGroup volume. If one member volume is a little out of balance from the others, then it's likely that no adjustments will be made to ingest. But if that member volume starts to approach 90% capacity, or has a capacity discrepancy that exceeds 512GB, then ONTAP makes more aggressive placement choices for new data to correct the imbalance. This adjustment means more remote file placement to other member volumes, which can have a negative (but potentially unnoticeable—roughly 5% to 10%) effect on FlexGroup volume performance.

In some cases, a FlexGroup volume might appear to be perfectly balanced in usage and load, but it has had to resort to more frequent remote placement frequently to maintain that state. This situation can occur when FlexGroup member volumes get closer to being 100% used.

Best practice 2: Stop worrying about capacity imbalances.

A FlexGroup with a capacity imbalance across member volumes is not in itself a problem and should not be treated as such. Instead, look at capacity imbalances as a potential cause if a FlexGroup is not performing as expected or if the capacity imbalances are so extreme that you're running out of space on your cluster. Be sure to engage NetApp support if you feel that your FlexGroup volume capacity imbalance is the source of a performance issue.

Workloads determine the degree to which a FlexGroup volume behaves optimally. Most workloads can be used with FlexGroup volumes in ONTAP 9.8, but some workloads (such as EDA/software development) perform more optimally than others.

Optimal workloads

A FlexGroup volume works optimally when it is under heavy ingest load—that is, when there is a high rate of file and directory creations. ONTAP makes its placement decisions as new files and directories are created, so the more often this action occurs, the more frequently ONTAP has an opportunity to correct existing imbalances in load and usage. If a workload is a heavy read or write-append to existing files, then the FlexGroup placement doesn't really factor in as much; once the files are placed, they remain where they landed. As mentioned in the section "Ingest algorithm improvements," each new ONTAP release adds improvements and adjustments to FlexGroup volumes that can address more variant workloads.

Generally speaking, the following represent attributes of the most optimal FlexGroup workloads.

- **FlexGroup volumes work best with numerous small subdirectories.** This means dozens to hundreds of files per directory, because they allow the FlexGroup volume to place new child subdirectories remotely while keeping individual files local to their parent directories for best performance. Directories containing more files experience more remote placement to other member volumes in an attempt to balance capacity and file counts.
- **A FlexGroup volume responds well to heavy concurrent traffic.** Bringing more workloads—especially traffic from multiple clients that are doing different things at the same time—to bear against a FlexGroup volume simultaneously can improve its overall performance. In other words, don't expect to push a FlexGroup volume to its limit and achieve the performance possibilities mentioned in this document with one to a few clients.
- **A FlexGroup volume works best when there is plenty of free space.** When constituents begin to fill up, the FlexGroup volume begins to employ remote placement more frequently so that no one constituent becomes full before its peers do. This increased usage of remote placement comes with a metadata performance penalty.
- **FlexGroup volumes work best with high rates of write metadata operations.** ONTAP FlexVol volumes already process read and write I/O in parallel, and metadata read operations (such as `GETATTR`). However, ONTAP processes write metadata (such as `SETATTR` and `CREATE`) serially, which can create bottlenecks on normal FlexVol volumes. FlexGroup volumes provide a parallel processing option for these types of workloads, which results in performance that is two to six times better for these types of workloads.

Performance and capacity considerations

For best performance, keep plenty of free space on the FlexGroup volume (at least 10% free space available) when it is under heavy load.

To better manage FlexGroup volume capacity, use ONTAP 9.8 or later. That release contains a number of features to manage capacity, including volume autogrow, elastic sizing and, specifically for ONTAP 9.8, proactive resizing.

Free space for FlexVol member or constituent volumes can be monitored at the admin privilege level with the following command:

```
cluster::> vol show -vserver SVM -volume [flexgroupname__]* -is-constituent true -fields
available,percent-used
```

Note: You can also monitor free space by using GUI utilities such as NetApp Active IQ Unified Manager and/or by configuring ONTAP to generate alerts. See the section "Capacity monitoring and alerting" for more details.

Good workloads

Even if a workload does not conform to the preceding parameters, odds are good that a FlexGroup volume can accommodate it with ease. Remember that the section “Optimal workloads” describes situations that can help a FlexGroup volume perform optimally, but even a suboptimal one provides good throughput, scaling, and load distribution for most use cases.

Nonideal workloads — large files

A few activities can make a FlexGroup volume work harder to maintain its balance of load and usage among constituents. Most of these activities relate to large files in one way or another. Although these workloads are able to use FlexGroup volumes, you should strive to understand the average file size and largest file size of the workload before implementing. The following considerations should be made when deploying workloads that don’t fit into the “ideal” or “good” workload definitions.

Consideration #1: ONTAP cannot predict the future size of your files

One of the key challenges of large file workloads is that storage systems are not aware of how large a file will become over time. Clients often do not have this information either; instead, a file starts out as a small inode in the storage system, and then data is written to it until the file creation/write is completed. This is exacerbated by the FlexGroup volume’s tendency to keep file placement local to the parent folder for performance considerations. It’s equally possible that a folder with 100 files that are 500MB will all land in the same member volume as it is that a folder with 100 files that are 4K in size, depending on how fast the files are written and how many clients are involved in creating the files. As a result, in that scenario, one member volume might end up with 50GB of used space and the other member volume might only have 400KB used.

As mentioned before, this isn’t inherently a problem, but it is a noticeable discrepancy to the storage administrator and can present problems if the FlexGroup volume isn’t sized appropriately. For example, what if the member volumes are all 100GB in size? Therefore, in this example, one member is 50% full, and the others have 0% capacity.

Generally speaking, workloads like this balance themselves out over time, and the FlexGroup volume maintains even distribution and good performance. The benchmark for concern with these workloads should not be “My member volume capacities are uneven,” but rather, “My FlexGroup volume is not performing as well as I expect.”

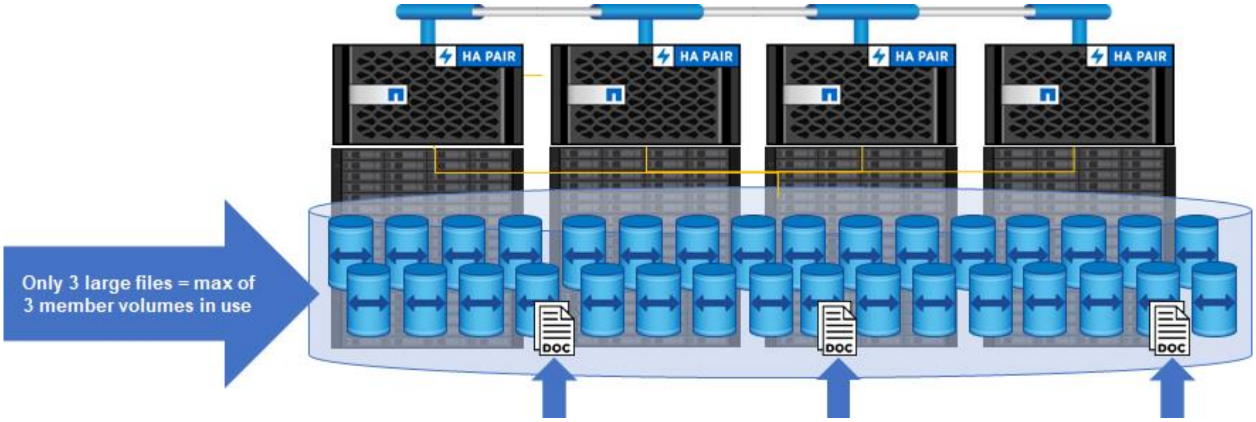
Consideration #2: Large-file workloads are generally low-file-count workloads

Large files are marginally more difficult for the FlexGroup volume to process than small files are, primarily because using large files typically means using fewer of them overall. As previously mentioned, the FlexGroup volume performs best when new files and directories are being created frequently. If the working set consists of many large files that are roughly the same size, the FlexGroup volume should not have trouble maintaining usage and load distribution among constituents. Performance with large file workloads act more like that of a FlexVol, as the benefits of parallel ingest don’t come into play with workloads that are not workloads that ingest many files at a time.

Consideration #3: Large-file workloads are not guaranteed to distribute evenly

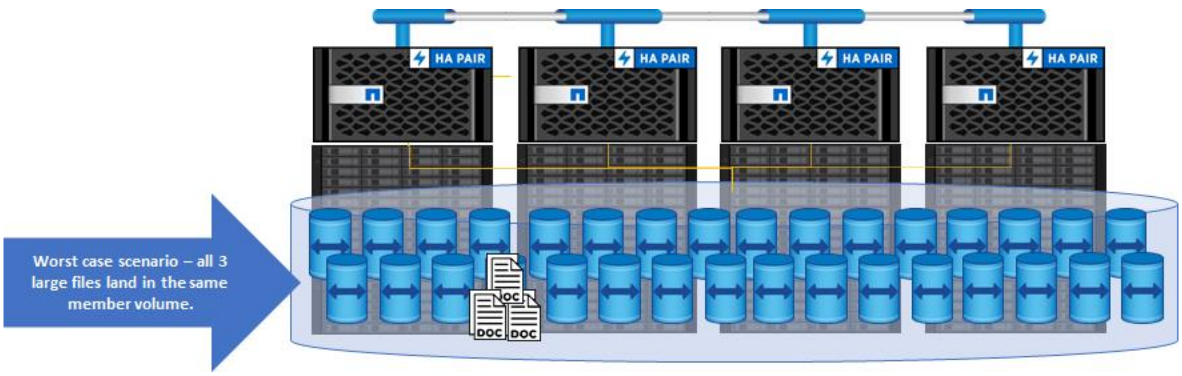
Large files also have the property of holding a great deal of information. Reading or writing that much information can take a long time. If the workload concentrates on only a few of those large files (say, reading or writing a large single-file database), then all that traffic is handled exclusively by the constituents that host those files. Because other constituents are not participating in the workload at the time, this situation can result in suboptimal usage of the FlexGroup volume. In general, expect roughly the same performance for large-file or streaming workloads in a FlexGroup volume as you would see in a FlexVol volume.

Figure 37) FlexGroup volume with a few large files; why usage can be suboptimal.



In addition, there is no guarantee that the large files are all distribute evenly across the FlexGroup volume. In the above graphic, the three large files landed in three different member volumes. This scenario might happen if we write one of those files at a time, and they finish writing the entire file before the next file is written. But if all three files are written at the same time, then we run the possibility that all three of the files land in the same member volume.

Figure 38) Potential worst case scenario for large files; all land in the same member volume.



Again, if there is enough capacity in the FlexGroup volume, it isn't necessarily a problem, but it does mean we have to factor in a few things when dealing with large file workloads.

- Total FlexGroup size
- Member volume count
- Member volume size (as compared to largest file size)

Having larger member volumes offsets potential issues large files might create, and features such as proactive resizing and elastic sizing mitigates potential capacity issues affecting data availability.

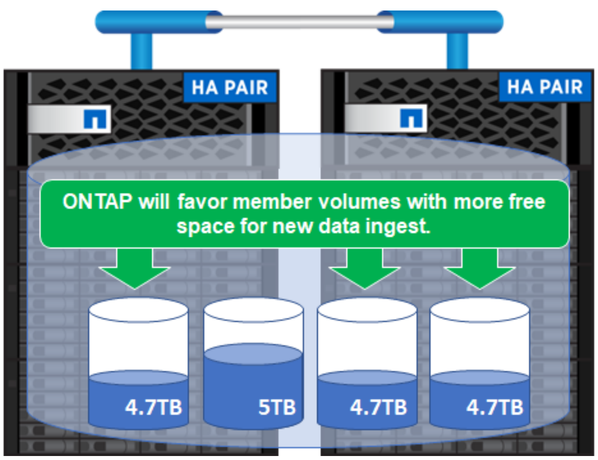
Consideration #4: Large files create imbalances that potentially affect performance

Another concern with large files is that a single file can consume enough space on the constituent to substantially affect the balance of usage among constituents. Sometimes a few files grow to a size that is orders of magnitude above the average file size. The result is that some constituents (the ones that happen to hold the aberrantly large files) end up with much more data than their peers have. In response, the FlexGroup volume begins to divert other new content creations onto the underused constituents. As a result, a subset of constituents can end up servicing most of the traffic. This problem is not typically severe; it simply represents suboptimal behavior. ONTAP 9.7 and later versions make substantial strides in handling the placement of these types of files and workloads so that they are better balanced across member volumes.

Best practice 3: Large file size considerations
Before sizing a FlexGroup volume, perform an analysis to determine the largest possible file size in a workload. Then, the member volume sizes should reflect those large file sizes, so that a large file cannot consume more than 1% to 5% of a FlexGroup member volume. Following this best practice helps avoid "out of space" concerns. Also, running ONTAP 9.6 and later can help avoid "out of space" concerns by way of the elastic sizing functionality. Running the latest patched version of ONTAP is always a good practice for FlexGroup volumes.

One other concern relates to running with the FlexGroup volume continually very close to full. As the FlexGroup volume becomes full, ONTAP becomes proactive in placing new content on those constituents that still have free space. If the working set consists primarily of small files, this behavior is adequate to prevent clients from receiving `Volume Full` errors until the collective FlexGroup volume is indeed full. However, before ONTAP 9.6, when large files were in the workload, those files continued to grow until they completely filled their constituent/member volume, resulting in `Volume Full` errors (`ENOSPC`) even if other constituents/members still had free space. Beginning with ONTAP 9.6, elastic sizing provides a way for ONTAP to borrow space from less-full member volumes and allow file writes to complete in member volumes. ONTAP 9.8 also introduces proactive resizing, which further improves the capacity management for full member volumes.

Figure 39) Capacity imbalance example.



Best practices when using large files with FlexGroup volumes

FlexGroup volumes operate best when dealing with lots of smaller files. However, they can also be effective when storing larger files as long as the FlexGroup volume is configured to account for that workload up front. When you're sizing a FlexGroup volume for large files, it's important to consider what a large file is, and what the largest and average file sizes in a workload are.

File sizes must be factored in when you design a FlexGroup volume so that member volumes are sized appropriately. ONTAP 9.6 and later versions make this process unnecessary with the addition of elastic sizing, and ONTAP 9.8 simplifies capacity management even more with proactive resizing. In general, you can apply the following best practices for large file workloads:

- For large file sizes, consider deploying larger member volumes with fewer members per FlexGroup volume. See Initial Volume Size Considerations for details.
- If you are running ONTAP 9.5 or earlier, enable volume autogrow on the FlexGroup volume to avoid running out of space in a member volume that contains large files. When running ONTAP 9.6 or

ONTAP 9.7, elastic sizing is enabled by default. Using volume autogrow disables elastic sizing for a FlexGroup volume in those releases, so decide how you want to manage capacity.

- Preferably, use ONTAP 9.8 or later to gain the benefits of proactive resizing. ONTAP 9.8 and later allow elastic sizing and volume autogrow to co-exist.
- Use quota enforcement to limit and monitor the capacity in qtrees or by user (ONTAP 9.5 and later).
- Before deploying a FlexGroup volume, use NetApp XCP to scan the file system and analyze the file sizes to understand average file size, largest file size, and so on.
- You should size a FlexGroup volume so that member volumes are less likely to become imbalanced. The largest file size should not exceed 1% to 5% of the member volume's capacity, but keep in mind that a smaller member volume size means that, for 1% to 5% of volume capacity, the largest files must be smaller, relatively speaking. Smaller files finish writing to the storage system faster and do not create enough discrepancy to shift the ingest algorithm very much. When possible, avoid member volume sizes under 1TB when file sizes are 50GB or less (minimum member volume size is 100GB). Since the notion of a "large file" is 1% to 5% of the member volume space, that "large file" size value is much smaller in a 100GB member volume (1GB to 5GB) than it is in a member volume that is 1TB (10GB to 50GB).

Performance expectations: Read-heavy workloads

Performance in a FlexGroup volume can greatly exceed that of a FlexVol volume or competitor systems for write-metadata-heavy workloads (high `CREATE` and `SETATTR` calls) that ingest many files. However, other workloads, such as file streams, file appends or read-heavy workloads, don't see the same extreme performance gains over FlexVol volumes that the ingest-heavy workloads see. This is because a FlexGroup volume is designed to overcome the bottleneck of serial processing of write metadata workloads by providing more volume affinities to those workloads. Basic read and writes don't face this serial processing bottleneck.

In some cases (especially with all local traffic), a set of multiple FlexVol volumes might perform slightly better than a FlexGroup volume for random and sequential read/write workloads. However, the complexity involved with creating and managing multiple FlexVol volumes versus a single FlexGroup volume might outweigh the slight performance gains.

For read-heavy workloads, using FlexGroup volumes has some benefits over using single FlexVol volumes, such as the following:

- Scaling across multiple CPUs and nodes to load balance reads to multiple files
- Single namespace for a large-capacity bucket

When deciding whether to use a FlexGroup volume, consider support for specific features. See the earlier section on what is and is not currently supported with FlexGroup volumes.

In addition, when using read-heavy workloads, also consider deploying FlexCache volumes attached to a FlexGroup origin volume to distribute the workload across more volume affinities in the cluster, or even across other SVMs in the same cluster or other clusters across multiple sites or in the cloud.

Data imbalances in FlexGroup volumes

In rare cases, a FlexGroup workload might have an imbalance of capacity in the member volumes. On its own, this does not indicate a problem; this is only a problem if performance is noticeably suffering, or capacity imbalances are causing clusters to run out of available space. In most cases, a capacity imbalance does not need to be addressed. ONTAP performs the work to balance out the workload if there are consistent new file creations. This is especially true in ONTAP 9.8 and later due to proactive resizing. In cases where the data is static, the imbalance does not resolve. However, the performance should remain roughly the same as if there was no data imbalance.

Data imbalances can occur in the following scenarios:

- A mix of large and small files are written to a FlexGroup volume.
- A file is written to a FlexGroup and then appended later, thus growing and increasing used capacity.
- Multiple large files are written at once to the same folder; ONTAP does not know how large these files will get, so they are placed to the local member volume for performance considerations.
- Large datasets get deleted and you happen to delete more files on one member volume than other member volumes.
- A user creates a very large file (such as a zip file of many existing files).

Each ONTAP release adjusts the ingest algorithms to try to address wider ranges of workload scenarios, so use the latest ONTAP release available. If issues are present, open a technical support case to isolate and remediate the issue.

Post-placement rebalance

Currently, ONTAP has no method to natively and nondisruptively rebalance the files that have already been ingested. ONTAP 9.10.1 introduces a disruptive file relocation feature for FlexGroup volumes that should be used with the guidance of NetApp support. The only way to rebalance data in releases earlier than ONTAP 9.10.1 is to copy the data from a FlexGroup volume to a new, empty FlexGroup volume all at once. This process is disruptive, because clients and applications must point to the new FlexGroup volume after the data has been moved. Also, this process is performed at a file level, so it could take a considerable amount of time. Rebalancing the files should be considered only if the imbalance of volumes creates an issue that affects production. As mentioned, capacity imbalances are usually imperceptible to client activity. Most customers don’t notice an imbalance until they are alerted to a capacity threshold. If data rebalance is necessary, consider using NetApp XCP to speed up the file copy process.

Initial FlexGroup design considerations

This section covers initial NetApp ONTAP FlexGroup volume design considerations. In presenting this information, NetApp assumes that no previous FlexGroup volumes have been created on the cluster. NetApp also assumes that you have experience with and knowledge about managing ONTAP through the CLI and the GUI and that you have administrator-level access to the storage system.

Cluster considerations

An ONTAP cluster that uses only NAS functionality (CIFS/SMB and NFS) can expand to up to 24 nodes (12 HA pairs). Each HA pair is a homogenous system (that is, two NetApp AFF nodes, two FAS8080 nodes, and so on), but the cluster itself can contain mixed system types. For example, a 10-node cluster could have a mixture of four AFF nodes, four NetApp FAS spinning disk systems, and two hybrid nodes for storage tiering functionality.

A FlexGroup volume can potentially span an entire 24-node cluster. However, keep the following considerations in mind.

- **FlexGroup volumes should ideally span only hardware systems that are identical.**

  Because hardware systems can vary greatly in terms of CPU, RAM, and overall performance capabilities, the use of only homogenous systems helps promote predictable performance across the FlexGroup volume. Data is balanced anywhere a FlexGroup volume has member volumes deployed; the storage administrator does not control this placement.

- **FlexGroup volumes should span only disk types that are identical.**

  Like hardware systems, disk type performance can vary greatly. Since a FlexGroup volume can span multiple nodes in a cluster and the storage administrator has no control over where the data is placed, you should make sure that the aggregates that are used are either all SSD, all spinning, or all hybrid. Mixing disk types can lead to unpredictable performance.

- **Disk sizes are not hugely important.**

  While it is important to use similar disk types on aggregates a FlexGroup might span, disk sizes are less important. For instance, if your aggregates have 3TB disks but you bought a set of new 16TB disks, feel free to deploy a FlexGroup across them, provided they are the same media type. The main caveat here is that the member volumes you deploy must be equivalent in size to the others.

- **FlexGroup volumes can span portions of a cluster.**

  You can configure a FlexGroup volume to span any combination of nodes in the cluster, from a single node to an HA pair, to all 24 nodes. The FlexGroup volume does not have to be configured to span the entire cluster. However, doing so can take advantage of all the hardware resources that are available.

ONTAP version considerations

Each release of NetApp ONTAP includes new features and improvements for FlexGroup volumes. Although NetApp recommends using the latest available patched release of ONTAP, many storage administrators are unable or unwilling to do that.

If you must run an older ONTAP version, familiarize yourself with the feature gaps in that release in the FlexGroup feature support and maximums section, and if possible, test the workload on a FlexGroup volume before deploying in production.

Failure domains

A failure domain is an entity that, if failure occurs, can negatively impact workloads. For example, in an ONTAP cluster, if a both nodes of an HA pair fail (a rare occurrence), the volumes on those nodes become unavailable because there is nowhere for them to fail over to. As a result, the HA pair is considered a failure domain in the cluster. However, a single node in an HA pair can fail in a cluster without disruption because its partner can take it over. In this situation, a single node would not be considered a failure domain.

Errors within a failure domain (such as RAID errors, losing a disk, multipath configuration errors, and metadata inconsistencies) are handled in ONTAP and do not negatively affect the FlexGroup volume.

FlexGroup volumes can span multiple nodes and HA pairs, and thus, multiple failure domains. However, even if a FlexGroup volume spans an entire 10-node cluster, the failure domain is still the HA pair. If you lose access to members in a FlexGroup volume (such as in the rare instance of failure of the HA pair), write access is disabled until all those members are repaired and reintroduced into the FlexGroup volume. The more HA pairs a FlexGroup volume spans, the higher the probability for failure is, because you are now spanning more failure domains. The fewer HA pairs that are used, the lower the probability for failure, but you see less overall performance for the FlexGroup because fewer hardware resources are available for the workload.

Therefore, when planning deployment, consider how many nodes to span in a FlexGroup volume and what SLAs are acceptable, and weigh those considerations against the capacity required and performance needed.

Aggregate layout considerations

An aggregate is a collection of physical disks that are laid out into RAID groups and provide the back-end storage repositories for virtual entities such as FlexVol and FlexGroup volumes. Each aggregate is owned by a specific node and is reassigned during storage failover events.

Since ONTAP 9, aggregates have dedicated NVRAM partitions for consistency points to avoid scenarios in which slower or degraded aggregates cause issues on the entire node. These consistency points are also known as per-aggregate consistency points and allow mixing of disk shelf types on the same nodes for more flexibility in the design of the storage system.

Best practice 4: Aggregate usage with FlexGroup
For consistent performance when using NetApp FlexGroup volumes or multiple FlexVol volumes, make sure that the design of the FlexGroup volume or FlexVol volumes spans only aggregates with the same disk type and RAID group configurations for active workloads. For tiering of cold data, predictable performance is not as crucial, so mixing disk types or aggregates should not have a noticeable effect.

Table 9 shows the best practices that NetApp recommends for aggregate layout when you use FlexGroup volumes. Keep in mind that these practices are not hard requirements. The one-aggregate-per-node recommendation for AFF systems originates from disk cost associated with NetApp RAID Triple Erasure Coding (RAID-TEC), because you might not want to use up expensive SSD space for the additional parity drives required for more than one RAID group. However, with ADP, partitions are spread across data disks, so in those cases, two aggregates per node on AFF systems are better because there are more available volume affinities per node with two aggregates present.

Table 9) Best practices for aggregate layout with FlexGroup volumes.

Spinning disk or hybrid aggregates	AFF
Two aggregates per node	One aggregate per node (without ADP) Two aggregates per node (with ADP)

Note: For consistent performance, aggregates should have the same number of drives and RAID groups across the FlexGroup volume.

For more information about aggregate layouts when dealing with existing FlexVol volumes, see "Failure Domains" in this document.

Deploying a FlexGroup volume on aggregates with existing FlexVol volumes

Because a FlexGroup volume can span multiple aggregates in a cluster and can coexist in the same SVM as normal FlexVol volumes, it is possible that a FlexGroup volume might have to share an aggregate with preexisting FlexVol volumes. Therefore, it is essential to consider the factors described in this section when you're deploying a FlexGroup volume.

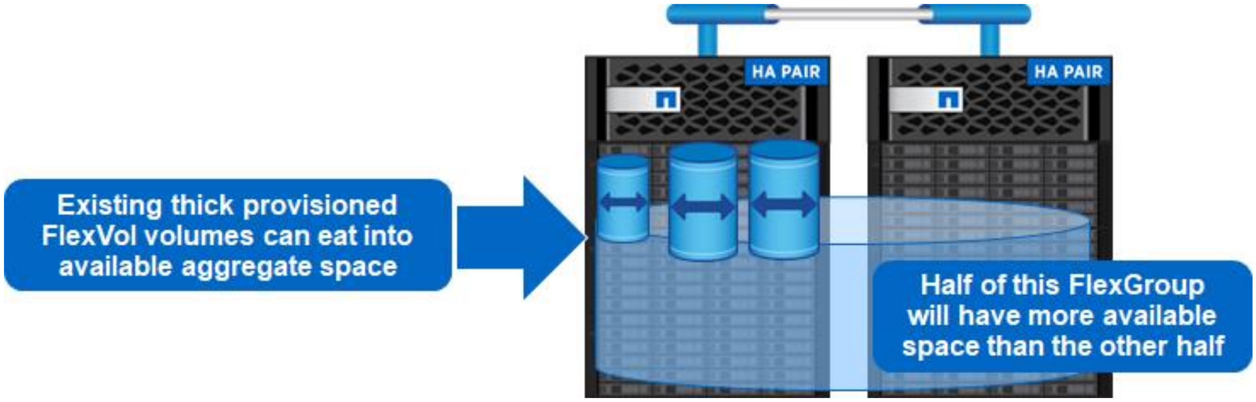
Consider the capacity footprint of the existing FlexVol volumes

A FlexGroup volume can span multiple aggregates and each of those aggregates might not have the same number of FlexVol volumes on them. Therefore, the aggregates might have disparate free space that can affect the ingest distribution of a FlexGroup volume that has space guarantees disabled, because the existing FlexVol volume capacity might eat into the FlexGroup volume's capacity.

For example, if aggr1 on node1 has four FlexVol volumes at 1TB each and aggr2 on node2 has two FlexVol volumes at 1TB each, then node1's aggregate would have 2TB less space than node2. If you deploy a FlexGroup volume that spans both nodes and is overprovisioned to fill both aggregates, then node1's member volumes already have "space used" in their capacity reports, which would cause node2's members to absorb most of the ingest of data until the capacity used is even across all member volumes.

In ONTAP versions earlier than 9.6, keep in mind that when thick-provisioned volumes are present in any aggregate that the FlexGroup attempts to span, the aggregates must have at least 3% free space available. For more information, see "Aggregate Layout Considerations" earlier in this document.

Figure 40) How FlexVol capacity can affect FlexGroup load distribution.



Note: This is an issue only if the FlexGroup volume is thin provisioned. Space-guaranteed FlexGroup volumes would not have other volumes eating into the space footprint. However, space-guaranteed FlexGroup volumes might not be created as large as desired if other volumes in the system prevent the space from being allocated.

Consider the performance effect of the existing FlexVol volumes

When you deploy a FlexGroup volume, it is also important to consider the amount of work the existing FlexVol volumes are doing. If a set of FlexVol volumes on one node is being hit heavily at given times, that can negatively affect the performance of a FlexGroup volume that spans the same nodes and aggregates as the existing FlexVol volumes. This is similar to the effect that can be seen with FlexVol volumes, but because a FlexGroup volume can span multiple nodes, the performance effect might appear to be intermittent from the client perspective, depending on which node the data I/O is occurring.

One way to mitigate this effect is to make use of storage QoS policies to help limit IOPS and throughput to those volumes or guarantee performance with QoS minimums on the FlexGroup volume. Alternately, you can use nondisruptive volume move to redistribute the volumes across nodes to balance the performance effect.

Consider the volume count limits

ONTAP has volume count limits per node that depend on the type of node in use (either AFF or FAS) and the personality of the node. For instance, an A800 node has higher volume count limits than a FAS8XXX series. A system with the data protection personality allows more total volumes than a system without. Additionally, there is a cluster-wide volume limit of 12,000 regardless of the node type in use that can affect how many FlexVol volumes can be provisioned per ONTAP cluster.

Because FlexGroup volumes generally contain multiple FlexVol member volumes, these member volumes count against this total limit. In addition, many ONTAP features also leverage FlexGroup volumes for their architectures. For instance, a single FlexGroup volume might use 16 member FlexVol volumes. If you use FlexClone for that volume, then you’ve used 32 FlexVol volumes. If you create a FlexCache volume in that cluster (which is also a FlexGroup), then you use whatever number ONTAP selects for that cache volume.

In most cases, you should not need to create multiple small FlexGroup volumes. Instead, provision a larger FlexGroup volume and use qtrees to separate workloads.

Best practice 5: Deploying FlexGroup volumes with existing FlexVol volumes in place
Before deploying a FlexGroup volume, note the following:

Best practice 5: Deploying FlexGroup volumes with existing FlexVol volumes in place

- If you have existing FlexVol volumes, be sure to verify that adding multiple FlexGroup volumes and their corresponding features to the mix do not exceed the volume count limits.
- Be sure to use the performance headroom features in NetApp Active IQ Unified Manager and ONTAP System Manager to review which nodes are being more heavily utilized.
- If there is an imbalance, use nondisruptive volume moves to migrate "hot" volumes to other less-utilized nodes to achieve as balanced a workload across nodes as possible.
- Be sure to evaluate the free space on the aggregates to be used with the FlexGroup volume and make sure that the available space is roughly equivalent.
- If the effect of volume count limit is a potential factor, create the FlexGroup volumes across nodes that have room to add more new volumes, or use nondisruptive volume moves to relocate volumes and balance out volume counts.
- Alternately, create FlexGroup volumes with fewer member volumes if volume count limits are a concern.

Flash Cache and Flash Pool

NetApp Flash Cache cards and NetApp Flash Pool aggregates are supported with FlexGroup volumes, but, if you choose to use them, be sure to have one on any node that participates in a FlexGroup volume for consistent performance results. Flash Cache cards are expected to provide the same performance benefits for FlexGroup volumes that they provide for FlexVol volumes.

Advanced Disk Partitioning

FlexGroup volumes have no bearing on the use of Advanced Disk Partitioning (ADP). No special considerations need to be made.

NetApp SyncMirror (mirrored aggregates)

FlexGroup volumes can reside on aggregates that participate in a NetApp SyncMirror® configuration, which is a way to replicate aggregates internally for extra data protection functionality. The FlexGroup should reside entirely on SyncMirror aggregates (for example, all member volumes are on SyncMirror aggregates, or none are). Otherwise, the SyncMirror is not useful.

For more information about SyncMirror, see this NetApp Support site page.

Note: SyncMirror does not provide the same functionality as StrictSync (NetApp SnapMirror Synchronous), which was new in ONTAP 9.5. FlexGroup volumes currently do not support StrictSync or SnapMirror Synchronous. SyncMirror is more akin to MetroCluster. For the latest data protection information regarding FlexGroup volumes, see TR-4678.

NetApp MetroCluster

ONTAP 9.6 introduced support for FlexGroup volumes on MetroCluster deployments (FC and IP).

MetroCluster software is a solution that combines array-based clustering with synchronous replication to deliver continuous availability and zero data loss at the lowest cost. There are no stated limitations or caveats for FlexGroup volumes with MetroCluster.

For more information about MetroCluster, see TR-4705: NetApp MetroCluster Solution Design and Architecture.

Note: When using NetApp Volume Encryption (NVE) or NetApp Aggregate Encryption (NAE) with MetroCluster, be sure to complete the MetroCluster configuration before enabling NVE/NAE. See bugs 1360199 and 1360229 for details.

NetApp Cloud Volumes ONTAP

ONTAP 9.6 introduced official support for Cloud Volumes ONTAP—an ONTAP solution running in the cloud. You can now deploy a FlexGroup volume using Cloud Volumes ONTAP (CVO).

FlexGroup volumes running in Cloud Volumes ONTAP can use the same feature sets available in the ONTAP version deployed to the Cloud Volumes ONTAP instance. Some common use cases seen for Cloud Volumes ONTAP and FlexGroup include the following:

- Data lake for analytics
- EDA repositories for use with Amazon Elastic Compute Cloud (Amazon EC2) instances
- Data backup and archive for use with on-premises SnapMirror

Although FlexGroup volumes can support multiple petabytes in a single namespace for on-premises deployments, Cloud Volumes ONTAP instances max out at 368TB per instance and FlexGroup volumes cannot span more than one cluster instance. FlexGroup volumes in CVO can only be created by using the CLI or ONTAP System Manager. Currently, you cannot use Cloud Manager to create FlexGroup volumes.

For more information about Cloud Volumes ONTAP, see Cloud Volumes ONTAP Enterprise Data Management Solution.

Capacity considerations

Although FlexGroup allows massive capacity and file count possibilities, the FlexGroup volume itself is still limited to the physical maximums of the underlying hardware. The current stated maximums (20PB, 400 billion files) are only tested maximums; the theoretical maximums could go a bit higher, but the official supported member volume count in a FlexGroup volume currently stands at 200. If you require more than 200 member volumes in a FlexGroup volume, contact your NetApp sales representative or email flexgroups-info@netapp.com.

Also, there are node-specific aggregate size limitations that allow only a set number of 100TB FlexVol volumes. Be sure to review your hardware's physical capacity limitations for more information.

For example, the FAS8080 EX allows 400TB aggregates before ONTAP 9.2TB and 800TB aggregates after ONTAP 9.2, which means that we would see a maximum of four 100TB volumes allowed per aggregate or eight 100TB volumes per aggregate, depending on the ONTAP version being used. However, NetApp recommends not reaching the 100TB limit for member volumes, because doing so would make it impossible to expand member volumes further in the future in the event a member volume runs out of space (you would have to add new 100TB member volumes to increase capacity in that case). Instead, aim to leave a cushion of no less than 10% to 20% of the total maximum FlexVol member space to provide for emergency space allocation features such as autogrow, elastic sizing, and proactive resizing to take effect.

These numbers are raw capacities before features such as NetApp Snapshot reserve, NetApp WAFL® reserve, storage efficiencies and FabricPool cloud tiering are factored in. To correctly size your FlexGroup solution, use the proper sizing tools, such as the System Performance Modeler (requires a NetApp login), or contact your NetApp sales representative to get assistance.

Maximums and minimums

This section covers the maximums and minimums that are specific to NetApp ONTAP FlexGroup volumes. Table 10 lists the maximum values and shows whether the maximum is hard-coded/enforced or a recommended/tested value.

Table 10) FlexGroup maximums.

	Value	Value Type
FlexGroup volume size	20PB	Tested/recommended*
File count	400 billion	Tested/recommended*
Cluster node count	24 (12 HA pairs)	Hard-coded/enforced
NetApp FlexVol member volume size	100TB	Hard-coded/enforced
FlexVol member volume file count	2 billion	Hard-coded/enforced
NetApp SnapMirror volume count (member per FlexGroup)	32 (ONTAP 9.4 and earlier) 200 (ONTAP 9.5 and later)	Hard-coded/enforced
SnapMirror volume count (FlexGroup total per cluster)	100 (ONTAP 9.4 and earlier) 6,000 (ONTAP 9.5 and later)	Hard-coded/enforced
File size	16TB	Hard-coded/enforced
FlexVol member constituent count	200	Tested/recommended*
Aggregate size/count	Same as NetApp ONTAP limits	Hard-coded/enforced

Table 11) FlexGroup minimums.

	Value	Value Type
FlexVol member size	100GB	Tested/recommended*
Data aggregate count	1	Hard-coded/enforced
SnapMirror schedule	30 minutes	Tested/recommended*
NetApp Snapshot copy schedule	30 minutes	Tested/recommended*

*Limits described as tested/recommended are tested limits based on a 10-node cluster. If allowed by the platform, actual limits are not hard-coded and can extend beyond these limits up to 24 nodes. For more information, see the section "Theoretical or absolute maximums." However, official support for the number of member volumes is 200. If you need to exceed this limit, contact your NetApp sales representative to start the qualification process for more member volumes.

Maximum number of FlexGroup volumes in a cluster

A FlexGroup volume can consist of a single FlexVol member volume or hundreds of FlexVol member volumes. The maximum number of FlexVol member volumes is physically constrained only by the total volume count in a cluster. As a result, a FlexGroup volume could theoretically have up to ~12,000 member volumes in a 24-node cluster.

The total number of FlexGroup volumes is similarly constrained by the total volume count in a cluster. Each FlexGroup volume's member volumes are part of the volume count, so the number of FlexGroup volumes allowed in a cluster depends on the number of member volumes.

For example, a two-node cluster may have 2,000 volumes to work with. As a result, you could have one of the following configurations (although others are possible), all of which add up to 2,000 volumes:

- 10 FlexGroup volumes with 200 member volumes
- 20 FlexGroup volumes with 100 member volumes
- 40 FlexGroup volumes with 50 member volumes
- 200 FlexGroup volumes with 10 member volumes

Keep in mind that the existence of other FlexVol volumes in the cluster (including SVM root volumes) affects the total number of available member volumes. FlexVol member volume limits also can be constrained by the FlexGroup volumes participating in NetApp SnapMirror relationships. For the latest details on those limitations, see TR-4678.

ONTAP, in most cases, creates multiple member volumes by default. For details on FlexGroup creation methods, see the section “FlexVol member volume layout considerations.” If you create multiple FlexGroup volumes, you might unknowingly begin to use up the volume count in your cluster. In general, use the automated `volume create -auto-provision-as` CLI command to create new FlexGroup volumes rather than getting bogged down in the details of member volume counts. Overall, it is better to create fewer, larger FlexGroup volumes and divide workloads using qtrees. Qtrees in ONTAP 9.8 offer quota enforcement, granular statistics, and QoS policies (currently, qtree QoS is only supported for NFS).

Theoretical or absolute maximums

The stated supported limits for a FlexGroup volume are 200 constituent volumes, 20PB, and 400 billion files. However, these are simply the tested limits in a 10-node cluster. When you factor in the maximum volumes that are allowed per node in a cluster, the limits can potentially expand dramatically.

Ultimately, the architectural limitation for a FlexGroup volume is the underlying hardware capacities and the total number of allowed volumes in a single cluster.

Table 12) Theoretical maximums for FlexGroup based on allowed volume count in ONTAP.

Maximum cluster size	Current architectural maximum member volumes per cluster (ONTAP 9.8)	Theoretical maximum capacity per FlexGroup volume	Theoretical maximum inodes per FlexGroup volume
24 nodes	12,000 (Note: SVM root, node root volumes and LS mirror volumes count against this value.)	~1195PB (based on 100TB per member volume * ~11,950 FlexGroup member volumes)	~23.9 trillion inodes (based on 2 billion inodes * ~11,950 FlexGroup member volumes)

Note: The main limiting factor in the number of 100TB member volumes allowed in a cluster is the underlying physical hardware limitations, which vary depending on platform.

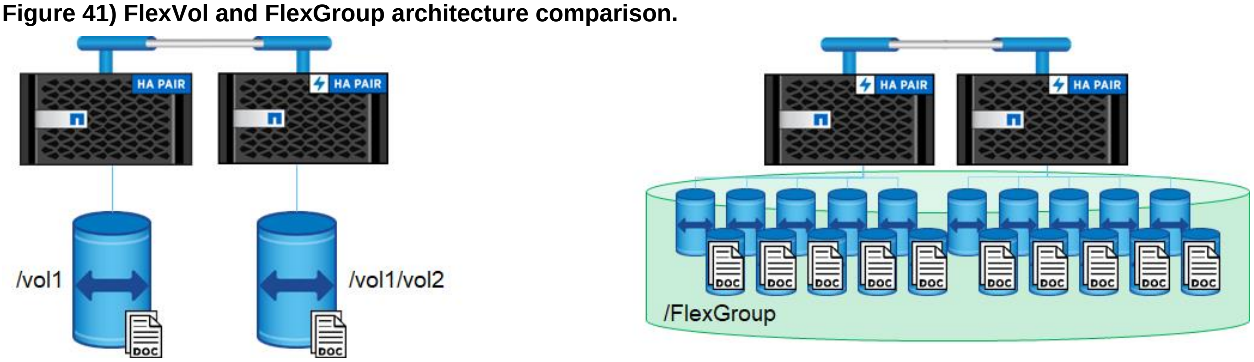
If you want to exceed the stated 20PB, 400 billion file, and 200-member volume limits, contact your NetApp sales representative to begin a qualification process.

FlexVol member volume layout considerations

FlexVol volumes are the building blocks of a FlexGroup volume. Each FlexGroup volume contains several member FlexVol volumes to provide concurrent performance and to expand the capacity of the volume past the usual 100TB limits of single FlexVol volumes.

Standard FlexVol volumes are provisioned from the available storage in an aggregate. FlexVol volumes are flexible and can be increased or decreased dynamically without affecting or disrupting the environment. A FlexVol volume is not tied to any specific set of disks in an aggregate and exists across all the disks in the aggregate. However, individual files themselves are not striped; they are allocated to individual FlexVol member volumes.

Figure 41) FlexVol and FlexGroup architecture comparison.



Because of this architecture and the potential for large files to affect FlexGroup operations, there are some considerations you should keep in mind when provisioning a FlexGroup volume.

When designing a FlexGroup volume, consider the following for the underlying FlexVol member volumes:

- When you use automated FlexGroup creation methods such as `volume create -auto-provision-as flexgroup` (introduced in ONTAP 9.2) or ONTAP System Manager, the default number of member FlexVol volumes in a FlexGroup volume depends on several factors covered in this section.

  Note: For nearly all use cases, NetApp recommends that you let ONTAP determine the member volume count per node, provided you are creating larger FlexGroup volumes (10TB or greater). For smaller FlexGroup volumes, closer attention should be paid to the file sizes in the workload and the percentage of capacity per member volume they would potentially use.

- If a node with spinning disks does not contain two aggregates, the automated FlexGroup creation method might fail in some earlier ONTAP versions. If this happens, continue with manual creation.

FlexVol member volumes are deployed in even capacities, regardless of how the FlexGroup volume was created. For example, if an eight-member, 800TB FlexGroup volume was created, each member is deployed with 100TB. If a larger or smaller quantity of FlexVol member volumes is required at the time of deployment, use the `volume create` command with the `-aggr-list` and `-aggr-list-multiplier` options to customize the number of member volumes deployed per aggregate. Refer to the section called "When do I need to manually create a FlexGroup volume?"

Deployment method #1: Command line

Using the ONTAP command line is the currently recommended way to deploy a FlexGroup volume. However, there are two different ways to do this from the CLI: manually and automatically.

Both commands us the `volume create` command set.

Automated FlexGroup creation (auto-provision-as) – CLI

This is the preferred method for FlexGroup creation, as it combines ease of use with predictable deployment logic and warnings during creation to help prevent misconfigured FlexGroup volumes. To use the automated CLI method, run the `volume create -auto-provision-as flexgroup` command. By default, this command provisions a FlexGroup volume with the following parameters:

- N number of member volumes (100GB each; four per aggregate in the cluster, up to two aggregates per eight member volumes)
- Total size = 100GB per member * number of member volumes (16 member volumes = 1.6TB)
- All nodes and data aggregates in the cluster used (regardless of node or hardware type)

When you run the command, you see a warning that informs you of the configuration. Be sure to review that before typing `Y`. Are the member volumes the size you want? Are the listed aggregates correct and using the same media types?

In most cases, the default values specified with no options are not adequate. You likely want to specify aggregates or nodes to use in the deployment.

As such, there are some additional configuration option flags for FlexGroup auto provisioning.

```
-support-tiering                    -use-mirrored-aggregates
-encryption-type                    -nodes
-size                               -state
-policy                             -user
-group                              -security-style
-unix-permissions                   -junction-path
-comment                            -max-autosize
-min-autosize                       -autosize-grow-threshold-percent
-autosize-shrink-threshold-percent  -autosize-mode
```



```
-space-guarantee                  -type
-percent-snapshot-space           -snapshot-policy
-language                         -foreground
-nvfail                           -qos-policy-group
-qos-adaptive-policy-group        -caching-policy
-encrypt                          -is-space-reporting-logical
-is-space-enforcement-logical     -tiering-policy
-tiering-object-tags              -analytics-state
```

Use those flags to customize your FlexGroup volume's size, tiering policies, space guarantees, nodes, aggregates and much more.

The following is the general behavior of the automated CLI commands:

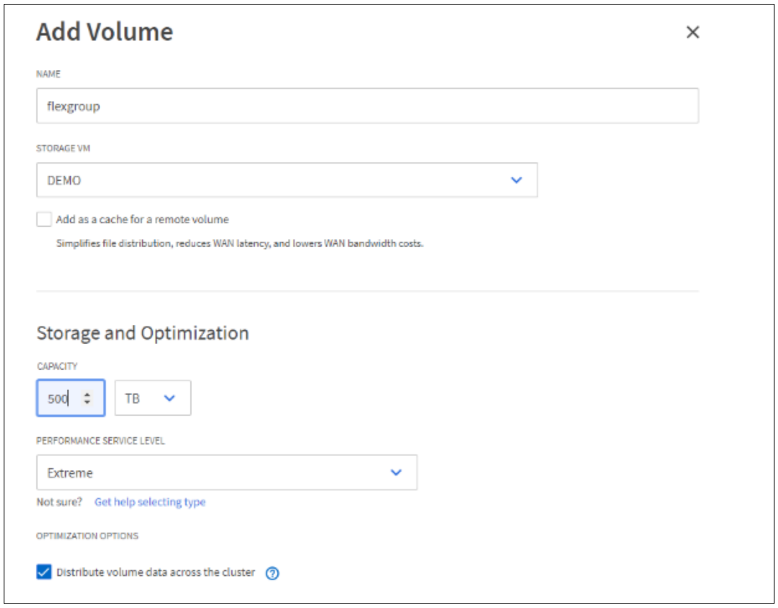
- Uses two aggregates per node, if possible. If not, use one aggregate per node.
- Uses the same number of aggregates on each node.
- The automated commands choose the aggregates that have the most free space.
- The automated commands create eight constituents per node if there are eight or fewer nodes.
- In clusters with more than eight nodes, scale back to four member volumes per node.
- Uses the fastest aggregates that you can. First try SSD, then Hybrid, and then spinning disk.
- CPU utilization, node performance, aggregate capacity, etc. are not currently considered.

Manual FlexGroup creation—CLI

The CLI also provides a more manual approach for creating FlexGroup volumes. In most cases, use the automated command, because it covers most use cases. However, if you need to customize the number of member volumes per aggregate, specify the aggregates to be used or have other reasons covered in the section "When do I need to manually create a FlexGroup volume?", you still use the `volume create` command. However, instead of the `–auto-provision-as` option, you must specify `–aggr-list` along with it. Specifying `–aggregate` creates a normal FlexVol volume and does not allow you to specify `–aggr-list`. To control the number of member volumes per aggregate, use `–aggr-list-multipler`. Your member volume count is the number of aggregates you specified multiplied by `-aggr-list-multiplier`.

Deployment method #2: ONTAP System Manager

Figure 42) ONTAP System Manager FlexGroup volume creation.



ONTAP System Manager has an easy-to-use GUI for volume creation. However, there are some caveats to consider when deploying with the System Manager GUI that make using the CLI a better choice when provisioning FlexGroup volumes.

To create a FlexGroup volume in ONTAP System Manager, the only thing you need to do to ensure the volume is a FlexGroup and not a FlexVol is to click More Options and select the Distribute Volume Data Across the Cluster box.

This tells System Manager to create a FlexGroup volume that spans multiple nodes; no aggregate or node specification is required.

System Manager deploys a FlexGroup volume according to the following rules.

- Member volumes are never smaller than 100GB
- The smallest allowed FlexGroup is 100GB (one 100GB member volume)
- Smaller FlexGroup volumes deploy fewer member volumes when necessary to adhere to the 100GB rule; for example, a 200GB FlexGroup deploys two 100GB member volumes.
- Aggregate and node selection for the FlexGroup is performed automatically. To specify nodes or aggregates, use the CLI or REST API.
- When a FlexGroup volume is large enough to accommodate, then the member volume count is capped to the number of volume affinities available per node.
- System Manager limits your initial FlexGroup size to the total space available as if space guarantees were enabled. You can go back into System Manager and grow the volume larger with the Edit functionality.
- System Manager only uses similar aggregates for the FlexGroup. In other words, it does not mix SSD and HDD aggregates.

Deployment method #3: REST APIs

Support for REST APIs was added to ONTAP 9.6. With REST APIs, you can create, monitor, and manage FlexGroup volumes. To use REST APIs to provision a FlexGroup volume, use the same guidance as described in the section “When do I need to manually create a FlexGroup volume?”

Deployment method #1: Command line.” For example, you can decide whether to let ONTAP choose the configuration or whether you should manually specify options.

You can find REST API documentation at `https://[your_cluster_IP_or_name]/docs/api`. This site provides examples and an interactive Try It Out feature that enables you to generate your own REST APIs.

For example, to create a FlexGroup volume, you can use the `POST` REST API under `/storage/volumes`. What makes a FlexGroup a FlexGroup (and not a FlexVol) in this call are one or a combination of the following values:

- **Aggregates.** If you specify more than one, then the REST API creates a FlexGroup volume. This is the same behavior as `-aggr-list` in the CLI.
- **constituents_per_aggregate.** Specifies the number of times to iterate over the aggregates listed with `aggregates.name` or `aggregates.uuid` when a FlexGroup volume is created or expanded. If a volume is being created on a single aggregate, the system creates a flexible volume if the `constituents_per_aggregate` field is not specified. If this field is specified, it creates a FlexGroup volume. If a volume is being created on multiple aggregates, the system always creates a FlexGroup volume. This is the same behavior as `-aggr-list-multiplier` in the CLI.
- **Style.** If you specify `style` as `flexgroup` and don’t set the `constituents_per_aggregate` value or more than one aggregate, ONTAP automatically provisions a FlexGroup volume of four members per aggregate. This is the same behavior as `–auto-provision-as` in the CLI.

In the REST API documentation, the Try It Out functionality helps guide you as you try to create the correct REST API strings. When you make a mistake, the interface delivers error messages and a list of error codes. Also, a job string URL is given if the REST API command is correct, but the job fails for another reason (such as creating a FlexGroup volume that has members that are too small). You can access the job string through the browser with the following URL:

```
https://[your_cluster_IP_or_name]/api/cluster/jobs/job_uuid]
```

This is what a failure message looks like:

```
{
  "uuid": "b5b04f0b-82ea-11e9-b3aa-00a098696eda",
  "description": "POST /api/storage/volumes/b5b02a66-82ea-11e9-b3aa-00a098696eda",
  "state": "failure",
  "message": "Unable to set parameter \"-min-autosize\" to specified value because it is too
small. It must be at least 160MB (167772160B).",
  "code": 13107359,
  "start_time": "2019-05-30T10:53:39-04:00",
  "end_time": "2019-05-30T10:53:39-04:00",
  "_links": {
    "self": {
      "href": "/api/cluster/jobs/b5b04f0b-82ea-11e9-b3aa-00a098696eda"
    }
  }
}
```

This is what a successful job looks like:

```
{
  "uuid": "ac2155d1-82ec-11e9-b3aa-00a098696eda",
  "description": "POST /api/storage/volumes/ac2131c5-82ec-11e9-b3aa-00a098696eda",
  "state": "success",
  "message": "success",
```



```
  "code": 0,
  "start_time": "2019-05-30T11:07:42-04:00",
  "end_time": "2019-05-30T11:07:46-04:00",
  "_links": {
    "self": {
      "href": "/api/cluster/jobs/ac2155d1-82ec-11e9-b3aa-00a098696eda"
    }
  }
}
```


For a sample REST API string that creates a FlexGroup volume, see the section “Command examples.”

When do clients experience out of space errors?

Generally, when a NAS client sees an `out of space` error, that means the volume is actually out of space and resulting `df` and `volume show` commands will confirm that.

However, getting an `out of space` error from a NAS storage system is not always straightforward, because it is a generic error from the server telling the client that there are no more available resources. There is no concept of an `out of inodes` error or `reached maximum directory size` in NFS or SMB, so ONTAP resorts to using the standard `out of space` error to let clients know that they cannot write more data.

With a FlexGroup volume, clients can also receive `out of space` errors when a member volume fills to 100%. However, in ONTAP 9.6 and later, this scenario is virtually nonexistent.

The following table shows situations in which `out of space` errors are seen, their causes, and how to address them.

Table 13) Situations in which you see out of space errors.

Situation	How to identify and resolve
Volume or aggregate has no available space to honor writes.	• `df` or `volume show-space` output from cluster CLI • View capacity from ONTAP System Manager • Active IQ Unified Manager alerts • EMS messages Resolution: Add more capacity to volume. Add more disks to aggregate. Use FlexGroup volumes to scale across nodes.
Quota limit reached	• `df` or `volume show-space` output from cluster CLI • View capacity from ONTAP System Manager • `quota report` output from cluster CLI • Quota report from ONTAP System Manager • Active IQ Unified Manager alerts • EMS messages Resolution: Increase the quota limit or notify the client that they need to delete data to stay under their quota.
Out of inodes	• `df` or `volume show-space` output from cluster CLI • View capacity from ONTAP System Manager • `df -i` or `volume show -fields files,files-used` command from cluster CLI • Active IQ Unified Manager alerts • EMS messages Resolution: • Increase the total files value in the volume. See the section “High file count considerations” for more information.

Situation	How to identify and resolve
Maxdirsize exceeded	• `df` or `volume show-space` output from cluster CLI • View capacity from ONTAP System Manager • `df -i` or `volume show -fields files,files-used` command from cluster CLI • Active IQ Unified Manager alerts • EMS messages • Client-side commands to view directory sizes (as shown in "Querying for used maxdirsize values") Resolution: Use the `volume file show-inode` command from the cluster CLI to find the affected file path. Reduce the file count in the offending directory or contact support to verify if the `maxdirsize` value is safe to increase. For more information on `maxdirsize`, see the section "Directory size considerations: maxdirsize."
Member volume at 100%	• `df` or `volume show-space` output from cluster CLI • Active IQ Unified Manager alerts • EMS messages • ONTAP version information Resolution: Member volumes at 100% generally only create `out of space` errors in ONTAP 9.5 and prior. ONTAP 9.6 introduces Elastic sizing as a safeguard against file write failures when a member volume fills. ONTAP 9.8 introduces Proactive resizing to proactively resize member volumes to maintain an even balance of free space in member volumes. For best capacity usage results, upgrade to ONTAP 9.8 or later.

When do I need to manually create a FlexGroup volume?

In most cases, letting ONTAP choose the member volumes is the best option when creating a FlexGroup volume. In other words, don't worry about CPU count and volume affinity best practices for FlexGroup creation—let ONTAP do that for you. Trying to manipulate volume counts can lead to confusion and issues that may affect your FlexGroup later.

However, in some use cases, manual creation might be needed. The following sections describe scenarios in which you might need to manually create FlexGroup volumes.

Concern regarding overprovisioning volume counts

In ONTAP, each node and cluster has a finite number of FlexVol volumes allowed. The limits are dependent on platform and ONTAP version, but, because a FlexGroup volume is composed of multiple FlexVol volumes, those limits also apply to FlexGroup volumes, as each FlexVol member volume counts against the total volume count limit. If you have FlexGroup volumes with many member volumes or you want to create many FlexGroup volumes in a cluster, then you would need to consider the overall volume limits per node. You might also need to manually create the FlexGroup volumes to modify the default volume counts or aggregate placement to keep total FlexVol volume numbers below the node limits.

Large files with limited capacity

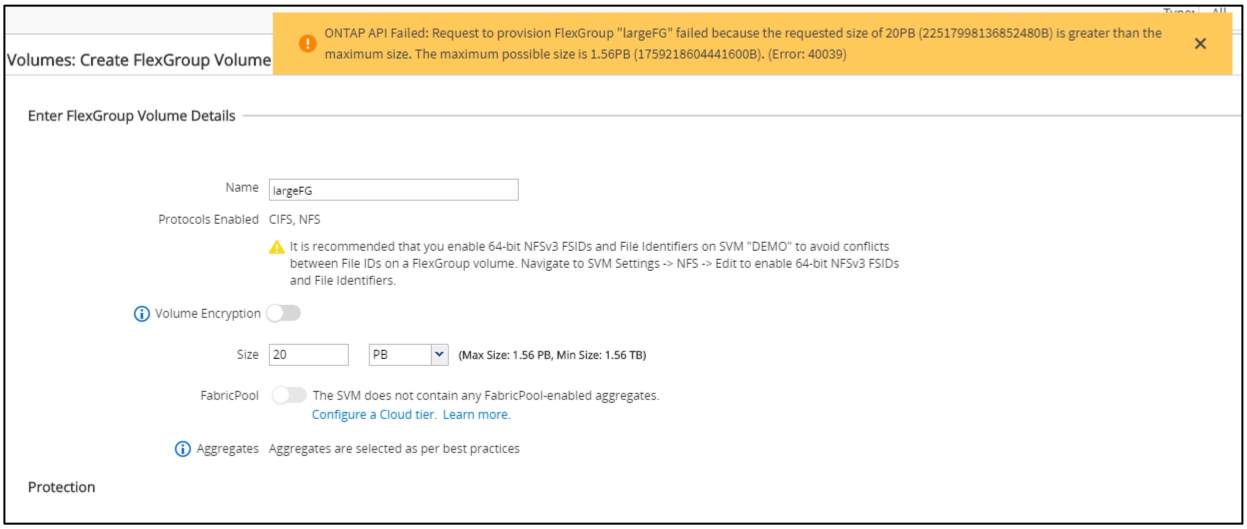
If you have a workload with larger files but cannot provision volumes that are tens or hundreds of TBs and you want to comply with best practices for large files, you might need to adjust the member volume count to create fewer members at larger individual capacities.

For example, if you want to create a 16TB FlexGroup volume across four nodes, using the ONTAP default methods would create a minimum of 32 member volumes that are 500GB in size each. If your average file size is 250GB, then 500GB member volumes are not large enough to distribute the data effectively. Manually creating a FlexGroup volume with fewer, larger member volumes works better for those use cases.

A need for a large amount of capacity or high file counts

FlexVol volumes are limited to 100TB in size and can contain up to two billion files. If you have a two-node cluster and you let ONTAP create a FlexGroup volume, you get at most 16 member volumes in a single FlexGroup volume in some cases, because it is code-limited to the best practice of eight per node. In the following example, the two-node cluster can only create a FlexGroup volume with a maximum of 1.56PB of capacity (eight members per node; 16 members in total; 100TB per member volume).

Figure 43) Error when creating a FlexGroup volume beyond the allowed maximum in System Manager.



The `auto-provision-as` option gives the same error:

```
cluster::*> vol create -vserver DEMO -volume largeFG -auto-provision-as flexgroup -size 2PB

Error: command failed: Request to provision FlexGroup volume "largeFGT" failed because the
requested size of 2PB (2251799813685248B) is greater than the maximum size. The maximum possible
size is 1.56PB (1759218604441600B).
```

If you desire a larger FlexGroup volume than what the automated tools allow, you need to create the FlexGroup manually to allow a higher number of member volumes by using the `-aggr-list-multiplier` option. For a 20PB FlexGroup volume, you need at least 200 member volumes. Ideally, you should size the Flexgroup member volumes to a value less than 100TB in case you later need room for those volumes to grow. 80-90TB should be the target maximum member volume value.

Similar considerations should be made if the file count needs to exceed the maximum files allowed. In the 16-member FlexGroup example, a maximum of 32 billion files is allowed. If more files are needed, increase the maxfiles value first. If that is not possible (for example, the maxfiles value is at the limit), then add more member volumes.

For an example of how to create a FlexGroup volume from the CLI and specify the number of members, see "Command examples" later in this technical report.

Avoiding the cluster network

A less common scenario is the desire to avoid the cluster network by creating a FlexGroup volume across a single node or to reduce exposure to failure domains. In this use case, use the ONTAP CLI to manage which aggregates are specified using the `-aggr-list` and `-aggr-list-multiplier` options with the `volume create` command.

Do I need a large number of member volumes?

Usually, you do not need to exceed the best practice volume count for a FlexGroup volume. However, if you need more capacity or higher file counts, you can increase the number of member volumes at initial deployment, or you can do so later by using the `volume expand` command. In general, it is better to increase member volume counts sooner than later before data starts to fill the existing member volumes. Adding member volumes later creates an imbalance, which ONTAP must adjust for, and it might affect workloads. For more information about when you might need to stray from ONTAP best practices for member volume counts, see the section "When do clients experience out of space errors?".

Member count considerations for large and small files

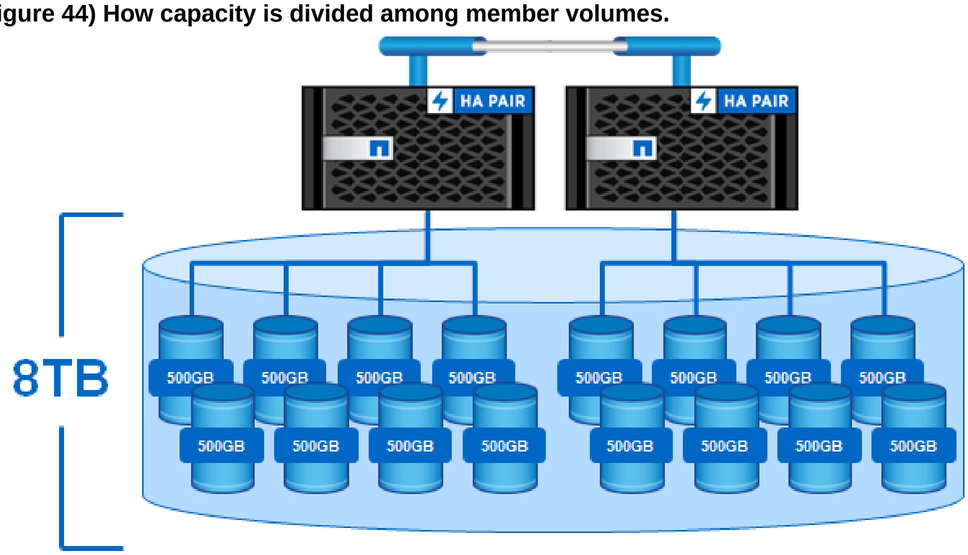
FlexGroup volumes work best in a high-file-count environment of many small files. However, they also work well with larger files. As mentioned in the section "What Are Large Files?," large files should be considered in terms of percentage of the total space allocated to a member volume.

When larger files are present in a workload, the initial deployment size of a FlexGroup should be kept in mind. By default, a FlexGroup deploys eight-member volumes per node, so any capacity footprint that is defined at the FlexGroup level effectively gets divided into [total space/n number of member volumes].

For example, if an 8TB FlexGroup is deployed across two nodes in a cluster and the member count is 16, then each member volume is about 500GB in size.

In many workloads, the distribution shown in Figure 44 would work well. However, if larger files in a workload would potentially fill in member volumes' large chunks of capacity used, then performance or even accessibility could be affected. If a member volume fills before other member volumes, the FlexGroup volume could report an "out of space" error in ONTAP versions earlier than 9.6. ONTAP 9.6 mitigates this "out of space" scenario with "Elastic sizing" and in ONTAP 9.8, "Proactive resizing" helps make these "member volume full" scenarios even less frequent.

Figure 44) How capacity is divided among member volumes.



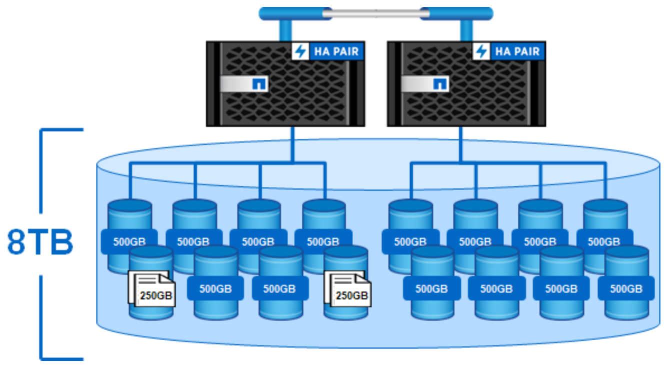
Best practice 6: Best ONTAP version to run for large file workloads

In general, the latest ONTAP release available is the best version to run when using FlexGroup volumes, but we understand that not everyone can upgrade to the latest ONTAP release. NetApp recommends using the latest patch release of ONTAP 9.8. If that version is not possible, then use the latest patch release of ONTAP 9.7.

For example, if some files in a workload are 250GB, then each time a file is written to a FlexGroup volume with 500GB members, 50% of the total capacity of a member volume is filled.

If a second 250GB file attempts to write to that 500GB member volume, then the member volume runs out of available capacity before the file completes its write. In versions of ONTAP prior to 9.6, that would result in the file not being able to complete its write (unless volume autogrow is enabled).

Figure 45) Effect of larger files in a FlexGroup member volume.



Remember, files in a FlexGroup volume do not stripe; they always write to a single FlexVol member volume. Therefore, there must be enough space in a single member volume to honor the write.

Elastic sizing in ONTAP 9.6 and later provides some relief by pausing before sending the client and `out of space` error and instead borrowing free space from other member volumes in the same FlexGroup if available. However, if volume autosize is enabled, elastic resizing is disabled for that volume. Elastic resizing is not intended to be a way to avoid capacity management, but instead is a reactive way to reduce the effect of capacity issues. It's still imperative that the member volume capacities remain below 80-90% for best results.

Proactive resizing in ONTAP 9.8 combines the benefits of elastic resizing and volume autosize. Rather than waiting for a file creation to run out of space, ONTAP increases member volume sizes at a free space threshold proactively to reduce the effect of capacity imbalances and reduce the need to manage capacity from individual member volumes. In addition, volume autogrow can be used in conjunction with proactive resizing, so, if the total FlexGroup capacity is at a threshold, ONTAP automatically increases the size to a specified value.

Eventually, more storage must be added when physical space is exhausted. Also, adding nodes or disks to a FlexGroup volume is nondisruptive, easy, and fast.

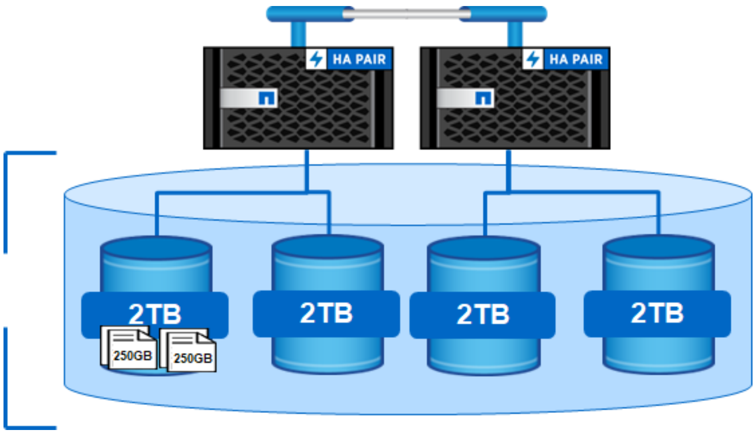
A better approach to sizing a FlexGroup volume is to analyze your workload and average file sizes before deploying a new FlexGroup volume or before allowing new workloads to access existing FlexGroup volumes. NetApp offers XCP, which can quickly analyze files and report on sizes. For more information about XCP, see the section "Migrating to NetApp FlexGroup."

After you have a good idea of what size files are going to land in a FlexGroup volume, you can make design decisions about how the volume should be sized at initial deployment.

Options include, but are not limited to the following:

- **Leave the member volume count at the defaults and grow the FlexGroup volume.** Size the total FlexGroup volume to a value large enough to accommodate member volume sizes that can handle the workload. In our example, the FlexGroup volume is 80TB, which provides 16-member volumes at 5TB per volume. However, this approach requires more physical capacity (unless you utilize thin provisioning).

Figure 46) Fewer, larger member volumes.



- **Manually reduce the member volume count and leave the FlexGroup capacity as is.** Rather than accept the default values from the automated commands, you can use the CLI to create a FlexGroup volume that is identical in total capacity but contains fewer (but larger) member volumes. In our example, reducing the member volume count to two per node in an 8TB FlexGroup would provide member volume sizes of 2TB each. This would reduce the number of volume affinities available (and could reduce the overall performance of the FlexGroup volume for file ingest), but it would allow larger files to be placed.

After the large files are placed in member volumes, performance should be similar to what you would see from a FlexVol volume or a FlexGroup volume with more member volumes.

Capacity management features

The following table shows which capacity management features are available depending on your ONTAP release.

Table 14) Capacity Management Decision Matrix

ONTAP version	Capacity management features
ONTAP 9.2 and prior	• Thin provisioning • Capacity alerting • Storage efficiencies
ONTAP 9.3 – ONTAP 9.4	• Thin provisioning • Capacity Alerting • Storage efficiencies • Volume autosize (autogrow/autoshrink) • Qtrees and monitoring quotas
ONTAP 9.5	• Thin provisioning • Capacity Alerting • Storage efficiencies • Volume autosize (autogrow/autoshrink) • Qtrees and monitoring quotas

ONTAP version	Capacity management features
	• Quota enforcement • FabricPool autotiering
ONTAP 9.6 - ONTAP 9.7	• Thin provisioning • Capacity Alerting • Storage efficiencies • Volume autosize (autogrow/autoshrink) • Qtrees and monitoring quotas • Quota enforcement • FabricPool autotiering • Elastic sizing
ONTAP 9.8 and later	• Thin provisioning • Capacity Alerting • Storage efficiencies • Volume autosize (autogrow/autoshrink) • Qtrees and monitoring quotas • Quota enforcement • FabricPool autotiering • Elastic sizing • Proactive resizing

Aggregate free space considerations

When you create a FlexGroup volume, it is ideal for the aggregate (or aggregates) that the FlexGroup is deployed on to have the following characteristics:

- A roughly even amount of free space across multiple aggregates (especially important when using thin provisioning)
- Roughly 3% free space available for aggregate metadata after creation of the FlexGroup volume (ONTAP 9.5 and earlier)
- At least 10GB or 0.6% free space (whichever is less) in ONTAP 9.6 and later

ONTAP 9.5 and earlier versions prevent a FlexGroup volume from filling an aggregate past 97% when using thick provisioning or space guarantees. Attempts fail with the error `request size is greater than maximum size`. It is possible to overcommit an aggregate by using thin provisioning, but if one aggregate has more space than the other, you run the risk of affecting performance or running out of space in members on one aggregate before the other aggregate runs out of space. ONTAP 9.6 and later versions no longer check for deduplication metadata.

Why is member volume capacity important?

The goal of a FlexGroup volume is to manage it from the FlexGroup level, while not having to pay much attention to the underlying member volumes. In most cases, this is how FlexGroup volumes operate. However, in releases prior to ONTAP 9.8, member volume capacity had to be considered more often when creating and managing FlexGroup volumes.

Available free space in a member volume affects how often new files are ingested locally or remotely in a FlexGroup volume, which in turn can affect performance and capacity distribution in the FlexGroup volume for new file creation.

Average and largest file sizes in a workload are important to consider when you are designing an initial FlexGroup volume, because large files can fill up individual member volumes faster, causing more remote allocation of new file creations, or even causing member volumes to run out of space before the other member volumes do. Files that are already placed in the member volume remain in place, which means that, as they grow, the used capacity in that volume increases. However, already placed files generally do not see the same performance effect in unbalanced FlexGroup volumes as new file creations do.

Capacity imbalances are not a problem in and of themselves, but they can be the root cause for performance issues or capacity utilization issues. If you have a performance issue and suspect capacity imbalance, open a NetApp support case for assistance in analyzing the performance data. In some cases, capacity imbalances might appear to be an issue, but the actual problem has a different root cause.

Best practice 7: Member volume size recommendations
NetApp recommends sizing a member volume such that the largest file does not exceed 1% to 5% of the member volume's capacity. Avoid creating FlexGroup volumes with fewer than two members per node.

Initial volume size considerations

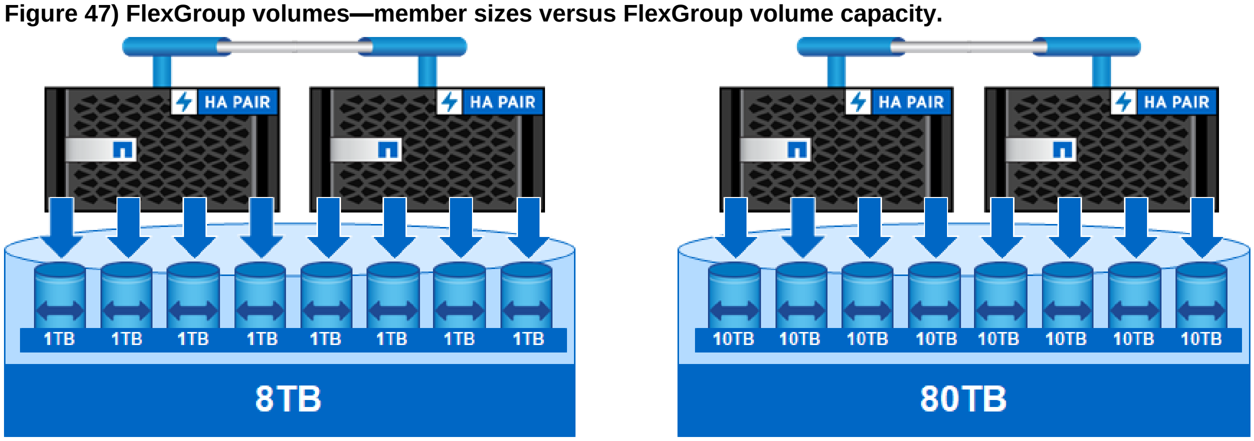
A common deployment issue is under sizing a FlexGroup volume's member volume capacity. This is often done unbeknownst to the storage administrator because all they care about is the total capacity; they don't usually stop to think about underlying member volumes. To them, 80TB should be 80TB. But in a FlexGroup, 80TB is actually 80TB divided by the total number of member volumes.

FlexGroup volumes can be created at almost any capacity, but it is important to remember that several FlexVol member volumes make up the total size of the FlexGroup volume. By default, automated FlexGroup commands create a default number of member volumes, depending on the deployment method used (see the section "FlexVol member volume layout considerations" for details).

Best practice 8: Simplifying FlexGroup deployment
If you want to stop worrying so much about member volumes, upgrade to ONTAP 9.8 to get the benefits of proactive resizing, which makes member volume sizes less of a concern.

For example, in an 80TB FlexGroup volume with eight member volumes, each member volume is 10TB in size. These member volume sizes are intended to be inconsequential in most workload cases since 10TB is a pretty large size to work with, but it's important to know what the file sizes of the workload are to help plan the capacity accordingly. For example, if you know your workload has 500GB files, then 10TB member volumes are fine, whereas 1TB member volumes would be problematic.

Figure 47) FlexGroup volumes—member sizes versus FlexGroup volume capacity.



Initial space consumption on new FlexGroup volumes

Each FlexGroup member volume sets aside a small amount of space (around 50MB) for internal use. When a member volume is sized to the minimum of 100GB, the used space is around 0.05%, which is negligible to ONTAP. However, used space still shows up in the output of empty FlexGroup volumes, so this is something to keep in mind as a nonissue when deploying a FlexGroup volume.

For example:

```
cluster::*> vol show -vserver DEMO -volume fgautogrow* -fields used
vserver volume           used
------- ---------------- -------
DEMO    fgautogrow__0006 57.48MB
DEMO    fgautogrow__0008 57.48MB
DEMO    fgautogrow__0001 57.50MB
DEMO    fgautogrow__0004 57.50MB
DEMO    fgautogrow__0005 57.52MB
DEMO    fgautogrow__0007 57.52MB
DEMO    fgautogrow__0002 57.57MB
DEMO    fgautogrow__0003 57.57MB
DEMO    fgautogrow          460MB
```

Shrinking a FlexGroup volume

Before ONTAP 9.6, FlexGroup volumes did not support shrinking of the total FlexGroup volume footprint, but volume autosize was able to shrink individual member volumes. Even with volume shrink support, avoid oversizing the volumes at the initial creation. If you make them too large, your administration options might be limited when you need to grow capacity later and you have to add new member volumes, because they would need to be added in identical member volume sizes.

Snapshot copies and Snapshot reserve

ONTAP Snapshot copies are designed to create a point-in-time copy of a filesystem without using any space until data is overwritten. When data is changed or deleted, the space is marked as removed from the active file system (AFS) and ONTAP redirects pointers to the Snapshot copy. After this is done, the Snapshot copy shows the space used.

Snapshot reserve

By default, volumes assign a 5% Snapshot copy reservation. This means that if you provision a volume that is 100TB, then 5% of that volume (5TB) is allocated for Snapshot copies. As a result, the volume size output (such as `volume show` and `df`) on the storage system show 95TB of usable space in that scenario, which is also what clients see as available space. In a FlexGroup volume, the Snapshot reserve is set at the FlexGroup level, but it is applied to each member volume. Although the Snapshot reserve on a 100TB FlexGroup volume might be 5TB, the individual member volumes share that 5TB evenly. If there are eight member volumes, then each has 640GB of capacity reserved for Snapshot copies.

Snapshot spill

If Snapshot copy use grows beyond the size of the snapshot reserve, then that space starts to use capacity from the AFS. ONTAP reports the space used by Snapshot copies as being greater than 100%, and the total used space in the volume increases, even if no physical data exists in the volume.

For an example, see the section “Snapshot spill example.”

Snapshot spill and snapshot scanners

ONTAP performs periodic scans of volumes to check for used and changed blocks in WAFL. This is used to properly calculate the used space in the volume. You can learn more about these scanners in the Knowledge Base article [Individual snapshot size doesn't match the DF output for that volume.]

These scanners are low priority jobs that defer to production workloads, so the speed which they complete can depend on the load on the system. You can view these scanners as follows:


```
Mon Dec 21 19:24:30 -0600 [CLUSTER: scan_ownblocks_calc_wkr: wafl.scan.ownblocks.done:info]:
Completed block ownership calculation on volume vol__0003@vserver:ebf4370e-208a-11eb-921e-
d039ea2020c8. The scanner took 202 ms.
```

If volumes are small (for example, 100GB), they can fill faster. In these cases, it is possible for incoming data writes to perform faster than the scanners, especially on faster AFF systems. As a result, the capacity reporting doesn't react in time for storage administrators to address capacity needs by adding space or deleting snapshots. In these cases, the volume might report `out of space` if volume autogrow is not enabled, because there is no space left to borrow from other member volumes.

Snapshot spill remediation tips

Snapshot spill is a normal function of how Snapshot copies work in ONTAP when the snapshot reserve is overrun. The effect of snapshot spill and how quickly it can grow depends on the total size of the volume. Smaller volumes have lower total snapshot reserve space and are more susceptible to snapshot spill. To minimize the effect of snapshot spill, you can do one or more of the following:

- Increase the total FlexGroup volume size, which also increases the total available snapshot reserve and makes snapshot spill less common.
- Avoid creating small FlexGroup volumes if you are using Snapshot copies.
- Use larger snapshot reservation percentages if you have more data churn; clients only see available space in the Active File System and do not see reserved snapshot space.
- Delete larger snapshots when possible. FlexGroup volumes do not currently support snapshot autodelete, so you need to delete Snapshot copies manually or via script.
- Set snap reserve to 0; this causes the snapshot used space to be reflected in the AFS immediately.

Volume autosize (autogrow and autoshrink)

In ONTAP 9.3, support for volume autogrow was added for 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. Applying volume autogrow to a FlexGroup volume is done in the same way as with a FlexVol volume; you specify thresholds and configure different options. Details can be found in the product documentation.

The configuration options are the same as a FlexVol and include the following:

```
[-max-autosize {<integer>[KB|MB|GB|TB|PB]}] - Maximum Autosize
This parameter allows the user to specify the maximum size to which a volume can grow. The
default for volumes is 120% of the volume size. If the value of this parameter is invalidated by
manually resizing the volume, the maximum size is reset to 120% of the volume size. The value for
-max-autosize cannot be set larger than the platform-dependent maximum volume size. If you
specify a larger value, the value of -max-autosize is automatically reset to the supported
maximum without returning an error.

[-min-autosize {<integer>[KB|MB|GB|TB|PB]}] - Minimum Autosize
This parameter specifies the minimum size to which the volume can automatically shrink. If the
volume was created with the grow_shrink autosize mode enabled, then the default minimum size is
equal to the initial volume size. If the value of the -min-autosize parameter is invalidated by a
manual volume resize, the minimum size is reset to the volume size.

[-autosize-grow-threshold-percent <percent>] - Autosize Grow Threshold Percentage
This parameter specifies the used space threshold for the automatic growth of the volume. When
the volume's used space becomes greater than this threshold, the volume will automatically grow
unless it has reached the maximum autosize.

[-autosize-shrink-threshold-percent <percent>] - Autosize Shrink Threshold Percentage
This parameter specifies the used space threshold for the automatic shrinking of the volume. When
the amount of used space in the volume drops below this threshold, the volume will shrink unless
it has reached the specified minimum size.

[-autosize-mode {off|grow|grow_shrink}] - Autosize Mode
This parameter specifies the autosize mode for the volume. The supported autosize modes are:
```



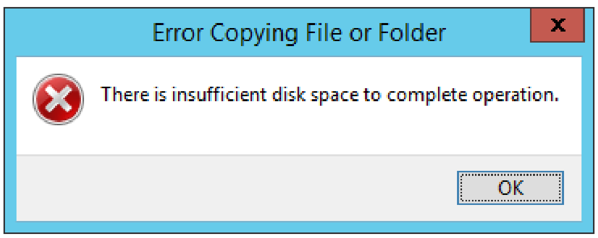
```
o   off - The volume will not grow or shrink in size in response to the amount of used space.
o   grow - The volume will automatically grow when used space in the volume is above the grow
threshold.
o   grow_shrink - The volume will grow or shrink in size in response to the amount of used space.

By default, -autosize-mode is off for new volumes, except for DP mirrors, for which the default
value is grow_shrink. The grow and grow_shrink modes work together with Snapshot autodelete to
automatically reclaim space when a volume is about to become full. The volume parameter -space-
mgmt-try-first controls the order in which these two space reclamation policies are attempted.

[-autosize-reset [true]] } - Autosize Reset
This allows the user to reset the values of autosize, max-autosize, min-autosize, autosize-grow-
threshold-percent, autosize-shrink-threshold-percent and autosize-mode to their default values.
For example, the max-autosize value will be set to 120% of the current size of the volume.
```

How volume autosize works in a FlexGroup volume

When a FlexGroup volume has a member that cannot honor a write, ONTAP returns an Insufficient Space (`ENOSPC`) error to the client in versions before ONTAP 9.6.



Starting in ONTAP 9.6, instead of sending an error to clients, ONTAP instead pauses the operation briefly while it searches other member volumes for available free space. If there is available free space, then the member volume grows, while shrinking another member volume by the same amount; this maintains the same total FlexGroup volume size. This is known as Elastic sizing and is covered in more detail later.

If volume autosize is enabled, rather than pausing and borrowing space from another member volume and keeping the same total capacity (which adds some latency to the workload), volume autogrow instead grows member volumes by a configured capacity when a capacity threshold has been reached. This increases the total FlexGroup volume size by the amount the member volume grew.

For example, if you had a 10TB FlexGroup volume and a member volume automatically grew by 1TB, then you now have an 11TB FlexGroup volume with volume autogrow.

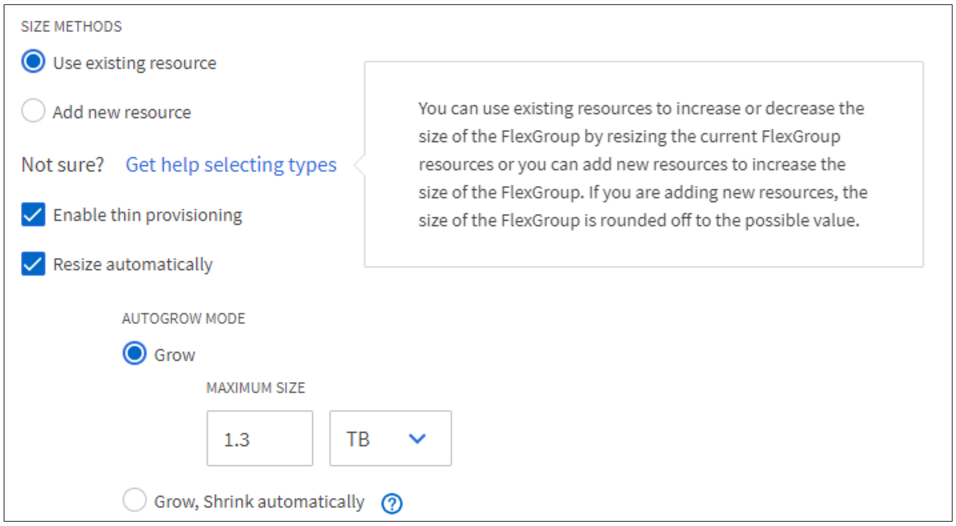
Volume autoshrink

In addition to autogrow, the volume autosize feature also has an autoshrink functionality. This can be enabled or disabled via the `-autosize-mode` option. Autoshrink allows ONTAP to shrink a member volume back to a normal size if the capacity used reaches the configured `-autosize-shrink-threshold-percent` value. If that 11TB FlexGroup volume has 1TB free up in the member volume that grew, it would shrink by whatever you have configured, but no smaller than the original volume size by default.

How to enable volume autosize

Enabling volume autosize can be done in several different ways.

1. You can use ONTAP System Manager during volume creation or use Edit:



2. You can use the `volume autosize` command via the CLI:

```
cluster::> volume autosize -vserver DEMO -volume fgautogrow -maximum-size 100g -grow-threshold-
percent 80 -autosize-mode grow
```

- You can check that it's enabled with the following commands:

```
cluster::> vol autosize -vserver DEMO -volume fgautogrow
Volume autosize is currently ON for volume "DEMO:fgautogrow".
The volume is set to grow to a maximum of 100g when the volume-used space is above 80%.
Volume autosize for volume 'DEMO:fgautogrow' is currently in mode grow.
```

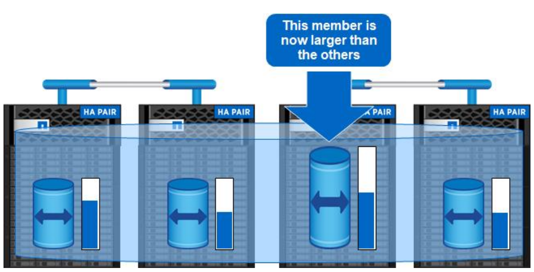
3. You can use the `volume modify` command via the CLI:

```
cluster::> volume modify -vserver DEMO -volume fgautogrow -autosize-mode grow_shrink -autosize-
grow-threshold-percent 95% -autosize-shrink-threshold-percent 50% -max-autosize 1.20PB -min-
autosize 1PB
```

After a member volume has been grown through autogrow, there is an imbalance of member volume available size/allocation. This is by design.

In addition, the total FlexGroup size is now larger due to the larger member volume size. If you do not want the total FlexGroup volume size to grow, you can leave volume autogrow disabled and instead use the other capacity management feature in ONTAP, such as elastic sizing (ONTAP 9.6 and later) and proactive resizing (ONTAP 9.8 and later).

Figure 48) Member volume size allocation after a `volume autosize` operation.



Volume autosize interaction with elastic sizing

Starting in ONTAP 9.6, elastic sizing provides a way for file writes to complete in nearly filled member volumes by borrowing space from other member volumes. This takes place without growing the total size of the FlexGroup volume. As space is freed up in the filled member volume, elastic sizing begins to normalize the member volume sizes back to their original capacities.

Volume autosize on the other hand adds space to the total size of the FlexGroup volume by automatically growing a member volume when it reaches a space threshold.

Elastic sizing is enabled for FlexGroup volumes by default. If you enable volume autosize in ONTAP 9.6 and ONTAP 9.7, elastic sizing no longer takes effect for that volume. ONTAP 9.8 and later enables the use of volume autosize with elastic sizing.

Volume autosize interaction with proactive resizing

Proactive resizing is available in ONTAP 9.8 and later and is covered in the section "Proactive resizing."

Volume autosize works in conjunction with proactive resizing. Proactive resizing adjusts member volume capacities, and, if a capacity threshold for autosize is reached, ONTAP applies volume autosize. If volume autosize is disabled, then proactive resizing works on its own. For more detail on how proactive resizing works with autosizing enabled, see the section "Proactive resizing behavior — volume autosize enabled."

Differences between FlexVol and FlexGroup volume storage efficiency

FlexVol and FlexGroup volumes both leverage the same storage efficiency features in ONTAP, but due to the way these features operate and the architectures of the volume types, there are distinct differences in how efficiencies operate and how much space savings you might see.

Storage efficiency feature comparison: FlexVol and FlexGroup volumes

Table 15 shows a list of storage efficiency features and the volume style support (how it is applied) as of ONTAP 9.9.1 and later. Most of the storage efficiency features are applied at the volume level for FlexGroup volumes except for inactive data compression, which is applied at the member volume level. For more information, see the product documentation for using storage efficiencies in ONTAP.

Table 15) ONTAP storage efficiency support matrix: FlexVol and FlexGroup volumes.

Efficiency feature	FlexVol support?	FlexGroup support?
Inline deduplication*	Yes	Yes
Inline data compaction	Yes	Yes
Inline data compression	Yes	Yes
Inline cross-volume deduplication*	Yes	Yes
Post-process volume deduplication	Yes	Yes
Background cross-volume deduplication*	Yes	Yes
Inline adaptive compression	Yes	Yes
Compression algorithm modification (background only)	Yes	Yes
Temperature Sensitive Storage Efficiency (TSSE)*	Yes	Yes
Inactive data compression	Yes	Yes**
Post-process data compression***	Yes	Yes

*AFF/Flash Pool supported feature only.

**Applied at the member volume level.

***Not supported for AFF.

Storage efficiency operations

ONTAP performs storage efficiency operations in two main ways: inline (such as during the initial data ingest) and post-process (after the data has already been written). Inline efficiencies are designed to be opportunistic — meaning, ONTAP applies efficiencies when it does not affect performance. If efficiencies are not applied inline, then they are applied in the post-process phase.

Inline efficiencies are only supported for AFF platforms, Flash Pool aggregates, and systems with the data protection optimization licenses. FAS systems rely on post-process storage efficiencies for space savings.

Storage efficiency domains

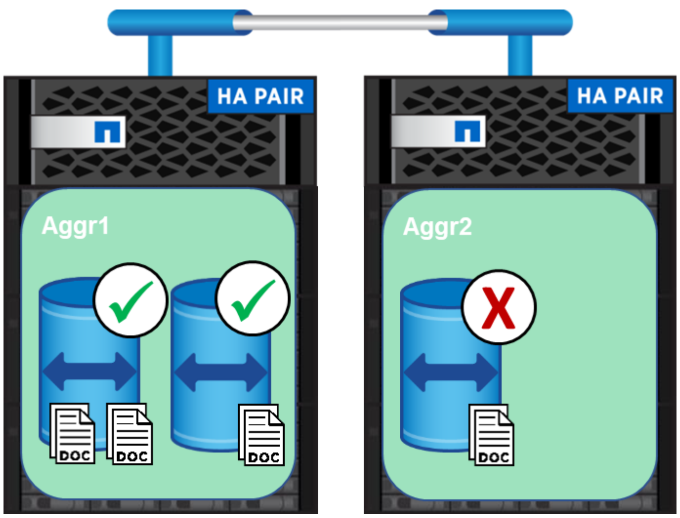
Storage efficiency operates in several different domains in ONTAP:

- File-level (compression—inline and post-process)
- Volume-level (deduplication—inline and post-process)
- Aggregate-level (deduplication—inline and post-process; inline data compaction)

This means that efficiency operations can have different behaviors based on which storage efficiency technologies are in use. For deduplication, ONTAP shows space savings at the volume level and the aggregate level. Therefore, if you have two copies of the same files in the storage system, deduplication reduces the space usage if both files either live in the same volume or the same aggregate (depending on the storage efficiency policy applied to the volume).

If the files reside in different volumes and/or different aggregates (again, depending on the volume efficiency policy in use), then ONTAP is not able to deduplicate the data because the files reside outside of the storage efficiency domains that ONTAP uses.

Figure 49) Storage efficiency domains—when will deduplication be effective?



- **Inline data compression.** Takes place at the file level; therefore, compression savings should be similar for both FlexVol volumes and FlexGroup volumes because there is no concept of a storage efficiency domain to consider.
- **Data compaction.** Occurs at the WAFL level in ONTAP, where data that is written to storage that doesn't completely fill a 4K block is compacted into a single 4K block. For instance, if two 2K blocks are written to ONTAP, they combine to a single 4K block. Data compaction savings are seen at the aggregate level.

FlexVol volume efficiency

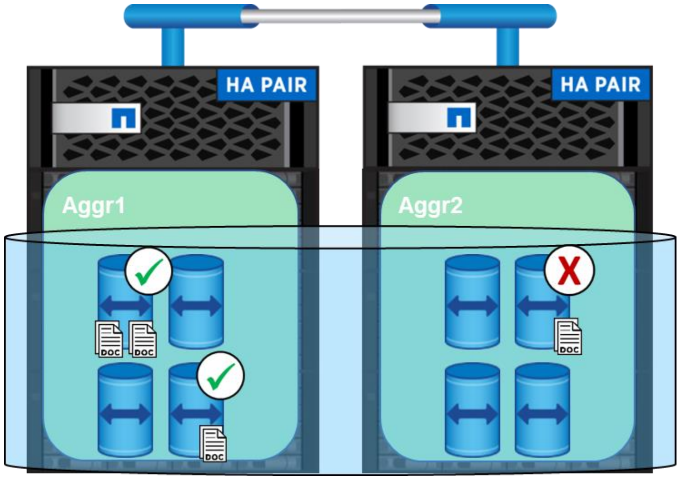
- A FlexVol volume only resides on a single node and/or aggregate at any given time so it is always in an eligible storage efficiency domain. Any file that exists within the FlexVol volume has volume-level efficiencies always applied.

- If two copies of the same file reside in different FlexVol volumes in the same aggregate, then storage efficiencies are applied at the aggregate level, provided cross-volume deduplication is enabled on the system.
- If two copies of the same file reside in different FlexVol volumes in different aggregates, then storage efficiency savings are not applied for those files.
- If the FlexVol volumes in use have NVE enabled, then the identical copies of the files are no longer identical, because there is an encryption layer that uniquifies the data, which negates deduplication effects.
- If the FlexVol volumes in use are in an aggregate with NVE, then all volumes in that aggregate share the same encryption keys and storage efficiencies will be effective.

FlexGroup volume efficiency

- A FlexGroup volume can contain one or more FlexVol member volumes and span one or more aggregates in an ONTAP cluster. As a result, the effectiveness of storage efficiencies in a FlexGroup volume depends solely on the layout of the FlexGroup volume. In general, storage efficiency savings in a FlexGroup volume that spans multiple aggregates are less than a FlexVol volume's savings.
- Because data ingest and layout (where the files land when created) is controlled by ONTAP, storage efficiency effectiveness can vary greatly depending on which data lives where in the volume. After data is placed in a FlexGroup volume, it stays where it lands and does not move retroactively unless a file-based copy is performed back in to the FlexGroup volume.
- If a FlexGroup volume has a single member volume and lives on one aggregate, then, effectively, it is a FlexVol volume, and the storage efficiencies are roughly the same as a FlexVol volume because the data is still considered to be in the volume storage efficiency domain.
- If a FlexGroup volume has more than one member volume and identical copies of a single file reside on the same member volume, then it has similar storage efficiencies as a FlexVol volume because the files are in the volume storage efficiency domain.
- If a FlexGroup volume has more than one member volume in the same aggregate and identical copies of a single file reside on different member volumes, then there are only storage efficiencies for aggregate-level deduplication and no efficiency savings for the volume storage efficiency domain. Reporting of savings are at the aggregate level rather than the FlexGroup volume level.
- If a FlexGroup volume has more than one member volume in the different aggregates and identical copies of a single file reside on different member volumes, then storage efficiency effectiveness depends solely on where those files live in the FlexGroup volume. If there are commonalities in the member volume locations or aggregate locations, you will see some storage efficiency savings. If each identical copy of the file happens to live in a unique member volume in unique aggregates, you will not see any storage efficiency savings for deduplication.
- Inline efficiencies do not gain extra efficiencies for FlexGroup volumes compared to post-process efficiencies because they must follow the same rules for efficiency domains (volume and aggregate).
- The same volume encryption rules apply for a FlexGroup volume; volume level encryption uniquifies data and negate efficiencies. Aggregate-level encryption maintains a common key and maintains cross-volume deduplication.

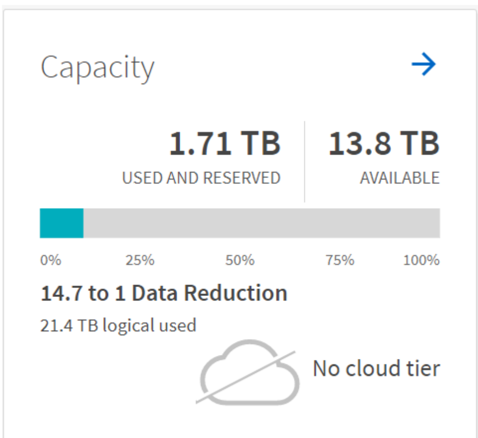
Figure 50) Storage efficiency domains with FlexGroup volumes



Viewing storage efficiency savings

There are several ways to view storage efficiency savings in ONTAP. The easiest, most common method is through ONTAP System Manager. The initial dashboard displays savings, and you can click the arrow to navigate to a more granular view of the efficiencies.

Figure 51) Storage efficiency savings — ONTAP System Manager



In addition, you can view storage efficiencies in the ONTAP CLI by using the following commands:

- `volume show -vserver SVM -volume volname`.

  Use an asterisk with the volume name in diagnostic privilege mode to include all FlexGroup member volumes or use `-is-constituent true`; and look for `sis-space-saved` and `sis-space-saved-percent` for total efficiency savings.
- `aggregate show -aggregate aggrname1,aggrname2`

  Shows cross-volume deduplication and data compaction savings for all volumes in the aggregate.
- `aggregate show-efficiency -aggregate aggrname`

  Shows the summary of efficiencies and ratios for the aggregate.
- `volume efficiency show -volume volname`

  Shows efficiency progress/status and savings for the volume.

- `aggregate show-space -aggregate-name aggrname -instance`

  Shows the summary of physical and logical used space.

  Note: `sis-space-saved` is an aggregation of all storage efficiency savings (compression plus deduplication plus data compaction). To see more granular views of how space is being saved, see the `dedupe-space-saved`, `compression-space-saved` and `data-compaction-space-saved` (aggregate only) fields in the CLI. Cross-volume deduplication is reported with `dedupe-space-saved` at the aggregate level.

FlexVol and FlexGroup efficiencies compared

In the following example, an approximately 36GB, 16,000 file source dataset (in an ONTAP 9.8 system) is copied to a FlexVol volume and a FlexGroup volume (eight member volumes spanning two aggregates on two different nodes in a cluster) on an ONTAP 9.9.1 system to show the difference in effectiveness of storage efficiencies. The dataset is a mix of text files and user data (Word documents, packet trace files, and so on) and has a roughly 2:1 data reduction ratio on the source volume.

Source dataset details: ONTAP 9.8 FlexVol volume

Both the FlexGroup and FlexVol volumes have volume efficiencies (inline and post-process) enabled.

```
cluster::*> vol show -vserver DEMO -volume files -fields used,files-used,dedupe-space-
saved,compression-space-saved,logical-used
vserver volume used    files-used dedupe-space-saved compression-space-saved logical-used
------- ------ ------- ---------- ------------------ ----------------------- ------------
DEMO    files  18.61GB 16152      15.62GB            1.51GB                  35.75GB

cluster::*> vol show -vserver DEMO -volume files -fields sis-space-saved-percent,compression-
space-saved-percent
vserver volume sis-space-saved-percent compression-space-saved-percent
------- ------ ----------------------- -------------------------------
DEMO    files  48%                     4%
```

Volume configuration details

The storage efficiency comparisons are performed on a few different destination volume types/configurations to illustrate how storage efficiency savings can vary.

The configurations include the following:

- Standard FlexVol volume
- FlexGroup volume (single member volume; single aggregate)
- FlexGroup volume (four member volumes; single aggregate)
- FlexGroup volume (eight member volumes; two aggregates)

The FlexGroup volume configurations used are to illustrate the differences in efficiencies compared to a FlexVol volume but are not always the ideal way to configure FlexGroup volumes. Keep in mind the following cautions/caveats.

FlexGroup volume: Single member volume configuration

A FlexGroup volume with a single member volume is not a common, nor recommended, deployment for FlexGroup volumes, but it is a way to illustrate how storage efficiency savings differ based on the volume layout. In this case, a FlexGroup volume with a single member volume gets many of the benefits of deduplication that you see with a FlexVol volume because it is using the volume-level efficiency domain.

FlexGroup volume: Four member volumes, one aggregate configuration

To ge better overall storage efficiency savings with a FlexGroup volume, it is possible to host all member volumes in the FlexGroup on the same aggregate. The following are the possible downsides to this approach:

- Limitation of possible FlexGroup volume size (can only be as large as the aggregate)
- Potential performance limitations (less CPU and disk that you can use for the workload)
- Overuse of disk/aggregate performance
- Less effective inline efficiencies due to disk backpressure

For these reasons, you should weigh the drawbacks of hosting the FlexGroup volume on a single aggregate against the benefits of added efficiency savings.

The following example shows the space savings from deduplication on a FlexGroup volume before running any post-process efficiencies—only inline deduplication is applied here. At the volume level, only 292.8MB is saved with deduplication. This is because ONTAP only reports savings from the volume-level efficiency domains here. Aggregate-level savings (such as with cross-volume deduplication) are reported at the aggregate level.

```
cluster::*> vol show -vserver DEMO -volume fgsingleaggr* -fields used,dedupe-space-saved,logical-used
vserver volume             used    dedupe-space-saved logical-used
------- ------------------ ------- ------------------ ------------
DEMO    fgsingleaggr       37.60GB 292.8MB            37.88GB
DEMO    fgsingleaggr__0001 4.91GB  7.43MB             4.92GB
DEMO    fgsingleaggr__0002 15.73GB 67.15MB            15.80GB
DEMO    fgsingleaggr__0003 16.17GB 212.3MB            16.38GB
DEMO    fgsingleaggr__0004 802.1MB 5.89MB             808.0MB

cluster::*> vol show -vserver DEMO -volume fgsingleaggr* -fields dedupe-space-saved-percent
vserver volume             dedupe-space-saved-percent
------- ------------------ -------------------------
DEMO    fgsingleaggr       1%
DEMO    fgsingleaggr__0001 0%
DEMO    fgsingleaggr__0002 0%
DEMO    fgsingleaggr__0003 1%
DEMO    fgsingleaggr__0004 1%
```

To see space savings when dealing with multiple FlexGroup member volumes in the same aggregate, look at the aggregate efficiency domain savings. In this example, before data is copied to the FlexGroup volume on the aggr named `tme_a300_efs02_02_SSD_1`, the following shows the space savings:

```
cluster::*> aggr show -aggregate tme_a300_efs02_02_SSD_1 -fields size,usedsize,physical-used,data-compaction-space-saved,sis-space-saved
aggregate               size   usedsize physical-used data-compaction-space-saved sis-space-saved
----------------------- ------ -------- ------------- --------------------------- ---------------
tme_a300_efs02_02_SSD_1 7.75TB 725.2GB  775.7GB       16.17GB                     16.17GB
```

After the files are copied, you can see that inline cross-volume deduplication savings are approximately 16GB.

```
cluster::*> aggr show -aggregate tme_a300_efs02_02_SSD_1 -fields size,usedsize,physical-used,data-compaction-space-saved,sis-space-saved
aggregate               size   usedsize physical-used data-compaction-space-saved sis-space-saved
----------------------- ------ -------- ------------- --------------------------- ---------------
tme_a300_efs02_02_SSD_1 7.75TB 763.0GB  792.5GB       32.26GB                     32.26GB
```

Note: Cross-volume deduplication is not allowed when NVE is in use.

In the preceding example, a volume with approximately 38GB of total space saves approximately 16.7GB. That is a savings of roughly 44%, which is the savings gained for the FlexVol volume.

FlexGroup volume: Eight member volumes, two aggregate configuration

This FlexGroup volume is created with the default settings across two aggregates (four member volumes per aggregate, eight member volumes total).

The following example shows the deduplication savings on the default, two-node FlexGroup volume after copying data. Note how much lower the total deduplication savings are compared to the FlexVol volume, single-member FlexGroup volume, and even the single aggregate FlexGroup volume (see Table 16) because it spans both volume-level and aggregate-level deduplication domains.

```
cluster::*> vol show -vserver DEMO -volume fgfiles* -fields used,dedupe-space-saved,logical-used
vserver volume        used    dedupe-space-saved logical-used
------- ------------- ------- ------------------ ------------
DEMO    fgfiles       38.09GB 1.80GB             39.89GB
DEMO    fgfiles__0001 8.88GB  275.1MB            9.15GB
DEMO    fgfiles__0002 4.08GB  195.9MB            4.27GB
DEMO    fgfiles__0003 4.09GB  187.8MB            4.27GB
DEMO    fgfiles__0004 4.85GB  290.2MB            5.13GB
DEMO    fgfiles__0005 4.03GB  243.0MB            4.27GB
DEMO    fgfiles__0006 4.07GB  204.3MB            4.27GB
DEMO    fgfiles__0007 4.03GB  238.0MB            4.27GB
DEMO    fgfiles__0008 4.06GB  206.2MB            4.26GB

cluster::*> vol show -vserver DEMO -volume fgfiles* -fields dedupe-space-saved-percent
vserver volume        dedupe-space-saved-percent
------- ------------- --------------------------
DEMO    fgfiles       5%
DEMO    fgfiles__0001 3%
DEMO    fgfiles__0002 4%
DEMO    fgfiles__0003 4%
DEMO    fgfiles__0004 6%
DEMO    fgfiles__0005 6%
DEMO    fgfiles__0006 5%
DEMO    fgfiles__0007 5%
DEMO    fgfiles__0008 5%
```

Cross-volume deduplication helps but you are still limited to the aggregate deduplication domain. The following output shows the aggregate space usage and deduplication savings for each aggregate that this FlexGroup volume spans, before the data copy:

```
cluster::*> aggr show -aggregate tme_a300_efs02* -fields size,usedsize,sis-space-saved
aggregate               size   usedsize sis-space-saved
----------------------- ------ -------- ---------------------------
tme_a300_efs02_01_SSD_1 7.75TB 1.01TB   10.49GB
tme_a300_efs02_02_SSD_1 7.75TB 726.8GB  10.58GB
```

And following is the comparison after the data copy. Note that approximately 2GB per aggregate (4GB total) is saved. This is a long way from the efficiency savings achieved in the FlexGroup volume spanning the single aggregate but better than simply relying on volume-level deduplication.

```
cluster::*> aggr show -aggregate tme_a300_efs02* -fields size,usedsize,sis-space-saved
aggregate               size   usedsize sis-space-saved
----------------------- ------ -------- ---------------
tme_a300_efs02_01_SSD_1 7.75TB 1.03TB   12.36GB
tme_a300_efs02_02_SSD_1 7.75TB 743.6GB  12.59GB
```

Note: `sis-space-saved` does not only show deduplication savings; it also shows all combined savings on the aggregate, so some of the savings might be due to auto adaptive data compression.

FlexGroup volume (two aggregates, default member volume count): Compression savings only

On a FlexGroup volume (across two aggregates) using adaptive compression, this is the result of the inline compression savings.

```
cluster::*> vol show -vserver DEMO -volume fgfiles* -fields used,compression-space-saved
vserver volume        used    compression-space-saved
------- ------------- ------- -----------------------
```



```
DEMO     fgfiles        36.91GB  1.53GB
DEMO     fgfiles__0001  6.78GB   1.53GB
DEMO     fgfiles__0002  4.18GB   344KB
DEMO     fgfiles__0003  4.56GB   0B
DEMO     fgfiles__0004  4.20GB   0B
DEMO     fgfiles__0005  4.69GB   304KB
DEMO     fgfiles__0006  4.17GB   0B
DEMO     fgfiles__0007  4.18GB   0B
DEMO     fgfiles__0008  4.15GB   0B
```

Note that most of the compression savings takes place in a single member volume; this is because the file (or files) that are most compressible are located there. In this case, a 4GB packet trace file is the compressed file. No other files in that volume exceed a few megabytes in size.

The key takeaway here is that, unlike deduplication, compression is unaffected by the layout of a FlexGroup volume and is instead associated with how compressible the individual files in the volume are. You can expect roughly the same compression rates on a FlexGroup volume that you see with a FlexVol volume.

Storage efficiency comparison: Deduplication

Table 16 offers a side-by-side comparison of deduplication savings on the different volume configurations.

Table 16) Storage efficiency comparisons: Deduplication.

Volume configuration	Total deduplication space saved*	Deduplication percentage saved*
FlexVol	15.63GB	44%
FlexGroup; single member volume	15.62GB	44%
FlexGroup; 4 member volumes, single aggregate**	16.17GB	44%
FlexGroup; 8 member volumes, two aggregates, two nodes**	1.8GB	5%

*After inline and post-process efficiencies are applied.

**Aggregate level savings (includes compression and compaction).

Note: For a complete summary of storage efficiency savings (compaction, compression, and deduplication), view storage efficiencies at the aggregate level.

FlexVol and FlexGroup storage efficiencies: Conclusion

Although you can apply storage efficiency policies at the volume level, the space savings that you achieve in the FlexVol and FlexGroup volumes mostly depend on the FlexGroup volume layout, data types, data placement, and other factors listed in the preceding sections. As a result, in nearly all cases, a FlexGroup volume that spans multiple aggregates in a cluster show less space savings than a single FlexVol volume due to how deduplication works in ONTAP. Additionally, some of the space savings are reported at the aggregate level, so if space savings do not align between FlexVol and FlexGroup volumes, compare the aggregate level efficiencies as well.

In summary, for FlexGroup volumes be aware of the following:

- In general, deduplication savings are effective on FlexGroup volumes compared to FlexVol volumes.
- Volume-level deduplication only achieves savings when identical blocks are in the same member volume (which is controlled by ONTAP data placement).
- Aggregate-level deduplication achieves savings when identical blocks are in the same aggregate, which bodes well for FlexGroup volumes until member volumes span different aggregates. These savings are achieved with aggregate storage efficiency commands rather than volume-level commands.

- Inline data compaction and data compression savings are nearly identical for FlexGroup and FlexVol volumes, but you must review both the volume and aggregate savings to see the space reduction.

Elastic sizing

Files written to a FlexGroup volume live in individual member volumes. They do not stripe across member volumes, so, if a file is written and grows over time or a large file is written to a FlexGroup volume in ONTAP 9.5 and prior, that write might fail because of lack of space in a member volume.

There are a few reasons why a member volume might fill up:

- You try to write a single file that exceeds the available space of a member volume. For example, a 10GB file is written to a member volume with 9GB available.
- If a file is appended over time, it eventually fills up a member volume—for example, if a database resides in a member volume.
- Snapshot copies eat into the active file system space available.

FlexGroup volumes do a good job of allocating space across member volumes, but, if a workload anomaly occurs, it can have a negative effect. For example, a volume is composed of 4,000 files but then a user zips some files up and creates a giant single tarball file.

One solution is to grow volumes, either manually or by using volume autogrow. Another solution is to delete data. However, administrators often don't see capacity issues until it's too late.

For example, a FlexGroup volume can be hundreds of terabytes in size, but the underlying member volumes and their free capacities are what determine the space available for individual files. If a 200TB FlexGroup volume has 20TB remaining (10% of the volume), the amount of space available for a single file to write is not 20TB; instead, it is closer to 20TB/[number of member volumes in a FlexGroup], provided all member volumes in the FlexGroup volume have evenly distributed capacities.

In a two-node cluster, a FlexGroup volume that spans both nodes is likely to have 16 member volumes. That means if 20TB are available in a FlexGroup volume, the member volumes would have 1.25TB available. Before ONTAP 9.6, any single file that exceeded 1.25TB could not write to a FlexGroup volume without volume autogrow enabled.

Starting in ONTAP 9.6, the elastic sizing feature helps avoid "out of space" errors in this scenario. This feature is enabled by default and does not require administrator configuration or intervention.

Elastic sizing is not a panacea; it is intended to be reactive to prevent file writes from failing. Capacity management to keep adequate member volume space available is still necessary, even with elastic sizing enabled.

Elastic sizing: an airbag for your data

One of our FlexGroup developers refers to elastic sizing as an airbag: it is not designed to stop you from getting into an accident, but it does help soften the landing when it happens. In other words, it's not going to prevent you from writing large files or running out of space, but it is going to provide a way for those writes to complete. In fact, in some cases, the peace of mind you get from elastic sizing can cause you to ignore capacity issues until the entire FlexGroup is out of space or until a performance issue occurs.

Here's how it works at a high level:

1. When a file is written to ONTAP, the system has no idea how large that file will become. The client doesn't know. The application usually doesn't know. All that's known is "hey, I want to write a file."
2. When a FlexGroup volume receives a write request, it is placed in the best available member based on various factors, such as free capacity, inode count, time since last file creation, and so on.

3. When a file is placed, since ONTAP doesn't know how large a file will become, it also doesn't know if the file is going to grow to a size that's larger than the available space. So, the write is allowed as long as we have space to allow it.
4. If/when the member volume runs out of space, right before ONTAP sends an `out of space` error to the client, it queries the other member volumes in the FlexGroup volume to see if there's any available space to borrow. If there is, ONTAP adds 1% of the volume's total capacity in increments (in a range of 10MB to 10GB) to the volume that is full (while taking the same amount from another member volume in the same FlexGroup volume), and then the file write continues.
5. During the time ONTAP is looking for space to borrow, that file write is paused. This appears to the client as a performance issue, usually as latency. But the overall goal here isn't to finish the write fast—it's to allow the write to finish at all. Usually, a member volume is large enough to provide the 10GB increment (1% of 1TB is 10GB), which is often more than enough to allow a file creation to complete. In smaller member volumes, the effect on performance could be greater, because the system needs to query to borrow space more often due to smaller increments, and files don't have to be as large to fill the volume.
6. The capacity borrowing maintains the overall size of the FlexGroup volume. For example, if your FlexGroup volume is 40TB in size, it remains 40TB.
7. After files are deleted or volumes are grown and space is available in that member volume again, ONTAP re-adjusts the member volumes back to their original sizes to maintain an evenness in space, but only when a member volume's capacity is within 75% of the average free space of the other member volumes in the FlexGroup.

Figure 52) File write behavior before elastic sizing.

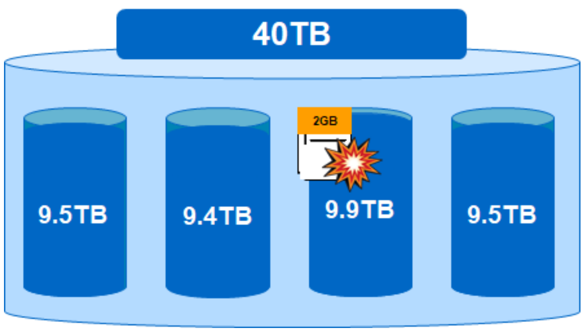
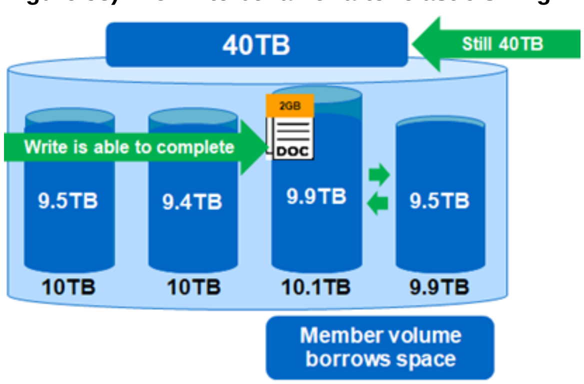


Figure 53) File write behavior after elastic sizing.



Ultimately, elastic sizing helps mitigate file write failures in full member volumes and removes some the administrative overhead of managing capacity, because a full member is no longer an urgent event. ONTAP can still write files as long as there is available free space in other member volumes.

However, due to its reactive nature, it's best to upgrade to ONTAP 9.8 to make use of proactive resizing.

When to use volume autogrow versus elastic sizing

When volume autogrow is enabled on a FlexGroup volume, elastic sizing is disabled for that volume in ONTAP 9.6 and ONTAP 9.7, because the two features are essentially redundant.

However, there are some differences in how they work and when you'd want to use one over the other.

- Volume autogrow should be used when the total capacity of the FlexGroup volume can be grown to accommodate new data being written to it.
- Elastic sizing is enabled by default and should be used when the total size of the FlexGroup volume should not be allowed to grow past the specified capacity.

Performance effect of elastic sizing

Each time a file write must pause for ONTAP to find more space in the FlexGroup volume, client latency occurs. The amount of latency seen for a write operation to a file depends on the number of times the write must pause to find more space. For example, if a member volume has just 10GB available, but a 100GB file is being written, then elastic sizing causes the write to pause a number of times to allow the write to complete. That number is determined by the member volume total size, which can be anywhere from 10MB to 10GB.

The following example shows a test in which a file was copied to a FlexGroup volume. In the first test, the FlexGroup constituent was not large enough to hold the file, so elastic sizing was used. The 6.7GB file took around 2 minutes to copy:

```
[root@centos7 /]# time cp Windows.iso /elastic/
real    1m52.950s
user    0m0.028s
sys     1m8.652s
```

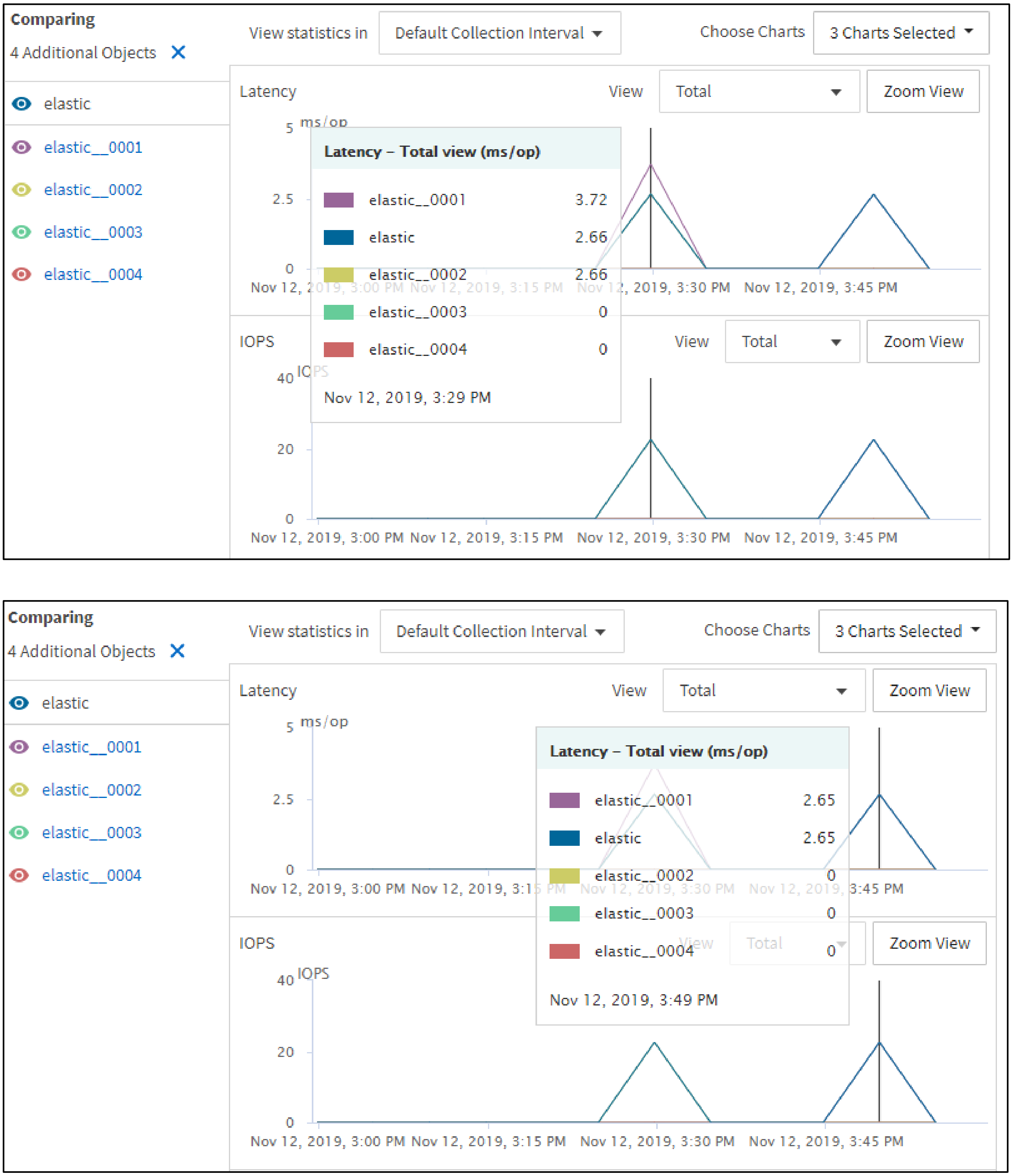
When the FlexGroup constituent volume was large enough to avoid elastic sizing, the same copy took 15 seconds less:

```
[root@centos7 /]# time cp Windows.iso /elastic/
real    1m37.233s
user    0m0.052s
sys     0m54.443s
```

That shows there can be a real latency effect with elastic sizing.

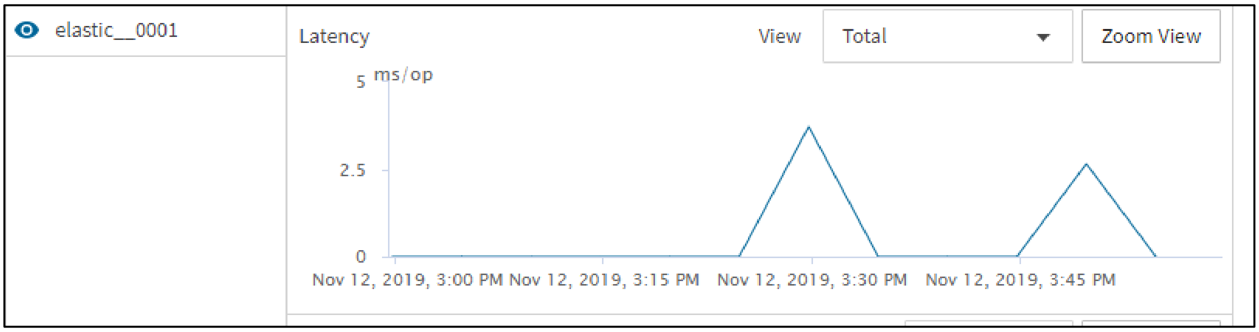
The following graphs illustrate the latency hit on the constituent volume:

Figure 54) Latency impact of elastic sizing.



The constituent volume 0001 has about .5ms more latency when elastic sizing is in effect:

Figure 55) Constituent volume latency increase from elastic sizing.



If you suspect that elastic sizing might be causing performance issues, you can do one of the following:

- Open a support case to confirm symptoms and logging.
- Grow the FlexGroup volume to make sure that there is enough space to remove elastic sizing from the equation.

ONTAP 9.8 and later introduces a new EMS event (`fg.member.elastic.sizing`) that lets you know that elastic resizing has occurred on a FlexGroup member volume.

However, with proactive resizing, every member volume resize event is considered elastic sizing and does not indicate a performance issue, but instead that the FlexGroup has some capacity issues that might need to be addressed by adding more space to the FlexGroup.

See an example of an elastic sizing EMS in the section "Event management system examples."

Keeping more than 20% available free space in a FlexGroup member volume is the ideal way to avoid the need for elastic sizing. This would require close management of capacity in ONTAP 9.6 and 9.7. However, ONTAP maintains free space for you in ONTAP 9.8 with proactive resizing.

Proactive resizing

ONTAP 9.8 introduces a new feature for capacity management, with the goal of removing capacity management tasks from the storage administrator and instead letting ONTAP manage FlexGroup member volume capacity.

The following issues should be considered regarding proactive resizing:

- Member volumes remain the same size if the member volume capacity is less than 60%, even if there is a large capacity disparity.
- Proactive resizing adjusts member volume sizes at between 60% to 80% used capacity in small increments so as to maintain a relatively even balance of available space.
- After 80% used capacity, the goal is to maintain even capacity usage by adjusting the total member volume sizes up or down.
- When a resize occurs, it is not large; the range is between 10M and 10GB. But it also does not affect performance the way elastic sizing does, because there is no pause needed to check for free space. Resizing occurs before any capacity issues appear.
- Volume autosize is implemented if a member volume reaches the autogrow threshold, provided you have enabled volume autosize.

This free space buffer helps maintain even file ingest across the volume, reduces capacity imbalance, and improves capacity management for FlexGroup volumes in ONTAP.

Proactive resizing behavior — volume Autosize disabled

In the example below, a 400GB FlexGroup volume with four 100GB member volumes is used. A client creates 32 10GB files in the FlexGroup volume across four folders. The FlexGroup volume has volume autosize disabled. This means that the FlexGroup volume size we've specified remains that size, even if we reach 100% capacity.

At the start of the job after the first files are written, this is how the capacity balance appears:

Figure 56) Initial FlexGroup data balance –– proactive resize, autosize disabled.

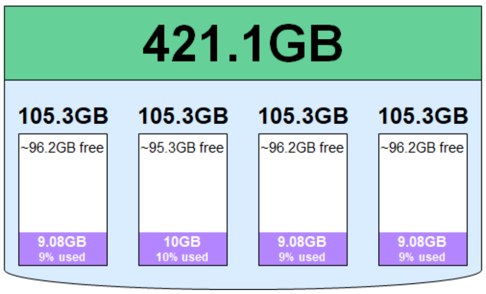
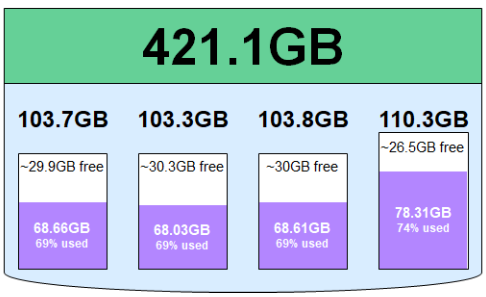
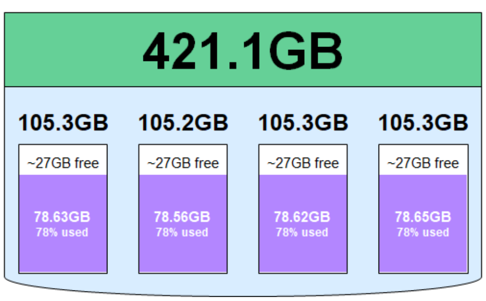


Figure 57) FlexGroup data balance, ~68% used — proactive resize, autosize disabled.



At around 70% capacity usage, we can start to see the member volumes resize a bit to maintain a balanced free space, but the total capacity remains the same.

Figure 58) FlexGroup data balance, job complete — proactive resize, autosize disabled.



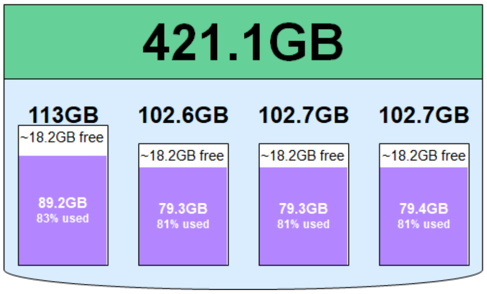
After the job finishes, ONTAP sees that the used space is even across all member volumes and proactive resizing shrinks the member volumes back down to their original sizes and makes them all the same because there is sufficient free space. The total FlexGroup size has not changed.

So, what happens when a new 10GB file is written after this?

When a file is written, it ends up in one of the member volumes. That creates a data imbalance, but ONTAP reacts accordingly by resizing the other member volumes to maintain an even amount of free space.

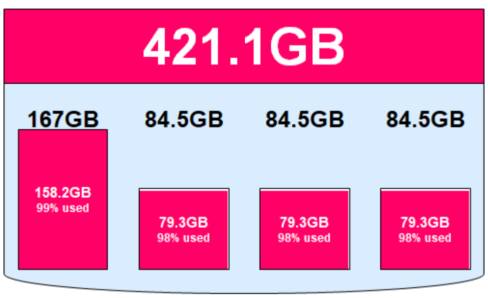
Here is the data balance after the new 10GB file is written:

Figure 59) FlexGroup data balance, new large file — proactive resize, autosize disabled.



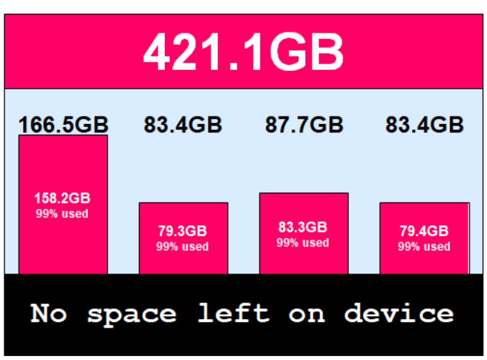
As you can see, the new file ended up in the first member volume. Elastic sizing grew that member volume to 113GB and shrunk the other member volumes while maintaining roughly the same amount of free space available and the total FlexGroup size.

Figure 60) FlexGroup data balance, 80GB file — proactive resize, autosize disabled.



Then we write a new file to the FlexGroup again. This time, the file is too big to fit into a single member volume (80GB) and is almost too large to fit into the FlexGroup itself. When that happens, ONTAP uses proactive resizing, but we're not in a situation where every member volume has just 1GB of free space remaining. That means the next 10GB file will fail, because the entire FlexGroup is out of space and autosize is disabled.

Figure 61) FlexGroup data balance, out of space — proactive resize, autosize disabled.



As a result, the next file creation fails, but proactive resizing becomes much more aggressive in adding free space to the member volume to avoid an out of space error. But when a FlexGroup volume itself is out of space, then the only remediation is growing the FlexGroup volume manually – or enabling volume autosize.

Proactive resizing behavior — volume autosize enabled

In the example below, a 400GB FlexGroup volume with four 100GB member volumes is used. A client creates 32 10GB files in the FlexGroup volume across four folders. The FlexGroup volume has volume autosize enabled with the default settings, which means the following:

- The FlexGroup volume maintains the same capacity, even if proactive resizing occurs, until the 92% used-space threshold is reached.
- After the used-space threshold is reached, the volume increases no more than 20%, as per the default settings. In this case, 566.7GB is the maximum size the volume grows, which is greater than 20% because this volume's size was increased and then later decreased.
- If the used capacity falls below 50%, then the volume shrinks back to the original size of 421.1GB.

These are the autosize settings for the FlexGroup:

```
cluster::> vol autosize -vserver DEMO -volume FG_SM_400G
Volume autosize is currently ON for volume "DEMO:FG_SM_400G".
The volume is set to grow to a maximum of 566.7g when the volume-used space is above 92%.
The volume is set to shrink to a minimum of 421.1g when the volume-used space falls below 50%.
Volume autosize for volume 'DEMO:FG_SM_400G' is currently in mode grow_shrink.
```

When a FlexVol or FlexGroup volume is smaller, the default growth threshold percentage is lower.

For example:

- A 100GB volume has a default grow threshold of 90% and a shrink threshold of 50%.
- A 10TB volume has a default grow threshold of 98% and a shrink threshold of 50%.

At the start of the job (after the first files are written), this is how the capacity balance appears:

Figure 62) Initial FlexGroup data balance — proactive resize, autosize enabled.

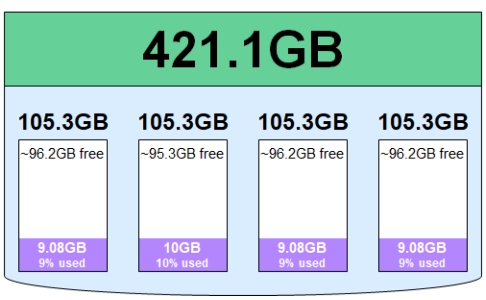
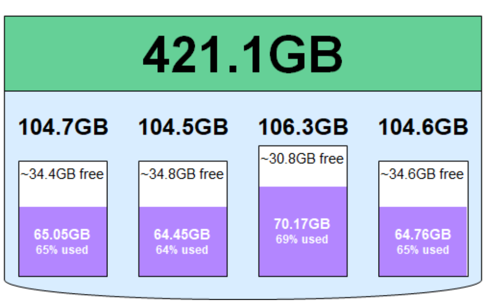
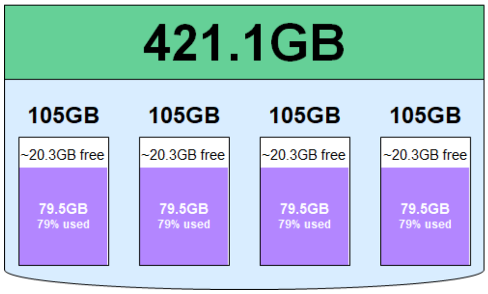


Figure 63) FlexGroup data balance, ~68% used — proactive resize, autosize enabled.



At around 66% capacity usage, we can start to see the member volumes resize a bit to maintain balanced free space, but the total capacity remains the same, just like when autosize is disabled.

Figure 64) FlexGroup data balance, job complete — proactive resize, autosize enabled.



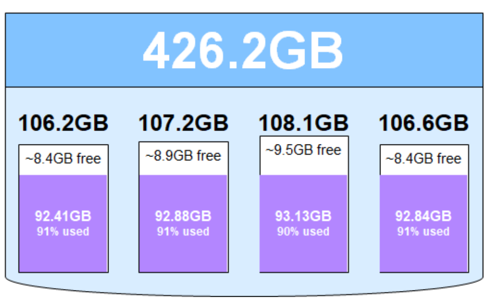
After the job finishes, ONTAP sees that the used space is even across all member volumes and proactive resizing shrinks the member volumes back down to their original sizes and makes them all the same size, because there is sufficient free space available. The total FlexGroup size has not changed.

As you can see, FlexGroup volumes with autosize enabled act just like FlexGroup volumes when autosize is disabled when the free space thresholds are below where autosize would kick in.

In the above graphic, we have roughly ~81GB free space available in the entire FlexGroup volume. If we keep writing 10GB files, we eventually reach the autosize threshold and ONTAP starts to react accordingly – this time with autosize growing the member volumes that need extra space, rather than by borrowing free space from other member volumes.

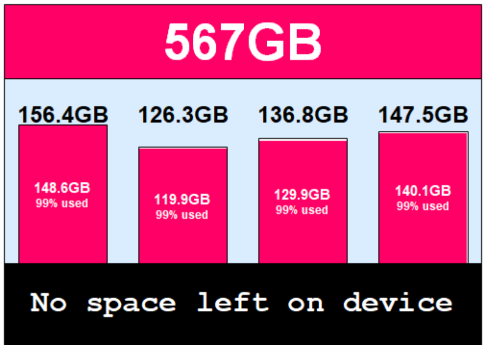
This results in the entire FlexGroup volume's capacity increasing. In the next test run, we created a new folder in the same FlexGroup and re-ran the test that creates 32 10GB files in the FlexGroup volume across four folders.

Figure 65) FlexGroup data balance, second test run — proactive resize, autosize enabled.



After one of the FlexGroup member volumes reaches the 92% used-space threshold, autosize grows *only* that member volume. If other member volumes also need to be grown when they hit 92%, then those are also increased. This increases the overall capacity of the FlexGroup volume. Proactive resizing also adjusts the other member volume capacities up or down so that a relatively even amount of free space is available per member volume.

Figure 66) FlexGroup data balance, autosize limit — proactive resize, autosize enabled.



Volume autosize defaults to only allow a volume to grow to 120% of the total volume size. Because the second test run needed 320GB of capacity to successfully complete and volume autogrow only allowed the volume to grow to 566.7GB total capacity, the job failed due to lack of space.

Autosize considerations: Smaller FlexGroup volumes

Since autosize capacity is based on the percentage of total size, smaller FlexGroup volumes (such as a 420GB FlexGroup volume) have less runway for growth by default than a larger FlexGroup volume would. The default autogrowth maximum is capped to 120% of the total volume size. If the volume is ever grown manually and shrunk back down, then the autogrow value reflects the larger volume size.

Table 17) Autosize maximum size examples.

FlexGroup volume size	Default maximum autosize	Default size delta
420GB	480GB	+80GB
100TB	120TB	+20TB
400TB	480TB	+80TB

As a result, if you're using volume autosize for FlexGroup volumes, use the following guidance:

Use larger FlexGroup volumes and maintain the default autosize values.

- If you use smaller FlexGroup volumes, modify the default `-max-autosize` value to avoid outages.
- If you don't want your end users to get more capacity than what you have provided, you can still use volume autosize if you use qtrees and quotas to limit the capacity seen and used by your end users.
- If you want to disable volume autosize, be aware that file writes fail when there is no more available space in the FlexGroup volume, even with proactive resizing in ONTAP 9.8.

Best practice 9: Combining ONTAP features for capacity management
The best way to approach capacity management involves a combination of larger FlexGroup volumes, volume autosize, ONTAP 9.8 or later, qtrees and quota enforcement. This story becomes more compelling when automatic tiering to cloud or S3 is performed by using NetApp FabricPool. Using these features minimizes the capacity management overhead for storage administrators.

Potential issues when using 100TB member volumes

The maximum FlexVol volume size in ONTAP for NAS is 100TB. As a result, the maximum size of member FlexVol volumes in a FlexGroup is also 100TB.

Features such as volume autosize, proactive resizing, and elastic sizing only function properly if FlexVol member volumes are able to grow when needed. If the volumes are at 100TB, then the aforementioned FlexGroup volume capacity management features won't be effective, because the member volumes cannot grow past 100TB. This introduces risk of member FlexGroup volumes reaching 100% capacity and creating an outage scenario.

Additionally, ONTAP 9.8 changed the 90% threshold for data placement urgency (where ONTAP forces more data to other less full member volumes much more frequently when a member volume hits 90% capacity) to a higher value (99%), so that FlexGroup member volumes that are growing in size won't adjust their data ingest rate by very much until they reach 99% capacity.

The following guidance should be followed for FlexGroup volumes, particularly in ONTAP 9.8 and later:

- Do not allow member volumes to grow beyond 80TB in size, if possible. This provides 20TB of buffer space before member volumes max out their capacity.
- Set alerts for free space usage to 80% on FlexGroup member volumes so that you are aware of when a member volume is nearly full.
- Set alerts for member volumes that exceed 80TB in size.
- If/when member volumes reach 80TB total size, consider adding more member volumes to the FlexGroup volume rather than growing the FlexGroup.
- If adding member volumes to a FlexGroup volume, add no less than two member volumes and preferably add the same number of member volumes the FlexGroup already has (for example, if the FlexGroup has eight member volumes, add eight member volumes).
- If you reach 80TB total size and 80% of used space and cannot add more member volumes, open a NetApp Support case for potential next steps.
- Use ONTAP 9.8P6 and ONTAP 9.9.1P1 or later for best results (as per bug 1391793).

Networking considerations

When you use CIFS/SMB or NFS, each mount point is made over a single TCP connection to a single IP address in the cluster. In ONTAP, these IP addresses are attached to data LIFs, which are virtual network interfaces in an SVM.

The IP addresses can live on a single hardware Ethernet port or multiple hardware Ethernet ports that participate in a Link Aggregation Control Protocol (LACP) or another trunked configuration. However, in ONTAP, these ports always reside on a single node, which means that they are sharing that node's CPU, PCI bus, and so on. To help alleviate potential bottlenecks on a single node, ONTAP allows TCP connections to be made to any node in the cluster, after which ONTAP redirects that request to the appropriate node through the cluster back-end network. This approach helps distribute network connections and load appropriately across hardware systems.

Best practice 10: Network design with FlexGroup

FlexGroup networking best practices are similar to FlexVol networking best practices. When you design a NAS solution in ONTAP, consider the following networking best practices regardless of the volume style:

- Create at least one data LIF per node, per SVM to confirm a path to each node.
- Present multiple IP addresses to clients behind a single fully qualified domain name (FQDN) by using some form of DNS load balancing. For DNS load balancing details, see TR-4523.
- When possible, use LACP ports to host data LIFs for throughput and failover considerations.
- When you manually mount clients, spread the TCP connections across cluster nodes evenly. Otherwise, allow DNS load balancing to handle the client TCP connection distribution.
- For clients that do frequent mounts and unmounts, consider using on-box DNS to help balance the load. If clients are not mounted and unmounted frequently, on-box DNS does not help much.

Best practice 10: Network design with FlexGroup

- If the workload is that of a "mount storm" (that is, hundreds or thousands of clients mounting at the same time), use off-box DNS load balancing and/or consider using NetApp FlexCache volumes. A mount storm to a single node can result in a denial of service to clients or performance issues.
- If you're using NFSv4.1, consider leveraging pNFS for data localization and parallel connections to files. pNFS works best with sequential I/O workloads; high metadata workloads might bottleneck over the single metadata server connection.
- If you have clients that support it, such as the latest SUSE and Ubuntu clients, the Nconnect mount option can provide even greater performance for NFS mounts on single clients.
- For SMB3 workloads, consider enabling the multichannel and large MTU features on the CIFS server.
- If you are using jumbo frames on your network, ensure jumbo frames are enabled at each endpoint in the network architecture; mismatched jumbo frame configurations can introduce hard-to-diagnose performance issues for any volume type.
- NFS clients can get greater performance with multiple mount points from the same client connected to the same volume in ONTAP across multiple network interfaces. However, this configuration can introduce complexity. If your NFS client supports it, use Nconnect.

LACP considerations

There are valid reasons for choosing to use an LACP port on client-facing networks. A common and appropriate use case is to offer resilient connections for clients that connect to the file server over the SMB 1.0 protocol. Because the SMB 1.0 protocol is stateful and maintains session information at higher levels of the OSI stack, LACP offers protection when file servers are in an HA configuration. Later implementation of the SMB protocol can deliver resilient network connections without the need to set up LACP ports. For more information, see TR-4100: Nondisruptive Operations with SMB File Shares.

LACP can provide benefits to throughput and resiliency, but you should consider the complexity of maintaining LACP environments when you are deciding. Even if LACP is involved, you should still use multiple data LIFs.

DNS load-balancing considerations

DNS load balancing (both off-box and on-box) provides a method to spread network connections across nodes and ports in a cluster. FlexGroup volumes do not change the overall thinking behind DNS load balancing. Storage administrators should still spread network connections across a cluster evenly, regardless of what the NAS container is. However, because of the design of FlexGroup volumes, remote cluster traffic is a near certainty (pNFS data locality is the exception) when a FlexGroup volume spans multiple cluster nodes. Therefore, network connection and data locality considerations are nullified in those configurations. As a result, some form of DNS load balancing fits in a bit better with a FlexGroup volume, because worrying about data locality is no longer a factor. Ultimately, the decision of which method of DNS load-balancing to use comes down to the storage and network administrators' goals. For more information about DNS load balancing, see TR-4523: DNS Load Balancing in ONTAP.

Best practice 11: Use some form of DNS load balancing

When possible, use some form of DNS load balancing with FlexGroup volumes on nodes that contain FlexGroup member volumes.

On-box DNS or off-box DNS?

ONTAP provides a method to service DNS queries through an on-box DNS server. This method factors in a node's CPU and throughput to help determine which available data LIF is the best one to service NAS access requests.

- Off-box DNS is configured by way of the DNS administrator creating multiple “A” name records with the same name on an external DNS server that provides round-robin access to data LIFs.
- For workloads that create mount-storm scenarios, the ONTAP on-box DNS server cannot keep up and balance properly, so it’s preferable to use off-box DNS.

NetApp recommends as a best practice creating at least one data LIF per node per SVM, especially when using a FlexGroup volume. Because of this, it might be prudent to mask the IP addresses behind a DNS alias through DNS load balancing. Then you should create multiple mount points to multiple IP addresses on each client to allow more potential throughput for the cluster and the FlexGroup volume.

- For more in-depth information on DNS load balancing, including a decision matrix, see [TR-4523: DNS Load Balancing in ONTAP].

Border Gateway Protocol (BGP): ONTAP 9.5 and later

Starting in version 9.5, ONTAP supports BGP to provide a more modern networking stack for your storage system. BGP support provides layer-3 (L3) routing, improved load-balancing intelligence, and virtual IPs (VIPs) for more efficient port utilization.

FlexGroup volumes need no configuration changes to use this new networking element.

Security and access control list style considerations

In ONTAP, you can access the same data through NFS and SMB/CIFS while preserving file ownership and honoring proper file permissions. This is known as multiprotocol NAS access. The same general guidance for multiprotocol NAS that applies to a FlexVol volume applies to a FlexGroup volume; these operate functionally the same for authentication and authorization. That guidance is covered in the product documentation in the CIFS, NFS, and Multiprotocol Express Guides and the CIFS and NFS Reference Guides, which can be found with the product documentation for the specific ONTAP version being used.

In general, for multiprotocol access, you need the following:

- Valid users (Windows and UNIX)
- Valid name-mapping rules or 1:1 name mappings through local files and/or servers such as LDAP or NIS. ONTAP uses name mappings to coordinate access for clients.
- Volume security style (NTFS, UNIX, or mixed). This can be configured for volumes or qtrees.
- A default UNIX user (pcuser, created by default for Windows to UNIX name mappings). Default Windows users (for UNIX to Windows name mappings) are not configured by default.

When a volume is created, a security style is applied. If you create a volume without specifying a security style, the volume inherits the security style of the SVM root volume. The volume security style determines the style of access control list (ACL) that is used for a NAS volume and affects how users are authenticated and mapped into the SVM. When a FlexGroup volume has a security style selected, all member volumes will have the same security style settings.

You can specify unique security styles in a FlexGroup volume by using qtrees.

Basic volume security style guidance

The following is some general guidance on selecting a security style for volumes:

- With the UNIX security style, Windows users must map to valid UNIX users. UNIX users only need to map to a valid user name if NFSv4.x is being used.
- In the NTFS security style, Windows users must map to valid UNIX users, and UNIX users must map to valid Windows users to authenticate. If a valid UNIX user name exists, NFS clients see proper ownership on files and folders. Authorization (permissions) is handled by the Windows client after the

initial authentication. If no valid UNIX user exists, the default UNIX user (pcuser) is used for authentication/UNIX ownership.

- The UNIX security style allows some Windows clients to modify basic mode bit permissions (ownership changes, rwx). However, it does not allow NFSv4.x ACL management over SMB, and it does not understand advanced NTFS permissions.
- A mixed security style allows permissions to be changed from any type of client. However, it has an underlying effective security style of NTFS or UNIX, based on the last client type to change ACLs.
- A mixed security style requires proper name mapping to function properly due to the changing effective security styles.
- If granularity of ACL styles in a FlexGroup volume is desired, consider deploying qtrees, which are available starting in ONTAP 9.3. Qtrees allow you to set security styles per logical directory in ONTAP. If you want other home directory features such as NetApp FPolicy, antivirus, native file auditing, and quota enforcement, then use the most recent patched release of ONTAP 9.5 or later.
- NFSv4.x and NFSv4 ACL support for FlexGroup volumes was added in ONTAP 9.7.

Best practice 12: Volume security style — mixed security style guidance

NetApp recommends a mixed security style only if clients need to be able to change permissions from both styles of clients. Otherwise, it's best to select either NTFS or UNIX as the security style, even in multiprotocol NAS environments.

More information about user mapping, name service best practices, and so on, can be found in the product documentation. You can also find more information:

- TR-4887: Multiprotocol NAS in NetApp ONTAP Overview and Best Practices
- TR-4835: How to Configure LDAP in ONTAP
- TR-4067: NFS Best Practice and Implementation Guide
- TR-4668: Name Services Best Practices Guide

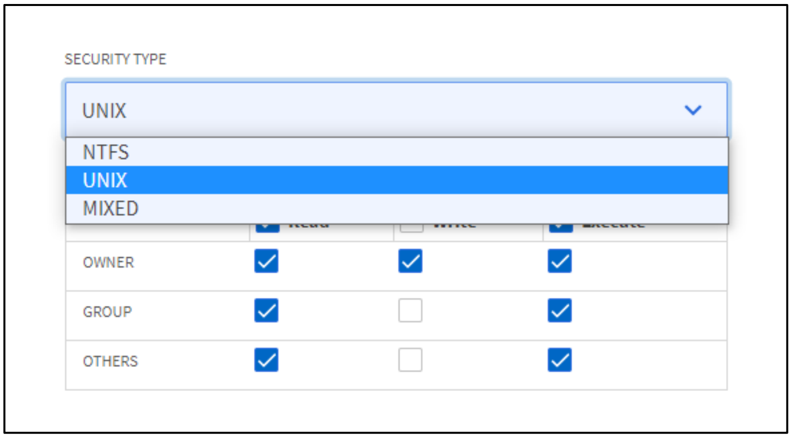
Changing the security style of a FlexGroup volume

FlexGroup volumes are intended to be managed similarly to FlexVol volumes; changing the security style of volumes is included in that philosophy. Volume security styles can be changed live, with no need for clients to remount. However, the subsequent change in ACL styles means that access permissions might become unpredictable. For the best possible results, NetApp recommends changing security styles in a maintenance window on production datasets.

To change the security style of a FlexGroup volume, do one of the following:

- Use `volume modify` from the command line.
- Use the Edit button or Advanced Features when initially creating the FlexGroup volume in NetApp ONTAP System Manager.

Figure 67) Modifying FlexGroup volume security styles in ONTAP System Manager.



NFSv4.x ACL workarounds for FlexGroup volumes: ONTAP 9.6 and earlier

Sometimes, storage administrators might want the extra, more granular security offered by NFSv4.x ACLs in addition to the performance characteristics and large-capacity footprint of FlexGroup volumes. However, prior to ONTAP 9.7, FlexGroup volumes did not support NFSv4.x. If you require native NFSv4.x ACL support for FlexGroup volumes, upgrade to ONTAP 9.7.

If you want/need to keep using ONTAP 9.6 or earlier, here are some possible workarounds, depending on the use case:

- **If the need for NFSv4.x ACLs is due to the NFSv3 group limitation (16 GIDs per user)**, then the NFS server options `-auth-sys-extended-groups enable` and `-extended-groups-limit [1-1024]` can be used to increase the number of supported GIDs per user. See TR-4067 for information about using this feature. NetApp highly recommends the use of a name service, such as LDAP, with this feature. See TR-4835 and TR-4668 for information about setting up name services.
- **If the need for NFSv4.x ACLs is due to the need for more granular permissions**, then consider adding a CIFS server to the SVM and using NTFS-style ACLs on the FlexGroup volume. NTFS-style ACLs provide the same level and granularity as NFSv4.x ACLs, with the added benefit of having an easy-to-manage GUI option. When you're using CIFS/SMB with NFS (also known as multiprotocol NAS), using a name service like LDAP to handle user authentication and mapping makes management easier. See TR-4835 and TR-4668 for information about setting up name services to apply multiprotocol NAS access.
- **If it's not possible to use a name service server like LDAP and/or set up CIFS in the environment**, then you need to upgrade to ONTAP 9.7 or later to get NFSv4 ACL support with FlexGroup volumes.

Using NFSv4.x ACLs with NFSv3 clients

In ONTAP, it is possible to leverage the benefits of NFSv4.x's granular ACL support even if your clients are only using NFSv3, provided you are using an ONTAP release that supports NFSv4 ACLs with FlexGroup volumes by using an administrative client that has mounted the export via NFSv4. By setting the NFSv4 ACLs from that client, clients accessing from NFSv3 honor those NFSv4 ACLs without needing to use NFSv4.x mounts. See TR-4067 for more information.

Labeled NFS 4.2

ONTAP 9.9.1 introduces support for the NFSv4.2 feature called labeled NFS, which is a way to manage granular file and folder access by using SELinux labels and Mandatory Access Control (MAC). Support for labeled NFS means that ONTAP now recognizes and understands the NFS client's SELinux label settings.

Labeled NFS is covered in RFC-7204.

Use cases include:

- MAC labeling of virtual machine images
- Data security classification for the public sector (secret, top secret, and so on)
- Security compliance
- Diskless Linux

FlexGroup administration considerations

This section covers general FlexGroup volume administration considerations, including tasks such as viewing FlexGroup volumes, volume moves, resizing FlexGroup volumes, renames, and so on.

Note: The ONTAP System Manager examples in this document use the new version of System Manager available in ONTAP 9.7 and later.

Initial deployment

In ONTAP 9.8 and later, it is best to let ONTAP automatically deploy FlexGroup volumes by using either System Manager or the `auto-provision-as` command option in the CLI. This is because FlexGroup features such as proactive resizing take care of most of the initial deployment considerations with larger files and automatic balance of capacity. In addition, each new release of ONTAP after ONTAP 9.8 makes further strides in reducing the amount of data management required for FlexGroup volumes through data ingest improvements, bug fixes, and new features for data balancing.

If you decide to use the CLI to manually specify aggregate lists for FlexGroup volumes (by using the `–aggr-list` option), note that ONTAP places the initial FlexGroup root member volumes (for example, FG___001) on the first specified aggregate, whereas automated FlexGroup creation distributes the FlexGroup root member volumes more evenly across the nodes. If a single aggregate has many FlexGroup root member volumes, then Snapshot and SnapMirror operations might run into problems, including timeouts and failures. Bug 1408116 covers this scenario.

If you find that a single node or aggregate contains many FlexGroup root member volumes, use nondisruptive volume moves to balance the volumes across different nodes and aggregates,

Viewing FlexGroup volumes

FlexGroup volumes can be created through the NetApp ONTAP GUI or through the command line and are designed to be managed, from a storage administrator's perspective, like a regular NetApp FlexVol volume. Things such as Snapshot copies, resize, and storage efficiency policies are all managed from the FlexGroup volume level.

However, the FlexGroup volume is not just a FlexVol volume; instead, it is made up of a series of FlexVol member volumes that act in concordance across the FlexGroup volume. NetApp ONTAP uses these member volumes for ingest of data to provide automated load balance and parallel operations across the file system, which provides capacity and performance gains.

In most cases, a FlexGroup volume can be managed at the FlexGroup level. For instance, when growing a FlexGroup volume, you can use the GUI or run the `volume size` command at the FlexGroup level to increase the total volume size. ONTAP makes sure that all underlying member volumes are given equivalent capacities, so the storage administrator doesn't have to think about how to distribute capacity. In ONTAP 9.8 and later, capacity management at the FlexGroup level is further simplified with the new proactive resizing functionality.

In rare cases, you may want to view individual FlexVol member volumes for capacity and performance concerns. These tasks are more commonplace in earlier releases of ONTAP.

Two scenarios where viewing member volumes might be useful:

- To view the member capacity usage (are we getting close to full in a single member volume?)
- To view individual member performance (do I need to use `volume move`?)

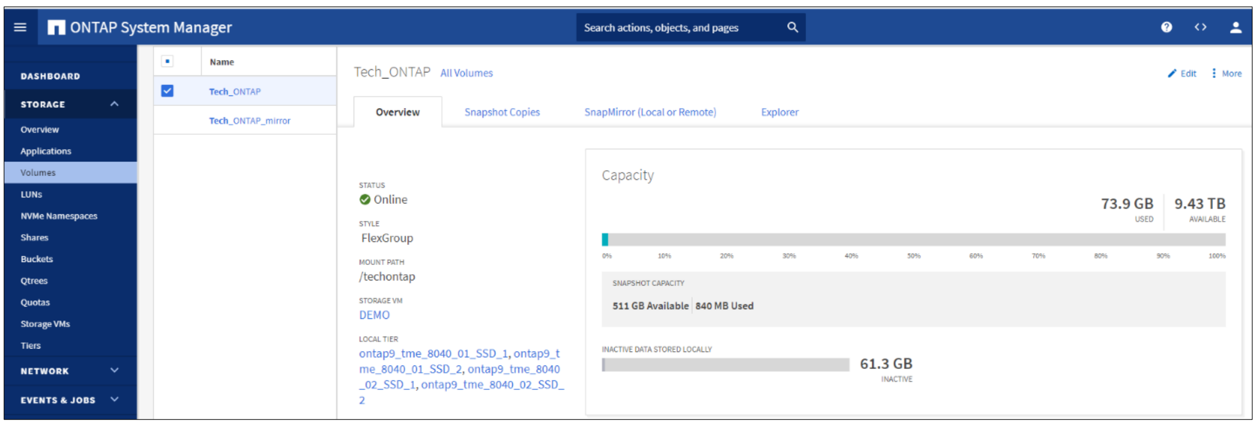
The following sections offer guidance on viewing FlexGroup volumes.

ONTAP System Manager

With ONTAP System Manager, you can view and manage a FlexGroup volume at the FlexGroup level through the FlexGroup tab; however, there are no views for member volumes. This is by design—a FlexGroup volume should be simple to manage. ONTAP System Manager provides useful information about the FlexGroup volume in these views, such as data protection information, real-time performance, and capacity information.

Note: Keep in mind that ONTAP System Manager cannot provide space allocation information for FlexGroup volumes that are thin-provisioned.

Figure 68) ONTAP System Manager FlexGroup volume view.



Like a FlexVol volume, you can manage basic tasks from the System Manager GUI, such as Snapshot copies, resizing, and data protection.

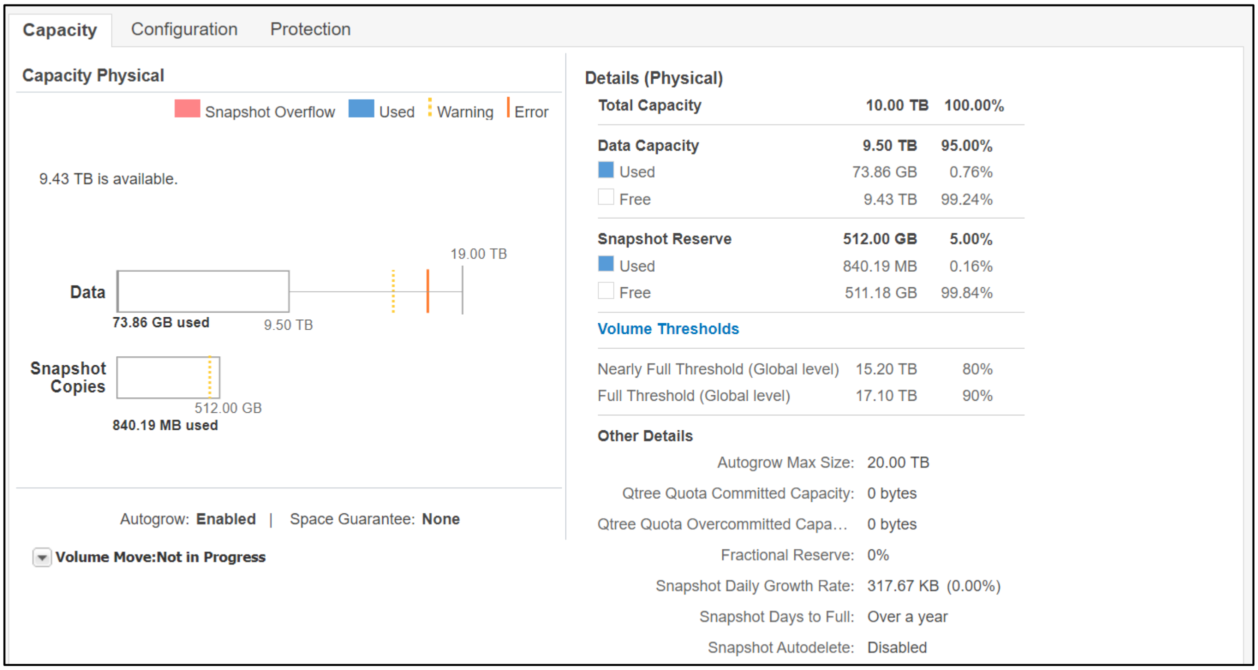
Active IQ Unified Manager

With [Active IQ Unified Manager](), storage administrators can use a single dashboard to review the health and performance of a NetApp ONTAP cluster.

With Active IQ Unified Manager, you can review FlexGroup volume capacity, configurations, and storage efficiencies in a graphical format. FlexGroup volume capacity in Active IQ Unified Manager is done from the FlexGroup level and does not show individual member volume capacities.

For member volume information about performance, use Active IQ Performance Manager. For capacity information about member volumes, use the command line.

Figure 69) Active IQ Unified Manager; FlexGroup capacity view.



Active IQ Unified Manager can also show capacity trends for your volumes under Workload Analysis.

Figure 70) Active IQ Unified Manager Capacity Trend.

Note: Currently, you cannot use Active IQ workloads with FlexGroup volumes.

Performance monitoring

NetApp Active IQ Unified Manager collects an archive of performance statistics for ONTAP, including the FlexGroup as a whole and its member volumes. A granular view of the FlexGroup volume allows storage administrators to evaluate individual member FlexVol volumes for performance anomalies and to take corrective actions as needed, such as the following:

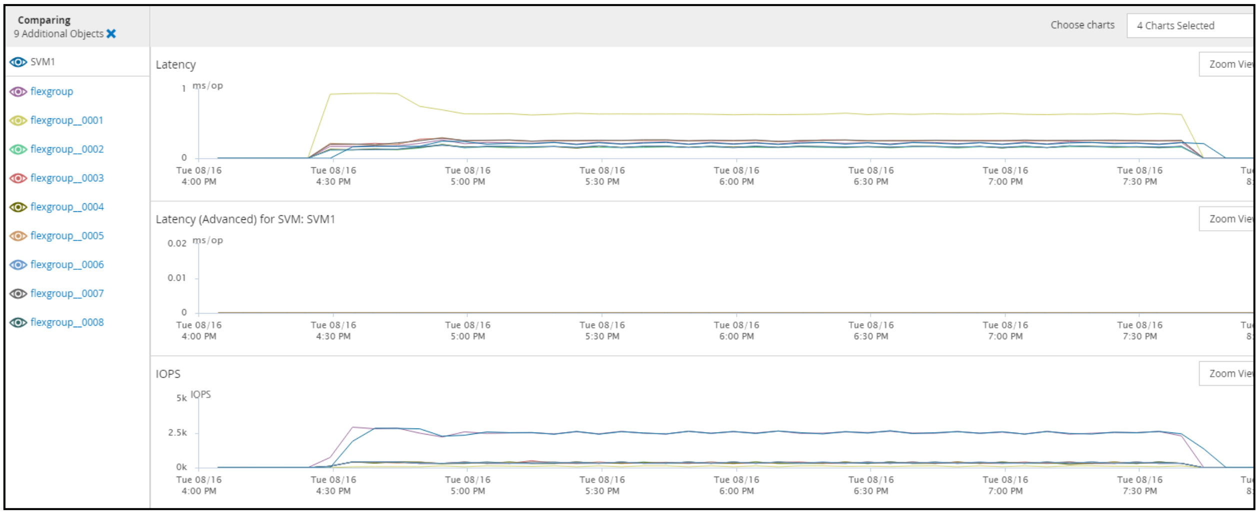
- Adding more space
- Adding more members (`volume expand`)

- Nondisruptive volume move to less allocated nodes

Note: These tasks cannot be carried out in Active IQ. Currently, only the command line and/or the ONTAP System Manager GUI can carry out these tasks.

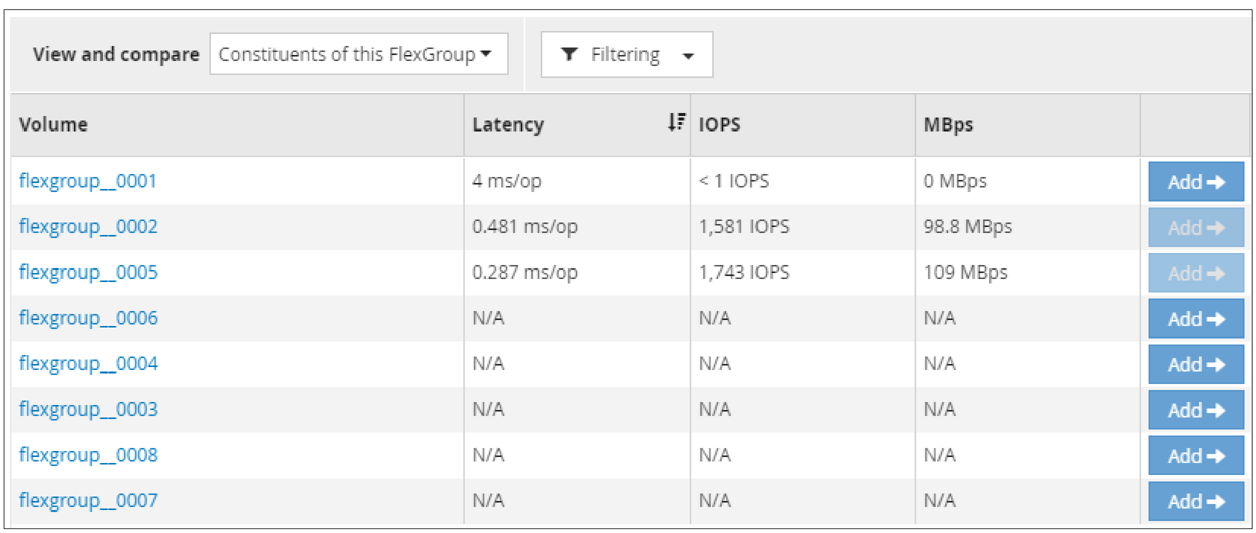
Figure 71 shows several FlexVol members and their corresponding performance. Each line represents a FlexVol member.

Figure 71) Active IQ Performance Manager FlexGroup volume view.



In Figure 72, two 1TB files are written to a FlexGroup volume. In the chart, you can see which member volumes took on that workload (members 2 and 5), and you see a summary of the workload performance. In Figure 73, you can see the IOPS and MBps graphs.

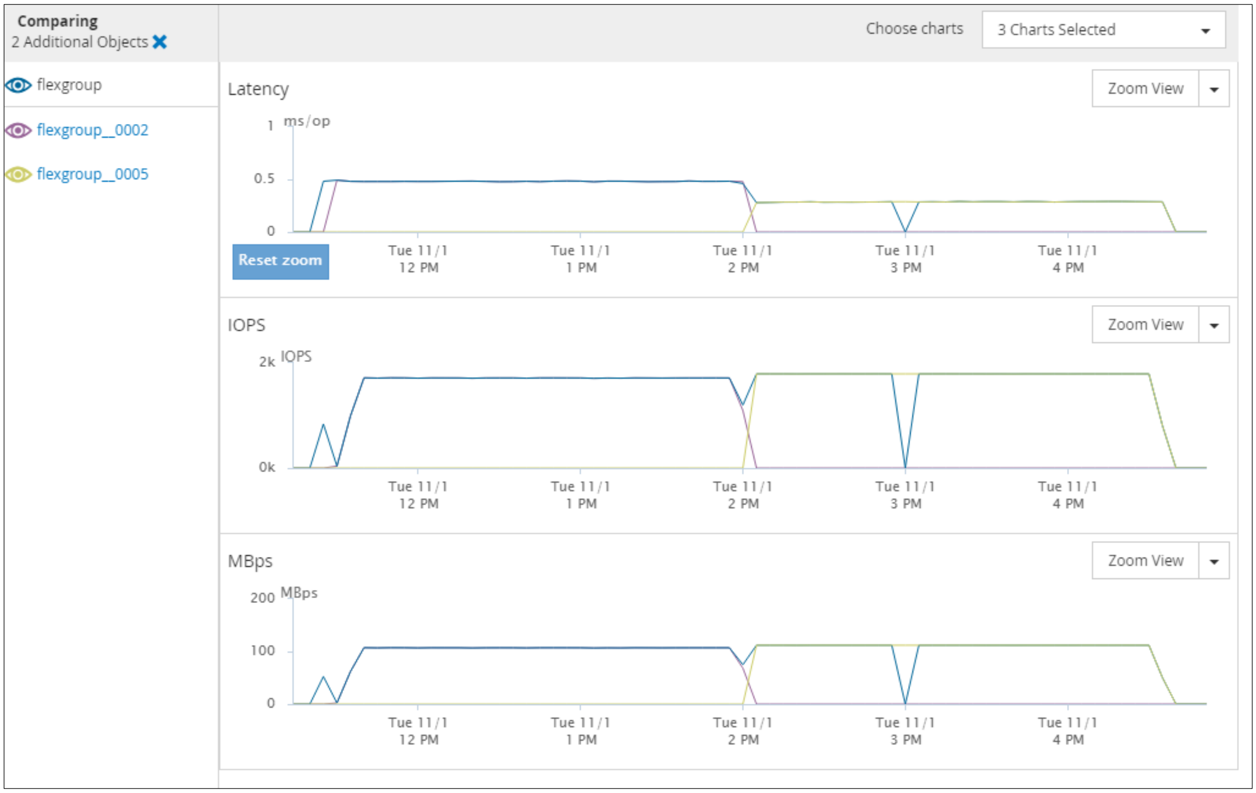
Figure 72) Member volume performance chart.



View and compare: Constituents of this FlexGroup | Filtering

Volume	Latency	IOPS	MBps	
flexgroup__0001	4 ms/op	< 1 IOPS	0 MBps	Add
flexgroup__0002	0.481 ms/op	1,581 IOPS	98.8 MBps	Add
flexgroup__0005	0.287 ms/op	1,743 IOPS	109 MBps	Add
flexgroup__0006	N/A	N/A	N/A	Add
flexgroup__0004	N/A	N/A	N/A	Add
flexgroup__0003	N/A	N/A	N/A	Add
flexgroup__0008	N/A	N/A	N/A	Add
flexgroup__0007	N/A	N/A	N/A	Add

Figure 73) Member volume graphs.



Command Line

The CLI is another way to view FlexGroup volume information. Each privilege level gives a different set of options for viewing the FlexGroup volume properties.

Admin privilege level

- Total capacity (total, available, and used: calculated from all the member volumes), storage efficiencies
- NetApp Snapshot reserve or Snapshot policy
- List of aggregates and nodes that the FlexGroup volume spans
- Volume style and extended volume style (tells us whether the volume is a FlexGroup volume)
- Security style, owner, or group
- Junction path
- Maximum files and inodes
- Member volume information (through `-is-constituent true` or `volume show-space`)

Advanced privilege level

- Maximum directory size value
- FlexGroup master set ID (MSID)
- Whether the volume was transitioned from 7-Mode (important for FlexVol to FlexGroup volume conversion)
- FlexGroup maximum member volume sizes

Diagnostic privilege level

- Detailed member volume information (capacity, used, and so on)
- FlexGroup ingest statistics (`flexgroup show`)

Note: Member volume space information can be seen in the admin privilege level by using the command `volume show-space`. For examples, see the section "Capacity monitoring and alerting with the command line."

Viewing FlexGroup volume capacity

This section covers various methods for monitoring a FlexGroup volume's capacity, including viewing total storage efficiency savings. Monitoring FlexGroup capacity is also possible with the NetApp FPolicy support introduced in ONTAP 9.4.

Note: In ONTAP 9.8 and later, capacity views should be focused to the total FlexGroup volume rather than the underlying member volumes, as proactive resizing handles member volume free space balancing. However, if you need to view member volumes, the following sections still apply.

Total FlexGroup capacity

The total FlexGroup capacity is a number that is derived from the following:

- **Total space.** Total combined allocated space for a FlexGroup volume (member volume capacity * number of members).
- **Available space.** The amount of space that is available in the most allocated member volume.

When you provision a 10TB FlexGroup volume, clients see 10TB. ONTAP sees 10TB divided by the number of member volumes created. In most cases, it is not necessary to think about this, especially with proactive resizing in ONTAP 9.8. However, with smaller FlexGroup volumes and/or larger files, these calculations can become more important.

You can view the total FlexGroup capacity in ONTAP System Manager, in Active IQ Unified Manager, or through the CLI at the admin privilege level.

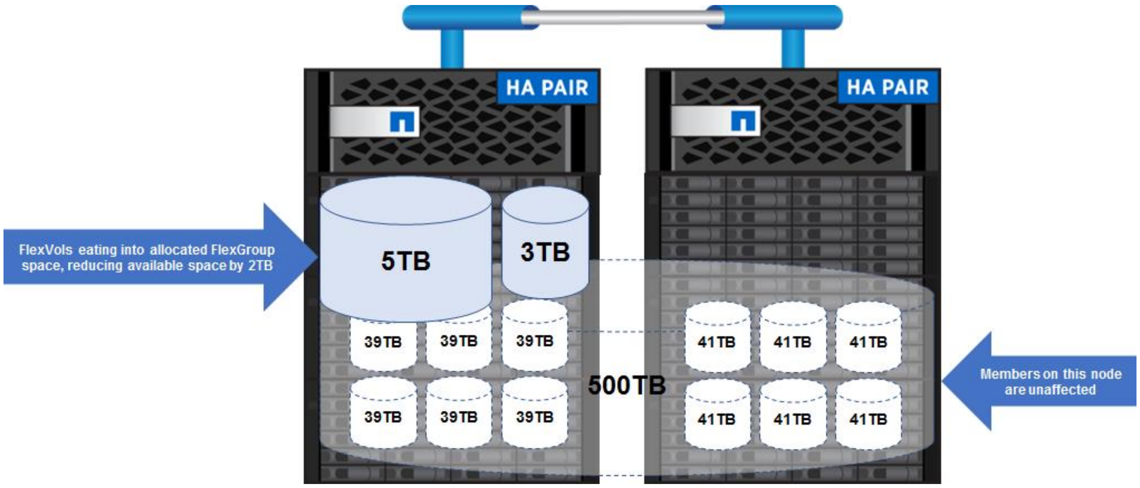
Overprovisioning or thin provisioning in a FlexGroup volume

Overprovisioning or thin provisioning with a FlexGroup volume can be useful in scenarios for which you want to minimize capacity management. You can create large FlexGroup volumes that don't take up actual capacity until data is written and instead rely on the physical available space in an aggregate.

Thin provisioning should be used with the following caveats in mind:

- When a volume is out of space, it is truly out of space because the physical capacity has been used. More disk space must be added to remediate space issues, whether on the same node or by adding new nodes in the cluster.
- The space allocated does not necessarily reflect the actual space available. It is possible to allocate volumes that are much larger than their physical space.
- If your system is sharing aggregates with other volumes (FlexGroup or FlexVol volumes) that are thick provisioned, you should use thin provisioning with caution. Existing FlexVol or FlexGroup volumes on the same aggregates as FlexGroup volumes can potentially affect how data is ingested. Existing volumes reduce the amount of space available for member volumes because used space eats into other volume allocations.

Figure 74) Capacity effect when thin-provisioned FlexGroup volumes exist with space-guaranteed FlexVol volumes.



Using volume space guarantees can protect against other datasets affecting volume capacities, but they do not offer the most efficient use of your storage capacity.

Best practice 13: Using thin provisioning
If you use thin provisioning with ONTAP, it is important to use tools such as Active IQ Unified Manager or to set up monitoring through the CLI to track the available space in your storage system.

Adding capacity to a FlexGroup volume

A FlexGroup volume can grow to immense capacities, but, as data grows, even a massive container such as a FlexGroup volume might require more capacity.

In some cases, FlexGroup performance can be adversely affected as the capacity of member volumes becomes closer to full due to an increase in remote hard link creation across member volumes. For more information about capacity, free space, and their effect on FlexGroup volumes, see the section “Capacity considerations.”

As a result, maintaining at least 10% free space in a FlexGroup member volume is generally a good practice. Naturally, that 10% has different meaning depending on the size of the member volumes and file sizes, so be sure to follow the recommendations on provisioning volumes that have large files or files that grow in this document.

Note: ONTAP 9.8 and later greatly reduces the management overhead for member volume capacities with proactive resizing and other features, so, if possible, move to that release when using FlexGroup volumes.

Recommendations for adding capacity

There are two main ways to add capacity to a FlexGroup volume, in order of preference:

- Grow existing member volumes by using the `volume size` command.
- Add new member volumes by using `volume expand`.

The preferred method for adding capacity to an existing FlexGroup volume is to grow the FlexGroup volume. Since data written to a FlexGroup is static and does not redistribute to other member volumes after it is placed in the FlexGroup, growing the FlexGroup in-place maintains a more consistent level of performance and capacity distribution than adding new member volumes to the workload.

If this approach is not possible because of physical aggregate limitations or the member FlexVol volumes approaching the 100TB limit, or if you are adding new nodes to the cluster and want to balance the workload out more, then you should add new member volumes using `volume expand` instead.

If you are adding new nodes or aggregates to use with the FlexGroup, it might make more sense to first use `volume move` to balance the member volumes across the nodes and then either grow them or add new member volumes in the same multiples per node. See the section "Adding disks, aggregates, and " for a visualization of the volume move approach.

Best practice 14: Increasing volume size

- If possible, increase capacity through `volume size` or resize from ONTAP System Manager rather than adding new members; this approach preserves the existing FlexGroup data balance.
- If you are adding new nodes or aggregates to use with the FlexGroup volume, use `volume move` to rebalance the members across the new hardware and then either grow FlexGroup by using `volume size` or add new members with `volume expand` in the same multiples per node. For example, if each node has four members, add four new members per node.
- Avoid running `volume size` commands on FlexGroup member volumes individually without the guidance of NetApp Support. Run `volume size` only on the FlexGroup volume itself.
- Upgrade to ONTAP 9.8 or later to gain the benefits of proactive resizing and use volume autosize to avoid needing to manage capacity manually.
- Use capacity monitoring to keep track of how full member volumes are becoming. Set warnings at threshold percentages based on the total member volume capacity to give you ample time to address any issues (for example, 80% may be fine to 10TB, but not for 100GB).
- Use thin provisioning to set higher virtual caps of space on volumes without affecting total space allocation.

Note: There currently is no way to reduce capacity by removing member volumes. Only shrinking a volume is allowed.

Growing the volume versus adding new members

If a FlexGroup volume requires capacity or increased file count, there are two main approaches.

Growing the FlexGroup volume (volume size)

You can add capacity to existing member volumes by growing the total size of the FlexGroup volume. You do this in the CLI by using the `volume size` command or in ONTAP System Manager. The added size is divided evenly across the member volumes in the FlexGroup volume. For instance, if you add 8TB to a FlexGroup volume with eight member volumes, each member volume grows by 1TB. Therefore, when you add space, it is important to know how many member volumes are in the FlexGroup volume. To find this number, use `volume show -name [flexgroup] -is-constituent true` from the CLI.

In deciding whether to grow the FlexGroup volume or add member volumes, consider your desired result and intent. Grow the FlexGroup volume if:

- You simply want more capacity on existing nodes or aggregates.
- You do not want to increase the total volume count in your cluster.
- Your FlexGroup member volumes are nowhere near the 100TB limit.
- You are not at the two billion file limit for the member volumes.
- You have available physical space where the member volumes currently live.
- You want to preserve the data balance across the member volumes.

These scenarios are by no means exhaustive; there might be other instances in which you want to grow a FlexGroup volume instead of adding new member volumes. If you are unsure, contact flexgroups-info@netapp.com.

Adding member volumes (volume expand)

Another way to add capacity or file count to a FlexGroup volume is by adding member volumes. Currently, NetApp officially supports up to 200-member FlexGroup volumes, which offers a maximum of 20PB capacity and 400 billion files. If you need more capacity or higher member volume counts than that, contact your NetApp sales representative to begin a qualification process for larger FlexGroup volumes.

To add more member volumes, you must currently use the `volume expand` command in the CLI. This command adds new, empty member volumes of exactly the same size as the existing FlexGroup member volumes. The number of new member volumes is determined by the `aggr-multiplier` and `aggr-list` options in the `volume expand` command. For an example of this command, see the end of this document in Command Examples.

Adding new member volumes is not the preferred way to add capacity or file counts in most cases, because the new member volumes will be empty and can throw off the ingest balance of new requests. If you are adding new member volumes, be sure to add them in multiples—preferably the same number as the volumes that already exist in the system.

However, in some use cases, adding member volumes is the best way to add capacity to a FlexGroup volume. For example:

- The FlexGroup member volumes are already at or near 100TB.
- New nodes or aggregates are being added to the cluster.
- More maxfiles are needed and the member volumes are already at their maximum values for their sizes.
- The cluster capacity is at a level where member volumes cannot be grown, and other nodes in the cluster have only enough space for member volumes of the same capacity.

These scenarios are by no means exhaustive; there might be other instances where you want to add members to a FlexGroup volume instead of growing the volume. If you are unsure, contact flexgroups-info@netapp.com.

Volume resizing

To grow or shrink the volume capacity as necessary, you can run the `volume size` command on the FlexGroup volume or use ONTAP System Manager. Adding capacity affects the ingest heuristics favorably because the member volumes have more available free space, so ONTAP favors local placement of files to a parent folder for better performance.

When you use this command, the member FlexVol volumes are each increased by the total capacity/total number of member volumes. For example, if a FlexGroup volume has eight member volumes and is grown by 80TB, then each member volume increases by 10TB automatically by ONTAP.

In releases prior to ONTAP 9.8, it was important to consider these individual increases when the total FlexGroup size increase is factored in. Proactive resizing with volume autogrow reduces the importance of considering individual member volume sizes, because ONTAP manages the free space available.

Note: FlexGroup volumes added manual volume shrink support in ONTAP 9.6 and autoshrink functionality in ONTAP 9.3.

Resizing a volume from ONTAP System Manager

1. To resize a volume in ONTAP System Manager 9.7 and later, select the volume you want to resize and click Edit. Then type in the new size.

Volume Expand

You can grow FlexGroup volumes nondisruptively with volume size or you can add more capacity dynamically by using the `volume expand` command, which is available at the admin privilege level. This command adds more FlexVol member volumes in the FlexGroup volume and should be used when one of the following occurs:

- The existing member volumes have reached their maximum capacities (100TB or two billion files)
- The physical limits of the node capacities have been reached and more aggregates or nodes are added to the cluster

Note: If you are simply trying to add more capacity or a higher file count to a FlexGroup volume, either resize the existing volume or increase the maxfiles value before adding new member volumes via volume expand.

Volume expand considerations

If you are adding new member volumes to a FlexGroup volume, you should take into consideration the following.

- Use `volume expand` only if increasing the existing volume size or file count is not an option or if the FlexGroup is being expanded across new hardware.
- `Volume expand` is currently done via the CLI only.
- If you must add new members, be sure to add them in the same multiples as the existing FlexGroup volume. That is, if the existing FlexGroup volume has 16 member volumes, eight per node, add 16 new members, eight per node, to promote consistent performance.
- If you add new nodes to an existing cluster and add new members to those nodes, try to maintain a consistent number of member volumes per node as in the existing FlexGroup volume.
- Adding new members to a FlexGroup volume changes the ingest heuristics to favor the new, empty member volumes for new data more often and can affect overall system performance for new data ingest while the new members catch up to the existing members.
- Add member volumes in multiples, preferably equal to the working set of member volumes, if possible. For example, if you have eight member volumes, add eight new member volumes when adding members.
- Adding member volumes adds FlexVols to the cluster, which counts against the maximum volume count allowed for the platform.
- When you add new members to a FlexGroup volume, the existing Snapshot copies and SnapMirror relationships are no longer valid for volume-level SnapRestore operations but can be used for client-driven file restore via previous versions and .snapshot directory access. For more information about snapshot restores, see TR-4678: Data Protection and Backup – FlexGroup Volumes.
- Existing data in the FlexGroup volume currently has no native way to nondisruptively rebalance, but this does not necessarily indicate a problem. For more information on data imbalances and their impact, see the section “Data imbalances in FlexGroup volumes.”

If you use `volume expand`, be sure to follow the guidance listed in the section "Recommendations for adding capacity."

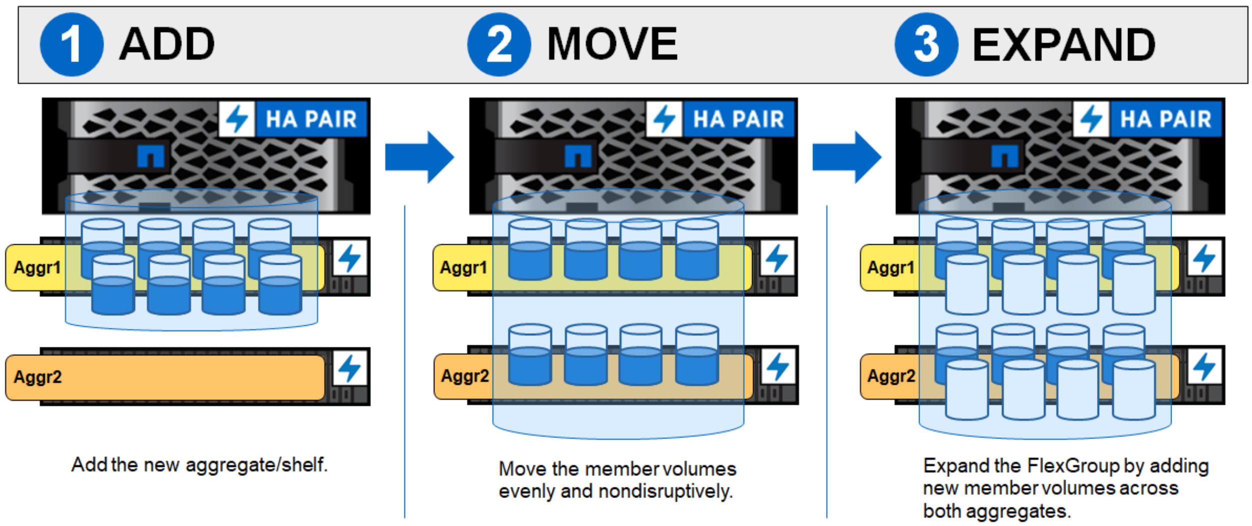
For an example of `volume expand`, see the section "Command Examples."

Adding disks, aggregates, and nodes

When adding disks to an existing aggregate that contains FlexGroup member volumes, no action is required unless you also want to increase the total volume size.

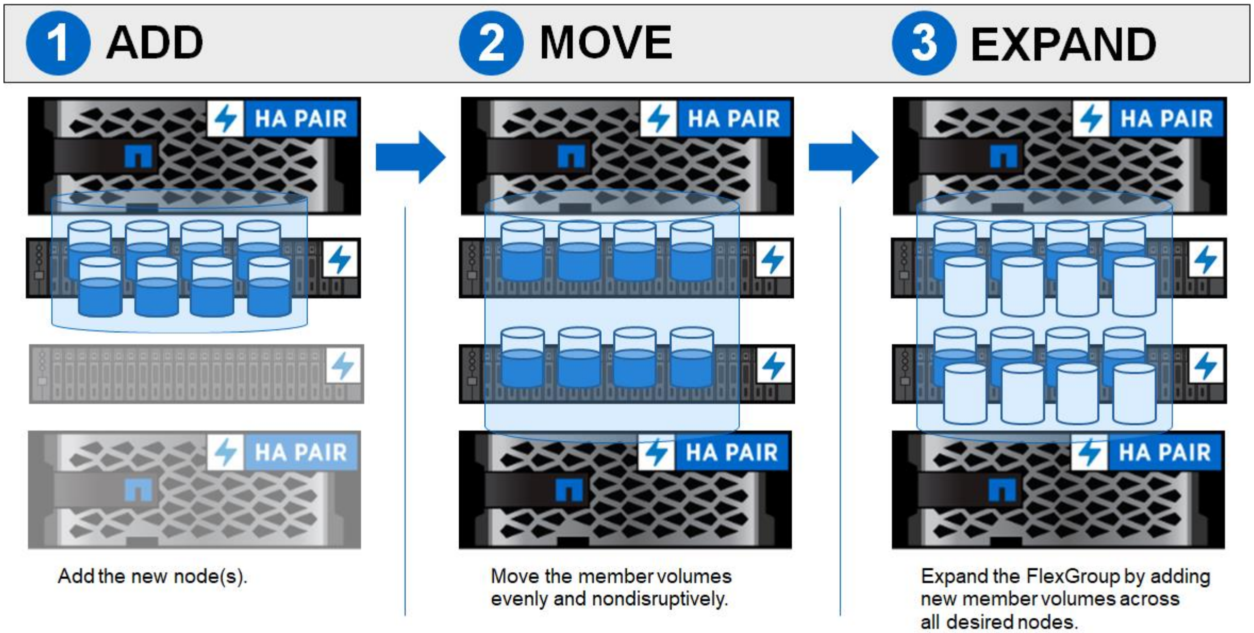
When adding aggregates to nodes, if the FlexGroup volume must span the new aggregates, you can use nondisruptive volume moves to move member volumes to the new aggregates without needing a maintenance window. Then you would create member volumes in the FlexGroup volume spanning new and old aggregates.

Figure 75) Adding aggregates with FlexGroup volumes.



When adding new nodes to a cluster, follow the same steps for adding aggregates to a cluster. Use `volume move` and `volume expand` commands to adjust the member volumes.

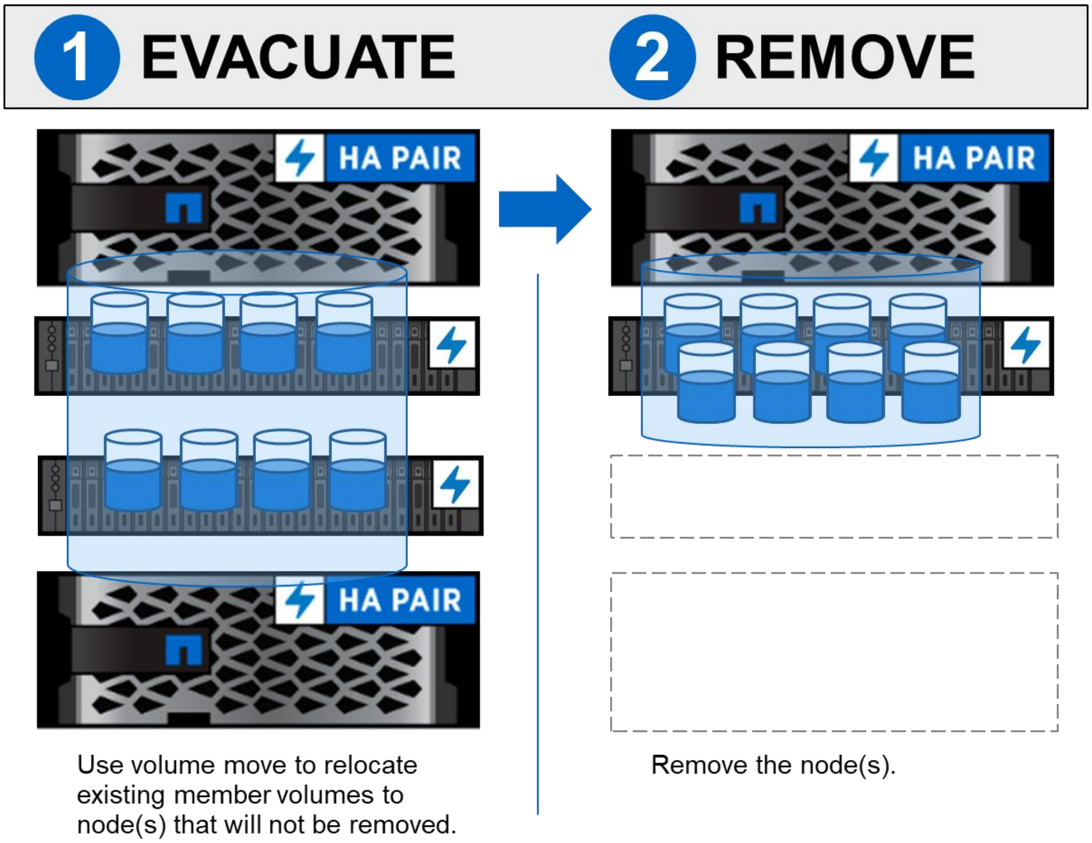
Figure 76) Adding nodes and expanding the FlexGroup volume.



Removing or evacuating nodes from a cluster

You can also remove nodes from a cluster that contain FlexGroup member volumes by using nondisruptive volume moves of member volumes. For example, if you want to remove two nodes from an eight-node cluster and each node has 16 member volumes, then you would use volume move to distribute 32 member volumes across the remaining six nodes. Because 32 volumes do not evenly divide across six nodes, use the next divisible node count to evenly distribute so that four nodes get eight member volumes per node. If there isn't enough space for four nodes to take on eight member volumes each, place six member volumes on two nodes (12) and five member volumes on four nodes (20).

Figure 77) Removing nodes that contain FlexGroup member volumes.



Note: Member volumes cannot be removed from FlexGroup volumes once they are added.

Nondisruptive volume move considerations

ONTAP enables you to perform nondisruptive volume moves between aggregates or nodes in the same cluster. This feature provides flexibility when you are dealing with maintenance windows or attempting to balance performance or capacity allocation in a cluster.

FlexGroup volumes also support this feature, but with even more granularity; you can move each member volume in a FlexGroup volume by using this functionality. Volume move does not move the entire FlexGroup in a single move.

Storage administrators therefore have a way to move workloads around in a cluster if capacity or performance concerns arise. With the ability of Active IQ to review individual member FlexVol volumes, you can quickly identify and resolve issues.

Note: Keep in mind that, although volume moves are nondisruptive, the amount of time that they take depends on the volume size and on the overall load on the node.

When to use nondisruptive volume moves

Nondisruptive volume moves can come in handy in the following scenarios for FlexGroup:

- The member volume is nearing capacity, and no physical storage is available on the current node to grow that volume in place.
- The member volume shares an aggregate with other FlexVol volumes and is being affected by the FlexVol volume's performance or capacity.

- A member volume is overworked in a FlexGroup volume and needs more node resources.
- You want to migrate FlexGroup volumes from spinning disk to SSD for performance or from SSD to spinning disk for archiving.
- New cluster nodes or data aggregates are added.
- You are performing a head swap or other planned maintenance operations (to provide for the least amount of downtime).

Using nondisruptive volume moves

Nondisruptive `volume move` for a FlexGroup member volume is available at the admin privilege level of the command line. Although you can use the ONTAP System Manager GUI to move FlexVol volumes, you currently cannot use it to move FlexGroup member volumes.

To move a FlexGroup member volume, complete the following steps:

1. Identify the volume that needs to be moved. Use Active IQ Unified Manager or the CLI to determine this information.
2. From the command line, run the `volume move start` command. This command can be run at the admin privilege level.

```
cluster::> volume move start -vserver SVM -destination-aggregate aggr1_node2 -volume
flexgroup4TB__000
    flexgroup4TB__0001 flexgroup4TB__0002 flexgroup4TB__0003
    flexgroup4TB__0004 flexgroup4TB__0005 flexgroup4TB__0006
    flexgroup4TB__0007 flexgroup4TB__0008

cluster::> volume move start -vserver SVM -volume flexgroup4TB__0003 -destination-aggregate
aggr1_node2
[Job 2603] Job is queued: Move "flexgroup4TB__0003" in Vserver "SVM" to aggregate "aggr1_node2".
Use the "volume move show -vserver SVM -volume flexgroup4TB__0003" command to view the status of
this operation.

cluster::> volume move show
Vserver   Volume     State    Move Phase Percent-Complete Time-To-Complete
--------- ---------- -------- ---------- ---------------- ----------------
SVM       flexgroup4TB__0003
                     healthy  replicating
                                         45%              Tue Dec 06 13:43:01 2016
```

Auto balance aggregate

The Auto Balance Aggregate feature is introduced in Data ONTAP 8.3. This feature provides ONTAP recommended nondisruptive volume moves when system performance or capacity reaches a point specified by the storage administrator. This feature is not currently supported with FlexGroup volumes.

Considerations when deleting FlexGroup volumes

In Data ONTAP 8.3, a feature called the volume recovery queue was added to help prevent accidental deletion of volumes by maintaining a recovery queue of deleted volumes for 12 hours. Although the volume is no longer accessible from clients and is hidden from administrator view at admin privilege levels, the space is still allocated, and the remnants of the volume remain in case an emergency recovery is needed. FlexGroup volumes use this recovery queue too, so space is not freed up until the recovery queue expires or is manually purged.

You can see deleted volumes from the command line by specifying `–type DEL` at the diagnostic privilege level. Neither volumes nor the recovery queue can be seen from the GUI.

```
cluster::*> volume show -vserver DEMO -type DEL
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
DEMO      flexgroup__0001_2321
```



```
                       aggr1_node1  offline    DEL         5TB          -     -
DEMO      flexgroup__0002_2322
                       aggr1_node2  offline    DEL         5TB          -     -
```

Deleted volumes can also be seen with the volume recovery-queue command, also with diagnostic privileges:

```
cluster::*> volume recovery-queue show
Vserver   Volume      Deletion Request Time     Retention Hours
--------- ----------- ------------------------  ---------------
DEMO      flexgroup__0001_2321
                      Tue May 01 17:14:14 2018               12
DEMO      flexgroup__0002_2322
                      Tue May 01 17:14:13 2018               12
2 entries were displayed.
```

To purge the volumes from the recovery queue manually, run the following commands:

```
cluster::*> volume recovery-queue purge -vserver DEMO -volume flexgroup__0001_2321
Queued private job: 4660

cluster::*> volume recovery-queue purge -vserver DEMO -volume flexgroup__0002_2322
Queued private job: 4661
```

To bypass the recovery queue when deleting a volume, use the `–force true flag` from the CLI with the `volume delete` command at the advanced privilege level. ONTAP System Manager does not support forced deletions or managing the volume recovery queue.

Volume rename considerations

ONTAP 9.6 and later supports FlexGroup volume renaming. FlexGroup volume names are meant only for identification of the volumes within the cluster by storage administrators. Client-facing names for volumes are exposed by way of CIFS/SMB shares and volume junction paths (export paths) for NFS, not by how you name a volume in the cluster.

For example, a volume could have an admin name of `vol1` but exported to clients as `/accounting`. Junction paths and SMB shares can be changed at any time in a FlexGroup volume, but doing so causes a disruption for clients, because they must reconnect to the new share or export path through a remount or SMB share reconnection. Volume renames, however, do not cause any client-side disruption.

Qtrees

ONTAP 9.3 introduced 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 provide flexibility in data management by enabling unique export policies, unique security styles, and granular statistics.

Qtrees have multiple use cases and are useful for home directory workloads because qtrees can be named to reflect the user names of users accessing data, and dynamic shares can be created to provide access based on a username.

The following bullets give more information regarding qtrees in FlexGroup volumes.

- Qtrees appear as directories to clients.
- Qtrees are able to be created at the volume level; you cannot currently create qtrees below directories to create qtrees that are subdirectories.
- Qtrees are created and managed the same way as a FlexVol qtree is managed.

- Qtrees cannot be replicated using SnapMirror. SnapMirror currently is only performed at the volume level. If you want more granular replication with a FlexGroup, use a combination of FlexGroup volume and junction paths.
- A maximum of 4,995 qtrees is supported per FlexGroup volume. Quota monitoring and enforcement (enforcement in ONTAP 9.5 and later) can be applied at the qtree or user level.

  Note: ONTAP 9.5 added quota enforcement support and qtree statistics. ONTAP 9.8 added qtree QoS support.

Qtrees and file moves

A qtree is considered a unique filesystem in ONTAP. While it looks like a directory from a NAS client perspective, some operations might behave differently than if it were an actual directory. One example of that is moving a file between qtrees in the same volume.

When a file move is done in a volume across directories, the file is simply renamed to a new name and happens within seconds because that is a move inside of the same filesystem.

When a file move occurs between two qtrees, the file is copied to the new location rather than being renamed. This causes the operation to take much longer.

This is a behavior that occurs whether the qtree lives in a FlexVol or a FlexGroup.

Qtree IDs and rename behavior

Once a non-inherited export policy is applied to a qtree, NFS file handles will change slightly when dealing with operations between qtrees. ONTAP will validate qtree IDs in NFS operations, which will impact things like file renames/moves when moving to or from a qtree in the same volume as the source folder or qtree. This is considered a security feature, which helps prevent unwanted access across qtrees, such as in-home directory scenarios. However, simply applying export policy rules and permissions can achieve similar goals.

For example, a move or rename to or from a qtree in the same volume will result in an Access Denied error. The same move or rename to or from a qtree in a different volume results in the file being copied. With larger files, the copy behavior can make it seem like a move operation is taking an unusually long time, where most move operations are near-instantaneous, as they are simple file renames when in the same file system and volume.

This behavior is controlled by the advanced privilege option and is covered in detail in the NetApp knowledge base article Permission denied while moving files between qtrees when NFS option '-validate-qtree-export' is enabled.

From that KB, these are the behaviors of different operations.

```
Assuming that file permissions allow and that client is allowed by export policies to access both
source and destination volume/qtree, these are the current permutations with the 'validate-qtree-
export' flag enabled or disabled:

Enabled:
- Rename in same volume and qtree: SUCCESS
- Rename in same volume, different qtrees: EACCESS
- Rename between volumes where qtree IDs differ: EACCESS
- Rename between volumes where qtree IDs match: XDEV

Disabled:
- Rename in same volume and qtree: SUCCESS
- Rename in same volume, different qtrees: SUCCESS
- Rename between volumes where qtree IDs differ: XDEV
- Rename between volumes where qtree IDs match: XDEV
```

Note: NFS3ERR_XDEV and NFS3ERR_ACCESS are defined in RFC-1813.

To change the behavior of renames/moves across qtrees, modify `-validate-qtree-export` to `disabled`. See the section Validating qtree IDs for qtree file operations for more information.

Note: There is no known negative impact to disabling the `-validate-qtree-export` option, outside of allowing renames across qtrees.

File handle effects for qtree exports

Normally, the NFS export file handles that are handed out to clients are 32 bits or less in size. However, with qtree exports, an extra few bits are added to create 40 bit file handles. In most clients, this is not an issue, but older clients (such as HPUX 10.20, introduced in 1996) might have problems mounting these exports. Be sure to test older client connectivity in a separate test SVM before enabling qtree exports, because there is currently no way to change file handle behavior after qtree exports have been enabled.

Managing quotas with FlexGroup

NetApp FlexGroup volumes support user/group and tree quotas. The level of support for these can be broken down into the following.

- Support for quota reporting in ONTAP 9.3.
- Support for FPolicy, which can provide quota enforcement from third-party vendors, such as DefendX (formerly NTP) in ONTAP 9.4.
- Enforcement of quotas (that is, setting hard and soft limits for capacity and file count) is supported in ONTAP 9.5 and later.

User and group quota considerations

To implement user or group quotas, the cluster must be able to resolve the specified username or group. This requirement means that the user or group must exist locally on the SVM or within a resolvable name service server, such as Active Directory, LDAP, or NIS. If a user or group cannot be found by the SVM, then the quota rule is not created. If a user or group quota fails to create because of an invalid user, the command line issues this error:

```
Error: command failed: User name user not found. Reason: SecD Error: object not found.
```

ONTAP System Manager delivers a similar message. Use the `event log show` command to investigate the issue further. For more information about configuring name services for identity management in ONTAP, see TR-4835: How to Configure LDAP in ONTAP and TR-4668: Name Services Best Practices Guide.

Creating a user or group quota

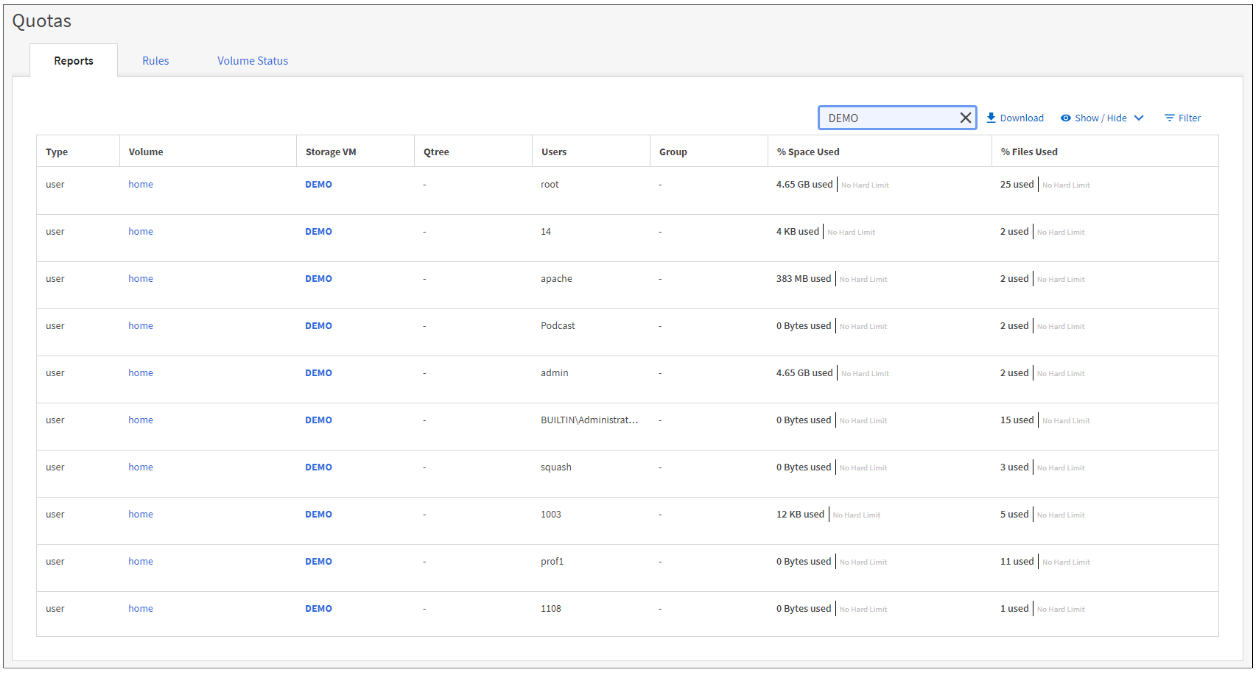
User and group quotas can be created to report or enforce capacity or file count limits on a per-user basis. These quotas would be used in scenarios where multiple users or groups share the same namespace or qtree. These steps are the same for FlexVol and FlexGroup volumes.

Creating a quota — ONTAP System Manager

To create user or group quota in ONTAP System Manager, navigate the left-hand menu to Storage > Quotas. That takes you to a page with three tabs: Reports, Rules, and Volume Status.

Reports show you the current quota tracking for users, groups, and qtrees.

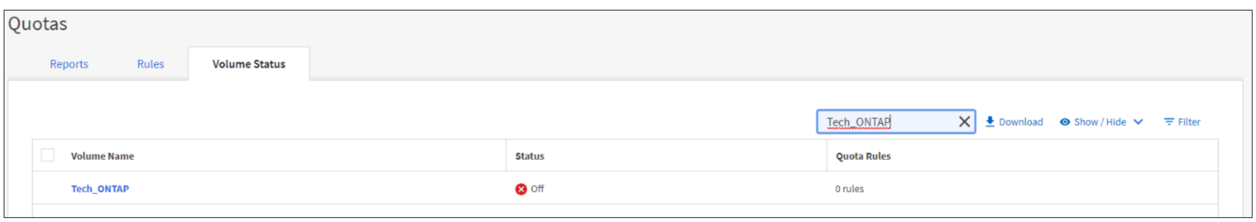
Figure 78) Quota reports – ONTAP System Manager.



Type	Volume	Storage VM	Qtree	Users	Group	% Space Used	% Files Used
user	home	DEMO	-	root	-	4.65 GB used | No Hard Limit	25 used | No Hard Limit
user	home	DEMO	-	14	-	4 KB used | No Hard Limit	2 used | No Hard Limit
user	home	DEMO	-	apache	-	383 MB used | No Hard Limit	2 used | No Hard Limit
user	home	DEMO	-	Podcast	-	0 Bytes used | No Hard Limit	2 used | No Hard Limit
user	home	DEMO	-	admin	-	4.65 GB used | No Hard Limit	2 used | No Hard Limit
user	home	DEMO	-	BUILTIN\Administrat...	-	0 Bytes used | No Hard Limit	15 used | No Hard Limit
user	home	DEMO	-	squash	-	0 Bytes used | No Hard Limit	3 used | No Hard Limit
user	home	DEMO	-	1003	-	12 KB used | No Hard Limit	5 used | No Hard Limit
user	home	DEMO	-	prof1	-	0 Bytes used | No Hard Limit	11 used | No Hard Limit
user	home	DEMO	-	1108	-	0 Bytes used | No Hard Limit	1 used | No Hard Limit

Volume status shows whether quotas are on or off for the volume.

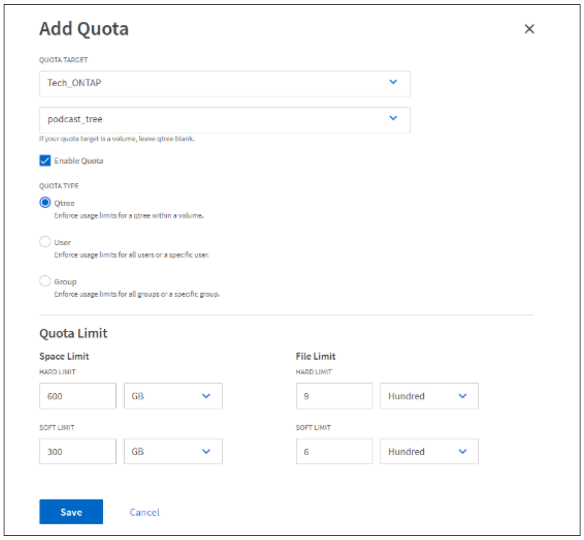
Figure 79) Quota volume status – ONTAP System Manager.

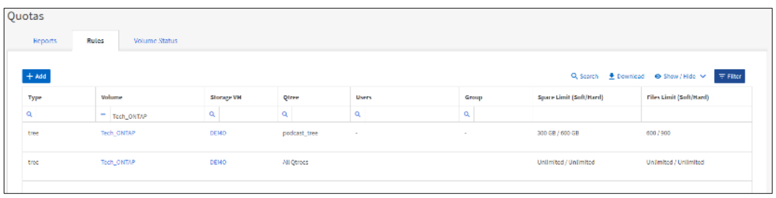


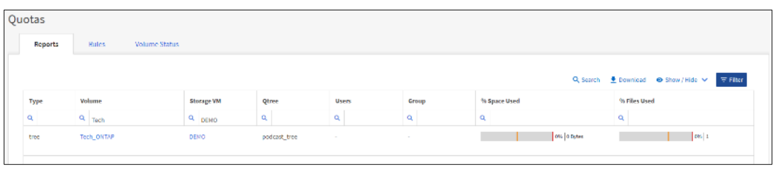
Volume Name	Status	Quota Rules
Tech_ONTAP	Off	0 rules

Rules is where you would create new quotas for users, groups, or qtrees. Click Add and enter the information for the user, group or qtree quota in the dialog screen. After the rule is created, ONTAP System Manager performs all of the necessary steps to enable and activate the quota.

Figure 80) Quota rules – ONTAP System Manager.







Creating a user or group quota — command line

To create a user or group reporting quota with the command line for a specific user or group, use the following command at the admin privilege level:

```
cluster::> quota policy rule create -vserver SVM1 -policy-name default -volume flexgroup -type
[user|group] -target [username or groupname] -qtree ""
```

To create a user or group reporting quota with the command line for all users or groups, use the following command at the admin privilege level. The target is provided as an asterisk to indicate `all`:

```
cluster::> quota policy rule create -vserver SVM1 -policy-name default -volume flexgroup -type
[user|group] -target * -qtree ""
```

Before ONTAP 9.5, quota enforcement was unsupported for use with FlexGroup volumes. As a result, you could not set limits for files or disk space usage. ONTAP 9.5 lets you set hard limits for files (`-file limit`) and capacity (`-disk-limit`) with quotas.

This example shows the `quota report` command in ONTAP 9.5 and later with FlexGroup volumes and quota enforcement:

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



```
                                    ----Disk----  ----Files-----   Quota
Volume   Tree      Type    ID        Used  Limit    Used   Limit   Specifier
-------  --------  ------  -------  -----  -----  ------  ------   ---------
flexgroup_local
         qtree     tree    1
                                   1.01GB    1GB       5      10   qtree
flexgroup                user    student1, NTAP\student1
                                      4KB    1GB      10      10   student1
```

Creating a tree reporting quota from the command line

To create a tree reporting quota with the command line for a specific user or group, use the following command at the admin privilege level:

```
cluster::> quota policy rule create -vserver DEMO -policy-name tree -volume flexgroup_local -type
tree -target qtree
```

To enable quotas, use `quota on` or `quota resize`.

```
cluster::> quota on -vserver DEMO -volume flexgroup_local
[Job 9152] Job is queued: "quota on" performed for quota policy "tree" on volume
"flexgroup_local" in Vserver "DEMO".

cluster::> quota resize -vserver DEMO -volume flexgroup_local
[Job 9153] Job is queued: "quota resize" performed for quota policy "tree" on volume
"flexgroup_local" in Vserver "DEMO".

cluster::> quota show -vserver DEMO -volume flexgroup_local

              Vserver Name: DEMO
               Volume Name: flexgroup_local
               Quota State: on
               Scan Status: -
          Logging Messages: -
          Logging Interval: -
          Sub Quota Status: none
  Last Quota Error Message: -
Collection of Quota Errors: -
       User Quota enforced: false
      Group Quota enforced: false
       Tree Quota enforced: true
```

The following example shows a `quota report` command on a FlexGroup volume with a tree quota specified:

```
cluster::> quota report -vserver DEMO -volume flexgroup_local
Vserver: DEMO

                                    ----Disk----  ----Files-----   Quota
Volume   Tree      Type    ID        Used  Limit    Used   Limit   Specifier
-------  --------  ------  -------  -----  -----  ------  ------   ---------
flexgroup_local
         qtree     tree    1           0B      -       1       -   qtree
```

Files used and disk space used are monitored and increment as new files are created:

```
cluster::> quota report -vserver DEMO -volume flexgroup_local
Vserver: DEMO

                                    ----Disk----  ----Files-----   Quota
Volume   Tree      Type    ID        Used  Limit    Used   Limit   Specifier
-------  --------  ------  -------  -----  -----  ------  ------   ---------
flexgroup_local
         qtree     tree    1
                                  13.77MB      -       4       -   qtree
```


Quota enforcement example

When quota enforcement is enabled on a qtree or for a user/group, ONTAP disallows new file creations or writes after a quota is exceeded. This helps storage administrators have greater control over how much data is being written to a volume or qtree.

In addition, when a quota is exceeded, an event management system message is logged at the DEBUG severity level to notify storage administrators of the quota violation. You can configure these 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.

```
cluster::*> quota policy rule show -vserver DEMO

Vserver: DEMO              Policy: tree              Volume: flexgroup_local

                                                Soft             Soft
                        User          Disk      Disk   Files    Files
Type   Target    Qtree   Mapping     Limit     Limit   Limit    Limit  Threshold
-----  --------  ------- -------   --------  -------  ------  -------  ---------
tree   qtree     ""      -            1GB        -      10        -          -
```

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 is 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     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.

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

                                    ----Disk----  ----Files-----   Quota
Volume   Tree      Type    ID        Used  Limit    Used   Limit   Specifier
-------  --------  ------  -------  -----  -----  ------  ------   ---------
flexgroup_local
         qtree     tree    1
                                   1.01GB    1GB       5      10   qtree
```

The same behavior occurs for file count limits. In this example, the file count limit is 10 and the qtree already has five files. An extra five files meet 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

                                    ----Disk----  ----Files-----   Quota
```



```
Volume   Tree      Type   ID        Used  Limit    Used   Limit   Specifier
-------  --------  ------ -------  -----  -----  ------  ------   ---------
flexgroup_local
         qtree     tree   1
                               1.01GB    1GB       5      10   qtree
home               user   student1, NTAP\student1
                                  4KB    1GB      10      10   student1
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:7e3cc08e-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:7e3cc08e-d9b3-11e6-85e2-00a0986b1210",
limit_item="disk", limit_value="1048576", user="", qtree="treeid=1", vfiler=""
```

You can use quotas to control the file counts allowed for FlexGroup volumes if you plan on leaving 64-bit file IDs disabled. For more information, see the section “Using quota enforcement to limit file count.”

Performance effect of using quotas

Performance effects are always a concern when enabling a feature. To alleviate performance concerns when using quotas, we ran a standard NAS benchmark test against FlexGroup volumes in ONTAP 9.5 with and without quotas enabled. We concluded that the performance effect for enabling quotas on a FlexGroup volume is negligible, as shown in Figure 81 and Figure 82

Figure 81) ONTAP 9.5 performance (operations/sec)—quotas on and off.

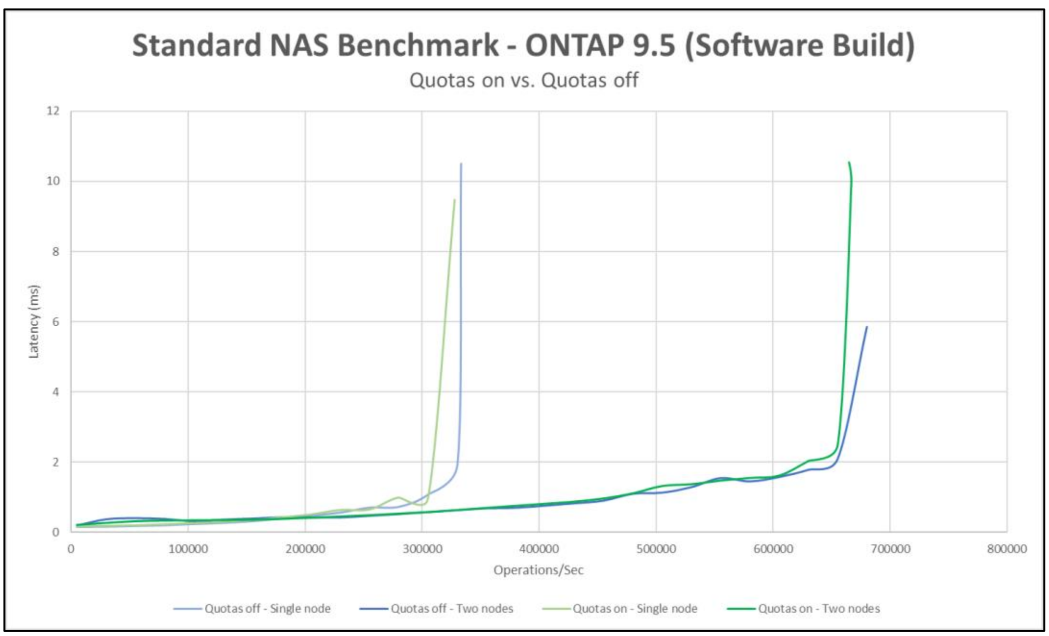
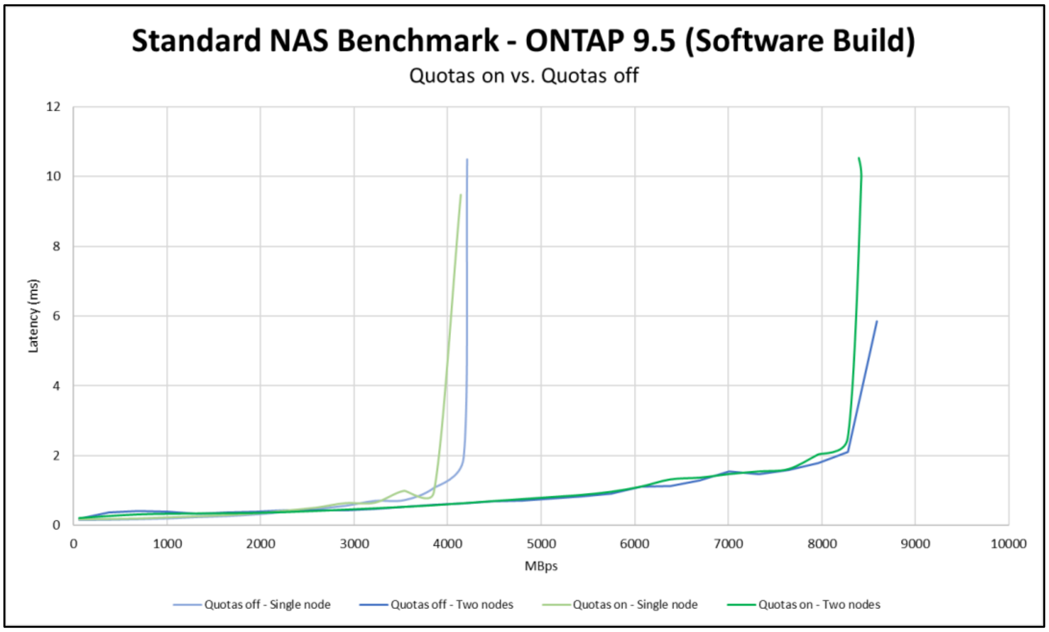


Figure 82) ONTAP 9.5 performance (MBps)—quotas on and off.



Quota scan completion times

When a quota initialization or resize takes place, ONTAP must perform some background tasks to complete the necessary work to reflect quota usage accurately. These tasks take time, which depends on a number of factors covered below.

Initialization completion time

The time it takes for quotas to initialize on a volume or qtree depends on the following factors:

- The number of files and folders in a volume. More files mean a longer initialization, while file size does not affect initialization time.
- Type of volume. FlexVol scans can take longer than FlexGroup scans, because FlexGroup quota scans are performed in parallel across the nodes a FlexGroup resides.
- Type of hardware and load on system. Heavily loaded systems with many files can result in scans that take hours.

You can check quota initialization status with the command `quota show -volume volname -instance`.

Quota resize completion time

Quota resize is used when a quota policy is changed. The resize performs a scan with the new limits. This process also has some considerations for time to completion.

- The resize only scans using the newly added rules, so it completes faster than an initialize.
- Resize typically completes in a matter of seconds, since it has to do less than quotas on/off.
- Use resize instead of toggling quotas on/off, because resize completes faster.
- Quota resize can run up to 100 concurrent jobs; after 100 jobs, resize operations must wait in a queue.
- More concurrent scans can impact resize performance and add time to the job completion.

User-mapping considerations with quotas

User mapping in multiprotocol environments (data access from both SMB and NFS) for quotas occurs at the member volume level. Eventually, all member volumes agree on the user mapping. However, sometimes there might be a discrepancy, such as when user mapping fails or times out when doing a name mapping that succeeded on another member. This means that at least one member considers the user to be part of a user-mapped pair, and at least one other member considers it to be a discrete record.

At worst, enforcement of the quota rules can be inconsistent until the issue is resolved. For instance, a user might be able to briefly overrun a quota limit.

An event management system message is sent when user mapping results are coordinated.

```
cluster::*> event route show -message-name fg.quota.usermapping.result -instance

                                     Message Name: fg.quota.usermapping.result
                                         Severity: NOTICE
                                Corrective Action: (NONE)
                                      Description: This message occurs when the quota mapper
decides whether to map the Windows quota record and the UNIX quota record of a user into a single
multiuser record.
```

Tree quota considerations

SVMs in ONTAP can have a maximum of five quota policies, but only one policy can be active at a time. To see the active policy in an SVM, use the following command:

```
cluster::> vserver show -vserver DEMO -fields quota-policy
vserver quota-policy
------- ------------
DEMO    default
```

Note: Currently, you cannot view this information in ONTAP System Manager.

The default policy is adequate in most cases and does not need to be changed. When `quota on` is issued, the active policy is used—not the policy that was assigned to a volume. Therefore, it's possible to get into a situation where you think you have applied a quota and rules to a volume, but `quota on` fails.

The following example applies a quota policy to a volume:

```
cluster::*> quota policy show -vserver DEMO -policy-name tree

      Vserver: DEMO
  Policy Name: tree
Last Modified: 10/19/2017 11:25:20
    Policy ID: 42949672962

cluster::*> quota policy rule show -vserver DEMO -policy-name tree -instance

                 Vserver: DEMO
             Policy Name: tree
             Volume Name: flexgroup_local
                    Type: tree
                  Target: tree1
              Qtree Name: ""
            User Mapping: -
              Disk Limit: -
             Files Limit: -
Threshold for Disk Limit: -
         Soft Disk Limit: -
        Soft Files Limit: -
```

Turning on quotas produces an error because the SVM has `default` assigned for quotas and does not contain any rules.

```
cluster::*> quota on -vserver DEMO -volume flexgroup_local -foreground true
```



```
Error: command failed: No valid quota rules found in quota policy default for volume
flexgroup_local in Vserver DEMO.
```

When you add a rule to `default`, the `quota on` command works, but the SVM does not use the new tree policy.

```
cluster::*> quota policy rule create -vserver DEMO -policy-name default -volume flexgroup_local -
type tree -target ""

cluster::*> quota on -vserver DEMO -volume flexgroup_local -foreground true
[Job 8063] Job succeeded: Successful

cluster::*> vserver show -vserver DEMO -fields quota-policy
vserver quota-policy
------- ------------
DEMO    default
```

To use the necessary policy, you must modify the SVM and then turn quotas off and back on.

```
cluster::*> vserver modify -vserver DEMO -quota-policy tree

cluster::*> quota off -vserver DEMO *

cluster::*> quota policy rule delete -vserver DEMO -policy-name default *
1 entry was deleted.

cluster::*> quota on -vserver DEMO -volume flexgroup_local -foreground true
[Job 8084] Job succeeded: Successful
```

This behavior is not unique to FlexGroup volumes; this would happen with FlexVol volumes as well.

How clients see space when quotas are enabled

When quotas are enabled for a qtree in ONTAP, the clients only see the available space as reported by that quota.

For example, this is a quota for qtree1:

```
cluster::*> quota report -vserver DEMO -volume flexgroupDS -tree qtree1
Vserver: DEMO

                                    ----Disk----  ----Files-----   Quota
Volume   Tree      Type    ID        Used  Limit    Used   Limit   Specifier
-------  --------  ------  -------  -----  -----  ------  ------   ---------
flexgroupDS
         qtree1    tree    1           0B  500GB       1       -   qtree1
```

This is how much space that volume actually has:

```
cluster::*> vol show -vserver DEMO -volume flexgroupDS -fields size
vserver volume      size
------- ----------- ----
DEMO    flexgroupDS 10TB
```

This is what the client sees for space for that volume:

```
# df -h /mnt/nas2
Filesystem         Size  Used Avail Use% Mounted on
demo:/flexgroupDS  9.5T  4.5G  9.5T   1% /mnt/nas2
```

This is what is reported for that qtree:

```
# df -h /mnt/nas2/qtree1/
Filesystem         Size  Used Avail Use% Mounted on
demo:/flexgroupDS  500G     0  500G   0% /mnt/nas2
```


General NAS and high-file-count considerations

This section covers general NAS and high-file-count environment considerations.

High file count considerations

An inode in ONTAP is a pointer to any file or folder within the file system, including Snapshot copies. Each FlexVol volume has a finite number of inodes and has an absolute maximum of 2,040,109,451.

You can increase inodes after a FlexVol volume has been created and decrease them only to a number that has not already been allocated.

Default and maximum inode counts

Default and maximum inode counts for volumes (both FlexVol and FlexGroup) are dependent on the total allocated capacity of the volume. For example, a 100GB FlexVol volume would not be able to hold as many inodes as a 8TB FlexVol volume.

Table 18 shows a sample of FlexVol sizes, inode defaults, and maximums.

Table 18) Inode defaults and maximums according to FlexVol size.

FlexVol volume size	Default inode count	Maximum inode count
20MB*	566	4,855
1GB*	31,122	249,030
100GB*	3,112,959	24,903,679
1TB	21,251,126	255,013,682
7.8TB	21,251,126	2,040,109,451
100TB	21,251,126	2,040,109,451

*FlexGroup member volumes should not be any smaller than 100GB in size.

Increasing maximum files: Considerations

If you would like to avoid monitoring and reacting to out of inode conditions, you can immediately configure high file count FlexGroup volumes and FlexVol volumes with the maximum supported files value, with the following considerations in mind.

The default or maximum number of inodes on a FlexVol volume depends on the volume size and has a ratio of one inode to 4KB of capacity. This means that for every 4KB of allocated space to a volume, you can allocate one inode. Examples of these values are seen in Table 18.

In addition, each inode uses 288 bytes of capacity – this means that having many inodes in a volume can also use up a non-trivial amount of physical space in addition to the capacity of the actual data as well. If a file is less than 64 bytes, it is stored in the inode itself and does not use additional capacity.

This used space counts against the 10% aggregate reserve in ONTAP. Two billion files can use as much as ~585GB of space, and if you have many volumes set to the maximum files limit, then each volume's inode capacity is allocated to that aggregate reserve. This capacity is only used when files are actually allocated to the volume and not by simply setting the maximum files value.

As a result, if you increase the files value to the maximum, you should pay attention to the both the used inodes as well as the used aggregate space. Keeping both values in the 80% range gives the best results for high file count environments.

Other considerations:

- FlexGroup volumes are ideally the volume of choice to use for high file count environments due to their ability to nondisruptively scale when a limit has been reached.
- An approximate maximum of one inode per 4KB of allocated size can be configured, so a FlexVol or FlexGroup member volume must be approximately 7.8TB or larger in size in order to configure it with the maximum possible files setting of two billion.
  - In a FlexGroup volume, this means each member volume must be 7.8TB or greater.
- You should still monitor for out of inode conditions in case your environment hits the maximum supported values, and you may need to revisit the files setting any time that you grow or shrink a FlexVol or FlexGroup.
- If you choose to set the maximum files value in your volume, you should also consider setting your monitoring thresholds to 80% of the allocated inodes to give yourself ample time to plan and react before you run out of inodes.
- If the files value is set to the maximum amount on a FlexVol or on the individual member volumes in a FlexGroup and you run out of inodes, you cannot increase them further unless you are using a FlexGroup volume and add new member volumes. Therefore, avoid setting FlexVol volumes to two billion if possible. Rather, use FlexGroup volumes so there is at least the option of adding member volumes in case you hit the two billion maximum.
- Finally, keep in the mind that inode metadata is stored in the underlying aggregate, so aggregate free space should be monitored to ensure that the aggregate does not run out of space.

Increasing the files value to the maximum easily

In many cases, you might be unsure of the maximum number of files for a volume. In ONTAP 9.9.1, a new volume option was introduced to make setting the maximum file value simpler.

When you set the `files-set-maximum` value on a volume to “true,” ONTAP automatically adjusts maxfiles to the largest possible value for you. You can only set this value to “true” – after it’s set, you cannot unset it. Set the value to “true” only if you want to set the maxfiles to the largest possible value.

```
[-files-set-maximum {true|false}] - Set Total Files (for user-visible data) to the Highest Value
that the Volume can Hold (privilege: advanced)
This optionally specifies whether the volume's total number of files will be set to the highest
possible value. If true, the volume's total number of files is set to the highest value that the
volume can hold. Only <true> is a valid input. <false> is not permitted. To modify the total
number of files to a specific value, use option files.
```

See “Increasing maximum files: Considerations” for information about the implications of setting the maximum files value.

Default and maximum inode counts – FlexGroup volume considerations

When a default volume inode count reaches 21,251,126, it remains at that default value, regardless of the size of the FlexVol volume. This feature mitigates potential performance issues, but it should be considered when you design a new FlexGroup volume. The FlexGroup volume can handle up to 400 billion files (two billion files x 200 FlexVol member volumes), but the default inode count for 200 FlexVol members in a FlexGroup volume is just 4,250,225,200.

This count is based on the following formula:

```
200 member volumes * 21,251,126 default inodes per member = 4,250,225,200 total default inodes
```

If the FlexGroup volume requires more inodes than what is presented as the default value, the inodes must be increased by using the `volume modify -files` command. As mentioned, this value can be increased to the absolute maximum value allowed if desired, but the guidance in section “Increasing maximum files: Considerations” should be followed.

Best practice 15: Inode count in a FlexGroup volume (before ONTAP 9.3)

When deciding where to place files, ONTAP versions earlier than 9.3 did not consider inode counts in the ingest calculations for data that is written into a FlexGroup volume. Thus, a member FlexVol volume could run out of inodes before other members run out of inodes, which would result in an overall `out of inodes` error for the entire FlexGroup volume. ONTAP 9.3 introduced inode count consideration for ingest of files to help avoid member volumes running out of inodes prematurely. If you're running ONTAP 9.2 or earlier, NetApp strongly recommends increasing the default inode count in the FlexGroup volume before using it in production. The recommended value depends on workload, but do not set the value to the maximum at the start. If you do, you won't have room to increase later without adding member volumes. If possible, upgrade to ONTAP 9.3 or later to take advantage of the new ingest calculations for high-file-count environments.

When you use a FlexGroup volume, the total default inode count depends on both the total size of the FlexVol members and the number of FlexVol members in the FlexGroup volume.

Table 19 shows various examples of FlexGroup configurations and the resulting default inode counts.

Table 19) Inode defaults resulting from FlexGroup member sizes and member volume counts.

Member volume size	Member volume count	Default inode count (FlexGroup)
100GB	8	24,903,672
100GB	16	49,807,344
1TB	8	170,009,008
1TB	16	340,018,016
100TB	8	170,009,008
100TB	16	340,018,016

High file counts, low-capacity needs

As mentioned, ONTAP allocates a default inode and maximum inode count based on volume capacity. In Table 20, member volumes smaller than 7.8TB are not able to achieve the maximum of two billion inodes. To get two billion inodes per member volume, the member volume capacity needs to be 7.8TB or greater. A FlexGroup volume with eight member volumes and space guarantees enabled supports up to 16 billion files, but it also provisions ~62.4TB of reserved storage.

If your dataset consists of very small files, you might never come close to that reserved capacity and would be wasting space that could be used for other workloads. For example, if all files in a workload are 288 bytes each in size, 16 billion files consume only ~4.6TB, which is well below the amount of capacity you'd need to get 16 billion files.

When deploying high file counts that use up little capacity, there are two main options for deploying the FlexGroup volume.

- **Deploy the FlexGroup volume with 7.8TB or greater member volumes with thin provisioning.** Thin provisioning a volume simply means that you are telling ONTAP a volume is a certain size, but that the size is not guaranteed in the file system. This provides flexibility in the file system to limit storage allocation to physical space. However, other volumes in the aggregate can affect the free capacity with their used space and if they have enabled space guarantees, so it's important to monitor available aggregate space when using thin provisioning. For more information, see the section "Overprovisioning or thin provisioning in a FlexGroup volume".
- **Manually create the FlexGroup volume with more member volumes than the default.** If you want to keep space guarantees for the FlexGroup volume, another option for high-file-count and small capacity environments is to create more member volumes in a FlexGroup volume.

Because inode counts are limited per FlexVol member volume according to capacity, adding more smaller member volumes can provide for higher file counts at the same capacity. The following table shows some possible configurations. For more information about manual creation of FlexGroup volumes, see the section “When do I need to manually create a FlexGroup volume?”

Table 20) High-file-count/small capacity footprint examples—increasing member volume counts.

Total FlexGroup size	Member volume count (size)	Maximum inode count (entire FlexGroup)
80TB (no space guarantee)	8 (10TB)	16,320,875,608
64TB (space guarantee enabled)	32 (2TB)	16,320,875,608
64TB (space guarantee enabled)	64 (1TB)	16,320,875,608

Planning for high file counts in ONTAP

With utilities like NetApp XCP (using the scan feature), you can evaluate your file count usage and other file statistics to help you make informed decisions about how to size your inode counts in your new FlexGroup volume. For more information about using XCP to scan files, contact ng-xcp-support@netapp.com.

Viewing used and total inodes

In ONTAP, you can view inode counts per volume by using the following command in advanced privilege:

```
cluster::*> volume show -volume flexgroup -fields files,files-used
vserver volume    files     files-used
------- --------- --------- ----------
SVM     flexgroup 170009008 823
```

You can also use the classic `df -i` command. To show all member volumes, use an asterisk with the volume name in diagnostic privilege:

```
cluster::*> df -i Tech_ONTAP*
Filesystem                iused      ifree  %iused  Mounted on                 Vserver
/vol/Tech_ONTAP/          10193  169998815      0%  /techontap                 DEMO
/vol/Tech_ONTAP__0001/      923   21250203      0%  /techontap                 DEMO
/vol/Tech_ONTAP__0002/     4177   21246949      0%  ---                        DEMO
/vol/Tech_ONTAP__0003/      878   21250248      0%  ---                        DEMO
/vol/Tech_ONTAP__0004/      848   21250278      0%  ---                        DEMO
/vol/Tech_ONTAP__0005/      750   21250376      0%  ---                        DEMO
/vol/Tech_ONTAP__0006/      972   21250154      0%  ---                        DEMO
/vol/Tech_ONTAP__0007/      879   21250247      0%  ---                        DEMO
/vol/Tech_ONTAP__0008/      766   21250360      0%  ---                        DEMO
```

What happens when you run out of inodes

When a volume runs out of inodes, no more files can be created in that volume until the number of inodes is increased or existing inodes are freed and the cluster triggers an EMS event (`callhome.no.inodes`). Additionally, a NetApp AutoSupport® message is triggered. Starting in ONTAP 9.3, a FlexGroup volume takes per-member inode numbers into account when deciding which member volumes are most optimal for data ingest. For examples, see the section “Inode-related EMS examples.”

EMS messages can be used for monitoring or for triggering scripts that automatically increase inode counts to help avoid space errors before they create production workload problems.

For information on increasing maximum files, see the section “Increasing maximum files: Considerations.”

Async delete

ONTAP 9.8 introduces a new feature that allows storage administrators to delete entire directories from the cluster CLI, rather than needing to perform deletions from NAS clients. This provides a way to remove high file count folders much faster than via NAS protocols, as well as removing network and client performance contention. This command works for both FlexVol and FlexGroup volumes.

In testing, async-delete performed almost 10 times faster than single threaded rm commands and is slightly faster on FlexVol volumes.

Table 21) Async-delete performance.

A300 (24,000 files/folders)	rm -rf * seconds	async-delete seconds	Speed increase
FlexVol	18.3	2	9.1x
FlexGroup	32.1	3	10.7x

When a directory deletion occurs with async delete, a job runs and creates several tasks that run in parallel to delete the directory. By default, the job throttles to 5,000 concurrent tasks, but that amount can be decreased to a minimum of 50 or increased to a maximum of 100,000.

When a delete command is issued, ONTAP scans the specified directory. If subdirectories are found, the contents of those directories are deleted first.

The following caveats apply:

- CLI only
- SVM and volumes must be valid
- Volume must be online and mounted
- Directory path must be valid
- Only one async-delete can be run at a time
- Must be run on a directory; cannot be run on single files

To run a delete job:

```
cluster::*> async-delete start -vserver DEMO -volume FlexGroup1 -path /files
[Job 34214] Job is queued: Asynchronous directory delete job.
```

To check progress:

```
cluster::*> async-delete show -vserver DEMO -instance
```

64-bit file identifiers

By default, NFS in ONTAP uses 32-bit file IDs. File IDs are unique identifiers in the file system that allows ONTAP to keep track of which files are which. 32-bit file IDs are limited to 2,147,483,647 maximum signed integers, which is where the two billion inode limit for FlexVols comes from.

FlexGroup volumes are able to support hundreds of billions of files in a single namespace by linking multiple member volumes together, but to get safely beyond the 32-bit signed integer limit of two billion (and remove the possibility of file ID collisions), 64-bit file IDs must be enabled.

ONTAP can hand out up to 4,294,967,295 file IDs (the 32-bit unsigned integer) in a FlexGroup volume when 32-bit file IDs are used before file ID collisions are guaranteed to occur. File ID collisions are mathematically impossible when there are 2,147,483,647 files, which is why that is the safest file count to use with 32-bit file IDs. After that value is exceeded, the likelihood of file ID collisions grows the closer the file count gets to the unsigned 32-bit integer value of 4,294,967,295. ONTAP does not prevent you from creating more than two billion files in a FlexGroup volume if you set the maxfiles value to a higher value. To learn more about what happens with file ID collisions, see the section “Effect of file ID collision.”

With 64-bit file IDs, ONTAP can allocate up to 9,223,372,036,854,775,807 unique file IDs to files, although the stated supported limit for maximum files in a FlexGroup volume is 400 billion.

The 64-bit file identifier option is set to off/disabled by default. This was by design to make certain that legacy applications and operating systems that require 32-bit file identifiers were not unexpectedly affected by ONTAP changes before administrators could properly evaluate their environments.

Note: Check with your application or operating system vendor for their support for 64-bit file IDs before enabling them or create a test SVM and enable it to see how applications and clients react with 64-bit file IDs. Most modern applications and operating systems can handle 64-bit file IDs without issue.

This option can be enabled with the following advanced privilege level command and has NFSv3 and NFSv4 options.

```
cluster::> set advanced
cluster::*> nfs modify -vserver SVM -v3-64bit-identifiers enabled -v4-64bit-identifiers enabled
```

Alternately, you can use [ONTAP System Manager] to enable or disable these values.

What happens when I modify this option?

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 on existing mounts. For more information about how enabling or disabling FSID change options can affect SVMs in high-file-count environments, see the section "How FSIDs operate with SVMs in high-file-count environments."

Do I have to enable 64-bit file IDs?

You might notice that, when you create a new FlexGroup volumes on an SVM that does not have 64-bit file IDs enabled, you get a warning that you should enable the option. However, since enabling the option forces you to remount volumes (and take an outage) and since some applications don't support 64-bit file IDs, you might not want to enable that option.

If your FlexGroup volumes do exceed two 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.

```
cluster::*> vol show -vserver SVM -volume flexgroup -fields files
```

Note: This option does not affect SMB operations and is unnecessary with volumes that use only SMB.

If your environment has volumes that need 32-bit and other volumes that require more than two billion files, then you can use different SVMs to host those volumes and enable or disable 64-bit file IDs as needed.

Best practice 16: 64-bit file identifiers

NetApp strongly recommends enabling the NFS server option `-v3-64bit-identifiers` at the advanced privilege level before you create a FlexGroup volume, especially if your file system exceeds or might exceed the two billion inode threshold.

NFSv3 versus NFSv4.x: File IDs

NFSv3 and NFSv4.x use different file ID semantics. Now that FlexGroup volumes support NFSv4.x, ONTAP 9.7 provides two different options for enabling or disabling 64-bit file IDs.

When you use both NFSv3 and NFSv4.x in an SVM and you want the 64-bit ID option to apply to both protocols, you must set both options.

If only one option is set and volumes are accessed by both protocols, you might see undesired behavior between protocols. For instance, NFSv3 might be able to create and view more than two billion files, whereas NFSv4.x would send an error when a file ID collision occurs.

The options are:

```
-v3-64bit-identifiers [enabled/disabled]
-v4-64bit-identifiers [enabled/disabled]
```

Note: If you upgrade to ONTAP 9.7 (the first release to support NFSv4.x on FlexGroup volumes), upgrade to 9.7P7 or later to avoid exposure to bug 1336512.

Using quota enforcement to limit file count

Starting with ONTAP 9.5, it is possible to set up a quota policy that prevents a FlexGroup volume from exceeding two billion files if 32-bit file handles are still being used by way of quota enforcement.

Because quota enforcement policies do not apply to files created below the parent volume (only monitoring/reporting policies), create a qtree inside the FlexGroup volume. Then create a quota tree rule for that qtree with two billion files as the limit to help reduce the risk of users overrunning the 32-bit file ID limitations. Alternately, you can create specific user or group quota rules if you know the user names and group names that will be creating files in the volume.

```
cluster::*> qtree create -vserver DEMO -volume FG4 -qtree twobillionfiles -security-style unix -
oplock-mode enable -unix-permissions 777
cluster::*> quota policy rule create -vserver DEMO -policy-name files -volume FG4 -type tree -
target "" -file-limit 2000000000
cluster::*> quota on -vserver DEMO -volume FG4
[Job 15906] Job is queued: "quota on" performed for quota policy "tree" on volume "FG4" in
Vserver "DEMO".
cluster::*> quota resize -vserver DEMO -volume FG4
[Job 15907] Job is queued: "quota resize" performed for quota policy "tree" on volume "FG4" in
Vserver "DEMO".
cluster::*> quota report -vserver DEMO -volume FG4
Vserver: DEMO

                                    ----Disk----  ----Files-----   Quota
Volume   Tree      Type    ID        Used  Limit    Used   Limit   Specifier
-------  --------  ------  -------  -----  -----  ------  ------   ---------
FG4      twobillionfiles
                   tree    1           0B      -       1
                                                    2000000000   twobillionfiles
FG4                tree    *           0B      -       0
                                                    2000000000   *
2 entries were displayed.
```

After that is done, use file permissions and/or export policy rules to limit access and prevent users from creating files at the volume level. Apply SMB shares to the qtree rather than the volume, and NFS mounts should occur at the qtree level.

Then, as files are created in the qtree, they count against the limit.

```
[root@centos7 home]# cd /FG4/twobillionfiles/
[root@centos7 twobillionfiles]# ls
[root@centos7 twobillionfiles]# touch new1
[root@centos7 twobillionfiles]# touch new2
[root@centos7 twobillionfiles]# touch new3
[root@centos7 twobillionfiles]# ls
new1  new2  new3
cluster::*> quota report -vserver DEMO -volume FG4
Vserver: DEMO

                                    ----Disk----  ----Files-----   Quota
Volume   Tree      Type    ID        Used  Limit    Used   Limit   Specifier
-------  --------  ------  -------  -----  -----  ------  ------   ---------
FG4      twobillionfiles
```

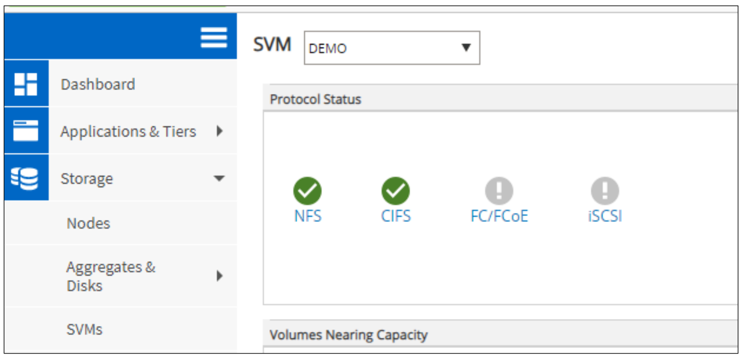


```
                  tree    1           0B      -      4
                                                   2000000000    twobillionfiles
FG4               tree    *           0B      -      0
                                                   2000000000    *
```

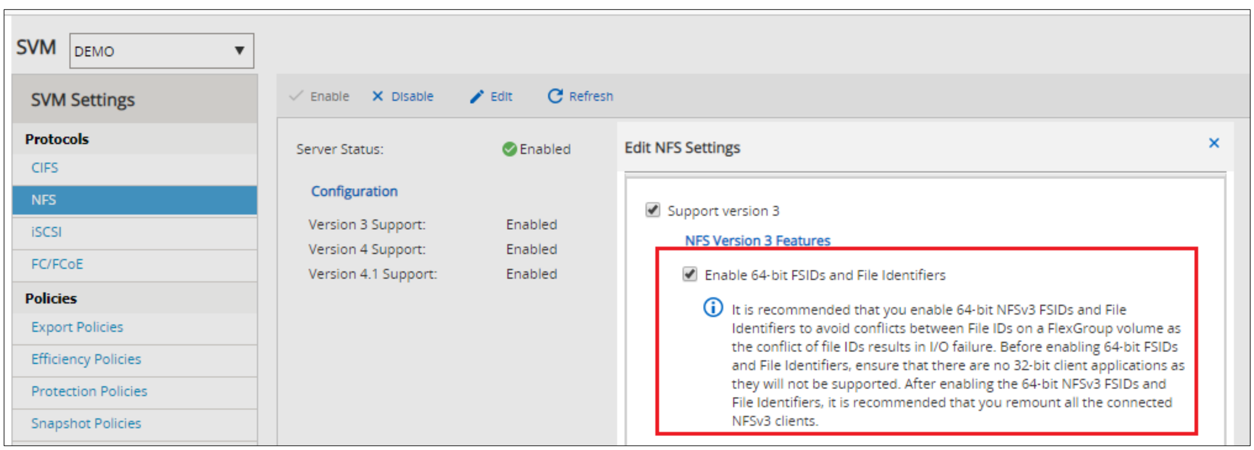
System Manager support for the 64-bit file ID option (classic view)

Starting in ONTAP 9.2, it is also possible to enable or disable the NFS server option from ONTAP System Manager.

1. Select Storage > SVMs. Select the desired SVM and then click NFS.



2. Click Edit. The dialog box that opens contains a checkbox that you can use to enable or disable 64-bit file identifiers.



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

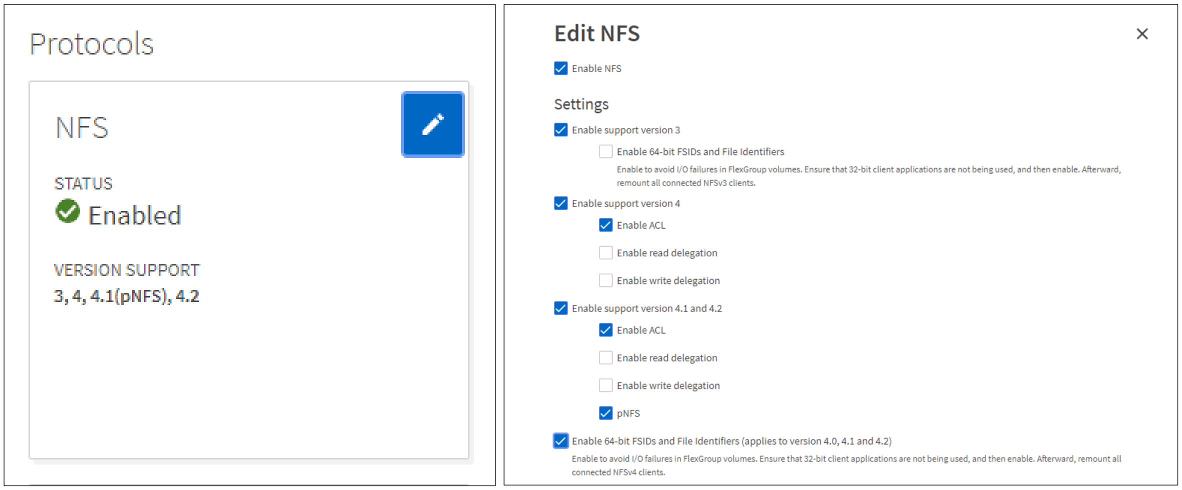
ONTAP System Manager: 9.7

ONTAP 9.7 introduced a new System Manager interface based on REST API capabilities. Because the 64-bit file ID option does not currently exist in the REST API, the only way to modify it in System Manager is to use the classic view. Alternately, use the CLI.

ONTAP System Manager: 9.8 and later

ONTAP System Manager 9.8 and later includes a GUI method for enabling or disabling the 64-bit file ID value from the Storage > Storage VMs menu option. Click on the desired SVM and select Edit from the NFS protocol menu.

Figure 83) 64-bit file IDs in ONTAP System Manager 9.8



Effect of file ID collision

If 64-bit file IDs are not enabled, the risk for file ID collisions increases. When a file ID collision occurs, the effect can range from a stale file handle error on the client, to the failure of directory and file listings, to the entire failure of an application. Usually, it is imperative to enable the 64-bit file ID option when you use FlexGroup volumes.

You can check a file's ID from the client using the `stat` or `ls -i` command. When an inode or file ID collision occurs, it might look like the following. The inode is `3509598283` for both files.

```
# stat libs/
  File: `libs/'
  Size: 12288           Blocks: 24         IO Block: 65536  directory
Device: 4ch/76d Inode: 3509598283  Links: 3
Access: (0755/drwxr-xr-x)  Uid: (60317/   user1)   Gid: (10115/      group1)
Access: 2017-01-06 16:00:28.207087000 -0700
Modify: 2017-01-06 15:46:50.608126000 -0700
Change: 2017-01-06 15:46:50.608126000 -0700

# stat iterable/
  File: `iterable/'
  Size: 4096            Blocks: 8          IO Block: 65536  directory
Device: 4ch/76d Inode: 3509598283  Links: 2
Access: (0755/drwxr-xr-x)  Uid: (60317/   user1)   Gid: (10115/      group1)
Access: 2017-01-06 16:00:44.079145000 -0700
Modify: 2016-05-05 15:12:11.000000000 -0600
Change: 2017-01-06 15:23:58.527329000 -0700

# ls -i libs
3509598283 libs

# ls -i iterable
3509598283 iterable
```

A collision can result in issues such as circular directory structure errors on the Linux client during `find` or `rm` commands and an inability to remove files. In some cases, you may even see `stale file handle` errors.

```
rm: WARNING: Circular directory structure.
This almost certainly means that you have a corrupted file system.
NOTIFY YOUR SYSTEM MANAGER.
The following directory is part of the cycle:
  '/directory/iterable'
```



```
rm: cannot remove '/directory': Directory not empty
```

Note: File ID collisions affect NFS only. SMB does not use the same file ID structure.

Effects of file system ID changes in ONTAP

NFS uses 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` and `chmod`.

An example of this issue can be found in [bug 671319](). If you disable the FSID change option (for NFSv3 or NFSv4), be sure to enable the 64-bit file ID option on the NFS server (see the section "64-bit file identifiers"), because the total number of file IDs are shared across volumes in the SVM and you run the risk of hitting file ID collisions sooner.

This FSID change option could also affect older legacy applications that require 32-bit file IDs. Perform the appropriate testing with your applications in a separate SVM before toggling FSID change.

How FSIDs operate with SVMs in high-file-count environments

The FSID change option for NFSv3 and NFSv4.x provides FlexVol and FlexGroup volumes with their own unique file systems, which means that the number of files allowed in the SVM is dictated by the number of volumes. However, disabling the FSID change options cause the 32-bit or 64-bit file identifiers to apply to the SVM itself, meaning that the file limits with 32-bit file IDs would apply to all volumes.

For example, if you have 10 billion files in 10 different volumes in your SVM, leaving the FSID change option enabled ensures that each volume can have its own set of unique file IDs. If you disable the FSID change option, then all 10 billion files share the pool of file IDs in the SVM. With 32-bit file IDs, you will likely see file collisions.

NetApp recommends leaving the FSID change option enabled with FlexGroup volumes to help prevent file ID collisions.

How FSIDs operate with Snapshot copies

When a Snapshot copy of a volume is created, 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 by 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 `-v3-fsid-change` option. The NFS RFC specification does not require FSIDs for a file to be identical across file versions.

Note: The `-v4-fsid-change` option does not apply to FlexGroup volumes prior to ONTAP 9.7, because NFSv4 is unsupported with FlexGroup volumes in those releases.

Directory size considerations: maxdirsize

In ONTAP, there are limitations to the maximum directory size on disk. This limit is known as [maxdirsize](). The `maxdirsize` value for a volume is capped at 320MB, regardless of platform. This means that the memory allocation for the directory size can reach a maximum of only 320MB before a directory can no longer grow larger. Directory sizes grow when file counts in a single directory increase. Each file entry in a directory counts against the allocated space for the directory. For information on how many files you can have in a single directory, see the section "Number of files that can fit into a single directory with the default maxdirsize."

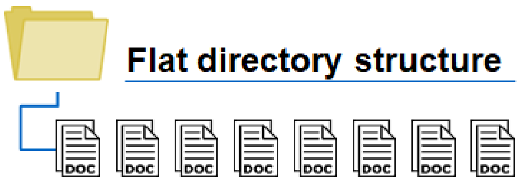
Best practice 17: Recommended ONTAP version for high file count environments

For high-file-count environments, use the latest ONTAP release available to gain the benefit of FlexGroup feature enhancements, WAFL enhancements, and performance improvements for high file count workloads.

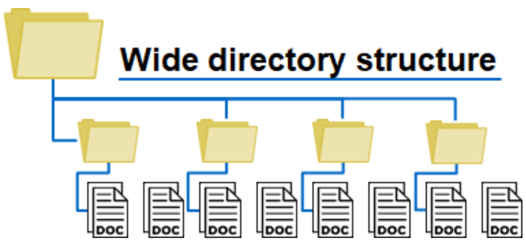
What directory structures can affect maxdirsize?

The `maxdirsize` value can be a concern when you are using flat directory structures, where a single folder contains millions of files at a single level. Folder structures where files, folders, and subfolders are interspersed have a low impact on `maxdirsize`. There are several directory structure methodologies.

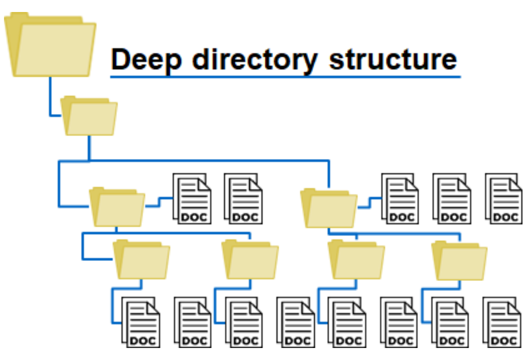
- **Flat directory structure.** A single directory with many files.



- **Wide directory structure.** Many top-level directories with files spread across directories.



- **Deep directory structures.** Fewer top-level directories, but with many subfolders; files spread across directories.



How flat directory structures can affect FlexGroup volumes

Flat directory structures (many files in a single or few directories) have a negative effect on a wide array of file systems, whether they are NetApp systems or not. Potential issues can include, but are not limited to:

- Memory pressure
- CPU utilization
- Network performance/latency (particularly during mass queries of files, `GETATTR` operations, `READDIR` operations, and so on)

FlexGroup volumes can also have an extra effect on `maxdirsize`. Unlike a FlexVol volume, a FlexGroup volume uses remote hard links inside ONTAP to help redirect traffic. These remote hard links are what allow a FlexGroup volume to deliver scale-out performance and capacity in a cluster.

However, in flat directories, a higher ratio of remote hard links to local files is seen. These remote hard links count against the total `maxdirsize` value, so a FlexGroup volume might approach the `maxdirsize` limit faster than a FlexVol will.

For example, if a directory has millions of files in it and generates roughly 85% remote hard links for the file system, you can expect `maxdirsize` to be exhausted at nearly twice the amount as a FlexVol would.

Best practice 18: Directory structure recommendation

- For the best performance, avoid flat directory structures in ONTAP if at all possible. Wide or deep directory structures work best, as long as the path length of the file or folder does not exceed NAS protocol standards.
- If flat directory structures are unavoidable, pay close attention to the `maxdirsize` values for the volume and increase them as necessary with the guidance of NetApp support.
- NFS path lengths are defined by the client OS.
- For information about SMB path lengths, see this Microsoft Dev Center link.

Querying for used maxdirsize values

It is important to monitor and evaluate `maxdirsize` allocation in ONTAP. However, there are no commands for this specific to ONTAP. Bug 1336142 has been filed to add this functionality, so if you need this added to ONTAP, open a support case and have it attached to the bug.

Instead, `maxdirsize` allocation would need to be queried from the client.

The following command from an NFS client is able to retrieve the directory size information for a folder inside a FlexGroup volume for the 10 largest directories in a given mount point, while omitting Snapshot copies from the search.

```
# find /mountpoint -name .snapshot -prune -o -type d -ls -links 2 -prune | sort -rn -k 7 | head
```

The following example took less than a second on a dataset in folders with millions of files:

```
[root@centos7 /]# time find /flexgroup/manyfiles/ -name .snapshot -prune -o -type d -ls -links 2
-prune | sort -rn -k 7 | head
787227871 328976 drwxr-xr-x   2 root     root     335544320 May 29 21:23
/flexgroup/manyfiles/folder3/topdir_8/subdir_0
384566806 328976 drwxr-xr-x   2 root     root     335544320 May 29 13:14
/flexgroup/manyfiles/folder3/topdir_9/subdir_0
3605793347 328976 drwxr-xr-x   2 root     root     335544320 May 29 21:23
/flexgroup/manyfiles/folder3/topdir_0/subdir_0
3471151639 328976 drwxr-xr-x   2 root     root     335544320 May 29 13:45
/flexgroup/manyfiles/folder3/topdir_4/subdir_0
2532103978 328976 drwxr-xr-x   2 root     root     335544320 May 29 14:16
/flexgroup/manyfiles/folder3/topdir_2/subdir_0
2397949155 328976 drwxr-xr-x   2 root     root     335544320 May 29 14:15
/flexgroup/manyfiles/folder3/topdir_1/subdir_0
1994984460 328976 drwxr-xr-x   2 root     root     335544320 May 29 13:43
/flexgroup/manyfiles/folder3/topdir_6/subdir_0
1860674357 328976 drwxr-xr-x   2 root     root     335544320 May 29 13:18
/flexgroup/manyfiles/folder3/topdir_5/subdir_0
1458235096 328976 drwxr-xr-x   2 root     root     335544320 May 29 14:25
/flexgroup/manyfiles/folder3/topdir_3/subdir_0
1325327652 328976 drwxr-xr-x   2 root     root     335544320 May 29 14:25
/flexgroup/manyfiles/folder3/topdir_7/subdir_0

real    0m0.055s
user    0m0.002s
sys     0m0.035s
```


Using XCP to check maxdirsize

NetApp XCP is mostly considered to be a rapid data mover, but it also derives value in its robust file scanning capabilities. XCP is able to run `find` commands in parallel as well, so the previous examples can be run even faster on the storage system as well as filter results to directories with specified file counts.

The following XCP command example allows you to run `find` only on directories with more than 2,000 entries:

```
# xcp diag find --branch-match True -fmt "'{size} {name}'.format(size=x.digest, name=x)"
localhost:/usr 2>/dev/null | awk '{if ($1 > 2000) print $1 " " $2}'
```

This XCP command helps you find the directory size values:

```
# xcp -match "type == d" -fmt "'{} {}'.format(used, x)" localhost:/usr | awk '{if ($1 > 100000)
print}' | sort -nr
```

When XCP looks for the directory size values, it scans the file system first. Here's an example:

```
[root@XCP flexgroup]#  xcp -match "type == d" -fmt "'{} {}'.format(used, x)"
10.193.67.219:/flexgroup_16/manyfiles | awk '{if ($1 > 100000) print}' | sort -nr

 660,693 scanned, 54 matched, 123 MiB in (24.6 MiB/s), 614 KiB out (122 KiB/s), 5s
 1.25M scanned, 58 matched, 234 MiB in (22.1 MiB/s), 1.13 MiB out (109 KiB/s), 10s
 …
 31.8M scanned, 66 matched, 5.83 GiB in (4.63 MiB/s), 28.8 MiB out (22.8 KiB/s), 7m52s

Filtered: 31816172 did not match
31.8M scanned, 66 matched, 5.83 GiB in (12.6 MiB/s), 28.8 MiB out (62.4 KiB/s), 7m53s.
336871424 10.193.67.219:/flexgroup_16/manyfiles/folder3/topdir_9/subdir_0
336871424 10.193.67.219:/flexgroup_16/manyfiles/folder3/topdir_8/subdir_0
336871424 10.193.67.219:/flexgroup_16/manyfiles/folder3/topdir_7/subdir_0
336871424 10.193.67.219:/flexgroup_16/manyfiles/folder3/topdir_6/subdir_0
336871424 10.193.67.219:/flexgroup_16/manyfiles/folder3/topdir_5/subdir_0
336871424 10.193.67.219:/flexgroup_16/manyfiles/folder3/topdir_4/subdir_0
336871424 10.193.67.219:/flexgroup_16/manyfiles/folder3/topdir_3/subdir_0
```

Number of files that can fit into a single directory with the default maxdirsize

To determine how many files can fit into a single directory with the default `maxdirsize` setting, use this formula:

- Memory in KB * 53 *25%

Since `maxdirsize` is set to 320MB by default on larger systems, the maximum number of files in a single directory is 4,341,760 for SMB and NFS on FlexVol volumes.

FlexGroup volumes use remote hard links to redirect I/O to member volumes. These hard links count against the total directory size, so the maximum number of files allowed with 320MB maxdirsize would depend on the number of hard links that were created. The file count per directory might be in the 2 to 2.6 million range for directories in a FlexGroup volume.

NetApp strongly recommends that you keep the `maxdirsize` value at the default value.

Event management system messages sent when maxdirsize is exceeded

The following event management system (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 with the ONTAP System Manager event section. Active IQ Unified Manager can be used to monitor `maxdirsize`, trigger alarms, and send a notification before the 90% threshold (see the section "Capacity monitoring and alerting"). These event management system messages also support SNMP traps.

```
wafl.dir.size.max
wafl.dir.size.max.warning
wafl.dir.size.warning
```

Effect of increasing the maxdirsize value

When a single directory contains many files, lookups (such as in a find operation) can consume large amounts of CPU and memory. Starting in ONTAP 9.2, 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. Usually, this helps large directory performance. However, for wildcard searches and `readdir` operations, indexing is not of much use. When possible, use the latest version of ONTAP for high file count environments to gain benefits from WAFL improvements.

Best practice 19: Maxdirsize maximums

Values for `maxdirsize` are hard coded not to be able to exceed 4GB. To avoid performance issues, NetApp recommends setting `maxdirsize` values no higher than 1GB.

Do FlexGroup volumes bypass maxdirsize limitations?

In FlexGroup volumes, each member volume has the same `maxdirsize` setting (which is configured at the FlexGroup level). Even though the files in a directory could potentially span multiple FlexVol member volumes and nodes, the directory itself resides on a single member volume. As a result, the same `maxdirsize` limitations you see in a FlexVol volume still come into play with a FlexGroup. This is because directory size is the key component, not the volume. In a FlexGroup volume, since a directory would reside in a single FlexVol member volume, there is no relief for environments facing `maxdirsize` limitations.

Best practice 20: Avoiding maxdirsize issues

Newer platforms offer more memory and CPU capacity, and AFF systems provide performance benefits for high-file-count environments. However, the best way to reduce the performance effect in directories with large numbers of files is to spread files across more directories in the file system.

Effect of exceeding maxdirsize

When `maxdirsize` is exceeded in ONTAP, an `out of space` error (`ENOSPC`) is issued to the client and an event management system message is triggered. This error can be misleading to storage administrators because they imply an actual capacity issue when the problem in this case has to do with file count. Always check the ONTAP event log to narrow down problems when clients report seeing capacity issues.

To remediate a directory size issue, a storage administrator must increase the `maxdirsize` setting or move files out of the directory. For more information about remediation, see KB 000002080 on the NetApp Support site. For examples of the `maxdirsize` event management system events, see the section “Example of maxdirsize message.”

File system analytics

ONTAP 9.8 introduced a new feature that provides a way called File System Analytics for storage administrators to get instant access to file and directory information from ONTAP System Manager.

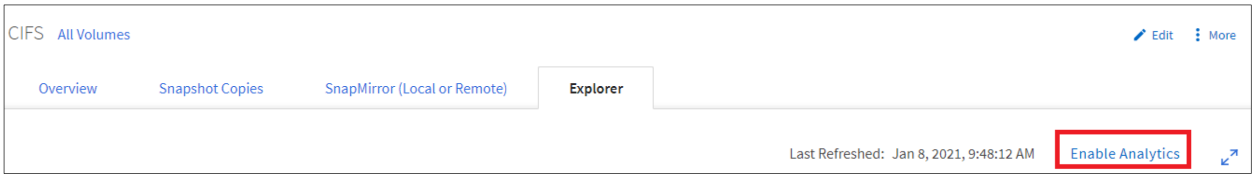
This initial release of FSA includes information such as:

- File sizes
- Folder sizes

- Atime and mtime histograms
- File and folder listings
- Inactive and active data reporting
- File and directory counts

This information is gathered by ONTAP as the file system is updated after an initial scan is performed and takes minimal system resources to use. File System Analytics are off by default and can be enabled (and disabled) via ONTAP System Manager from the new Explorer tab on the volume page for both FlexVol and FlexGroup volumes, regardless of the NAS protocol in use.

Figure 84) File System Analytics – enable.



After analytics are enabled and the initial scan completes (completion time depends on file and folder count), you can browse the entire directory structure by clicking through the directory trees in ONTAP System Manager's Explorer tab.

Figure 85) File System Analytics – directory and file information.

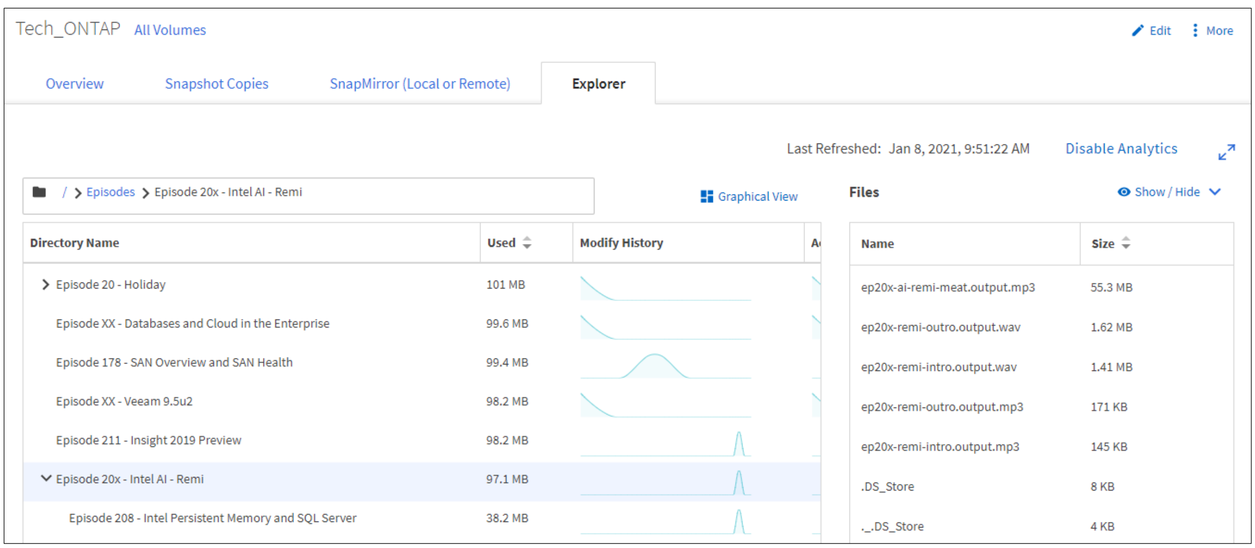
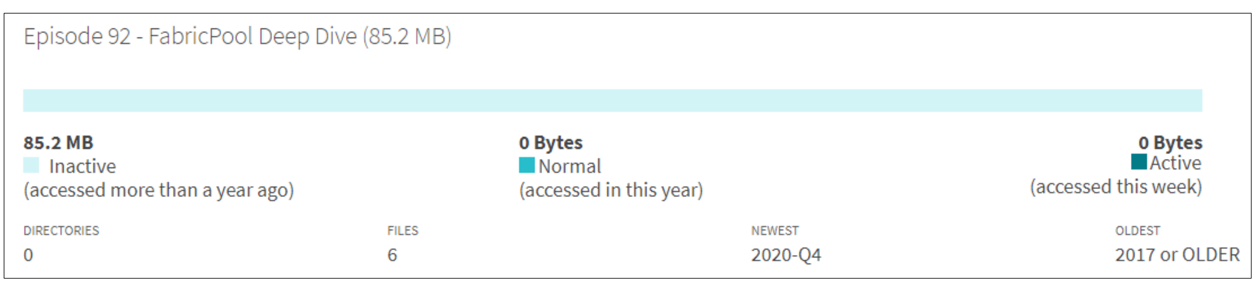


Figure 86) File System Analytics – inactive and active data.



When files and folders are created or deleted, File System Analytics updates the tree in seconds with the new information. File System Analytics allows storage administrators to get file and folder information without the need to use off-box utilities or commands such as `du`, `find` and `ls`, which can be time-intensive in high file count environments.

For more information on File System Analytics, including best practices and REST API examples, see the following resources:

- ONTAP File System Analytics: Better visibility for better management
- File System Analytics overview
- TR-4867: Best Practice Guide for File System Analytics
- TR-4863: Best Practice Guidelines for XCP
- Tech ONTAP Podcast Episode 270: NetApp ONTAP File System Analytics (audio podcast)

Special character considerations

Most common text characters in Unicode (when they are encoded with UTF-8 format) use encoding that is equal to or smaller than three bytes. This common text includes all modern written languages, such as Chinese, Japanese, and German. However, with the popularity of special characters such as the emoji, some UTF-8-character sizes have grown beyond three bytes. For example, a trophy symbol is a character that requires four bytes in UTF-8 encoding.

Special characters include, but are not limited to, the following:

- Emojis
- Music symbols
- Mathematical symbols

When a special character is written to a FlexGroup volume, the following behavior occurs:

```
# mkdir /flexgroup4TB/🏆
mkdir: cannot create directory '/flexgroup4TB/\360\237\217\206': Permission denied
```

In the preceding example, `\360\237\217\206` is hex `0xF0 0x9F 0x8F 0x86` in UTF-8, which is a trophy symbol.

ONTAP software did not natively support UTF-8 sizes that are greater than three bytes in NFS, as indicated in bug 229629. To handle character sizes that exceed three bytes, ONTAP places the extra bytes into an area in the operating system known as `bagofbits`. These bits are stored until the client requests them. Then the client interprets the character from the raw bits. FlexVol supports `bagofbits`, and FlexGroup volumes added support for `bagofbits` in ONTAP 9.2.

Best practice 21: Special character handling in FlexGroup volumes

For optimal special character handling with FlexGroup volumes, use ONTAP 9.5 or later and the utf8mb4 volume language.

Also, ONTAP has an event management system message for issues with `bagofbits` handling, which includes how to identify the offending file ID.

```
Message Name: wafl.bagofbits.name
Severity: ERROR

Corrective Action: Use the "volume file show-inode" command with the file ID and volume name
information to find the file path. Access the parent directory from an NFSv3 client and rename
the entry using Unicode characters.
```



```
Description: This message occurs when a read directory request from an NFSv4 client is made to a
Unicode-based directory in which directory entries with no NFS alternate name contain non-Unicode
characters.
```

To test `bagofbits` functionality in FlexGroup, use the following command:

```
# touch "$(echo -e "file\xFC")"
```

In ONTAP 9.1, this would fail:

```
# touch "$(echo -e "file\xFC")"
touch: cannot touch `file\374': Permission denied
```

In ONTAP 9.2 and later, this command succeeds:

```
# touch "$(echo -e "file\xFC")"
# ls -la
-rw-r--r--.  1 root root    0 May  9  2017 file?
```

Support for utf8mb4 volume language

As mentioned before, special characters might exceed the supported three bytes UTF-8 encoding that is natively supported. ONTAP then uses the `bagofbits` functionality to allow these characters to work.

This method for storing inode information is not ideal, so, starting in ONTAP 9.5, utf8mb4 volume language support was added. When a volume uses this language, special characters that are four bytes in size are stored properly and not in `bagofbits`.

Volume language is used to convert names sent by NFSv3 clients to Unicode, and to convert on-disk Unicode names to the encoding expected by NFSv3 clients. In legacy situations in which NFS hosts are configured to use non-UTF-8 encodings, you should use the corresponding volume language. Use of UTF-8 has become almost universal these days, so the volume language is likely to be UTF-8.

NFSv4 requires use of UTF-8, so there is no need to use non-UTF-8 encoding for NFSv4 hosts. Similarly, CIFS uses Unicode natively, so it works with any volume language. However, use of utf8mb4 is recommended because files with Unicode names above the basic plane are not converted properly on non-utf8mb4 volumes.

Volume language can only be set on a volume at creation by using the `-language` option. You cannot covert a volume's language. To use files with a new volume language, create the volume and migrate the files by using a utility like the NetApp XCP.

Best practice 22: UTF-8 or utf8mb4?

If you're running ONTAP 9.5 or later, it is best to use the utf8mb4 volume language to help prevent issues with filename translation unless clients are unable to support the language.

Managing slow directory listings via NFS in high-file-count environments

Some workflows in high-file-count environments include running `find`, `ls`, or other read metadata-heavy operation on an existing dataset. This type of workload is inefficient and can take a long time to complete. If it is necessary to run these operations, there are a few things you can try to help speed things along.

Generally speaking, the issue with these types of operations is client, protocol, or network related. The storage rarely is the bottleneck for read metadata slowness. ONTAP is able to multithread read metadata operations. With `ls` operations, `getattr` requests are sent one at a time, in serial, which means for millions of `getattr` operations, there might be millions of network requests to the storage. Each network request incurs n milliseconds of latency, which adds up over time.

As such, there are a few ways to speed these up:

- **Send more** getattr **requests at a time.** By itself, `ls` can't send requests in parallel. But with utilities like XCP, it is possible to send multiple threads across the network to greatly speed up `ls` operations. Using XCP scan can help with speed, depending on what the `ls` output is being used for later. For example, if you need the user permissions/owners of the files, using `ls` by itself might be a better fit. But for sheer listing of file names, XCP scan is preferable.
- **Add more network hardware (for example, 100GB instead of 10GB) to reduce round-trip time (RTT).** With larger network pipes, more traffic can be pushed over the network, thus reducing load, and potentially reducing overall RTT. With millions of operations, even shaving off a millisecond of latency can add up to a large amount of time saved for workloads.
- **Run** ls **without unnecessary options, such as highlighting/colors.** When running `ls`, the default behavior is to add sorting, colors, and highlighting for readability. These add work for the operation, so it might make sense to run `ls` with the `–f` option to avoid those potentially unnecessary features.
- **Cache** getattr **operations on the client more aggressively.** Client-side caching of attributes can help reduce the network traffic for operations, as well as bringing the attributes local to the client for operations. Clients manage NFS caches differently, but in general, avoid setting `noac` on NFS mounts for high-file-count environments. Also, keep `actimeo` to a level no less than 30 seconds.
- **Create FlexCache volumes.** NetApp FlexCache volumes are able to create instant caches for read-heavy workloads. Creating FlexCache volumes for workloads that do a lot of read metadata operations, such as `ls`, can have the following benefits:
  - For local clusters, it can help offload the read metadata operations from the origin volume to the cache volumes, and, as a result, frees the origin volume up for regular reads and writes.
  - FlexCache volumes can reside on any node in a cluster, so creating FlexCache volumes makes the use of cluster nodes more efficient by allowing multiple nodes to participate in these operations, in addition to moving the read metadata operations away from the origin node.
  - For remote clusters across a WAN, FlexCache volumes can provide localized NFS caches to help reduce WAN latency, which can greatly improve performance for read-metadata-heavy workloads.

  When using FlexCache volumes to help read metadata workloads, be sure to disable `fastreaddir` on the nodes that use FlexCache.

```
cluster::> node run "priv set diag; flexgroup set fastreaddir=false persist
```

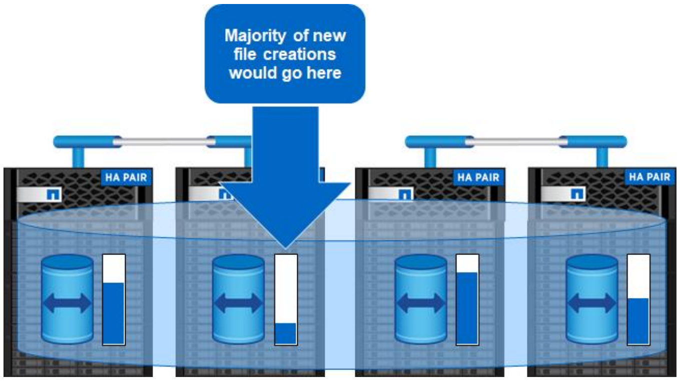
Note: For this to take effect, a reboot or storage failover is required.

Note: Starting in ONTAP 9.7, FlexGroup volumes can be origins for FlexCache volumes. For more information about FlexCache volumes, see TR-4743: FlexCache in NetApp ONTAP.

File deletions/FlexGroup member volume balancing

A FlexGroup volume spreads data across multiple member volumes relatively evenly on ingest of data. This data layout can help file deletions operate a bit more efficiently on a FlexGroup volume as compared to a FlexVol volume, as the system is able to use more hardware and WAFL affinities to spread out the delete load more efficiently and use less CPU per node for these operations.

Figure 87) Capacity imbalance after deletion of larger files.



However, overall performance of file deletions might be slower because of remote access across the FlexGroup volume as compared to FlexVol volumes. In rare cases, the deletion of files (especially sets of large files) can create artificial hot spots in a FlexGroup volume by way of capacity imbalances.

A FlexGroup volume's workload balance can be viewed with the following `diag-privilege-level` command:

```
cluster::*> set diag
cluster::*> node run * flexgroup show [flexgroup name]
```

This displays the following output:

- Member volume dataset ID (DSID)
- Member volume capacities (used and available, in blocks)
- Member volume used %
- Urgency, target, and probability percentages (used in ingest calculations)

For more information, see the section "flexgroup show."

Rebalancing data within a FlexGroup volume

Starting in ONTAP 9.10.1, new functionality is added to make it possible to use the ONTAP CLI to move larger files in a FlexGroup member volume to another member volume in the same FlexGroup configuration to allow for better data balance. This new command is available in diagnostic privilege mode and is disruptive. For more information about the command and how to use it, contact NetApp Support. In most cases, it is not necessary to rebalance data, however ONTAP generally does a good job of balancing the ingest load so that new writes redirect to less full member volumes, and, with the ONTAP 9.8 proactive resizing feature, ONTAP grows and shrinks member volumes as needed to maintain an even buffer of available free space, so a rebalance is not necessary. A data imbalance does not mean there will be a performance issue unless the data imbalance is also accompanied by very full member volumes. In those cases, performance issues are only seen during new file creation.

In the rare case in which a member volume grows significantly larger than other member volumes, you should analyze the workload to see if anything has changed (for example, the workload went from creating 1MB files to 100GB files). You can use XCP software to scan folders and files to identify file sizes and anomalies. One common scenario that can over allocate a single member volume is if an end user zips up a large amount of data in the FlexGroup. That single zip file might grow to be very large and can fill up a member volume.

For an example of scanning files with XCP, see the section "Using XCP to scan files before migration."

After the files are identified, either delete them, move them to other volumes, add space to the member volumes, or add new member volumes to help balance the ingest load in a FlexGroup volume. Ideally, upgrade the cluster to ONTAP 9.8 to gain the benefits of proactive resizing, which helps remove the management overhead for member volume capacity.

Why doesn't a FlexGroup volume automatically rebalance existing data?

As a FlexGroup volume ingests data, it has three goals:

- The volume should encourage all its member FlexVol volumes to participate in hosting the workload in parallel. If only a subset of member volumes is active, the FlexGroup volume should distribute more new data toward the underactive members.
- The FlexGroup volume should prevent any member FlexVol volume from running out of free space unless all other members are also out of free space. When one member has more data than others, the FlexGroup volume should align the underused members by placing new data on them at a higher-than-average rate.
- The FlexGroup volume must minimize the performance losses caused by pursuing the previous two goals. If the FlexGroup volume were to place each new file carefully and accurately where it could be most beneficial, then the previous two goals could be easily achieved. However, the cost of all that careful placement would appear as increased service latency. An ideal FlexGroup volume blends performance with capacity balance but favors performance.

Some of these goals are in conflict, so ONTAP employs a sophisticated set of algorithms and heuristics to maintain a balance in the FlexGroup volumes. However, in some scenarios, imbalances such as the following might occur:

- Large files or files that grow over time might be present in a FlexVol member volume.
- A workload changes from smaller files to large files (such as a change in how video surveillance cameras record from 4K resolution to 8K resolution).
- Many files might be zipped or tarred into a single file in the same FlexGroup volume as the files themselves.
- A large amount of data might be deleted, and most of that data could be from the same member volume (rare).

In scenarios where FlexGroup member volumes have an imbalance of capacity or files, ONTAP takes extra measures to help the less-allocated member volumes catch up to the filled members. As a result, performance can be affected for new file creations. Existing data should see little to no effect.

Performance issues when member volumes reach 80% used capacity

In versions earlier than ONTAP 9.5P4, the ingest heuristics could encounter an issue that might contribute to performance issues when a member volume capacity reaches 80%. If possible, when using FlexGroup volumes, upgrade to ONTAP 9.5P4 or later. For more information about the issue, check bug 1231125.

ONTAP 9.8 and proactive resizing further mitigates performance impact when member volumes reach a capacity threshold and is the preferred ONTAP version for FlexGroup volumes.

Listing files when a member volume is out of space

If a FlexGroup member volume runs out of space, the entire FlexGroup volume reports that it is out of space. Even read operations, such as listing the contents of a folder, can fail when a FlexGroup member is out of space.

Although `ls` is a read-only operation, FlexGroup volumes still require a small amount of writable space to allow it to work properly. ONTAP uses that storage to establish metadata caches. For example, suppose the name `foo` points to an inode with X properties, and the name `bar` points to an inode with Y properties. The amount of space used is negligible—a few kilobytes, or maybe a few megabytes on large systems—and this space is used and released repeatedly. Internally, this space is called the RAL reserve.

Under normal circumstances, even if you manage to fill up a member volume, a bit of space is left for the FlexGroup volume to use as it performs read-only operations like `ls`. However, ONTAP prioritizes other operations over the RAL reserve. If a member volume is 100% full, for example, and you create a Snapshot copy and then try to continue using the volume, the WAFL Snapshot reserve is used as you overwrite blocks and therefore consumes more space. ONTAP prioritizes the Snapshot space and takes space from things like the RAL reserve. This scenario rarely occurs, but it explains why an operation like `ls` might fail because of lack of space.

File rename considerations

FlexGroup volumes handle most high-metadata workloads well. However, with workloads that do a large amount of file renames at a time (for example, hundreds of thousands), performance of these operations suffers in comparison to FlexVol volumes. This is because a file rename does not move the file in the file system; instead, it just moves the file name to a new location. In a FlexGroup volume, moving this name would likely take place as a remote operation and create a remote hard link. Subsequent renames would create more remote hard links to the file's location, which would keep adding latency to operations that occur on that file. If an application's workflow is mostly file renames, you should consider using FlexVol volumes instead of FlexGroup volumes. If the desired final landing spot is a FlexGroup volume after the rename occurs, consider moving the files from the FlexVol volume to the FlexGroup volume after the rename process.

Symlink considerations

If your workload contains many symlinks (that is, symlink counts in the millions) in a single FlexGroup volume, attempts to resolve that many symlinks might have a negative effect on performance. The negative effect is caused by creating remote hard links artificially in addition to the remote hard links ONTAP creates.

Best practice 23: Symlinks in FlexGroup volumes
Try to keep the number of symlinks below a few thousand per FlexGroup if possible.

NFS version considerations

When a client using NFS attempts to mount a volume in ONTAP without specifying the NFS version (for example, `-o nfsvers=3`), a protocol version negotiation between the client and server takes place. The client asks for the highest versions of NFS supported by the server. If the server (in the case of ONTAP, an SVM serving NFS) has NFSv4.x enabled, the client attempts to mount with that version.

In ONTAP 9.6 and earlier, FlexGroup volumes did not support NFSv4.x, so the mount request fails in those releases. Usually, this error manifests as `access denied`, which can mask what the actual issue is in the environment:

```
# mount demo:/flexgroup /flexgroup
mount.nfs: access denied by server while mounting demo:/flexgroup
```

In ONTAP 9.7 and later, NFSv4.x is supported. This can create a different set of issues, however. Clients still mount the latest NFS version advertised by the NFS server (in this case, the ONTAP SVM). If NFSv4.x versions are enabled, clients might mount through NFSv4.x when NFSv3 is desired or expected. When NFSv4.x mounts, performance and access permissions behave differently than in NFSv3.

Network connection concurrency: NFSv3

In addition to the preceding considerations, it is worth noting that ONTAP has a limit of 128 concurrent operations per TCP connection for NFSv3 operations. This limit means that for every IP address, the system can handle only up to 128 concurrent operations. Therefore, it's possible that an NFSv3 client would not be able to push the storage system hard enough to reach the full potential of the FlexGroup technology. Clients can be configured to control the number of concurrent operations (by using RPC slot tables) that are sent through NFSv3, which can help avoid hard-to-track performance issues.

Identifying potential issues with RPC slot tables

Many modern NFSv3 clients use dynamic values for RPC slot tables, which means that the client sends as many concurrent operations on a single TCP thread as possible—up to 65,336. However, ONTAP allows only 128 concurrent operations per TCP connection, so if a client sends more than 128, ONTAP enacts a form of flow control on NFSv3 operations to prevent rogue clients from overrunning storage systems by blocking the NFS operation (exec contexts in ONTAP) until resources free up. This flow control may manifest as performance issues that cause extra latency and slower job completion times that might not have a readily apparent reason from the general storage system statistics. These issues can appear to be network related, which can send storage administrators down the wrong troubleshooting path.

To investigate whether RPC slot tables might be involved, use the ONTAP performance counter. You can check whether the number of exec contexts blocked by the connection being overrun is incrementing.

To gather those statistics, run the following command:

```
statistics start -object cid -instance cid
```

Then, review the statistics over a period of time to see if they are incrementing.

```
statistics show -object cid -instance cid -counter execs_blocked_on_cid
```

In ONTAP 9.8 and later, a new EMS message (`nblade.execsOverLimit`) has been added to help identify RPC slot table issues. This EMS triggers when the execs_blocked_on_cid counters exceed a certain amount over a set period of time. If you see this message in your events, contact NetApp support, or look into reducing the number of slot tables used on your NFSv3 clients.

For more detailed information about network connection concurrency for NFS, see TR-4067 NFS Best Practices and Implementation Guide and TR-4617 EDA Best Practice Guide.

NFS write appends

In some cases, older releases of ONTAP experience performance issues with file write appends over NFS based on how large the appends were. Later ONTAP releases provide parallel processing of these write appends to improve performance on write appends regardless of the file sizes involved.

See bug 1256520 for more information and guidance on which ONTAP releases have fixed this problem.

Nconnect

Nconnect is a mount option available in some Linux distributions. This option specifies how many TCP connections should be used per mount and offers substantial performance benefits in some workloads per client—generally only when the network threads are the bottleneck in a workload. This also provides benefits to ONTAP by allowing clients to leverage more RPC slot tables per mount. See the section "Network connection concurrency: NFSv3" for details on RPC slot tables.

ONTAP 9.8 offers support for the use of nconnect with NFS mounts, provided the NFS client also supports it. If you wish to use nconnect, check to see if your client version provides it and use ONTAP 9.8 or later.

Table 22 shows results from a single Ubuntu client using different nconnect thread values.

Table 22) nconnect performance results.

Nconnect Value	Threads per process	Throughput	Difference
1	128	1.45GB/s	-
2	128	2.4GB/s	+66%
4	128	3.9GB/s	+169%
8	256	4.07GB/s	+181%

For more detailed information about nconnect, see [TR-4067 NFS Best Practices and Implementation Guide]() and [TR-4617 EDA Best Practice Guide]().

Mapping NFS connected clients to volume names

To check what version of NFS is being mounted from the cluster, use the `nfs connected-clients show` command available in ONTAP 9.7:

```
cluster::> nfs connected-clients show -node * -vserver DEMO

     Node: node1
  Vserver: DEMO
  Data-Ip: 10.x.x.x
Client-Ip       Volume-Name       Protocol Idle-Time     Local-Reqs Remote-Reqs
--------------- ----------------- -------- ------------- ---------- -----------
10.x.x.x        CIFS              nfs4.1   2d 0h 9m 3s   153        0
10.x.x.x        vsroot            nfs4.1   2d 0h 9m 3s   0          72
10.x.x.x        flexgroup_16__0001
                                  nfs3     0s            0          212087
10.x.x.x        flexgroup_16__0002
                                  nfs3     0s            0          192339
10.x.x.x        flexgroup_16__0003
                                  nfs3     0s            0          212491
10.x.x.x        flexgroup_16__0004
                                  nfs3     0s            0          192345
10.x.x.x        flexgroup_16__0005
                                  nfs3     0s            212289     0
```

To avoid issues with mounting a FlexGroup volume in environments in which NFSv4.x is enabled, either configure clients to use a default mount version of NFSv3 through `fstab` or explicitly specify the NFS version when mounting.

For example:

```
# mount -o nfsvers=3 demo:/flexgroup /flexgroup
# mount | grep flexgroup
demo:/flexgroup on /flexgroup type nfs (rw,nfsvers=3,addr=10.193.67.237)
```

Also, if a FlexGroup volume is junctioned to a parent volume that is mounted to a client with NFSv4.x, traversing to the FlexGroup volume in ONTAP 9.6 and earlier fails, because no NFSv4.x operations are allowed to FlexGroup volumes.

For example, FlexGroup volumes are always mounted to the vsroot (vserver root), which operates as (/) in the NFS export path. If a client mounts vsroot with NFSv4.x, attempts to access a FlexGroup volume in ONTAP 9.6 or earlier from the NFSv4.x mount fail. This includes `ls -la` operations because they require the ability to do NFSv4.x `GETATTR` operations.

Note in the following example that the information for the FlexGroup volumes is incorrect because of the lack of NFSv4.x support:

```
# mount demo:/ /mnt
# mount | grep mnt
```



```
demo:/ on /mnt type nfs (rw,vers=4,addr=10.193.67.237,clientaddr=10.193.67.211)
# cd /mnt/flexgroup
-bash: cd: /mnt/flexgroup: Permission denied
# ls -la
ls: cannot access flexgroup_4: Permission denied
ls: cannot access flexgroup_local: Permission denied
ls: cannot access flexgroup_8: Permission denied
ls: cannot access flexgroup_16: Permission denied
drwx--x--x. 12 root root   4096 Mar 30 21:47 .
dr-xr-xr-x. 36 root root   4096 Apr  7 10:30 ..
d??????????  ? ?    ?         ?            ? flexgroup_16
d??????????  ? ?    ?         ?            ? flexgroup_4
d??????????  ? ?    ?         ?            ? flexgroup_8
```

Compare that to the NFSv3 mount:

```
# ls -la
drwx--x--x. 12 root root 4096 Mar 30 21:47 .
dr-xr-xr-x. 36 root root 4096 Apr  7 10:30 ..
drwxr-xr-x.  6 root root 4096 May  9 15:56 flexgroup_16
drwxr-xr-x.  5 root root 4096 Mar 30 21:42 flexgroup_4
drwxr-xr-x.  6 root root 4096 May  8 12:11 flexgroup_8
```

As a result, be sure to avoid using NFSv4.x in any path where a FlexGroup volume resides in ONTAP 9.6 or earlier. If NFSv4.x is desired, upgrade ONTAP to 9.7 or later.

Enabling and using NFSv4.x with FlexGroup volumes

FlexGroup volumes function identically to FlexVol volumes when you configure NFSv4.x in your environment. TR-4067: NFS Best Practice and Implementation Guide covers NFS in detail for use with ONTAP and applies to FlexGroup volumes as well. Rather than focusing on performance, the benefits of using NFSv4.x with workloads include:

- **Security.** NFSv4.x greatly improves security with NFS through integration of ancillary protocols (such as NLM, NSM, mountd, and portmapper) into a single port over 2049. Fewer firewall ports being open helps reduce the threat vectors available.

  Additionally, NFSv4.x includes Kerberos encryption (krb5, krb5i, and krb5p) as part of its RFC requirements, meaning that a client/server is not compliant with the RFC unless it includes Kerberos support.

  NFSv4.x also provides better masking of UID/GID information by requiring the client and server matching domain IDs in their configurations, which helps make spoofing users harder—particularly when using Kerberos encryption.

  Finally, NFSv4.x offers granular ACL support that mimics the functionality of Windows NTFS ACLs. This provides the ability to add more users and groups to an ACL than NFSv3 offered with mode bits, as well as allowing more ACL functionality beyond basic read/write/execute (rwx). NFSv4.x ACLs can even be applied to datasets that will mount only NFSv3, which can offer granular security on files and folders even if NFSv4.x isn't being used.

- **Improved locking.** NFSv3 locking was performed outside the NFS protocol, using ancillary protocols like NSM and NLM. This often resulted in stale locks when clients or servers had outages, which prevented access to files until those stale locks were cleared.

  NFSv4.x provides locking enhancements by way of a leasing mechanism that holds a lease for a specified time and keeps that lease if the client/server communication is intact. If there are any issues with that communication (whether network or server outage), the lease expires and releases the lock until it is reestablished.

  Additionally, locking in NFSv4.x is integrated within the NFS packets, providing more reliable and efficient locking concepts than NFSv3.

- **Data locality and parallel access.** NFSv4.x offers data locality functionality for scale-out NAS environments, such as NFSv4.x referrals, which can redirect mount requests to volumes in ONTAP according to the location on a node to ensure local access to the mount.

  NFSv4.1 also offers parallel NFS support, which establishes a metadata server on mount and then redirects data I/O across the namespace. To do this, it uses a client/server communication that keeps track of data according to node and data LIF location. This concept is similar to that of asymmetric logical unit access (ALUA) for SAN. For more information, see the section on pNFS with FlexGroup volumes.

NFSv4.x performance enhancements in ONTAP

In general, NFSv4.x is less performant than NFSv3 because NFSv4.x is stateful, so it has more to do for each protocol operation. NFSv4.x overhead comes in the form of locking and leasing, ACLs, compound calls, and communication of state IDs between the client and server, as well as the processing of each packet.

ONTAP 9.2 and later brought NFSv4.x performance somewhat closer to NFSv3 performance for streaming I/O workloads, such as SAP HANA and databases. ONTAP 9.6 and later added some metadata workload performance enhancements for NFSv4.x that improved results on EDA benchmarks to be more in line with NFSv3.

One of the weak points for performance with NFSv4.x includes workloads with high metadata ingest. FlexGroup volumes work best with these types of workloads, so if you're considering NFSv4.x for these workloads, NetApp strongly recommends using FlexGroup volumes.

One of the benefits of using NFSv4.x is that it does not use RPC slot tables in its operations, so it is not susceptible to RPC slot exhaustion.

If you are using Kerberos with NFS, there is also a small performance effect to operations for processing overhead of the encrypted packets. The effect varies depending on several factors, including:

- ONTAP version
- Hardware being used
- Network latency, WAN latency, and cloud region
- Performance headroom on the cluster
- Kerberos encryption being used (krb5, krb5i, or krb5p)
- ONTAP 9.2 and later versions offer AES-NI offloading of NFS Kerberos packets, and ONTAP 9.2 is the minimum recommended version if you are considering Kerberized NFS. For more information about configuring and managing Kerberos with NFS, see TR-4616: NFS Kerberos in ONTAP.

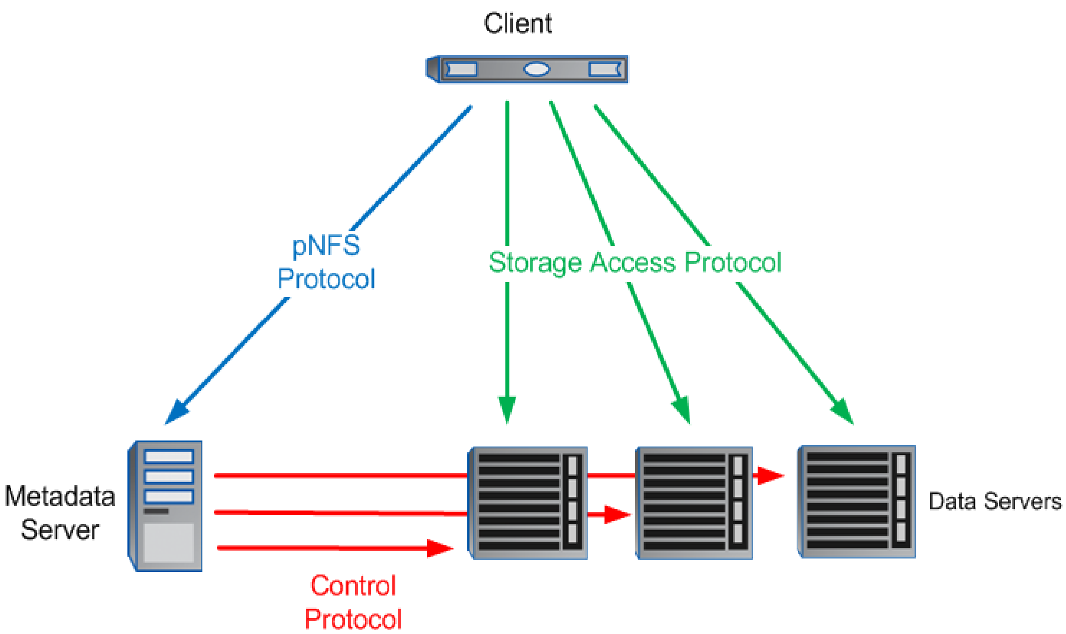
pNFS with FlexGroup volumes

ONTAP 9.7 offers support for NFSv4.x, which includes NFSv4.1 and its RFC mandatory features. Included in those features is parallel NFS (pNFS), which provides localization of reads and writes across multiple volumes and nodes in a cluster. ONTAP provides the file version of pNFS and does not use the striping or block versions of the feature.

How pNFS works in ONTAP

If pNFS has been enabled on the NFS server in an SVM, clients that support pNFS and mount by using NFSv4.1 will first connect to a specific node in the cluster with a single TCP connection that acts as a metadata server. This connection will service pNFS operations, such as client/server communications for data layout, LIF location, and pNFS mappings to help redirect I/O traffic to the local volumes and data LIFs in the cluster. The metadata server also services NFS metadata operations such as `getattr` operations and `setattr` operations.

Figure 88) pNFS diagram.



The pNFS architecture includes three main components:

The metadata server that handles all non-data I/O traffic. It is responsible for all metadata operations, such as `GETATTR`, `SETATTR`, `LOOKUP`, `ACCESS`, `REMOVE`, and `RENAME` operations. The metadata server also provides information about the layout of files.

- **Data servers that store file data and respond directly to client read and write requests.** Data servers handle pure `READ` and `WRITE` I/O.
- **One or more clients that are able to access data servers directly.** This access is based on metadata received from the metadata server.

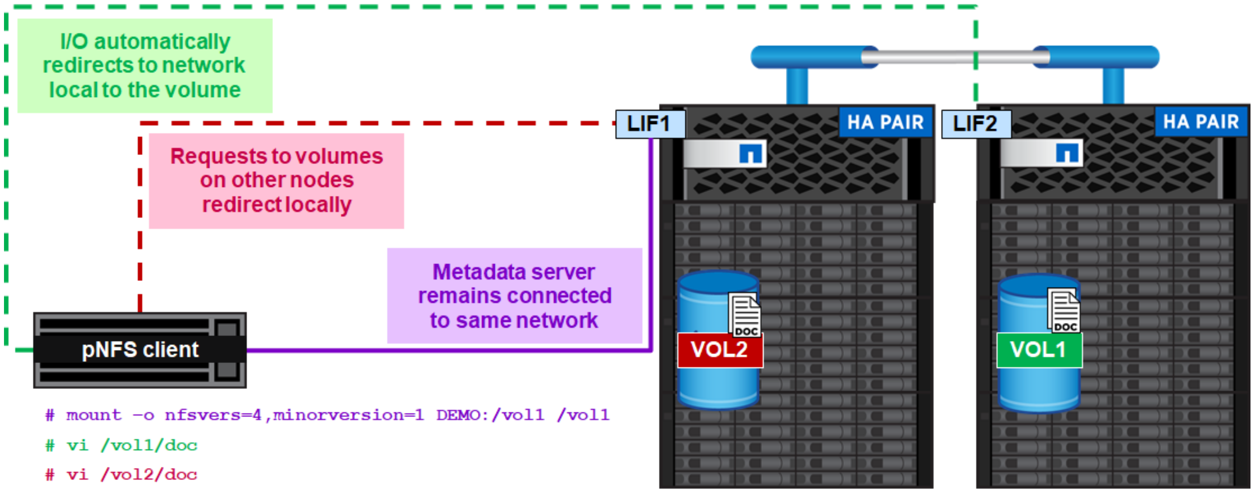
There are three types of protocols used between the clients, metadata server, and data servers:

- **A control protocol used between the metadata server and data servers.** This protocol synchronizes file system data.
- **The pNFS protocol, used between clients and the metadata server.** This is essentially the NFSv4.1 protocol with a few pNFS-specific extensions. It is used to retrieve and manipulate layouts that contain the metadata that describes the location and storage access protocol required to access files stored on numerous data servers.
- **A set of storage access protocols used by clients to access data servers directly.** The pNFS specification currently has three categories of storage protocols: file based, block based, and object based. Data ONTAP 8.1 and later support file-based storage protocol and access the data servers over NFSv4.1.

When a read or write request is performed by a client over pNFS, the client and server negotiate where to send those requests by using the data layout mappings. For example, if a file lives on volume1 (which lives on node1) in a cluster, but the metadata server is connected to node2, then the data layout mapping informs the client to perform the reads/writes over a network connection local to node1.

If a volume is moved (for example, with a nondisruptive volume move operation), the data layout table is updated and ONTAP redirects local traffic to the volume on the next request. This process is similar to how ALUA works in SAN environments, where a path can switch based on locality of the block device.

Figure 89) pNFS operations diagram.



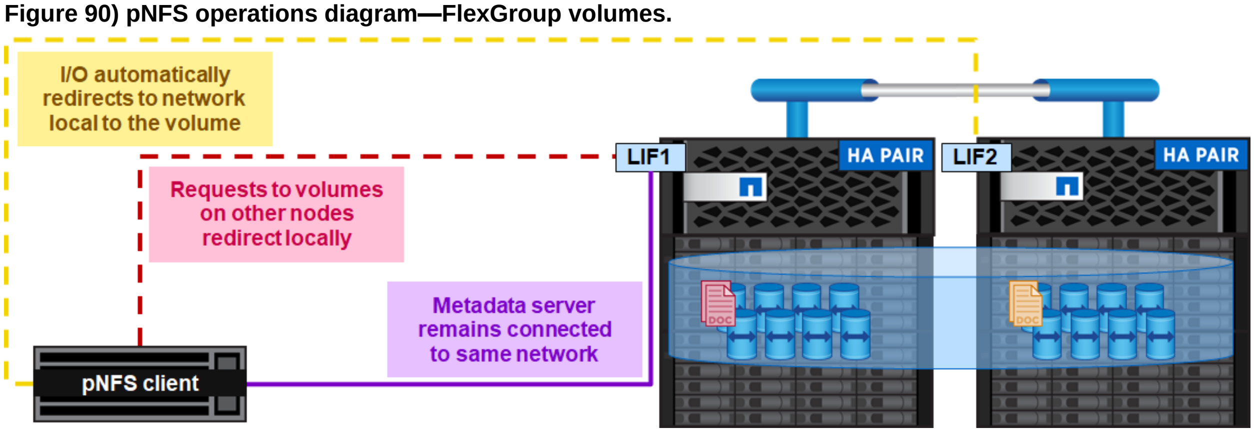
How pNFS works with FlexGroup volumes

A FlexGroup volume operates as a single entity but is constructed of multiple FlexVol member volumes. Each member volume contains unique files that are not striped across volumes. When NFS operations connect to FlexGroup volumes, ONTAP handles the redirection of operations over a cluster network.

With pNFS, these remote operations are reduced, because the data layout mappings track the member volume locations and local network interfaces; they also redirect reads/writes to the local member volume inside a FlexGroup volume, even though the client only sees a single namespace. This approach enables a scale-out NFS solution that is more seamless and easier to manage, and it also reduces cluster network traffic and balances data network traffic more evenly across nodes.

FlexGroup pNFS differs a bit from FlexVol pNFS. Even though FlexGroup load-balances between metadata servers for file opens, pNFS uses a different algorithm. pNFS tries to direct traffic to the node on which the target file is located. If multiple data LIFs per node are given, connections can be made to each of the LIFs, but only one of the LIFs of the set is used to direct traffic to volumes per network interface.

Figure 90) pNFS operations diagram—FlexGroup volumes.



pNFS best practices

pNFS best practices in ONTAP don't differ much from normal NAS best practices, but here are a few to keep in mind. In general:

- Use the latest supported client OS version.
- Use the latest supported ONTAP patch release.
- Create a data LIF per node, per SVM to ensure data locality for all nodes.
- Avoid using LIF migration on the metadata server data LIF, because NFSv4.1 is a stateful protocol and LIF migrations can cause brief outages as the NFS states are reestablished.
- In environments with multiple NFSv4.1 clients mounting, balance the metadata server connections across multiple nodes to avoid piling up metadata operations on a single node or network interface.
- If possible, avoid using multiple data LIFs on the same node in an SVM.
- In general, avoid mounting NFSv3 and NFSv4.x on the same datasets. If you can't avoid this, check with the application vendor to ensure that locking can be managed properly.
- If you're using NFS referrals with pNFS, keep in mind that referrals establish a local metadata server, but data I/O still redirect. With FlexGroup volumes, the member volumes might live on multiple nodes, so NFS referrals aren't of much use. Instead, use DNS load balancing to spread out connections.

NFSv4.x general considerations

When considering NFSv4.x for your SVM, be sure to factor in performance, client and application support, name services infrastructure, and locking mechanisms before deploying. Also consider whether applications can use both NFSv3 and NFSv4.x on the same datasets. For instance, VMware recommends against service datastores over both protocol versions.

If possible, set up a separate SVM to conduct functionality and performance testing before deploying in production.

NFSv4.x configuration generally requires the following to work properly:

- NFS clients that support NFSv4.x.
- NFS mounts that specify NFSv4.x.
- NFS server configuration (NFSv4.x and desired features enabled—such as referrals, pNFS, ACL support, NFSv4 ID domain configured to be identical on client and NFS server).
- Matching user names and groups on client and server (case sensitive; for example, user1@domain.com should exist on both server and client; USER1 and user1 are not considered matches).
- Optional: Name services for UNIX identities, such as NIS or LDAP, can greatly simplify NFSv4.x implementation and functionality.

For more detailed information, see TR-4067: NFS Best Practice and Implementation Guide.

NFS mount considerations: READDIRPLUS (READDIR+)

If you are running a version of ONTAP earlier than 9.1P4 and use the `READDIR+` functionality in NFS, you might experience some latency on rename operations. This is caused by bug 1061496, which is fixed in 9.1P4 and later. If you're running a release of ONTAP that is exposed to this bug and are experiencing latencies, consider mounting FlexGroup volumes with the option `-nordirplus` to disable `READDIR+` functionality.

NAS metadata effect in a FlexGroup volume

The overhead for metadata operations affects how a workload performs, which can be anywhere from a 10% to 30% performance hit for remote operations. Most of the metadata effect is related to write metadata. Most read metadata has little to no effect.

- `getattr, access, statfs, lock, unlock.` Little to no FlexGroup overhead.
- `Readdirplus.` Before ONTAP 9.3, mostly remote; after 9.3, little remote overhead.

- `nfs create, unlink, lookup.` Little to no FlexGroup overhead under heavy load.
- `nfs mkdir, rmdir, lookup dir.` 50% to 100% remote access, so high overhead.
- `CIFS open/close.` High overhead.

CIFS/SMB considerations

FlexGroup volumes support both NFS and SMB workloads. ONTAP SMB servers offer some features that can help improve the overall performance experience for SMB workloads on both FlexGroup and FlexVol volumes. The following section covers some of those features, as well as some caveats that apply to FlexGroup volumes or high-file-count environments.

SMB version considerations

FlexGroup volumes support SMB 2.x and SMB 3.x versions only. SMB 1 versions are not able to access CIFS/SMB shares pointing to FlexGroup volumes. As SMB 1 is deprecated by Microsoft, there are no future plans to add SMB 1 support to FlexGroup volumes. For full SMB support information, see the section “FlexGroup feature support and maximums.”

Before you migrate a CIFS/SMB workload to a FlexGroup, you should verify that no SMB 1 clients are connected to the existing workloads. In ONTAP, you can do that with the following command:

```
cluster::> cifs session show -protocol-version SMB1
```

If SMB 1 access is attempted to a FlexGroup volume, the `Nblade.flexgroupStatefulProtocolAccess` EMS event is logged.

Use of change notifications with SMB

SMB change notifications are how SMB clients are informed of a file’s existence in a SMB share without needing to close a session or refresh a window (such as pressing the F5 key). SMB clients are in constant communication with the SMB server during SMB sessions, and the SMB server sends periodic updates to the client regarding any file changes in the share. This feature is most useful for applications that must write files and then be able to immediately read the files in SMB shares. Change notifications are controlled through the `changenotify` share property. ONTAP 9.4 and later versions automatically set this share property on new SMB shares, even if change notifications are not needed.

Support for SMB change notifications was added for FlexGroup volumes in ONTAP 9.2. ONTAP 9.5 introduced inherited change notifications, which perform change notifications per folder rather than for the entire volume each time.

In ONTAP versions prior to ONTAP 9.7, the change notification processes performed in serial, which means that only one change notify operation at a time can be completed before the next operation can proceed. As a result, performance for the entire cluster can be affected and high latency for SMB operations can occur.

This issue is detailed in the KB article High FlexGroup CIFS latency due to CIFS ChangeNotify.

Best practice 24: SMB change notification recommendation

- In ONTAP 9.6 and earlier, disable SMB change notifications for CIFS shares if they are not needed. You can do this by using System Manager or the `cifs share properties remove` command in the CLI.
- If you require the use of SMB change notifications, use ONTAP 9.7 or later.

Large MTU

Large MTU was introduced in ONTAP 9.0 and allows SMB's maximum transmission unit (MTU) to be increased from 64KB to 1MB, significantly improving the speed and efficiency of large file transfers by reducing the number of packets that need to be processed. You can enable large MTU with the advanced privilege command `cifs options modify -is-large-mtu-enabled true`.

When this is enabled on the CIFS/SMB server in ONTAP, if the client and SMB protocol version support it (SMB 2.1 and later), then the negotiation for MTU size happens automatically.

You can check to see if your ONTAP SVM is using large MTU with the following command:

```
cluster::> cifs session show -is-large-mtu-enabled true
```

Note: Large MTU refers to large read/writes allowed by SMB 2.1 and later servers. It does not refer to MTU sizes by the network layer.

SMB multichannel

ONTAP 9.4 and later offered support for SMB multichannel, which is an SMB 3.0 protocol feature that enables an SMB 3.x client to establish a pool of connections over a single network interface card (NIC) or multiple NICs and use them to send requests for a single SMB session. This is similar to the nconnect functionality for NFS.

By doing this, single-client performance can be drastically improved over clients that don't make use of this functionality. Further details regarding SMB multichannel can be found in TR-4740: SMB 3.0 Multichannel, including how to set it up, failover behavior, and expected performance.

SMB multichannel can be enabled with the following advanced privilege command and takes effect on new SMB sessions:

```
cluster::*> cifs options modify -is-multichannel-enabled true
```

You can check to see if your ONTAP SVM's CIFS/SMB sessions are using SMB multichannel with the following command:

```
cluster::> cifs session show -connection-count >1
```

On a Windows client, you can see if multichannel is in use with the Get-SmbMultichannelConnection PowerShell cmdlet.

Continuously available shares (CA shares)

CA shares provide a way for SMB connections to survive storage failovers without disruption by using SMB 3.x functionality such as scale-out, persistent handles, witness, and transparent failover. CA shares were first supported in ONTAP 8.2 and are officially supported only for SQL and Hyper-V workloads.

CA shares are set at the CIFS/SMB share level with the following command:

```
cluster::*> cifs share properties add -share-name SQL -share-properties continuously-available
```

FlexGroup volumes first offered support for CA shares in ONTAP 9.6 and are only officially qualified for Hyper-V and SQL workloads. However, there are caveats to that support.

- SQL Server workloads with only a few large database files might not be a good fit for FlexGroup volumes. However, SQL Server workloads that have many files (logs or databases) are an appropriate use case for FlexGroup volumes and CA shares. Also, see the section "Databases on FlexGroup volumes."
- Hyper-V workloads are listed as officially supported for CA shares but as of ONTAP 9.8, only VMware virtualization workloads are officially supported with FlexGroup volumes. Hyper-V workloads can be

used on FlexGroup volumes with CA shares, but there has not been the same testing and qualification done as with VMware workloads.
- Virtual hard disk workloads (such as FSLogix VHDx profiles) can be used on FlexGroup volumes and work with CA shares but have not been officially tested or qualified. In some cases, CA shares aren't necessary to host these workloads, so testing should be performed before deploying in production.

In general, CA shares should not be used with metadata-intensive SMB workloads (such as home directories), as this can cause performance issues. If you are using CA shares, other share properties such as homedirectory, branchcache, access-based enumeration, and attribute caching should not be set.

You can see which CIFS/SMB sessions are using CA shares with the following command:

```
cluster::*> cifs session show -continuously-available Yes|Partial
```

Other considerations

There are a few other potential issues you might see in certain scenarios while performing specific tasks like renaming using SMB 8.3 short names or using CIFS symlinks. The following is a list of these issues. The links show which ONTAP release these issues are resolved. The general recommendation is to run the latest patch release of ONTAP for best results.

Table 23) Other Potential SMB Issues.

Other potential SMB issues
• 1143151 - An SMB client listing or reading FlexGroup directories might cause a controller disruption • 1187162 - Memory leak when operating on broken symbolic links in a FlexGroup from SMB • 1219744 - SMB rename operation on a FlexGroup volume might cause storage controller disruption • 1236696 - Incorrect status code when accessing a non-existent file using a FlexGroup path with a symlink • 1278427 - Renaming a file through the SMB protocol using a short filename (8.3 filename) on a FlexGroup volume might cause controller disruption

Virtualization workload considerations

ONTAP 9.8 is the first release that offers official support for VMware virtualization workloads. That means you can provision a VMware NFS datastore using FlexGroup volumes to scale across multiple nodes in a cluster and provide more than 100TB for virtual machines.

Scalable VMware datastores using FlexGroup volumes offer some advantages over FlexVol volumes.

- Up to 20PB and 400 billion files in a single NFS datastore (VMware limits may reduce that amount)
- Rapid VM cloning using NetApp ONTAP sis clone and template caches

However, VMware datastores on FlexGroup volumes do not currently support VMware vVols, nor the VMFS file system (FlexGroup volumes are NAS only). For more information on VMware workloads on FlexGroup volumes, see the following resources:

- TR-4597: VMware vSphere with ONTAP
- Best Practices for VMware vSphere and NetApp ONTAP (Insight video presentation)

ONTAP tools for VMware vSphere support (formerly Virtual Storage Console)

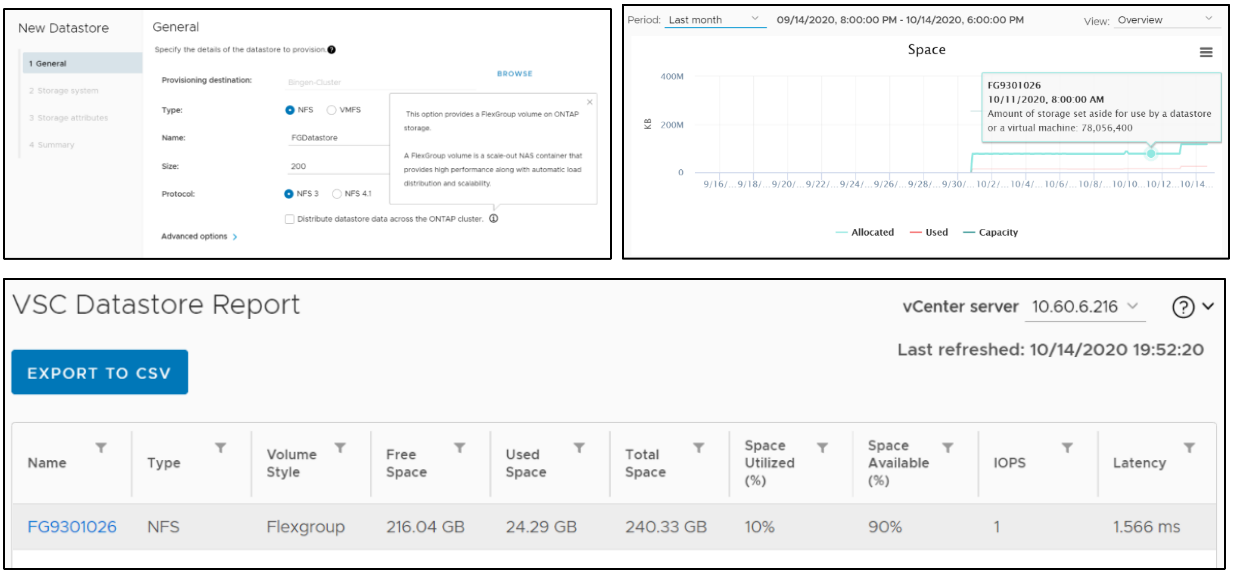
The new release of ONTAP tools for VMware vSphere provide ways to provision and manage datastores as well. Some of that functionality includes:

- Datastore provisioning as a FlexGroup or FlexVol
- QoS policy management down to the VM level

- Performance metrics at the VM level
- SnapCenter for vSphere support

The following images highlight some of the functionality available.

Figure 91) ONTAP tools for VMware vSphere Support – FlexGroup Datastores.



Copy offload

Although ONTAP 9.8 supports optimized copy offload (VAAI) for faster cloning, there are some limitations and considerations to keep in mind.

For more information on copy-offload in ONTAP, see [When does the NetApp NFS Plugin for VMware VAAI improve clone performance?]

- To enable the full copy offload functionality in a FlexGroup volume using the NetApp Spin Copy Engine, run the following command in diagnostic privilege:

```
::> set diag
::*> vserver copy-offload modify -vserver SVM -flexgroup-optimized-copy enabled -nfs enabled
```

- The first few clones in a FlexGroup volume are slower than a FlexVol initially, because the template cache is not populated right away. Subsequent clones have performance on par with FlexVol volume datastores. A larger number of clones perform better than a smaller number. FlexGroup volumes with more member volumes have slower clone performance in general because each member volume must populate the clone template with a file copy for copy offload to work properly. The more member volumes, the more caches there are to populate (and more file copies).
- Copy-offload operations take a variable amount of time (proportional to file size).
- There is a limit of 50 parallel on-demand jobs per node. If jobs exceed that limit, they are placed into a queue.
- There is additional space usage overhead because the template file has to be copied in each of the members. As a result, offload operations in FlexGroup volumes that don't have enough free space may fail.
- Replication is disallowed until the data pull scan finishes.
- When a copy offload operation is performed from the ONTAP VMware Tools, a special NFS RPC is triggered to inform ONTAP that the file is a VMDK and the FlexGroup volume ensure that the clones evenly distribute across all member volumes.

VAAI clone performance – FlexVol versus FlexGroup

We performed several tests to measure cloning time for FlexGroup datastores against a FlexVol counterpart. The following VM types were used as source templates with 100% of data populated. The FlexGroup volume used the default settings for ONTAP VMware Tools (16 member volumes across two nodes).

Windows Template: 40GB main disk, additional disks: 25GB, 50GB, and 25GB

Unix Template: 20GB main disk, additional disks: 25GB, 50aGB, and 25GB

Table 24) FlexGroup versus FlexVol VAAI results

VM type	FlexGroup (NFSv3)	FlexGroup (NFSv4.1)	FlexVol (NFSv3)	FlexVol (NFSv4.1)
Windows	39m 9s	38m 49s	3m 25s	5m 27s
Linux	17m 7s	15m 30s	3m 27s	4m 52s

As mentioned, the initial cloning time for a FlexGroup NFS datastore is longer than for a FlexVol because the cache is populated (those initial clones are actual file copies). These caches live on each member volume in the FlexGroup and are used if the template properties match the incoming clone request. If the member volume does not contain an entry, then a fresh local source copy is used. In FlexGroup volumes with many member volumes, it can take longer to populate each member volume's cache, so it might take longer with more clone operations to fully populate all member volume copy offload caches. You might want to create FlexGroup volumes with fewer member volumes than the default values set by the ONTAP VMware Tools to reduce clone times.

For example, ONTAP VMware Tools creates 16 member volumes in a two-node cluster when provisioning a datastore. Instead, use the CLI to create four to eight total member volumes for the FlexGroup to help reduce clone times. You can then use ONTAP VMware Tools to manage and mount that FlexGroup as a datastore. See the section "Manual FlexGroup creation—CLI" for details on creating a FlexGroup from the CLI.

To measure performance for VM clones in a FlexGroup volume, you can use the following command:

```
cluster::> set diag; statistics start -object vol_remote_clone
```

(perform clone operations)

```
cluster::> statistics show -object vol_remote_clone
```

Other virtualization workload considerations

Although ONTAP has tested and qualified virtualization workloads and ONTAP 9.8 offers features such as proactive resizing to better accommodate these types of datasets, there are a few considerations you should keep in mind.

- FlexGroup datastores can be provisioned using ONTAP tools for VMware vSphere (preferred) or manually created using the CLI/ONTAP System Manager and then mounted using VMware vSphere.
- Virtual machines and Snapshot copies start out as small files and grow over time. ONTAP resizes individual member volumes as they reach capacity thresholds to automatically maintain an even balance of available free space. This results in some member volumes being larger than others and is normal.
- Qualification testing was done up to 1500 VMs in a FlexGroup datastore. This is not a hard limit but going beyond the tested limit might create unpredictable results.
- When sizing a FlexGroup datastore, keep in mind that the FlexGroup consists of multiple smaller FlexVol volumes that create a larger namespace. As such, size the datastore to be at least 8x the size of your largest virtual machine. For example, if you have a 6TB VM in your environment, size the FlexGroup datastore no smaller than 48TB.

- FlexGroup volumes support VAAI starting in ONTAP 9.7, which is used to offload copy operations from vSphere to storage. Note that copy offload is not always faster than host copy, and vSphere only offloads operations on cold VMs for NFS storage.
- With virtualization workloads, VMDK files in the same FlexGroup datastore could live in multiple FlexVol member volumes across the cluster. As a result, use SnapCenter for vSphere to coordinate snapshots and replication.
- It is possible to use FlexGroup volumes with virtualization backup products such as Veeam or Rubrik. Check with these vendors for their level of support and interaction with NetApp SnapDiff 2.0 or later.
- FlexGroup volumes have only been tested and qualified for VMware datastores. Use of Hyper-V, Citrix Xen, RedHat KVM, and so on has not been tested or qualified and is not officially supported.
- Placing virtual hard disks (VHD) files on FlexGroup volumes is supported regardless of the virtualization provider.
- Sis clone support was added in ONTAP 9.8 but is only available for use via the VMware vSphere APIs.
- VMware support limitations apply (for example, no pNFS).
- VMware and NetApp do not currently support a common multipath networking approach. For NFSv4.1, NetApp supports pNFS, whereas VMware supports session trunking. NFSv3 does not support multiple physical paths to a volume. For FlexGroup with ONTAP 9.8, our recommended best practice is to let ONTAP tools for VMware vSphere make the single mount, because the effect of indirect access is typically minimal (microseconds). It is possible to use round-robin DNS to distribute ESXi hosts across LIFs on different nodes in the FlexGroup, but this would require the FlexGroup to be created and mounted without ONTAP tools for VMware vSphere, and then the performance management features would not be available.
- Use ONTAP tools for VMware vSphere 9.8 to monitor performance of FlexGroup VMs using ONTAP metrics (dashboard and VM reports), and to manage QoS on individual VMs. These metrics are not currently available through ONTAP commands or APIs.
- SnapCenter for vSphere release 4.4 supports backup and recovery of VMs in a FlexGroup datastore on the primary storage system. Although SnapMirror can be used manually to replicate a FlexGroup to a secondary system, SCV 4.4 does not manage the secondary copies.

Databases on FlexGroup volumes

Usually, databases (such as Oracle) create a few small files when they are deployed. In a FlexGroup volume, small numbers of small files tend to favor local placement to their parent folder. This means that an Oracle deployment of eight databases might all land inside the same FlexGroup member volume. Not only does this provide no benefits from load distribution across nodes in a cluster, it can also present a problem as the files grow over time. Eventually, the files start to fill the member volume to capacity, and there is a need for remediation steps to move around data.

Database workloads, in theory, work well in a single namespace that can span a cluster. However, because the files are likely to grow over time and latency-sensitive databases might run on volumes that traverse the cluster network, NetApp currently recommends placing database files in FlexVol volumes.

Note: ONTAP 9.8 provides proactive resizing, which makes hosting large files/files that grow less of a concern, so database workloads on FlexGroup volumes become more realistic.

FlexCache volume considerations

ONTAP 9.5 added support for NetApp FlexCache. This feature provides a sparse volume that can accelerate performance for NAS workloads and prevent volume hot spots in a cluster or across a WAN. The FlexCache cache volume is powered by FlexGroup volumes, and the underlying protocol that redirects the pointers and blocks is the remote access layer (RAL). The RAL is also what makes a FlexGroup volume a FlexGroup volume. ONTAP 9.6 increased the limit of maximum cache volumes per node to 100, so FlexCache has more scalability in current releases. ONTAP 9.7 added support for FlexGroup origin volumes for FlexCache.

ONTAP 9.8 adds additional functionality for FlexCache volumes, including:

- SMB and multiprotocol NAS support
- 1:100 origin to cache ratio
- SnapMirror secondary origins
- Block-level invalidation
- Pre-population of a FlexCache

For more information about FlexCache, see [TR-4743: FlexCache in NetApp ONTAP](https://www.netapp.com/pdf.html?item=/media/7336-tr4743pdf.pdf).

FlexClone

Starting in ONTAP 9.7, NetApp FlexClone is supported for use with FlexGroup volumes. This feature provides storage administrators with a way to create instant, space-efficient copies (backed by Snapshot technology) of volumes to use for testing, development, backup verification, and a variety of other use cases. There are no specific considerations for use with FlexGroup volumes, except that a FlexClone copy of a FlexGroup volume uses the same number of member volumes as the FlexGroup parent volume. As a result, the volume count on a node can start to add up as FlexClone copies are created.

For example, if you have a FlexGroup volume that contains 16 member volumes and then create a FlexClone copy of that FlexGroup volume, you now have used 32 volumes in the system. Each new clone of the volume uses 16 member FlexVol volumes as well.

```
cluster::*> volume clone create -vserver DEMO -flexclone FGclone -type RW -parent-vserver DEMO -
parent-volume flexgroup_16

cluster::*> vol show -vserver DEMO -volume flexgroup_16*,FGclone* -fields name -sort-by name
vserver volume             name-ordinal
------- ------------------ ------------
DEMO    FGclone            -
DEMO    flexgroup_16       -
DEMO    FGclone__0001      base
DEMO    FGclone__0002      base
DEMO    FGclone__0003      base
DEMO    FGclone__0004      base
DEMO    FGclone__0005      base
DEMO    FGclone__0006      base
DEMO    FGclone__0007      base
DEMO    FGclone__0008      base
DEMO    FGclone__0009      base
DEMO    FGclone__0010      base
DEMO    FGclone__0011      base
DEMO    FGclone__0012      base
DEMO    FGclone__0013      base
DEMO    FGclone__0014      base
DEMO    FGclone__0015      base
DEMO    FGclone__0016      base
DEMO    flexgroup_16__0001 base
DEMO    flexgroup_16__0002 base
DEMO    flexgroup_16__0003 base
DEMO    flexgroup_16__0004 base
DEMO    flexgroup_16__0005 base
DEMO    flexgroup_16__0006 base
DEMO    flexgroup_16__0007 base
DEMO    flexgroup_16__0008 base
DEMO    flexgroup_16__0009 base
DEMO    flexgroup_16__0010 base
DEMO    flexgroup_16__0011 base
DEMO    flexgroup_16__0012 base
DEMO    flexgroup_16__0013 base
DEMO    flexgroup_16__0014 base
DEMO    flexgroup_16__0015 base
```



```
DEMO     flexgroup_16__0016 base
```

FlexClone to different storage virtual machines (SVMs)

ONTAP enables you to create a FlexClone volume that spans different SVMs than the parent volume. You can do this by using the `-vserver` and `-parent-vserver` command options. This enables you to use the same export path for clients if you need to maintain mount paths.

Example:

```
cluster::> vol clone create -vserver NFS  -flexclone clone -type RW -parent-vserver DEMO -parent-
volume flexgroup -junction-path /flexgroup

cluster::> vol show -junction-path /flexgroup -fields junction-path,volume,size
vserver volume       size junction-path
------- ------------ ---- -------------
DEMO    flexgroup    1PB  /flexgroup
NFS     clone        1PB  /flexgroup
```

Volume rehost

ONTAP provides a method to quickly change the owning SVM for a volume via the volume rehost command. This is currently unsupported for use with FlexGroup volumes.

FlexClone deletion

When a FlexClone that is also a FlexGroup is deleted, that deletes multiple volumes in parallel. In most cases, this is not an issue, but bug 1368356 can potentially create an issue that causes new volume creations to fail until the condition is cleared. No data outages or access issues are caused by this, but for best results when using FlexClone volumes with FlexGroup volumes, check which ONTAP release the bug is fixed in and use that release. You can check for that in the bug 1368356 report.

Encryption at-rest

ONTAP 9.2 introduced support for NetApp Volume Encryption (NVE) for FlexGroup volumes. Implementing this feature with FlexGroup volumes follows the same recommendations and best practices as stated for FlexVol volumes. Re-keying an existing FlexGroup volume is possible in ONTAP 9.5 and later. See the section “Re-keying a FlexGroup volume or encrypting existing FlexGroup volumes” for details.

Generally speaking, NVE requires the following:

- A valid NVE license
- A key management server (on-box or off-box as of ONTAP 9.3)
- A cluster-wide passphrase (32 to 256 characters)
- AFF or FAS hardware that supports AES-NI offloading

For information about implementing and managing NVE with FlexGroup and FlexVol volumes, see the NetApp Encryption Power Guide and the Scalability and Performance Using FlexGroup Volumes Power Guide on the support site for your release of ONTAP.

ONTAP 9.6 added NetApp Aggregate Encryption (NAE), which allows you to encrypt at the aggregate level. FlexGroup volumes can use NAE, provided all aggregates that contain member volumes belonging to the same FlexGroup volume are encrypted.

Re-keying a FlexGroup volume or encrypting existing FlexGroup volumes

ONTAP 9.5 added support for both rekeying FlexGroup volumes and encrypting FlexGroup volumes that have not yet been encrypted. The process is the same as for a FlexVol volume. See the Managing NetApp encryption documentation for more information.

Drive-level encryption (NSE and SED)

FlexGroup volumes can use NSE and SED disks, provided the FlexGroup volume spans only encrypted drives.

MetroCluster considerations

As described in the section "NetApp MetroCluster," if you plan on using NVE/NAE on a MetroCluster, you must complete the MetroCluster configuration before setting NVE/NAE. See bugs 1360199 and 1360229 for details.

FlexGroup sample designs

A NetApp ONTAP FlexGroup offers multiple benefits and can be managed like a normal NetApp FlexVol volume. The following design variations are examples of what is allowed with a FlexGroup volume.

FlexGroup volumes can:

- Share SVM as a FlexVol volume
- Share the same physical disks and aggregates as a FlexVol volume
- Be mounted to other FlexGroup or FlexVol volumes
- Be mounted below the FlexGroup level, similar to FlexVol volumes
- Share export policies and rules with FlexVol volumes
- Enforce quotas (starting in ONTAP 9.5)

FlexGroup volumes ideally should not:

- Be configured to span mixed disk or aggregate types (for example, member volumes of the same FlexGroup volume on SATA and SSD)
- Span nodes of different hardware types
- Span aggregates with uneven free capacity

Volume affinity and CPU saturation

To support concurrent processing, ONTAP assesses its available hardware at startup and divides its aggregates and volumes into separate classes called affinities. In general terms, volumes that belong to one affinity can be serviced in parallel with volumes that are in other affinities. In contrast, two volumes that are in the same affinity often must take turns waiting for scheduling time (serial processing) on the node's CPU.

A node's affinities are viewed with the advanced privilege nodeshell command `waffinity_stats -g`.

In ONTAP 9.3 and earlier, a node had up to eight affinities available (four per aggregate).

```
cluster::> set -privilege advanced

cluster::*> node run * waffinity_stats -g

     Waffinity configured with:
                                              # AGGR affinities :    2
        # AGGR_VBN_RANGE affinities / AGGR_VBN affinity :    4
```



```
                        # VOL affinities / AGGR affinity :   4
          # VOL_VBN_RANGE affinities / VOL_VBN affinity :   4
             # STRIPE affinities / STRIPEGROUP affinity :   9
                # STRIPEGROUP affinities / VOL affinity :   1
                      # total AGGR_VBN_RANGE affinities :   8
                                  # total VOL affinities :   8
                       # total VOL_VBN_RANGE affinities :  32
                              # total STRIPE affinities :  72
                                     # total affinities : 149
                                              # threads :  19
```

The sample NetApp FAS8080 EX node above is reporting that it can support fully concurrent operations on eight separate volumes simultaneously. It also says that to reach that maximum potential, it works best with at least two separate aggregates hosting four constituents each. Therefore, when you are building a new FlexGroup volume that is served by this node, that new FlexGroup volume should include eight constituents on this node evenly distributed across two local aggregates. Provisioning tools such as ONTAP System Manager attempts to take these affinities into account when creating new FlexGroup volumes, provided the FlexGroup size is adequate to span the available affinities and stay above the minimum 100GB member volume size.

In ONTAP 9.4 and later, the number of available affinities increased to eight per aggregate (two aggregates, 16 per node) for high-end platforms like the AFF A700 and AFF A800:

```
cluster::*> node run * waffinity_stats -g

Waffinity configured with:
                                  # AGGR affinities :   2
   # AGGR_VBN_RANGE affinities / AGGR_VBN affinity :   8
                     # VOL affinities / AGGR affinity :   8
     # VOL_VBN_RANGE affinities / VOL_VBN affinity :   4
        # STRIPE affinities / STRIPEGROUP affinity :   3
           # STRIPEGROUP affinities / VOL affinity :   3
                 # total AGGR_VBN_RANGE affinities :  16
                             # total VOL affinities :  16
                  # total VOL_VBN_RANGE affinities :  64
                         # total STRIPE affinities : 144
                                # total affinities : 325
                                         # threads :  18
                                          # pinned :   0
                                # leaf sched pools :  18
                                     # sched pools :  21
```

However, storage administrators usually do not need to worry about volume affinities, because ONTAP deploys a FlexGroup volume according to best practices for most use cases. For guidance on when you might need to manually create a FlexGroup volume, see the section above.

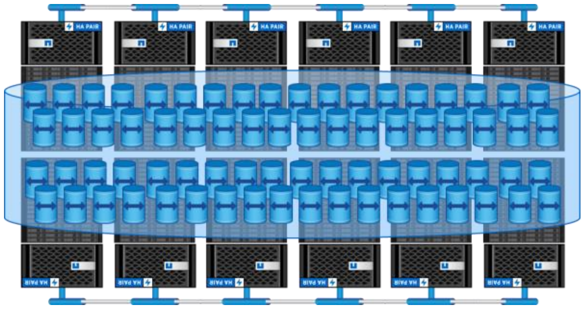
To simplify the experience, the `vol create -auto-provision-as flexgroup` command (introduced in ONTAP 9.2), the `flexgroup deploy` command, and the ONTAP System Manager GUI handle this setup for the storage administrator.

For more information about volume affinities, see: Volume Affinities: How ONTAP and CPU Utilization Has Evolved.

FlexGroup sample design 1: FlexGroup volume, entire cluster (24 nodes)

A FlexGroup volume can span an entire 24-node cluster, thus gaining the benefits of using all of the available hardware in the cluster with a single distributed namespace. In addition to using all your available hardware, you get the added benefit of gaining more potential capacity and more volume affinities in workloads.

Figure 92) FlexGroup volume, entire cluster (24 nodes).



Considerations

If you use an entire cluster to host a FlexGroup volume, note the information in the section “Cluster considerations.”

Use cases

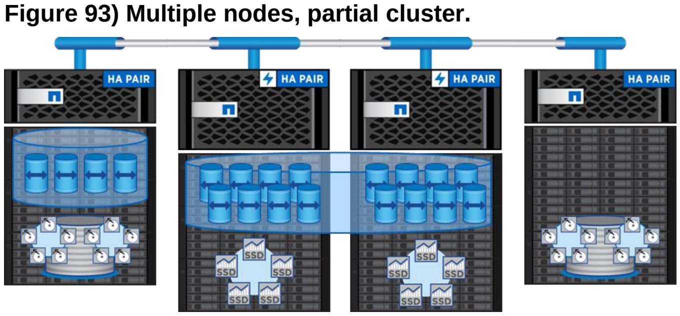
- Immense capacity (archives, scratch space, and media repositories)
- Workloads that require immense compute power in addition to storage (EDA)

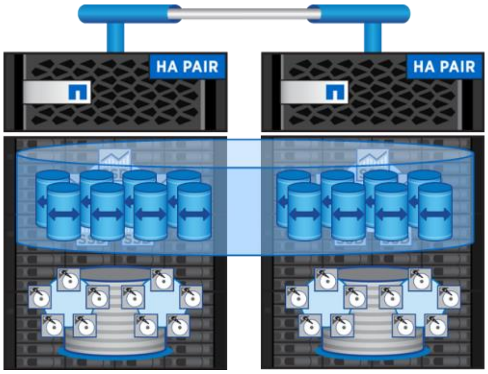
FlexGroup sample design 2: multiple nodes, aggregates, partial cluster

Sometimes, storage administrators might not want to span a FlexGroup volume across the nodes of an entire cluster. The reasons include, but are not limited to, the following:

- Mix of hardware or FAS (that is, some nodes are AFF)
- Mix of aggregate or disk types (that is, hybrid aggregates on the same node)
- Desire to dedicate nodes to specific tasks, storage tiers, or tenants
- In these scenarios, the FlexGroup volume can be created to use only specific aggregates, whether on the same node or on multiple nodes. If a FlexGroup volume has already been created, the member FlexVol volumes can be moved nondisruptively to the desired nodes and aggregates. For details, see the section “When to use nondisruptive volume moves.”

Figure 93) Multiple nodes, partial cluster.





Considerations

When you try to create a FlexGroup volume on a mix of nodes and aggregates, the automated commands are not of much use. Instead, use `volume create` or the GUI, where it is possible to specify aggregates on FlexGroup creation. For already-created FlexGroup volumes, the command line is the only option.

Use cases

- Mixed workloads (high performance and archive)
- Mixed cluster hardware
- Nodes with hybrid aggregates

FlexGroup sample design 3: FlexGroup, single node

An ONTAP cluster uses a robust back-end cluster network to pass reads and writes from a node that receives an I/O request on a data LIF to the node that owns the physical data. When traffic is remote, a small latency penalty is incurred (about 5% to 10%) for remote I/O as these packets are processed. When traffic is all local to the node that owns the data, no cluster back end is used. Also, NAS operations get special bypass consideration to direct requests to disk even faster, so there is a benefit to going locally to a node.

With FlexGroup, there is no manual intervention of control over where a data requests lands; ONTAP controls that portion for simplicity's sake. Because of this aspect, if a FlexGroup volume spans multiple nodes in a cluster, there is going to be indirect traffic over the cluster interconnects.

Although FlexGroup concurrency often more than outweighs any performance penalty for remote traffic, you can achieve some performance gains by isolating a FlexGroup volume to a single node. In addition, some deployments are performed on a single node to reduce the overall failure domains of the cluster.

Figure 94 shows a single FlexVol volume that is accessed 100% locally on an AFF A700 node versus a single FlexGroup volume with eight FlexVol members that is also accessed 100% locally. The test used was a Git clone during a compilation of the GCC library. The same testing equipment and data described in AFF A700 Testing in section 7.3 were used.

This test shows that a cluster-wide FlexGroup volume provides marginally better completion times because more hardware can be used. As extra threads are added to a local FlexGroup volume, the completion times start to get longer because the hardware cannot keep up as well. However, both FlexGroup volumes are two to three times faster than a local FlexVol volume and have a more gradual performance curve.

Figure 94) Git clone completion times comparison.

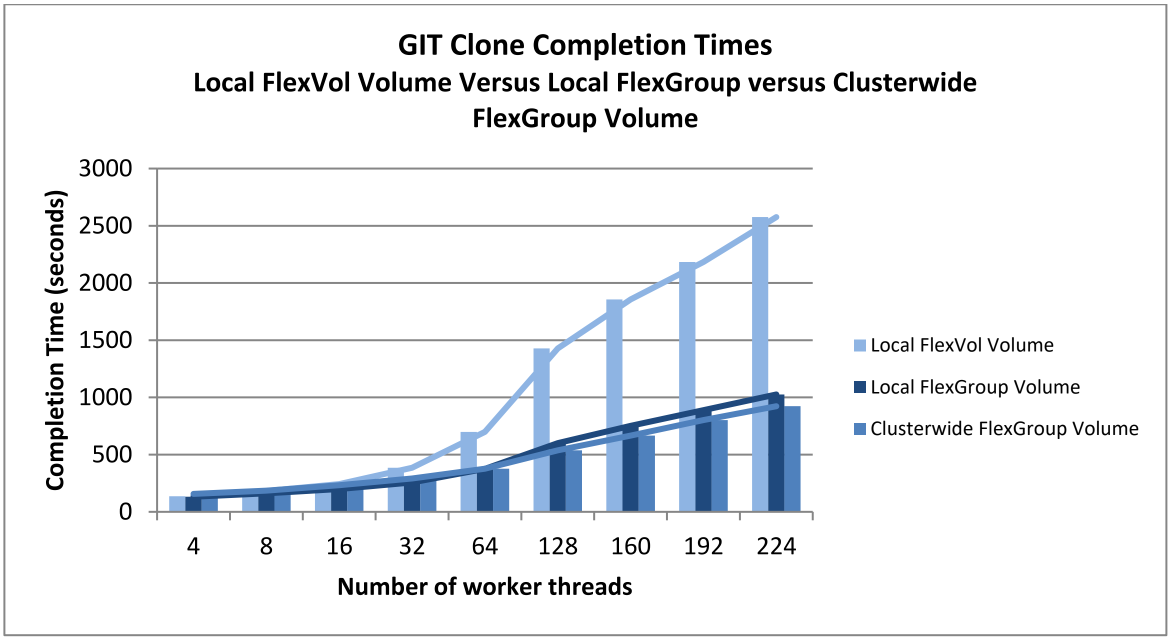
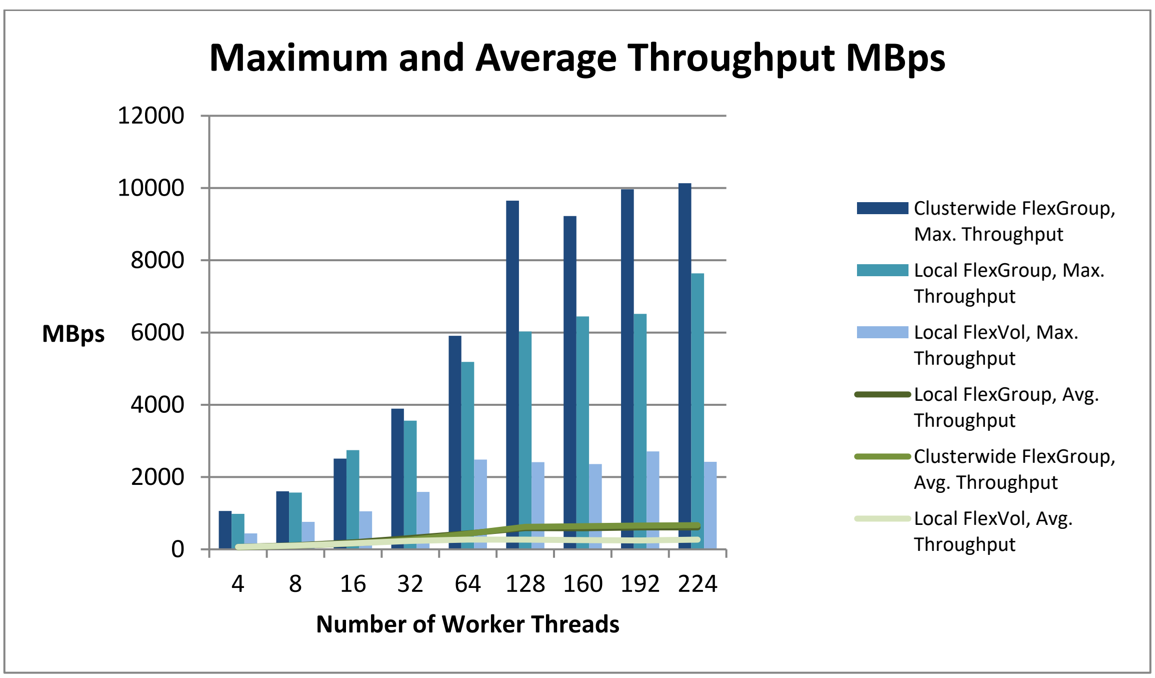


Figure 95 shows average and maximum throughput for the local FlexVol volume versus the local FlexGroup volume. For good measure, the cluster-wide FlexGroup volume was also added for comparison. The local FlexGroup volume shows better overall throughput than the cluster-wide FlexGroup volume until it reaches 16 threads. Then the all-local FlexGroup volume starts to lag behind slightly because the additional hardware allows the workload to push past the limits of a single node.

Figure 95) Average and maximum throughput comparison.



In Figure 96 and Figure 97, we compare read and write throughput, respectively, with the local and cluster-wide FlexGroup volumes. At the 64-thread tipping point, the local FlexGroup volume starts to show a shift. Read throughput increases, while write throughput decreases. The cluster-wide FlexGroup volume shows the opposite trend.

Figure 96) Maximum read throughput comparison.

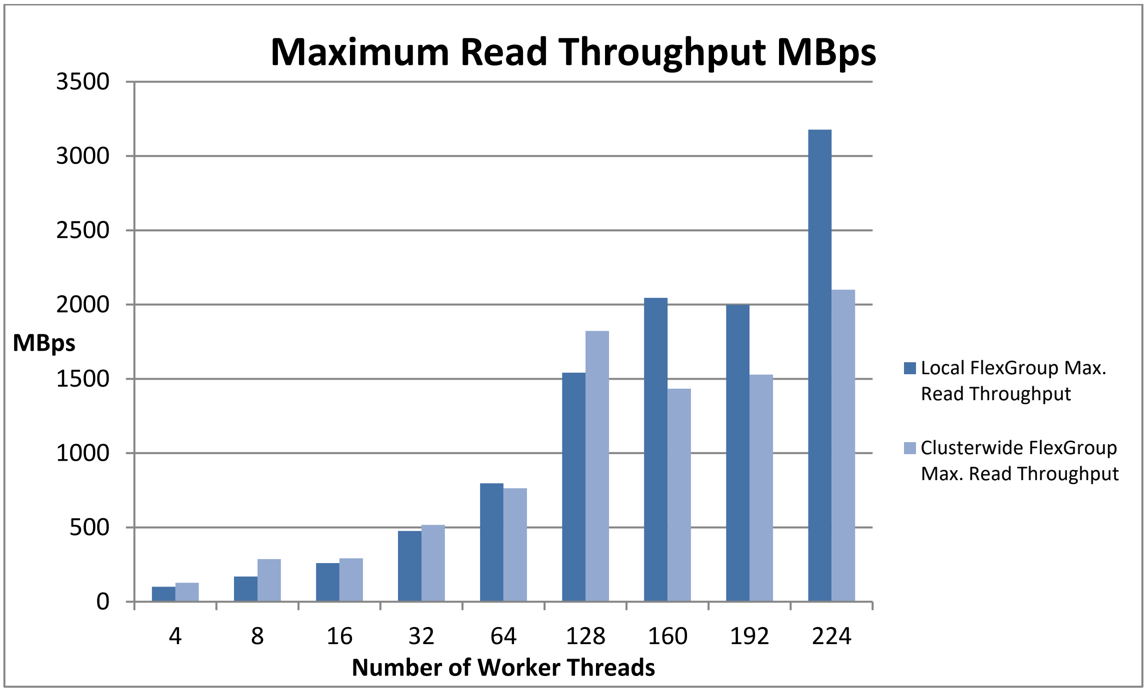


Figure 97) Maximum write throughput comparison.

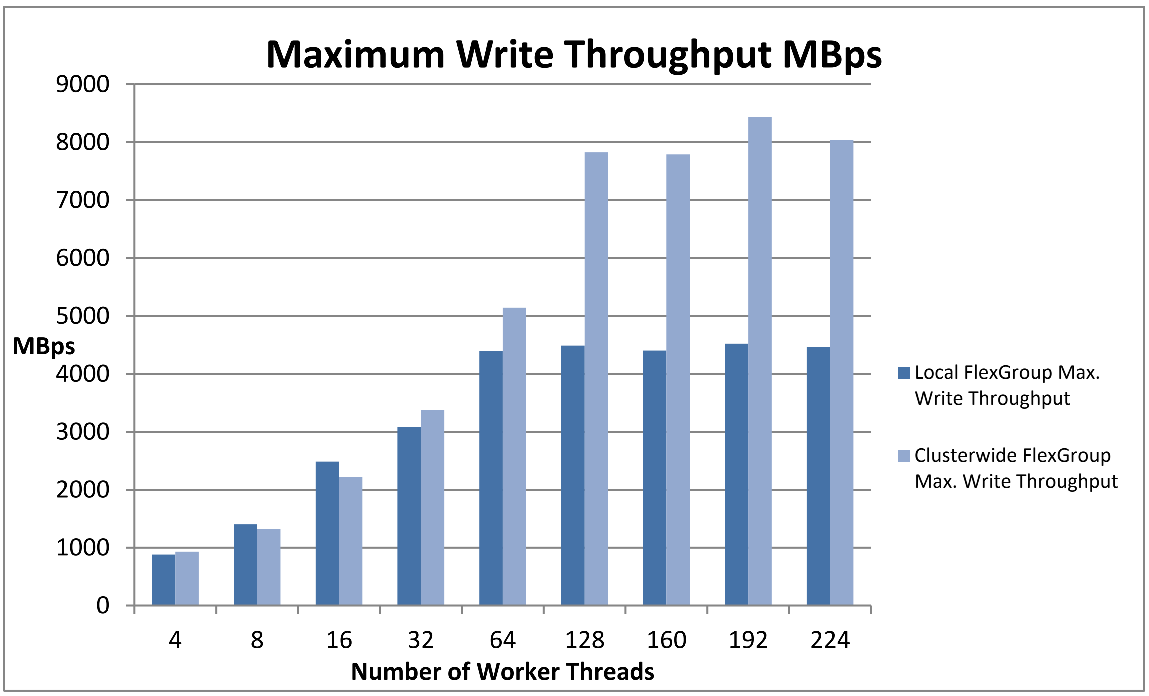
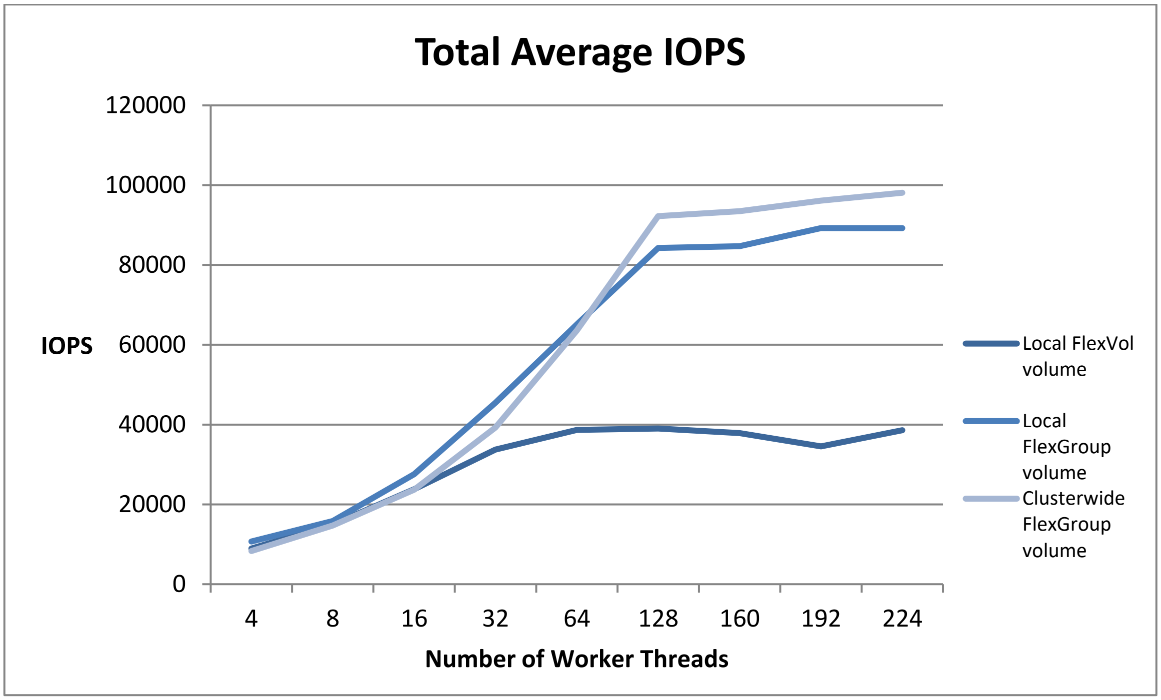


Figure 98 displays the total average IOPS for a local FlexVol volume versus the local and cluster-wide FlexGroup configurations. The FlexGroup configurations produce twice the IOPS that the FlexVol volume does, with the local FlexGroup volume outperforming the cluster-wide FlexGroup volume until the 64-thread tipping point.

Figure 98) Total average IOPS comparison.



In this test, 64 worker threads appear to be a sweet spot. Let's look at the average CPU utilization for a single-node FlexGroup volume versus a FlexGroup volume that spans the HA pair at just above 64 threads. Keep in mind that using more CPUs is a good thing; it means that work is being performed. That work is evidenced by the greater number of IOPS and the higher throughput for a FlexGroup volume that spans multiple nodes under the same workload.

Figure 99) Average CPU utilization, throughput, and IOPS for a FlexGroup volume—AFF A700 HA pair, 128 threads.

Figure 100) Average CPU utilization, throughput, and IOPS for a FlexGroup volume—single-node AFF A700, 128 threads.

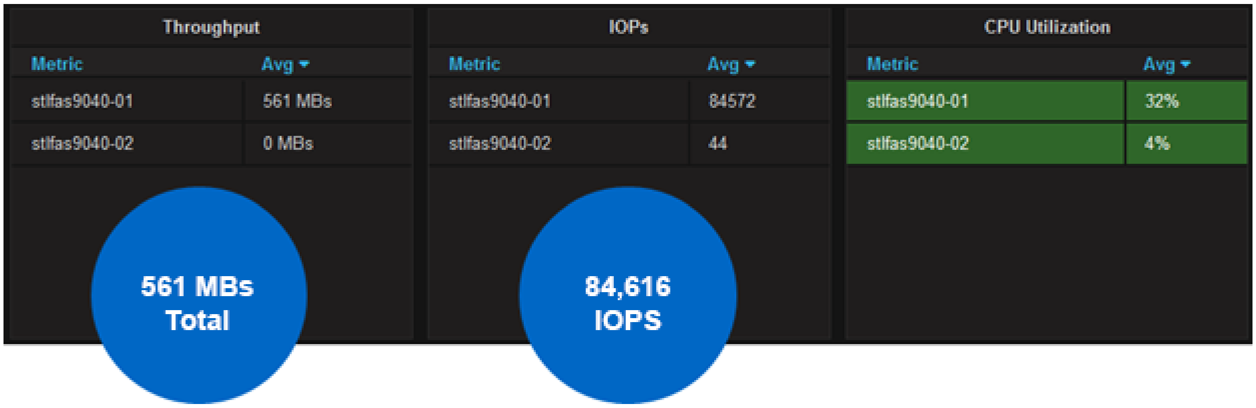
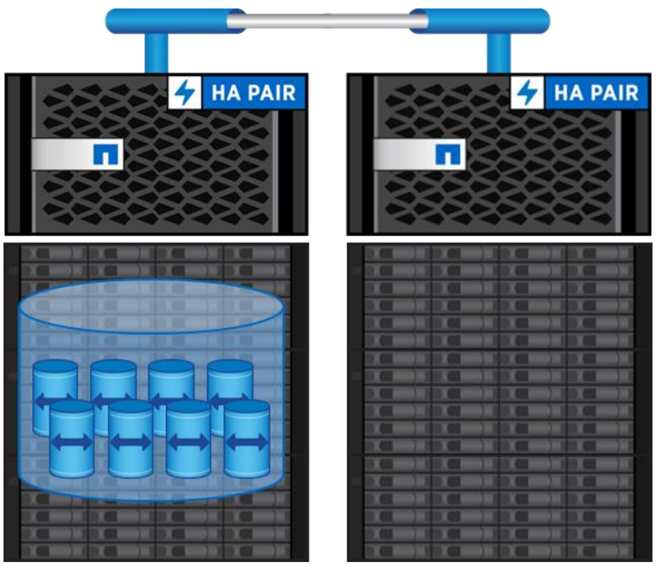


Figure 101) FlexGroup volume, single node.



Considerations

When you use a single node for a FlexGroup volume, the gains that are realized by removing cluster interconnect traversal from the equation disappear relatively quickly. They disappear as load is added to the node and CPU, RAM, network bandwidth, and disk utilization becomes an issue. Usually, it makes more sense to spread the FlexGroup volume across multiple nodes rather than trying to save minimal cluster interconnect bandwidth.

Use cases

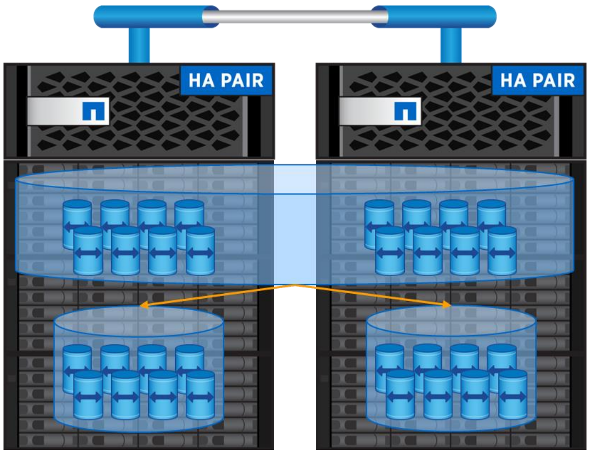
- High read workloads
- Need to isolate workloads to nodes
- Need to keep traffic off the cluster network

FlexGroup sample design 4: FlexGroup volumes mounted to FlexGroup volumes

With FlexVol volumes in ONTAP, you can mount volumes to other volumes to span the cluster and get >100TB in capacity, which was not possible with a single FlexVol volume. This method of designing a file system compares favorably with FlexGroup in terms of performance. However, the management overhead of creating multiple FlexVol volumes across multiple nodes and mounting them to each other in the namespace takes valuable personnel hours. In addition, scaling out capacity and performance can carry similar management headaches.

FlexGroup volumes can be managed like FlexVol volumes in the same way, by mounting a FlexGroup to another FlexGroup to create a folder structure with more granular data management.

Figure 102) FlexGroup volume mounted to FlexGroup volume.



Considerations

Mounting FlexGroup volumes to other FlexGroup volumes offers flexibility, but at the cost of management overhead and additional member volume counts.

Use cases

- More granular control over export policies and rules
- Greater control over the physical location of data
- Granular SnapMirror at the volume level for smaller datasets

FlexVol volumes mounted to FlexGroup volumes

NetApp FlexVol volumes can also mount to FlexGroup volumes, and conversely. This configuration is another possibility with a FlexGroup solution.

Use cases

- More granular control over export policies and rules
- Greater control over the physical location of data
- Features and functionality that are supported in FlexVol that aren't supported in FlexGroup use cases where a workload might occasionally create a large file or a small file that grows over time (for example, if a set of files gets zipped up to a larger zip file)

General troubleshooting and remediation

Failure scenarios

This section covers some failure scenarios and how a FlexGroup reacts.

Storage failovers

FlexGroup volumes are built on FlexVol volumes, so storage failover operations are functionally the same as for a FlexVol volume. Takeovers have no noticeable disruption. Nondisruptive upgrades, head swaps, rolling upgrades, and so on all perform normally. Givebacks of stateful protocols, such as SMB, are slightly disruptive, because of the transfer of locking states.

One caveat is that if an aggregate is not at home (not on the node that owns it, such as in a partial giveback state or if an aggregate has been relocated), FlexGroup volumes cannot be created until the aggregates are at home.

Note: In rare cases, if a node is powered off (dirty shutdown) in a cluster with four or more nodes, then the member volume cache entries on other nodes that host the FlexGroup volume might not flush properly and the FlexGroup appears to be hung because the surviving nodes are not made aware that the other member volumes have changed their locations. This issue is most commonly encountered during failover and resiliency testing. Bug 1358906 covers this problem and the workaround. This is fixed in ONTAP 9.7P9 and later.

Network failures

If a network connection that is accessing a FlexGroup volume has an interruption or failure, the behavior for a FlexGroup volume mirrors that of a FlexVol volume. The cluster attempts to migrate the data LIF to a port or node that can access the network successfully. Clients may experience a brief disruption, as expected with network issues and depending on the protocol version in use.

Snapshot failures

If a FlexGroup Snapshot copy fails, ONTAP considers that Snapshot copy to be partial and invalidates it for NetApp SnapRestore operations. The Snapshot set is cleaned up by ONTAP and an EMS event is logged (`mgmtgwd.snapshot.partCreate`).

Hardware failures

Disk failures on aggregates hosting FlexGroup volumes operate the same as with a FlexVol; ONTAP fails the disk and selects a spare to use in a rebuild operation. If more disks in an aggregate fail than are allowed in a RAID configuration, then the aggregate is considered offline and the member volumes that live on the offline aggregate are inaccessible.

The main difference between a FlexVol and a FlexGroup here is that in a FlexGroup volume that spans multiple aggregates, access to other member volumes is fenced off to prevent data inconsistencies until the hardware issue has been addressed and the other member volumes are back online.

Node failures result in a storage failover event, where aggregates owned by the node that fails transfer ownership to the HA partner node, and the FlexGroup volume continues operations normally. If two nodes fail in the same HA pair, then the FlexGroup volume has member volumes that are considered to be offline, and data access is fenced off until the nodes are repaired and back in working order.

Time synchronization

If there is a time skew between the nodes hosting the members of a NetApp ONTAP FlexGroup, disruption might occur on the FlexGroup volumes. A time skew is a relative difference in local time between the nodes hosting a FlexGroup.

SMB/NFS protocol operations such as renaming a directory or symlink or unlinking a symlink trigger an internal cache invalidation resulting in a time-based calculation between nodes. If the time skew is large enough, incorrect repeat triggering of cache invalidation operations might occur. The repeat cache invalidation operations prevent the completion of the rename and unlink operations. The messages continue to retry and might affect operations across the FlexGroup and cause disruption.

This issue is covered in bug 1252998 and is fixed in ONTAP 9.8. If you are running an ONTAP release prior to ONTAP 9.8, make sure that the cluster node times are in sync, ideally with a Network Time Protocol (NTP) configuration. ONTAP 9.6P10 and ONTAP 9.7P6 introduce an EMS notification to alert storage administrators if this issue is occurring. To obtain an ONTAP release, use the ONTAP 9 download page.

Incorrect mtime values on files and folders

In rare instances, a file or folder created on a FlexGroup volume might have an mtime value that does not match the actual time the file was modified (it might be off by several seconds), even though all node times are in sync. This issue is exacerbated in environments using CIFS/SMB and NFS on the same volume and is covered in bug 1380552. This bug is fixed in the ONTAP releases listed on the bug's public report page.

Capacity monitoring and alerting

This section covers various methods of monitoring a FlexGroup volume's capacity, including viewing total storage efficiency savings. Monitoring FlexGroup capacity is also possible with the NetApp FPolicy support introduced in ONTAP 9.4.

Capacity monitoring and alerting becomes less of a concern with the ONTAP 9.8 proactive resizing feature, because the total FlexGroup free space should more closely mirror the individual member volume free space.

Capacity monitoring and alerting with the command line

When you use thin provisioning, use the command `storage aggregate show-space` with `volume show -is-constituent true`, `volume show-space`, and `storage aggregate show` to get better total visibility into FlexGroup volume space usage overall. In the command line, you can also use the `-sort-by` option to organize the list.

Note: To get an accurate portrayal of the space that is being used, pay attention to the `Physical Used` portion of the `volume show-space` command. You can find an example in the section "Command Examples."

Event management system messages

Event management system messages alert storage administrators about the capacity of volumes in ONTAP. The messages are listed in this section. You can view them in the command line with the command `event route show -messagename [message] -instance`. For an example of these messages, see the section "Examples of capacity-related event management system messages."

- Unmodifiable values:
  - Severity level
  - Corrective actions
  - Description
  - SNMP support
- Modifiable values:
  - Destinations
  - Allowed drops or intervals between transmissions

When an event management system message that has SNMP support is triggered, an SNMP trap fires to the configured SNMP server. This action is specified through the `destinations` value. For more information about configuring event management system destinations, see the [Express Guide for your specific version of ONTAP](#).

The default values for Nearly Full (Warning) and Full (Error) are as follows:

```
cluster::*> vol show -vserver SVM -volume flexgroup -fields space-nearly-full-threshold-
percent,space-full-threshold-percent
vserver volume       space-nearly-full-threshold-percent space-full-threshold-percent
------- ------       ----------------------------------- ----------------------------
SVM     flexgroup    95%                                 98%
```

Event management system messages for `volume.full` look like the following:

```
11/28/2016 18:26:34 cluster-01
DEBUG         monitor.volume.full: Volume flexgroup@vserver:05e7ab78-2d84-11e6-a796-00a098696ec7
is full (using or reserving 99% of space and 0% of inodes).
```

In the preceding example, the following values are provided:

- The type of object
- The name of the volume
- The SVM (called `vserver` in the CLI) universal unique identifier (UUID)
- Percentage of space used
- Percentage of inodes used

You can use these values when testing event management system messages. When you look for which SVM is affected by the errors, use the UUID string at the advanced privilege level:

```
cluster::*> vserver show -uuid 05e7ab78-2d84-11e6-a796-00a098696ec7
                               Admin      Operational Root
Vserver     Type    Subtype    State      State       Volume     Aggregate
----------- ------- ---------- ---------- ----------- ---------- ----------
SVM         data    default    running    running     vsroot     aggr1_
                                                                 node1
```

Testing event management system messages

To test an event management system message, use the `event generate` command (available at the diagnostic privilege level). Each message has a unique string of values. The values for `volume.full` and `volume.nearlyFull` are listed in the preceding section. The following example shows how to construct a test message for a `volume.nearlyFull` event and the resulting event management system message:

```
cluster::*> event generate -message-name monitor.volume.nearlyFull -values Volume flexgroup
@vserver:05e7ab78-2d84-11e6-a796-00a098696ec7 95 0

cluster::*> event log show -message-name monitor.volume.nearlyFull
Time                Node             Severity      Event
------------------- ---------------- ------------- ---------------------------
11/28/2016 18:36:35 cluster-01
                                     ALERT         monitor.volume.nearlyFull: Volume
flexgroup@vserver:05e7ab78-2d84-11e6-a796-00a098696ec7 is nearly full (using or reserving 95% of
space and 0% of inodes).
```

Modifying the Volume Full and Nearly Full thresholds

With a FlexVol volume, the default values for `full` and `nearlyFull` are fine because the volume is isolated to a single container. With a FlexGroup volume, by the time a member FlexVol volume reaches the full or nearly full threshold, the application or end user might already be seeing a performance degradation. This decreased performance is due to increased remote file allocation or a FlexGroup volume that is already reporting to be out of space because of a full or nearly full member volume.

Aggressive capacity monitoring is necessary for versions earlier than ONTAP 9.3 because the volume autogrow functionality was unavailable for a FlexGroup volume until ONTAP 9.3. Later versions of ONTAP—especially ONTAP 9.8 and later—don't need the same level of aggressiveness, as features such as proactive resizing take much of the guesswork out of capacity management.

To help monitor for these scenarios, the volume full and nearly full thresholds might require adjustment to deliver warnings and errors before a volume fills up. Volumes have options to adjust these thresholds at the admin privilege level.

```
-space-nearly-full-threshold-percent
-space-full-threshold-percent
```

Use the `volume modify` command to adjust these thresholds.

Best practice 25: Volume space threshold recommendations for FlexGroup volumes

Generally speaking, the nearly full threshold of a FlexGroup volume should be set to 80%, and full should be set to 90%. With these settings, you have enough time to remediate space issues by increasing the FlexGroup volume size or adding more capacity in the form of additional member volumes through volume expand. These values can vary based on the average file size and the FlexGroup member volume size.

For instance, a 1TB FlexGroup member can reach 80% immediately with an average file size of 800GB, but a 100TB FlexGroup member would take longer to hit that threshold. See the section "Nonideal workloads — large files" for guidance on large file workloads.

For examples using Active IQ to monitor and alert for capacity and inodes, see Capacity Monitoring and Alerting Examples in Active IQ Unified Manager.

Client-side capacity considerations with thin provisioning

When using a FlexGroup volume, the client usually reports the available space, the used space, and so on in a way that reflects what the storage administrator has provisioned. This reporting is especially true when the volume space guarantee is set to `volume`, because ONTAP returns the expected capacities to the client.

However, when you use thin provisioning and overprovisioning for your physical storage, the client values do not reflect the expected used capacity of the FlexGroup volume. Instead, they reflect the used capacity in the physical aggregate. This approach is no different from the behavior of FlexVol volumes.

In the following example, there are three FlexGroup volumes:

- `flexgroup` has 80TB allocated and is thin provisioned across two aggregates with about 10TB available.
- `flexgroup4TB` has 4TB allocated with a space guarantee of `volume`.
- `flexgroup4TB_thin` has 4TB allocated and is thin provisioned across two aggregates with about 4TB available.

The following output shows that the cluster sees the proper used space in the volumes.

```
cluster::> vol show -fields size,used,percent-used,space-guarantee,available -vserver SVM
-volume flexgroup*,!*__0* -sort-by size
vserver volume            size available used     percent-used space-guarantee
------- ----------------- ---- --------- -------- ------------ ---------------
SVM     flexgroup4TB      4TB  3.77TB    30.65GB 5%            volume
SVM     flexgroup4TB_thin 4TB  3.80TB    457.8MB 5%            none
SVM     flexgroup         80TB 10.13TB   5.08GB  87%           none
3 entries were displayed.
```

However, the client sees the used capacity of the overprovisioned FlexGroup volume named `flexgroup` as 66TB, rather than the 5GB that is seen on the cluster. This total includes the total available size of the physical aggregate (5.05TB + 5.08TB = ~10TB) and subtracts that from the total size.

The volumes that are not overprovisioned report space normally.

```
# df -h
Filesystem                        Size  Used Avail Use% Mounted on
10.193.67.220:/flexgroup          76T   66T   11T  87% /flexgroup
10.193.67.220:/flexgroup4TB       3.9T   31G  3.8T   1% /flexgroup4TB
10.193.67.220:/flexgroup4TB_thin  3.9T  230M  3.8T   1% /flexgroup4TB_thin

cluster::*> aggr show -aggregate aggr1* -fields usedsize,availsize,percent-used,size
aggregate   availsize percent-used size   usedsize
----------- --------- ------------ ------ --------
aggr1_node1 5.05TB    36%          7.86TB 2.80TB
aggr1_node2 5.08TB    35%          7.86TB 2.78TB
2 entries were displayed.
```

The ~11TB of available space comes from the way that the Linux client calculates the space. This client does 1K blocks, so the number 10881745216 is divided into factors of 1,000. ONTAP uses factors of 1,024 to calculate space.

```
# df | grep flexg
10.193.67.220:/flexgroup         85899345920 75017600704 10881745216  88% /flexgroup
10.193.67.220:/flexgroup4TB       4080218944    32143296  4048075648   1% /flexgroup4TB
10.193.67.220:/flexgroup4TB_thin  4080218944      468736  4079750208   1% /flexgroup4TB_thin
```

Also, the `size` portion of the output considers the default 5% that is allocated for Snapshot space. That's why 80TB becomes 76TB in the preceding `df` output.

```
cluster::> vol show -fields size,percent-snapshot-space -vserver SVM -volume flexgroup*,!*__0*
-sort-by size
vserver volume            size percent-snapshot-space
------- ----------------- ---- ----------------------
SVM     flexgroup4TB      4TB  5%
SVM     flexgroup4TB_thin 4TB  5%
SVM     flexgroup         80TB 5%
3 entries were displayed.
```

When the Snapshot space allocation is reduced to 0, `df` reports a more normalized version of the actual size (but still has the strangeness of the `used` space).

```
cluster::*> vol modify -vserver SVM -volume flexgroup -percent-snapshot-space 0
[Job 2502] Job succeeded: volume modify succeeded

# df -h | grep flexgroup
Filesystem                        Size  Used Avail Use% Mounted on
10.193.67.220:/flexgroup           80T   70T   11T  88% /flexgroup
```

Windows capacity reporting

Windows reports in very much the same way as the Linux clients. The difference is that Windows uses a factor of 1,024, so the numbers are closer to the ONTAP values.

Viewing FlexVol member capacity from the ONTAP command line

When FlexGroup volumes are created, each member is evenly divided according to the total capacity and the number of FlexVol members. For example, in the case of an 80TB FlexGroup volume, the FlexVol members are 10TB apiece. To view member volume capacity, use the `volume show` command at the diagnostic privilege level; use `volume show -is-constituent true` or use the `volume show-space` command at the admin privilege level.

Viewing FlexVol member capacity is useful when you are trying to determine the true available space in a FlexGroup volume. When a FlexGroup volume reports total available space, it considers the total available space on all member volumes. However, when an individual member volume fills to capacity, the entire FlexGroup volume reports as out of space, even if other member volumes show available space. To mitigate this scenario, the FlexGroup ingest algorithms attempt to direct traffic away from a volume that becomes more heavily used than other volumes.

FlexGroup capacity viewer

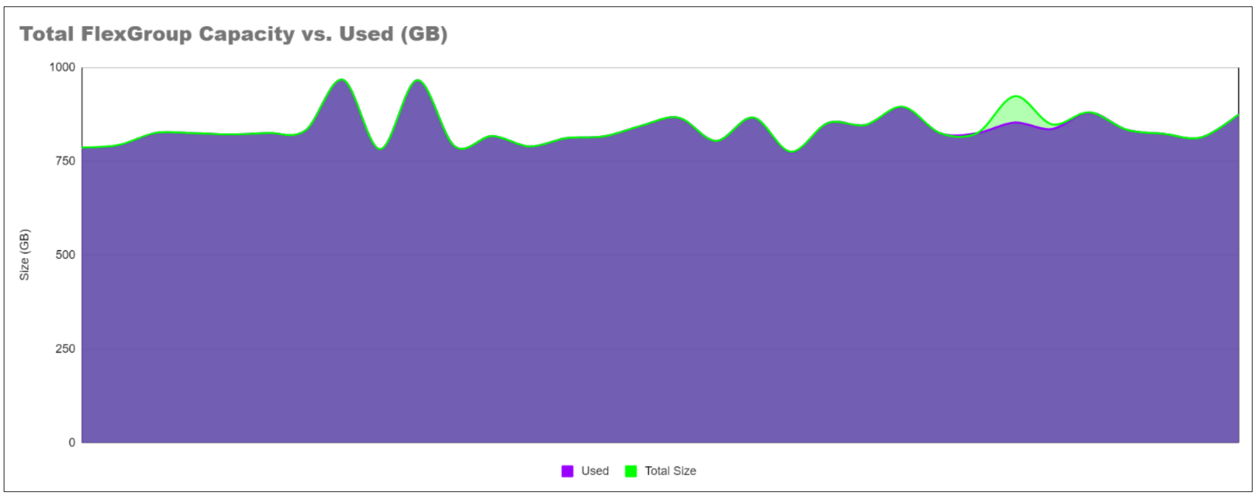
You can also view capacity and inode usage for a FlexGroup volume via custom Google sheets. You can request access for your own copy with an email to flexgroups-info@netapp.com and including your cluster serial number and the name of a FlexGroup volume you would like to see graphed out.

The Google Sheet graphs include:

- Holistic views of the FlexGroup used and total space
- Snapshot reserve and used graphs
- Member volume level views of used and total capacity
- Inode and used capacity trend lines

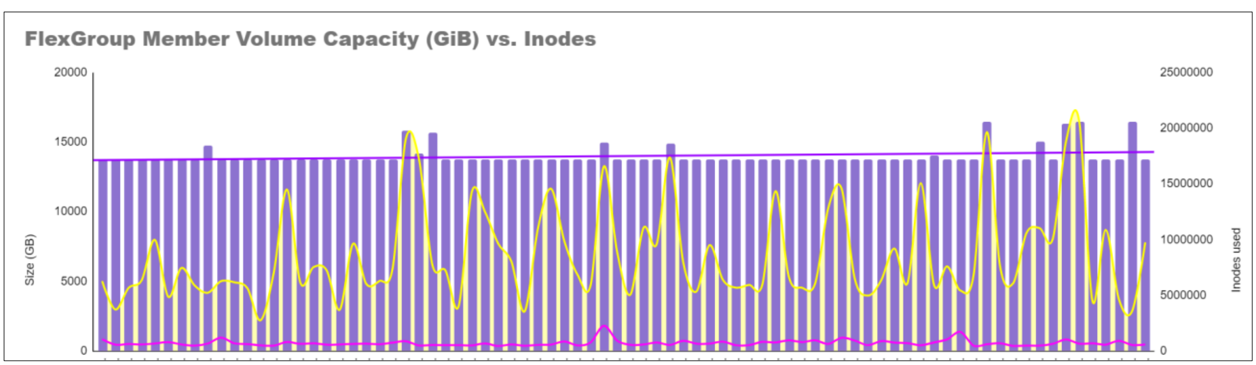
Here is a view of a FlexGroup volume with an uneven capacity usage that also has elastic resizing in action. In this case, the FlexGroup needs total capacity to be increased.

Figure 103) Google Sheet — FlexGroup capacity view.



This view shows a FlexGroup volume capacity at the member volume level and includes a trend line of used inodes. Ideally, the inodes used trend in parallel with the used capacity. In this case, we see spikes in capacity per member volume that do not correspond with counts for used inodes, meaning some member volumes might have larger files than others.

Figure 104) Google Sheet — FlexGroup member capacity view.

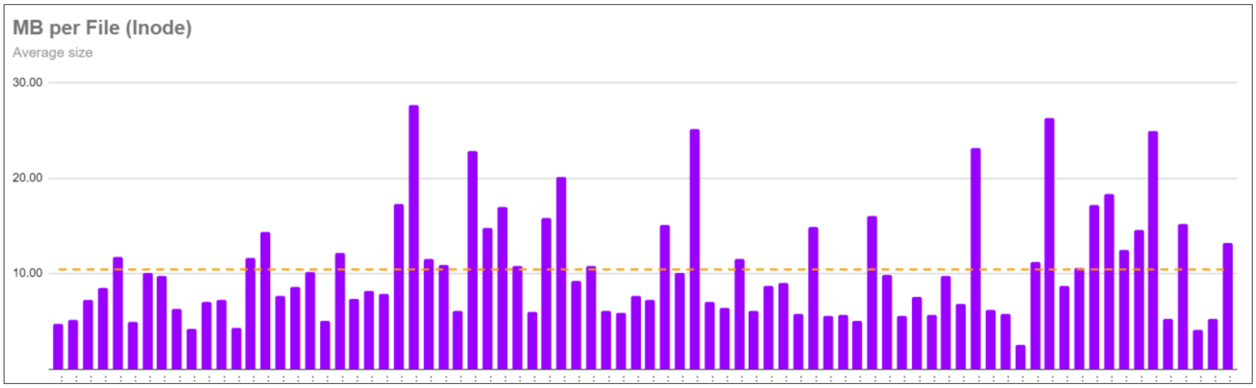


There are also graphs that attempt to take the total used capacity and divide by the used inodes to attempt to show an average file size. It's not an exact science but gets fairly close.

In the following example, we can see that the purple bar graphs have widely disparate values, which means that the average file sizes have a large range. The orange line is the average inode size for the entire FlexGroup volume. Further investigation may be needed for this FlexGroup to discover why there is such a wide range of average inode sizes.

Note: File size disparity and capacity imbalances do not necessarily indicate a problem, but you can use these graphs if there is a problem and it needs to be identified. See the section "Data imbalances in FlexGroup volumes" for more information.

Figure 105) Google Sheet – Average inode size.



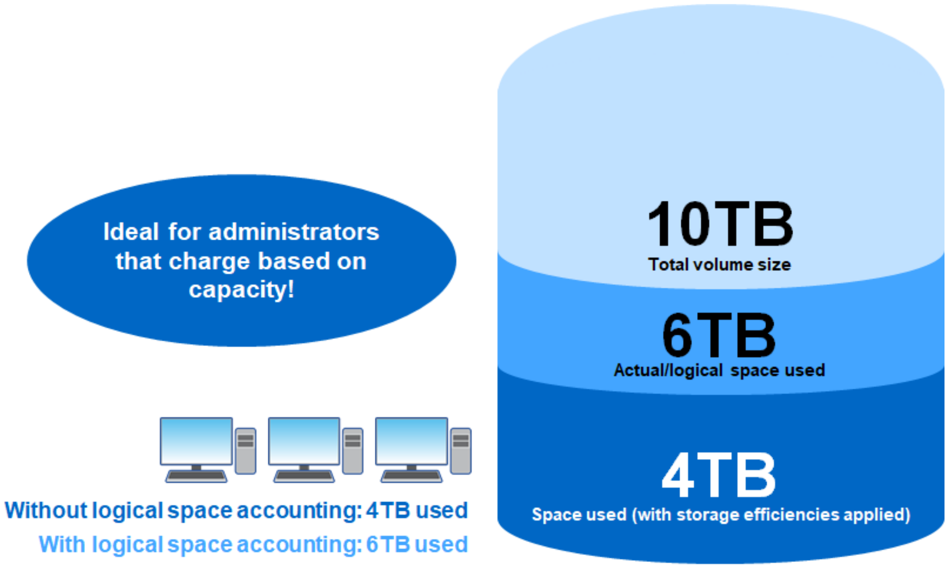
Note: Request a copy of the FlexGroup Capacity Viewer by emailing flexgroups-info@netapp.com with your cluster serial number and FlexGroup volume name.

Logical space accounting

Logical space accounting was introduced for FlexVol volumes in ONTAP 9.4. It enables storage administrators to mask storage efficiency savings so that end users avoid overallocating their designated storage quotas.

For example, if a user writes 6TB to a 10TB volume and storage efficiencies save 2TB, logical space accounting can control whether the user sees 6TB or 4TB.

Figure 106) How logical space accounting works.



ONTAP 9.5 enhanced this feature and added quota enforcement support to give more control to storage administrators by preventing new writes according to the logical space.

ONTAP 9.9.1 and later adds support for logical space reporting and enforcement.

Monitoring FlexGroup performance

FlexGroup performance can be monitored in many of the same ways that a normal FlexVol volume's performance can be monitored. The same concepts of CPU utilization, disk saturation, NVRAM bottlenecks, and other NetApp WAFL-related performance characteristics apply. Also, NAS performance monitoring doesn't change. You still use the basic CIFS/SMB and NFS statistics that you always have.

The main difference with monitoring FlexGroup performance is that you must consider multiple nodes, aggregates, member FlexVol constituent volumes, and the notion of remote placement of files and folders. These elements add another layer to consider when you want to monitor and isolate performance issues.

Monitoring performance from the command line

From the command line, you have several ways to view performance statistics.

Real-time performance monitoring

To monitor system performance in real time, use the `statistics show-periodic` command.

```
cluster::*> statistics show-periodic ?
  [[-object] <text>]           *Object
  [ -instance <text> ]         *Instance
  [ -counter <text> ]          *Counter
  [ -preset <text> ]           *Preset
  [ -node <nodename> ]         *Node
  [ -vserver <vserver name> ]  *Vserver
  [ -interval <integer> ]      *Interval in Seconds (default: 2)
  [ -iterations <integer> ]    *Number of Iterations (default: 0)
  [ -summary {true|false} ]    *Print Summary (default: true)
  [ -filter <text> ]           *Filter Data
```


This command provides an up-to-date glimpse into system performance. Leaving the default values alone gives you a cluster-wide view. Specifying an SVM gives you a more granular look, but mainly at the counters that would be specific to an SVM, such as NAS counters, rather than to CPU or disk. When you use SVM-specific statistics, defining the counters that are provided for the object helps reduce the noise on the CLI. You can also get real-time FlexGroup statistics for the ratios of local to remote top-level directories (`tld`), high-level directories (`hld`), regular directories, and files.

For examples of these commands, see the section "Command Examples."

The FlexGroup statistics also can show various other information and can be gathered over a period of time if you initiate a `statistics start -object flexgroup` command. This command collects statistics over time that can be captured in iterations through an automated tool such as [Perfstat or perfarchives](#).

```
cluster::*> statistics start -object flexgroup
Statistics collection is being started for sample-id: sample_69197
```

Use the following to view the statistics:

```
cluster::*> statistics show -object flexgroup -instance 0

Object: flexgroup
Instance: 0
Start-time: 11/30/2016 16:44:42
End-time: 11/30/2016 17:42:57
Elapsed-time: 3495s
Scope: cluster-01

    Counter                                                        Value
    -------------------------------- --------------------------------
    cat1_tld_remote                                                    2
    cat2_hld_local                                                   180
    cat2_hld_remote                                                 1292
    cat3_dir_local                                                146804
    cat3_dir_remote                                                  283
    cat4_fil_local                                                734252
    cat4_fil_remote                                                 1124
    groupstate_analyze                                             12232
    groupstate_update                                              86242
    instance_name                                                      0
    node_name                                             cluster-01
    process_name                                                       -
    refreshclient_create                                            5241
    refreshclient_delete                                            5241
    refreshserver_create                                            5244
    refreshserver_delete                                            5244
```

The statistics capture gives a nice summary of the percentages of remote file and directory placement in the FlexGroup volume when it spans multiple nodes. In the following example, the values are 14% remote directories and 1% remote files.

```
    remote_dirs                                                       14
    remote_files                                                       1
```

Protocol statistics

It is also possible to get a glimpse of how individual NAS protocols are influencing performance. Simply use the `statistics start` command to include NFS or SMB performance counters in the capture. You have more options with diagnostic privileges.

```
cluster::*> statistics start -object nfs
    nfs_credstore                      nfs_exports_access_cache
    nfs_exports_cache                  nfs_exports_match
    nfs_file_session_cache             nfs_file_session_cache:constituent
    nfs_generic                        nfs_idle_conn
```



```
    nfs_idle_total_conn                 nfs_qtree_export
    nfs_server_byname                   nfserr
    nfsv3                               nfsv3:constituent
    nfsv3:cpu                           nfsv3:node
    nfsv4                               nfsv4:constituent
    nfsv4:cpu                           nfsv4:node
    nfsv4_1                             nfsv4_1:constituent
    nfsv4_1:cpu                         nfsv4_1:node
    nfsv4_1_diag                        nfsv4_1_error
    nfsv4_diag                          nfsv4_error
    nfsv4_spinnp_errors

cluster::*> statistics start -object smb
    smb1          smb1:node     smb1:vserver  smb1_ctx      smb1_ctx:node
    smb2          smb2:node     smb2:vserver  smb2_ctx      smb2_ctx:node

cluster::*> statistics start -object cifs
    cifs                             cifs:node
    cifs:vserver                     cifs_cap
    cifs_cap:constituent             cifs_client
    cifs_client:constituent          cifs_ctx
    cifs_ctx:node                    cifs_shadowcopy
    cifs_unsupp_ioctl                cifs_unsupp_ioctl:constituent
    cifs_watch
```

flexgroup show

During FlexGroup I/O, you can also view the member constituent usage and balance through the nodeshell command `flexgroup show`. The command also provides other information that can be useful, such as how often a member volume might be avoided for new files. Be sure to capture this command output if you run into a FlexGroup issue and need to open a support case.

```
cluster::*> node run * flexgroup show Tech_ONTAP
FlexGroup 0x80F03868 (Tech_ONTAP)
* next snapshot cleanup due in 9334 msec
* next refresh message due in 334 msec (last to member 0x80F0386E)
* spinnp version negotiated as 10.13, capability 0x3F7F
* Ref count is 8
* ShouldEnforceQuotas true
* IsAnyMemberInNvfailedState false
* reaction +0.0, workload +0.0, activity level 0, cv 0%
Idx   Member L            Used      Avail Urgc  Target          Probabilities   D-Ingest Alloc
F-Ingest Alloc
--- -------- - --------------- ---------- ---- -------  ---------------------  --------- -----  -
-------- -----
  1     4503 L    2238647   0%  318698244   0%  12.50%  [100% 100%  87%  87%]     0+   0     0
0+   0     0
  2     4369 R    3239783   1%  318638088   0%  12.49%  [100% 100%  87%  87%]     0+   0     0
0+   0     0
  3     4674 L    2011415   0%  318697586   0%  12.50%  [100% 100%  87%  87%]     0+   0     0
0+   0     0
  4     4477 R    2334885   0%  318694396   0%  12.50%  [100% 100%  87%  87%]     0+   0     0
0+   0     0
  5     4329 L    2250619   0%  318697596   0%  12.50%  [100% 100%  87%  87%]     0+   0     0
0+   0     0
  6     4370 R    2255368   0%  318697148   0%  12.50%  [100% 100%  87%  87%]     0+   0     0
0+   0     0
  7     4675 L    2252390   0%  318697125   0%  12.50%  [100% 100%  87%  87%]     0+   0     0
0+   0     0
  8     4478 R    2201995   0%  318698611   0%  12.50%  [100% 100%  87%  87%]     0+   0     0
0+   0     0
```

Output breakdown for the flexgroup show command

The `flexgroup show` command has a series of values that might not be intuitive at first glance. Table 25 describes those values and how to interpret them.

Table 25) **`flexgroup show`** output column definitions.

Column	Definition
`Idx`	Index number of the member volume.
`Member`	DSID of the FlexGroup member.
`L`	Local or remote to the node.
`Used`	Number and overall percentage of 4K blocks used.
`Urgc`	Urgency: Probability of a file or directory creation being allocated to a remote member volume to avoid premature `ENOSPC` in a member volume. This value increases according to how close to 100% used a volume's capacity is.
`Targ`	Target: Percentage of what new content should be placed on a member volume as related to its peers. The total summation of all target percentages equal ~100%.
`Probabilities`	The likelihood that a member volume is avoided for use. This number increases according to how full a member volume becomes in relation to other member volumes (tolerance).
`D-Ingest` and `D-Alloc`	Directory ingest and directory allocation, respectively; how many directories have been allocated to a local member volume.
`F-Ingest` and `F-Alloc`	File ingest and file allocation, respectively; how many files have been allocated to a local member volume.

You should run the `flexgroup show` command during a period of I/O activity on a FlexGroup volume. This command gives the following useful information:

- How evenly the traffic is distributed across members
- How evenly distributed the space is on members
- How likely a member volume is to be used for ingesting
- The ratio of directory to file creation in a workload
- The member volume's node locality

Monitoring performance by using Perfstat (ONTAP 9.4 and earlier)

Perfstat is a tool on the NetApp Support site that can capture real-time performance statistics for benchmarking or to troubleshoot current performance issues. These statistics are captured at specified intervals and are crucial for NetApp Support to resolve performance cases.

Perfstat in ONTAP 9.1 and later supports the capture of FlexGroup statistics. When Perfstat runs, it collects data in raw text format and can be reviewed with any text-editing software. Perfstat even has a Windows GUI version that you can download.

When Perfstat finishes running, the tool zips the contents up for submittal to NetApp Support. The tool also creates a folder with the output files that can be read with a text editor. However, this version is in plaintext and isn't easy to read. If you are a NetApp original equipment manufacturer (OEM) partner or an internal employee, you can use the LaTeX tool to view these files.

For more information, see the KB article on how to use Perfstat.

Note: Perfstat is deprecated as of ONTAP 9.5 and later. For later ONTAP versions, use Active IQ, NetApp Harvest, or the Performance Archiver.

Performance archiver

Starting in ONTAP 9.5, performance data is captured for support issues through the performance archiver, which runs by default in ONTAP. These statistics are always on and can be sent to NetApp using the `autosupport invoke-performance-archive` command. This command can specify date ranges, so you don’t have to hope you catch the problem in action.

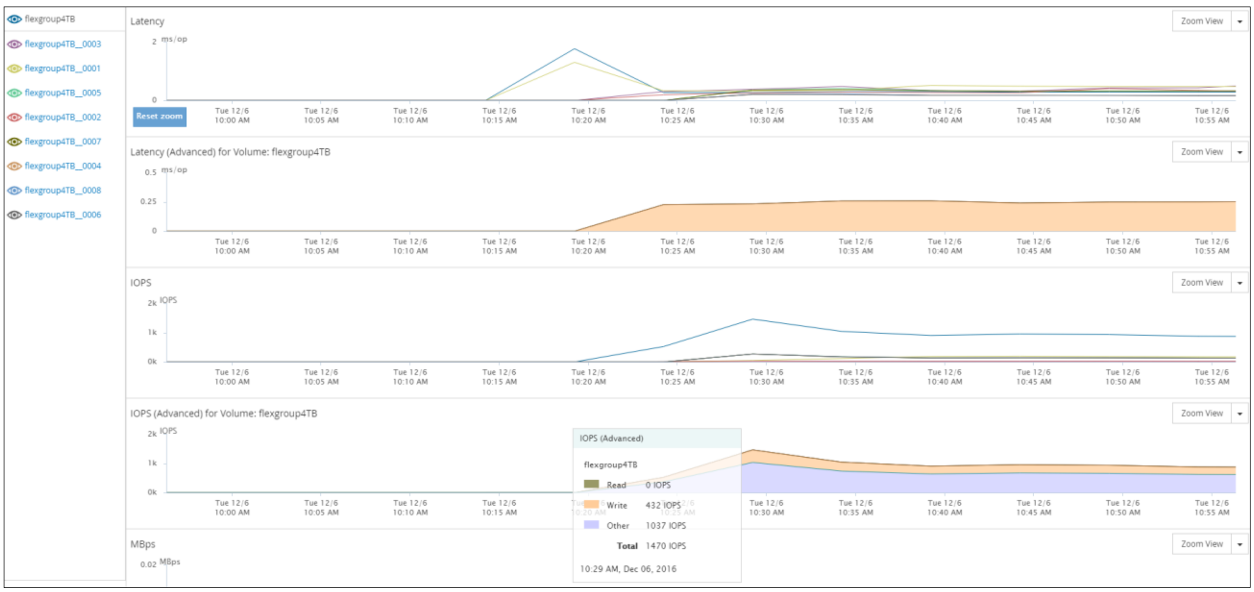
Monitoring performance (Active IQ Unified Manager)

A more palatable and widely available tool for monitoring the performance of a FlexGroup volume is NetApp Active IQ Unified Manager. This tool is available as a free .ova file or as a Linux installation from the [NetApp download page of the Support site](https://mysupport.netapp.com/site/downloads).

Active IQ Unified Manager offers both real-time and historical performance information to provide a single monitoring point. Active IQ Unified Manager can give granular performance views for the entire FlexGroup volume or for individual member constituent FlexVol volumes.

Figure 107 is a capture of a simple file creation script on a single Linux VM, so the performance benefits of FlexGroup are not seen here. However, the figure does provide a sense of what Unified Manager can deliver.

Figure 107) Active IQ Performance Manager graphs.



FlexGroup data protection best practices

For the FlexGroup data protection best practices, see [TR-4678, Data Protection and Backup: NetApp FlexGroup Volumes](https://www.netapp.com/pdf.html?item=/media/17064-tr4678pdf.pdf).

Migrating to NetApp ONTAP FlexGroup

One challenge of having many files or a massive amount of capacity is deciding how to effectively move the data as quickly and as nondisruptively as possible. This challenge is greatest in high-file-count, high-metadata-operation workloads. Copies of data at the file level require file-system crawls of the attributes and the file lists, which can greatly affect the time that it takes to copy files from one location to another. That duration does not account for other aspects such as network latency, WANs, system performance bottlenecks, or other things that can make a data migration painful.

With NetApp ONTAP FlexGroup, the benefits of performance, scale, and manageability are apparent.

Data migrations can take three general forms when dealing with FlexGroup:

- Migrating from non-NetApp (third-party) storage to FlexGroup
- Migrating from NetApp Data ONTAP operating in 7-Mode to FlexGroup
- Migrating from NetApp FlexVol volumes, SAN LUNs, or Infinite Volume in ONTAP to FlexGroup

Data migrations to FlexGroup volumes are the best way to migrate. FlexGroup volume migrations currently cannot be performed with the following methods:

- FlexVol to FlexGroup volume move
- NetApp SnapMirror or SnapVault between FlexVol and FlexGroup
- 7-Mode Transition Tool (CBT and CFT)

The following sections cover different migration use cases and how to approach them.

Migration using NDMP

In ONTAP 9.7 and later, FlexGroup volumes now support NDMP operations. These include the `ndmpcopy` command, which can be used to migrate data from a FlexVol to a FlexGroup volume. For information about setting up `ndmpcopy`, see [How to run ndmpcopy in Clustered Data ONTAP.](#)

In the following example, `ndmpcopy` is used to migrate around five million folders and files from a FlexVol to a FlexGroup volume. The process takes around 51 minutes:

```
cluster::*> system node run -node ontap9-tme-8040-01 ndmpcopy -sa ndmpuser:AcDjtsU827tputjN -da
ndmpuser:AcDjtsU827tputjN 10.x.x.x:/DEMO/flexvol/nfs 10.x.x.x:/DEMO/flexgroup_16/ndmpcopy
Ndmpcopy: Starting copy [ 2 ] ...
Ndmpcopy: 10.x.x.x: Notify: Connection established
Ndmpcopy: 10.x.x.x: Notify: Connection established
Ndmpcopy: 10.x.x.x: Connect: Authentication successful
Ndmpcopy: 10.x.x.x: Connect: Authentication successful
Ndmpcopy: 10.x.x.x: Log: Session identifier: 12584
Ndmpcopy: 10.x.x.x: Log: Session identifier: 12589
Ndmpcopy: 10.x.x.x: Log: Session identifier for Restore : 12589
Ndmpcopy: 10.x.x.x: Log: Session identifier for Backup : 12584
Ndmpcopy: 10.x.x.x: Log: DUMP: creating "/DEMO/flexvol/../snapshot_for_backup.1" snapshot.
Ndmpcopy: 10.x.x.x: Log: DUMP: Using subtree dump
Ndmpcopy: 10.x.x.x: Log: DUMP: Using snapshot_for_backup.1 snapshot
Ndmpcopy: 10.x.x.x: Log: DUMP: Date of this level 0 dump snapshot: Thu Jan  9 11:53:18 2020.
Ndmpcopy: 10.x.x.x: Log: DUMP: Date of last level 0 dump: the epoch.
Ndmpcopy: 10.x.x.x: Log: DUMP: Dumping /DEMO/flexvol/nfs to NDMP connection
… (output omitted for length)
Ndmpcopy: 10.x.x.x: Notify: dump successful
Ndmpcopy: 10.x.x.x: Log: RESTORE: RESTORE IS DONE
Ndmpcopy: 10.x.x.x: Notify: restore successful
Ndmpcopy: Transfer successful [ 0 hours, 50 minutes, 53 seconds ]
Ndmpcopy: Done
```

The same dataset using `cp` over NFS takes 316 minutes—six times as long as `ndmpcopy`:

```
# time cp -R /flexvol/nfs/* /flexgroup/nfscp/

real    316m26.531s
user    0m35.327s
sys     14m8.927s
```

Using NetApp XCP, that dataset takes just under 20 minutes—or around 60% faster than `ndmpcopy`:

```
#  xcp copy 10.193.67.219:/flexvol/nfs 10.193.67.219:/flexgroup_16/xcp
Sending statistics...
5.49M scanned, 5.49M copied, 5.49M indexed, 5.60 GiB in (4.81 MiB/s), 4.55 GiB out (3.91 MiB/s),
19m52s.
```

Note: This XCP copy was done on a VM with a 1GB network and not much RAM or CPU; more robust servers will perform even better.

FlexVol to FlexGroup volume conversion

In ONTAP 9.7 and later, you can convert a single FlexVol volume to a FlexGroup volume containing a single member volume, in place, with less than 40 seconds disruption. This is regardless of how much data capacity or number of files reside in the volume. There is no need to remount clients, copy data, or make any other modifications that could create a maintenance window. After the FlexVol volume is converted to a FlexGroup volume, you can add new member volumes to the converted FlexGroup volume to expand the capacity.

Why convert a FlexVol volume to a FlexGroup volume?

FlexGroup volumes offer a few advantages over FlexVol volumes, such as the following:

- Ability to expand beyond 100TB and two billion files in a single volume
- Ability to scale out capacity or performance nondisruptively
- Multi-threaded performance for high-ingest workloads
- Simplification of volume management and deployment

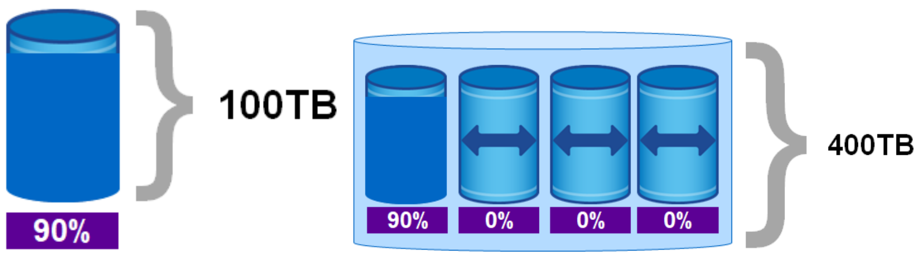
For example, perhaps you have a workload that is growing rapidly, and you do not want to have to migrate the data, but you still want to provide more capacity. Or perhaps a workload's performance is not good enough on a FlexVol volume, so you want to provide better performance handling with a FlexGroup volume. Converting to a FlexGroup volume can help in this case.

When not to convert a FlexVol volume

Converting a FlexVol volume to a FlexGroup volume might not always be the best option. If you require FlexVol features that are not available in FlexGroup volumes, then you should hold off. For example, NetApp SnapLock® and SnapMirror Synchronous are not currently supported for FlexGroup volumes, so if you need them, you should stay with FlexVol volumes.

Also, if you have a FlexVol volume that is already very large (80–100TB) and already very full (80–90%), you should copy the data rather than convert, because the converted FlexGroup volume has a very large, very full member volume. This might create performance issues and does not fully resolve your capacity issues, particularly if that dataset contains files that grow over time.

Figure 108) Converting a FlexVol volume that is nearly full and at maximum capacity.



If you convert this 90% full volume to a FlexGroup volume, you will have a 90% full member volume. If you add new member volumes, they are 100TB each and 0% full, so they take on a majority of new workloads. The data does not automatically rebalance and if the original files grow over time, you can still run out of space with nowhere to go (because 100TB is the maximum member volume size).

Things that can block a conversion

ONTAP blocks conversion of a FlexVol volume for the following reasons:

- The ONTAP version isn't 9.7 or later on all nodes.

- ONTAP upgrade issues are preventing conversion.
- A FlexVol volume was transitioned from 7-Mode using 7MTT (ONTAP 9.7).
  - Transitioned volumes can be converted as of ONTAP 9.8.
- Something is enabled on the volume that is not supported with FlexGroup yet (SAN LUNs, Windows NFS, SMB1, Snapshot naming/autodelete, `vmalign` set, SnapLock, space SLO, logical space enforcement/reporting, and so on).
- The SVM where the FlexVol volume to be converted lives is currently using SVM DR.
- NetApp FlexClone volumes are present, and the FlexVol volume is the parent volume (the volume being converted cannot be a parent or a clone).
- The volume is a NetApp FlexCache origin volume.
- NetApp Snapshot copies with Snap IDs greater than 255 (ONTAP 9.7).
  - ONTAP 9.8 adds support for 1023 snapshots, so this limit does not apply in that release.
- Storage efficiencies are enabled (can be reenabled after).
- The volume is a source of a SnapMirror relationship, and the destination has not been converted yet.
- The volume is part of an active (not quiesced) SnapMirror relationship.
- Quotas are enabled (they must be disabled first, then reenabled after).
- Volume names are longer than 197 characters.
- The volume is associated with an application.
  - ONTAP 9.7 only; ONTAP 9.8 removes this limitation.
- ONTAP processes are running (mirrors, jobs, wafliron, NDMP backup, inode conversion in process, and so on).
- SVM root volume.
- Volume is too full.

You can check for upgrade issues with the following commands:

```
cluster::*> upgrade-revert show
cluster::*> system node image show-update-progress -node *
```

You can check for transitioned volumes with the following commands:

```
cluster::*> volume show -is-transitioned true
There are no entries matching your query.
```

You can check for Snapshot copies with Snap IDs greater than 255 with the following command:

```
cluster::*> volume snapshot show -vserver DEMO -volume testvol -logical-snap-id >255 -fields
logical-snap-id
```

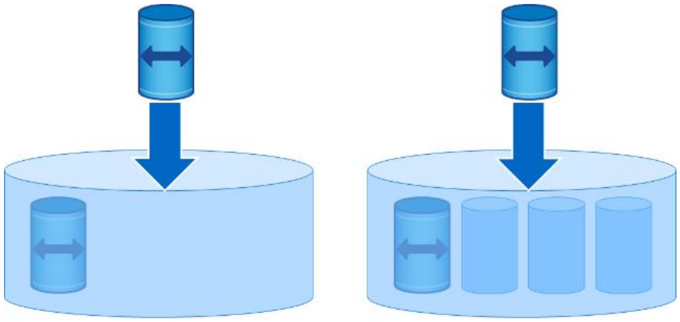
How it works

To convert a FlexVol volume to a FlexGroup volume in ONTAP 9.7 and later, you run a single, simple command at the advanced privilege level:

```
cluster::*> volume conversion start ?
-vserver <vserver name> *Vserver Name
[-volume] <volume name> *Volume Name
[ -check-only [true] ] *Validate the Conversion Only
[ -foreground [true] ] *Foreground Process (default: true)
```

When you run this command, ONTAP converts a single FlexVol volume into a FlexGroup volume with one member. You can even run a validation of the conversion before you do the real thing.

Figure 109) Converting a FlexVol volume to a FlexGroup and adding member volumes.



The process is 1:1, so you cannot currently convert multiple FlexVol volumes into a single FlexGroup volume. When the conversion is done, you have a single-member FlexGroup volume to which you can then add more member volumes of the same size to increase capacity and performance.

Other considerations and caveats

Although the actual conversion process is simple, there are some things to consider before converting. Most of these considerations will go away with future ONTAP releases as support is added for features, but it is still prudent to identify them here.

After the initial conversion is performed, ONTAP unmounts the volume internally and remounts it to get the new FlexGroup information into the appropriate places. Clients do not have to remount or reconnect, but they will see a disruption that last less than 1 minute while this takes place. For an example, see the section “Sample FlexVol to FlexGroup volume conversion”. Data does not change at all; file handles all stay the same.

- FabricPool does not need anything. It just works. No need to rehydrate data on premises.
- Snapshot copies remain available for clients to access data from, but you are not able to use them to restore the volume through `snaprestore` commands. Those Snapshot copies are marked as pre-conversion.
- SnapMirror relationships pick up where they left off without rebaselining, provided the source and destination volumes have both been converted. But there are no SnapMirror restores of the volume—just file retrieval from clients. SnapMirror destinations need to be converted first.
- FlexClone volumes need to be deleted or split from the volume to be converted.
- Storage efficiencies need to be disabled during the conversion, but your space savings are preserved after the conversion.
- FlexCache instances with an origin volume being converted must be deleted.
- Space guarantees can affect how large a FlexGroup volume can become if they are volume guarantees. New member volumes must be the same size as the existing members, so you need adequate space to honor them.
- Quotas are supported in FlexGroup volumes but are done a bit differently than in FlexVol volumes. So, while the conversion is being performed, quotas must be disabled (`quota off`) and then reenabled later (`quota on`).

Conversion to FlexGroup volumes is a one-way street after you expand it, so be sure you're ready to make the jump. If anything goes wrong during the conversion process, there is a rescue method that NetApp Support can help you use so that your data is safe even if you run into an issue.

When you expand the FlexGroup volume to add new member volumes, they are the same size as the converted member volume, so be sure there is adequate space available. Additionally, the existing data that resides in the original volume remains in that member volume. Data does not redistribute. Instead, the FlexGroup volume favors newly added member volumes for new files.

Are you nervous about converting?

If you do not feel comfortable about converting your production FlexVol volume to a FlexGroup volume right away, you have options.

First, ONTAP enables you to run a check on the conversion command with `-check-only true` that tells you what prerequisites you might be missing.

For example:

```
cluster::*> volume conversion start -vserver DEMO -volume flexvol -foreground true -check-only
true
Error: command failed: Cannot convert volume "flexvol" in Vserver "DEMO" to a FlexGroup. Correct
the following issues and retry the command:
* The volume has Snapshot copies with IDs greater than 255. Use the (privilege: advanced) "volume
snapshot show -vserver DEMO -volume flexvol -logical-snap-id >255 -fields logical-snap-id"
command to list the Snapshot copies with IDs greater than 255 then delete them using the
"snapshot delete -vserver DEMO -volume flexvol" command.
* Quotas are enabled. Use the 'volume quota off -vserver DEMO -volume flexvol' command to disable
quotas.
* Cannot convert because the source "flexvol" of a SnapMirror relationship is source to more than
one SnapMirror relationship. Delete other Snapmirror relationships, and then try the conversion
of the source "flexvol" volume.
* Only volumes with logical space reporting disabled can be converted. Use the 'volume modify -
vserver DEMO -volume flexvol -is-space-reporting-logical false' command to disable logical space
reporting.
```

For an example of a FlexVol to FlexGroup volume conversion, see the section “Sample FlexVol to FlexGroup volume conversion.”

Creating a conversion sandbox — migrating data

ONTAP can create multiple SVMs, which can be fenced off from network access. You can use this approach to test things such as volume conversion. The only trick is getting a copy of that data over—but it is really not that tricky.

Option 1: SnapMirror

You can use SnapMirror to replicate your to-be-converted volume to the same SVM or a new SVM. Then, break the mirror and delete the relationship. Now you have a sandbox copy of your volume, complete with Snapshot copies, to test out conversion, expansion, and performance.

Option 2: FlexClone and volume rehost

If you do not have SnapMirror or you want to try a method that is less taxing on your network, you can use a combination of FlexClone (an instant copy of your volume backed by a Snapshot copy) and `volume rehost` (an instant move of the volume from one SVM to another). Keep in mind that FlexClone copies cannot be rehosted, but you can split the clone and then rehost.

Essentially, the process is as follows:

1. Use `flexclone create`.
2. Use `flexclone split`.
3. Issue `volume rehost` to the new SVM (or convert on the existing SVM).

Note: Alternately, you can create FlexClone volumes from a source SVM to a destination SVM and then split the FlexClone as covered in the section “FlexClone to different storage virtual machines (SVMs)”.

Converting a FlexVol volume in a SnapMirror relationship

You can also convert FlexVol volumes that are part of existing SnapMirror relationships without disruption.

The basic steps are:

1. Break the SnapMirror relationship.
2. Convert the SnapMirror destination FlexVol volume to a FlexGroup volume.
3. Convert the source SnapMirror FlexVol volume to a FlexGroup volume.
4. Resync the SnapMirror.

If you expand the newly converted FlexGroup volume to add more member volumes, ONTAP automatically expands the destination volume without needing to return to baseline the SnapMirror relationship.

For an example of this process, see the section "Converting FlexVol volumes in existing SnapMirror relationships: Example."

Does a high file count affect the conversion process?

Short answer: No.

In the sample conversion shown in the section "Sample FlexVol to FlexGroup volume conversion," a volume with 300,000 files is converted. However, having 300,000 files in a volume is not a true high file count. For an example of converting a FlexVol volume with 500 million files, see the section "Sample FlexVol to FlexGroup volume conversion — 500 million files."

Note: For a video example, see Statistics show-periodic during FlexVol — FlexGroup convert.

Migrating from third-party storage to FlexGroup volumes

When migrating from non-NetApp storage (SAN or NAS), the migration path is a file-based copy. Various methods are available to perform this migration; some are free, and some are paid through third-party vendors.

For NFSv3-only data, NetApp strongly recommends NetApp XCP. XCP is a free, license-based tool that can vastly improve the speed of data migration of high-file-count environments. XCP also offers robust reporting capabilities. XCP 1.5 and later versions also offer NFSv4.x and NFSv4.x ACL support, as well as being officially supported by NetApp.

Note: XCP is supported only for migration to a NetApp storage system.

For CIFS/SMB data, XCP for SMB is available. Robocopy is a free tool, but the speed of transfer depends on using its multithreaded capabilities. Third-party providers can also perform this type of data transfer.

Migrating from NetApp Data ONTAP operating in 7-Mode

Migrate data from Data ONTAP operating in 7-Mode to FlexGroup in one of two ways:

- Full migration of 7-Mode systems to ONTAP systems by using the copy-based or copy-free transition methodology. When using copy-free transition, the process is followed by copy-based migration of data in FlexVol volumes to FlexGroup volumes.
- Copy-based transition from a FlexVol or host-based copy from a LUN by using the previously mentioned tools for migrating from non NetApp storage to FlexGroup.

If you wish to migrate to FlexGroup volumes from FlexVol, you can use FlexVol to FlexGroup volume conversion. In ONTAP 9.8 and later, this works on volumes transitioned from 7-Mode systems.

Migrating from SAN LUNs or Infinite Volume in ONTAP

When migrating from existing ONTAP objects such as SAN-based LUNs or Infinite Volumes, the current migration path is copy-based. The previously mentioned tools for migrating from non-NetApp storage to FlexGroup can also be used for migrating from ONTAP objects.

Deprecation of Infinite Volume

Starting in ONTAP 9.4, you can no longer create infinite volumes with admin privileges. This step prepared for the eventual removal of Infinite Volume support in ONTAP 9.5 and later. Starting in ONTAP 9.5, you can no longer create or modify infinite volumes, and an infinite volume cannot have protocol access.

If infinite volumes are present in an ONTAP cluster and you attempt to upgrade, the ONTAP compatibility checker prevents the upgrade from completing and warns of existing infinite volumes. Be sure to use the NetApp Upgrade Advisor when planning your ONTAP upgrade.

NetApp XCP

The NetApp XCP is free and was designed specifically for scoping, migration, and management of large sets of unstructured NAS data. The initial version was NFSv3 only, but a CIFS version is now available. To use the tool, download it and request a free license (for software tracking purposes only).

XCP addresses the challenges that high-file-count environments have with metadata operation and data migration performance by using a multicore, multichannel I/O streaming engine that can process many requests in parallel.

These requests include the following:

- Data migration
- File or directory listings (a high-performance, flexible alternative to `ls`)
- Space reporting (a high-performance, flexible alternative to `du`)

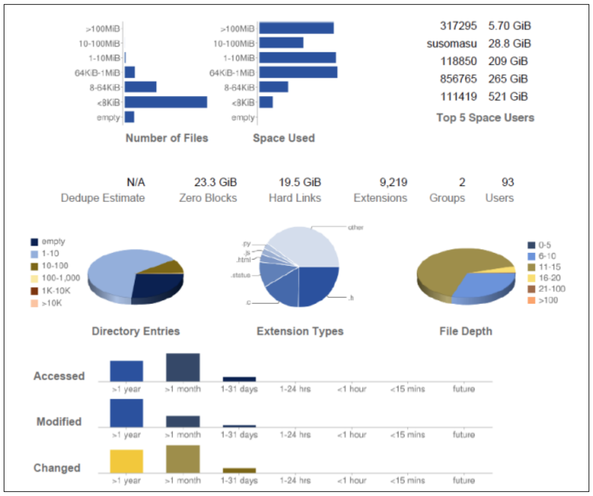
XCP has sometimes reduced the length of data migration by 20 to 30 times for high-file-count environments. In addition, XCP has reduced the file list time for 165 million files from 9 days on a competitor's system to 30 minutes on NetApp technology—a performance improvement of 400 times. As of XCP 1.5, the tool is officially supported by NetApp support.

XCP 1.6 also adds File Systems Analytics functionality. This is similar to the functionality added to ONTAP 9.8, but is able to scan systems that are not running ONTAP as well.

Note: For best results, use the latest XCP release available.

XCP also gives some handy reporting graphs, as shown in Figure 110.

Figure 110) XCP reporting graphs.



For more information, see the official XCP website at http://xcp.netapp.com.

Using XCP to scan files before migration

When deploying a FlexGroup volume, evaluate the file system and structure to help you determine initial sizing considerations and the best way to lay out member volumes. In high-file-count environments, this can be time consuming and tedious. XCP allows you to scan files and export to the CSV or XML format to easily review your file system.

The following example shows a FlexGroup volume with over a million files. Ideally, we don't want to spend much time analyzing these files.

```
Cluster::> vol show -vserver DEMO -fields files,files-used -volume flexgroup_16
vserver volume       files     files-used
------- ------------ --------- ----------
DEMO    flexgroup_16 318766960 1103355
```

To streamline this process, you can use `xcp scan` to get file information. Here's a sample command:

```
C:\> xcp scan -stats \\demo\flexgroup > C:\destination.csv
```

When you do this, the client scans the files and adds information to a comma-separated values (CSV) document. This document shows information such as the following:

- Maximum and average values for size, depth of directory, and dirsize

```
== Maximum Values ==
      Size       Depth     Namelen     Dirsize
    340MiB           9          86         500

== Average Values ==
      Size       Depth     Namelen     Dirsize
   1.61KiB           4           6          11
```

- Top file extensions

```
== Top File Extensions ==
                .docx        .png       .pptx        .pdf        .css       other
   1000038        260         175         128          91          33         219
```

- Number of files, broken down by size ranges

```
== Number of files ==
     empty       <8KiB     8-64KiB 64KiB-1MiB     1-10MiB  10-100MiB     >100MiB
```



```
       8    1000215        156        288        265         10          2
```

- Space used by size range

```
== Space used ==
     empty      <8KiB    8-64KiB 64KiB-1MiB    1-10MiB  10-100MiB    >100MiB
         0    28.7MiB    3.94MiB     124MiB     695MiB     272MiB     453MiB
```

- Directory entries, broken down by file counts

```
== Directory entries ==
     empty       1-10     10-100     100-1K     1K-10K       >10k
         7     100118         30        200
```

- Directory depth ranges

```
== Depth ==
       0-5       6-10      11-15      16-20     21-100       >100
   1100966        333
```

- Modified and created date ranges

```
== Modified ==
   >1 year   >1 month  1-31 days   1-24 hrs    <1 hour   <15 mins     future
       579    1100559         11                   150

== Created ==
   >1 year   >1 month  1-31 days   1-24 hrs    <1 hour   <15 mins     future
              1100210                             1089
```

- A summary of the file structure, including total file count, total directories, symlinks, junctions, and total space used

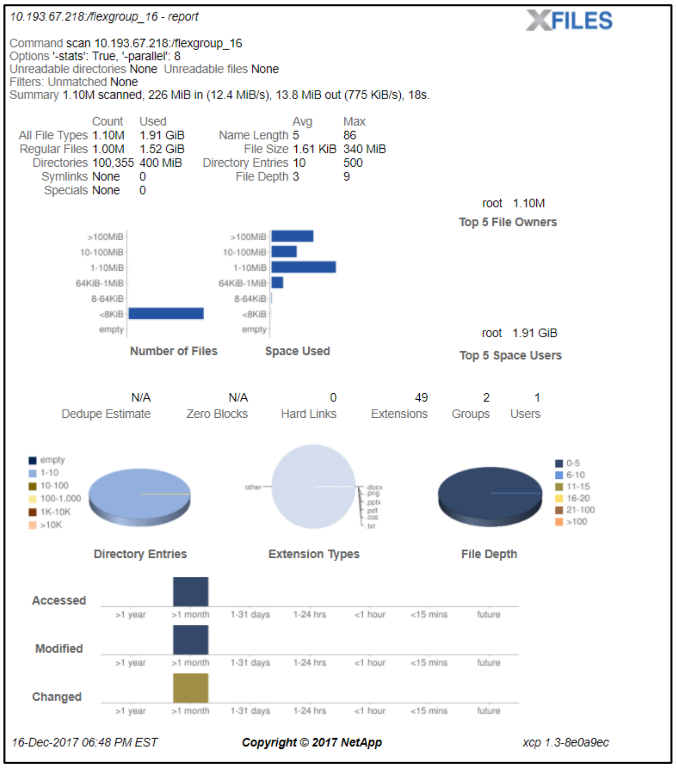
```
Total count: 1101299
Directories: 100355
Regular files: 1000944
Symbolic links:
Junctions:
Special files:
Total space for regular files: 1.54GiB
Total space for directories: 0
Total space used: 1.54GiB
1,101,299 scanned, 0 errors, 26m34s
```

You can also use XCP over NFS to scan CIFS volumes and get more robust reporting and the ability to export to HTML, which presents the data in graphical format.

For example, the following command creates the report shown in Figure 111:

```
xcp scan -stats -html demo:/flexgroup_16 > /flexgroup.html
```


Figure 111) XCP report.



Using XCP to scan file systems provides average file size information, largest file size, capacity, and file count measurements for the top five file owners, and much more. These statistics are available only in the NFS version of XCP, but you can still run NFS scans on datasets that only do SMB traffic by setting up a virtual machine that can use NFS.

Using XCP to run disk usage (du) scans

One common complaint is that, in high-file-count environments, running commands like `du` can take an exceedingly long time. For example, this `du` command ran on a FlexGroup volume with 1,101,002 files and folders and took 21 minutes and 22.600 seconds.

With XCP, this command scanned the same dataset in 22.852 seconds with the same client:

```
[root@centos7 ~]# xcp -duk DEMO:/FGlocal 2>/dev/null | egrep -v '.*?/.*?/'
```

Using XCP to map files to member volumes

You can also use XCP (with help from a Python plugin) to map out file locations to member volumes, list all files in a specific member volume, and even filter files by size to discover larger files.

In most cases, you do not need to worry about where in a cluster a specific file lives, but in some cases, you might need to locate the node or member volume that has a file or be able to figure out what files reside on a specific node, aggregate, or volume in cases where capacity is severely imbalanced and remediation measures are needed.

To use XCP to map files to member volumes, you will need the following:

- A host with XCP installed (latest version is preferred)
- Python3 installed on the client
- Network connectivity between the XCP host and the ONTAP SVM

- NFS server running and the XCP host exported to allow access to the data volume
- The [Python plugin]() copied to the XCP host

Scan all files/folders in a FlexGroup volume and map them all to member volumes

In this example, XCP scans all the files and folders in the FlexGroup volume, and the plugin identifies which member volumes the files live in by using the file handle information it gathers.

We use the following XCP command:

```
xcp diag -run fgid.py scan -fmt '"{} {}".format(x, fgid(x))' ONTAPNFS:/exportname
```

Note: In this example, fgid.py is located in the same directory in which the command is run. If the plugin lives in a different folder, use the absolute path, or set a PATH environment variable to ensure the plugin can be found.

If you want to avoid long console output, you can redirect to a text file with the following command:

```
xcp diag -run fgid.py scan -fmt '"{} {}".format(x, fgid(x))' ONTAPNFS:/exportname > file.txt
```

When we run this against a volume with 2 million files and folders, it completes in 37 seconds.

```
# xcp diag -run fgid.py scan -fmt '"{} {}".format(x, fgid(x))' ONTAPNFS:/FGNFS > FGNFS.txt
Xcp command : xcp diag -run fgid.py scan -fmt "{} {}".format(x, fgid(x)) ONTAPNFS:/FGNFS
Stats : 2.00M scanned
Speed : 351 MiB in (9.49 MiB/s), 1.79 MiB out (49.5 KiB/s)
Total Time : 37s.
STATUS : PASSED
```

The resulting text file is 120MB:

```
-rw-r--r--. 1 root root 120M Apr 27 15:28 FGNFS.txt
```

You can parse it out by using awk, sed, grep or any other tool of your choice.

Alternately, if you know a specific folder or file name that you want to scan to filter the results down, you can use the `-name` flag. In the following example, we search for all directories named “dir_33” This is an exact match query and might be useful in scenarios where you already know the file or folder name in question and want to quickly find the member volume in the FlexGroup volume.

```
# xcp diag -run fgid.py scan -fmt '"{} {}".format(x, fgid(x))' -match "name=='dir_33'"
ONTAPNFS:/FGNFS > dir_33_FGNFS.txt
```

The search finds two different entries — one in member volume #3 and one in member volume #4:

```
# cat dir_33_FGNFS.txt
x.x.x.x:/FGNFS/files/client1/dir_33 3
x.x.x.x:/FGNFS/files/client2/dir_33 4
```

If you want to do a pattern matching search (for example, you want all files with “moarfiles3” in the name), you can leverage regex and/or wildcards. The XCP documentation has some examples of this, but you can run that with the plugin as follows.

```
# xcp diag -run fgid.py scan -fmt '"{} {}".format(x, fgid(x))' -match "fnm('moarfiles3*')"
ONTAPNFS:/FGNFS > moarfiles3_FGNFS.txt
```

The search finds 440,000 files with the pattern “moarfiles3” in the name and takes just 27 seconds.

```
Filtered: 444400 matched, 1556004 did not match

Xcp command : xcp diag -run fgid.py scan -fmt "{} {}".format(x, fgid(x)) -match
fnm('moarfiles3*') 10.10.10.10:/FGNFS
Stats : 2.00M scanned, 444,400 matched
Speed : 351 MiB in (12.6 MiB/s), 1.79 MiB out (65.7 KiB/s)
```



```
Total Time : 27s.
```

Following is a sample output of some of the files. In this case, the files are in member volume #3:

```
x.x.x.x:/FGNFS/files/client1/dir_45/moarfiles3158.txt 3
x.x.x.x:/FGNFS/files/client1/dir_45/moarfiles3159.txt 3
```

Another way to filter files is to look only for files of a certain size or greater. This can help identify larger files in a volume while filtering out smaller files. In this case, we use `–match` with regex, wildcards, and so on.

In this example, we search for all files in the TechONTAP podcast volume that are over 100MB.

```
# xcp diag -run fgid.py scan -fmt '"{} {}".format(x, fgid(x))' -match "fnm('.wav') and size >
100*M" ONTAPNFS:/techontap > TechONTAP_ep.txt
```

The following is a sample of the files it finds:

```
ONTAPNFS:/techontap/Episodes/Episode 20x - Genomics Architecture/ep20x-genomics-meat.wav 4
ONTAPNFS:/techontap/archive/combine.band/Media/Audio Files/ep104-webex.output.wav 5
ONTAPNFS:/techontap/archive/combine.band/Media/Audio Files/ep104-mics.output.wav 3
ONTAPNFS:/techontap/archive/Episode 181 - Networking Deep Dive/ep181-networking-deep-dive-
meat.output.wav 6
ONTAPNFS:/techontap/archive/Episode 181 - Networking Deep Dive/ep181-networking-deep-dive-
meat.wav 2

Filtered: 16 matched, 7687 did not match

xcp command : xcp diag -run fgid.py scan -fmt "{} {}".format(x, fgid(x)) -match fnm('.wav') and
size > 100M ONTAPNFS:/techontap
Stats : 7,703 scanned, 16 matched
Speed : 1.81 MiB in (1.44 MiB/s), 129 KiB out (102 KiB/s)
Total Time : 1s.
STATUS : PASSED
```

Use a FlexGroup member volume number and find all files in that member volume

In some cases, you might want to find all files located in a specific member volume. The most common use case is if one member volume is much larger than another, or if that member volume is approaching the available capacity and you want to free up some space without growing the volume.

XCP with the plugin can do that for you as well.

The numbers seen with the XCP output using the plugin are index numbers that align with the member volume numbers. If you see a 6 in the index (or idx) column, then that is FGVol___0006. So, if you know member volume 6 is overused, you can run the following XCP command to find all files in that member volume using `–match` with fgid.

```
# xcp diag -run fgid.py scan -match 'fgid(x)==6' -parallel 10 -l ONTAPNFS:/FGNFS > member6.txt
```

This dumps the files in the member volume, but also the file sizes:

```
rw-r--r-- --- root root 4KiB 4KiB 18h22m FGNFS/files/client2/dir_143/moarfiles1232.txt
rw-r--r-- --- root root 4KiB 4KiB 18h22m FGNFS/files/client2/dir_143/moarfiles1233.txt
rw-r--r-- --- root root 4KiB 4KiB 18h22m FGNFS/files/client2/dir_143/moarfiles1234.txt
```

You can still filter the files by size, too. In this case, we filter for files that are 500MB or larger in size, but only in member volume #6:

```
# xcp diag -run fgid.py scan -match 'fgid(x)==6 and size > 500*M' -parallel 10 -l
ONTAPNFS:/techontap

rw-r--r-- --- 501 games 596MiB 598MiB 3y219d techontap/Episodes/Episode 1/Epidose 1
GBF.band/Media/Audio Files/Tech ONTAP Podcast - Episode 1 - AFF with Dan Isaacs v3_1.aif
```



```
rw-r--r-- --- 501 games 885MiB 888MiB 3y219d techontap/archive/Prod - old MacBook/Insight
2016_Day2_TechOnTap_JParisi_ASullivan_GDekhayser.mp4
rw-r--r-- --- 501 games 787MiB 790MiB 1y220d techontap/archive/Episode 181 - Networking Deep
Dive/ep181-networking-deep-dive-meat.output.wav

Filtered: 3 matched, 7700 did not match

Xcp command : xcp diag -run fgid.py scan -match fgid(x)==6 and size > 500*M -parallel 10 -l
10.10.10.11:/techontap
Stats : 7,703 scanned, 3 matched
Speed : 1.81 MiB in (1.53 MiB/s), 129 KiB out (109 KiB/s)
Total Time : 1s.
STATUS : PASSED
```

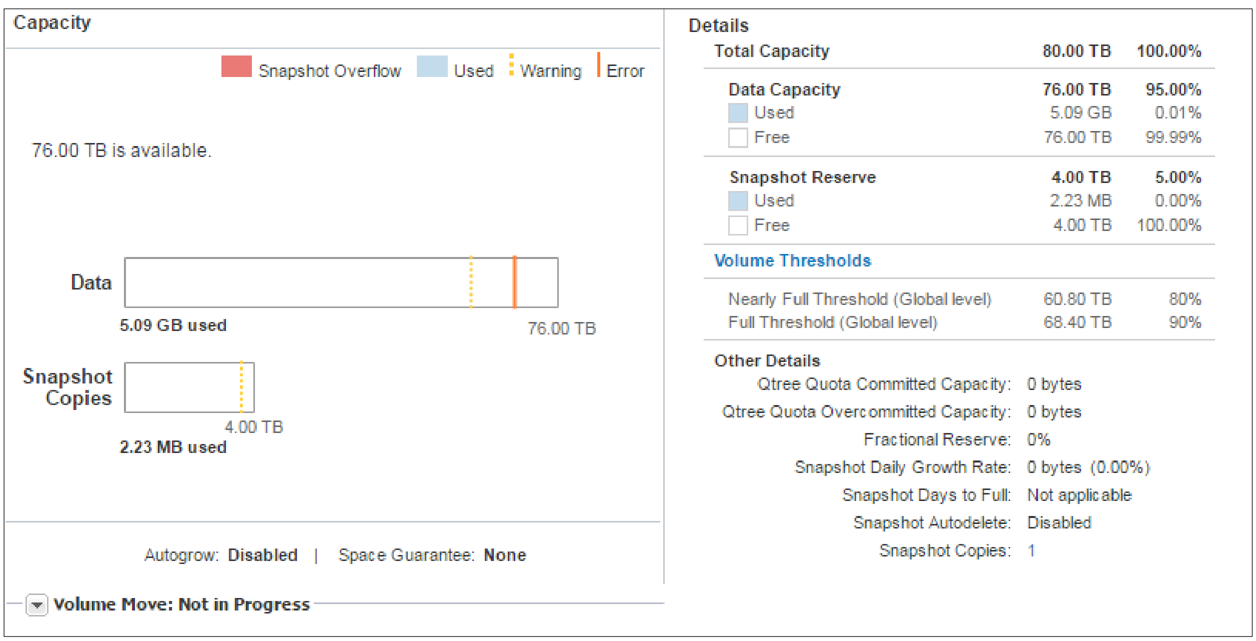
After you have this information, you can accurately move or delete files to free up space in a member volume in a FlexGroup volume.

Examples

Thin provisioning example

The following image shows that the FlexGroup volume has a total capacity of 80TB and 5GB used. Also, 4TB have been reserved for Snapshot copies (5%). The available space is 76TB.

Figure 112) FlexGroup capacity breakdown—Active IQ Unified Manager.



However, in the following CLI output, a few anomalies stand out:

- In Active IQ Unified Manager, the FlexGroup volume shows as having 76TB available, but, in the CLI, only 11.64TB is available.
- The FlexGroup volume shows as having 11.64TB available, but the member FlexVol volumes all show roughly 5.8TB available.
- The percentage used for the FlexGroup volume shows as 85%, even though we have used only 5GB, which is a negligible amount of space compared with 80TB (5GB of 81920GB is less than 1%).

- The FlexGroup volume shows as 85% used, but the member FlexVol volumes all show as 41% used, despite each having a different amount of space per FlexVol member.

Example:

```
cluster::> volume show -is-constituent true -fields size,used,percent-used,available -vserver SVM
-volume flexgroup* -sort-by volume
vserver volume             size  available used    percent-used
------- ------------------ ----- --------- ------- ------------
SVM     flexgroup          80TB  11.64TB   5.08GB  85%
SVM     flexgroup__0001    10TB  5.81TB    147.5MB 41%
SVM     flexgroup__0002    10TB  5.83TB    145.2MB 41%
SVM     flexgroup__0003    10TB  5.81TB    144.9MB 41%
SVM     flexgroup__0004    10TB  5.83TB    148.0MB 41%
SVM     flexgroup__0005    10TB  5.81TB    4.08GB  41%
SVM     flexgroup__0006    10TB  5.83TB    147.6MB 41%
SVM     flexgroup__0007    10TB  5.81TB    145.3MB 41%
SVM     flexgroup__0008    10TB  5.83TB    146.5MB 41%
9 entries were displayed.
```

The anomalies are due to ONTAP calculating against the aggregate's available space. The FlexVol member volumes show equivalent available values depending on the aggregates where they are located.

```
cluster::> volume show -is-constituent true -fields available,aggregate -vserver SVM -volume
flexgroup* -sort-by aggregate
vserver volume             aggregate   available
------- ------------------ ----------- ---------
SVM     flexgroup__0001    aggr1_node1 5.81TB
SVM     flexgroup__0003    aggr1_node1 5.81TB
SVM     flexgroup__0005    aggr1_node1 5.81TB
SVM     flexgroup__0007    aggr1_node1 5.81TB
SVM     flexgroup__0002    aggr1_node2 5.83TB
SVM     flexgroup__0004    aggr1_node2 5.83TB
SVM     flexgroup__0006    aggr1_node2 5.83TB
SVM     flexgroup__0008    aggr1_node2 5.83TB

cluster::> storage aggregate show -aggregate aggr1* -fields availsize
aggregate   availsize
----------- ---------
aggr1_node1 5.81TB
aggr1_node2 5.83TB
2 entries were displayed.
```

Using thin provisioning means that you must consider the aggregate capacity and the volume footprint when monitoring space.

Volume autosize example

In the following example, we'll attempt to show how volume autosize works when a member volume reaches a capacity threshold.

In this case, the member volumes are all 1GB in size (not recommended). This was done to make filling the volume with a single file easier.

```
cluster::*> vol show -vserver DEMO -volume fgautogrow* -sort-by used -fields available
vserver volume           size
------- ---------------- ----
DEMO    fgautogrow__0001 1GB
DEMO    fgautogrow__0002 1GB
DEMO    fgautogrow__0003 1GB
DEMO    fgautogrow__0004 1GB
DEMO    fgautogrow__0005 1GB
DEMO    fgautogrow__0006 1GB
DEMO    fgautogrow__0007 1GB
DEMO    fgautogrow__0008 1GB
```


Note: 1GB is not a recommended size for member volumes; the minimum member volume size should be no less than 100GB. ONTAP programmatically prevents creation of FlexGroup volumes that have member volumes smaller than 100GB with REST APIs and in ONTAP System Manager and warns you in the CLI.

With `volume autosize`, the write succeeds because the member volume in which the write lands grows to the appropriate size to honor the write. In this case, the file was written to member volume `fgautogrow__0003`.

```
cluster::*> vol show -vserver DEMO -volume fgautogrow* -sort-by used -fields available,size
vserver volume           size   available used
------- ---------------- ------ --------- -------
DEMO    fgautogrow__0004 1GB    915.6MB   57.23MB
DEMO    fgautogrow__0005 1GB    915.6MB   57.23MB
DEMO    fgautogrow__0006 1GB    915.6MB   57.23MB
DEMO    fgautogrow__0007 1GB    915.6MB   57.23MB
DEMO    fgautogrow__0008 1GB    915.6MB   57.23MB
DEMO    fgautogrow__0002 1GB    915.5MB   57.26MB
DEMO    fgautogrow__0001 1GB    915.5MB   57.27MB
DEMO    fgautogrow__0003 1.60GB 498.7MB   1.03GB
```

When this happens, an event is triggered in the event management system and can be seen with `event log show`.

```
INFORMATIONAL wafl.vol.autoSize.done: Volume Autosize: Automatic grow of volume
'fgautogrow__0003@vserver:7e3cc08e-d9b3-11e6-85e2-00a0986b1210' by 611MB complete.
```

This event can also be monitored with SNMP, by sending alerts through event destinations, or with Active IQ Unified Manager.

```
cluster::> event route show -message-name wafl.vol.autoSize.done -instance

                                     Message Name: wafl.vol.autoSize.done
                                         Severity: INFORMATIONAL
                                Corrective Action: (NONE)
                                      Description: This message occurs on successful autosize of volume.
                               Supports SNMP trap: true
                                     Destinations: -
            Number of Drops Between Transmissions: 0
Dropping Interval (Seconds) Between Transmissions: 0
```

Snapshot spill example

For example, if snap reserve is set to 5% on a 400GB FlexGroup volume, then that is a total of 20GB of snapshot reserve. If there are four member volumes, then the snapshot reserve per member volume is ~5GB.

```
cluster::*> vol show -vserver DEMO -volume FG_SM_400G* -fields size,used,size-used-by-
snapshots,snapshot-reserve-available
vserver volume     size    used   size-used-by-snapshots snapshot-reserve-available
------- ---------- ------- ------ ---------------------- --------------------------
DEMO    FG_SM_400G 420.9GB 2.01GB 3.16MB                 21.04GB
DEMO    FG_SM_400G__0001
                   105.2GB 513.7MB
                                  860KB                  5.26GB
DEMO    FG_SM_400G__0002
                   105.2GB 513.8MB
                                  432KB                  5.26GB
DEMO    FG_SM_400G__0003
                   105.2GB 513.8MB
                                  828KB                  5.26GB
DEMO    FG_SM_400G__0004
                   105.2GB 513.7MB
                                  1.09MB                 5.26GB
```


If I write a 4GB file to that volume, nothing gets used in the snapshot. That's because no blocks have been overwritten:

```
cluster::*> vol show -vserver DEMO -volume FG_SM_400G* -fields size,used,size-used-by-
snapshots,snapshot-reserve-available
vserver volume     size    used   size-used-by-snapshots snapshot-reserve-available
------- ---------- ------- ------ ---------------------- --------------------------
DEMO    FG_SM_400G 420.9GB 6.70GB 3.17MB                 21.04GB
DEMO    FG_SM_400G__0001
                   105.2GB 5.19GB 868KB                  5.26GB
DEMO    FG_SM_400G__0002
                   105.2GB 513.9MB
                                  432KB                  5.26GB
DEMO    FG_SM_400G__0003
                   105.2GB 513.9MB
                                  828KB                  5.26GB
DEMO    FG_SM_400G__0004
                   105.2GB 513.8MB
                                  1.09MB                 5.26GB
```

If I take a snapshot now, the existing blocks are locked into place in case an overwrite occurs later. But again, no space is used by the snapshot yet.

```
cluster::*> snapshot create -vserver DEMO -volume FG_SM_400G -snapshot file1

cluster::*> vol show -vserver DEMO -volume FG_SM_400G* -fields size,used,size-used-by-
snapshots,snapshot-reserve-available
vserver volume     size    used   size-used-by-snapshots snapshot-reserve-available
------- ---------- ------- ------ ---------------------- --------------------------
DEMO    FG_SM_400G 420.9GB 6.69GB 3.78MB                 21.04GB
DEMO    FG_SM_400G__0001
                   105.2GB 5.18GB 1.02MB                 5.26GB
DEMO    FG_SM_400G__0002
                   105.2GB 514.0MB
                                  580KB                  5.26GB
DEMO    FG_SM_400G__0003
                   105.2GB 514.0MB
                                  976KB                  5.26GB
DEMO    FG_SM_400G__0004
                   105.2GB 513.9MB
                                  1.25MB                 5.26GB
```

Space is used if we overwrite data. This can happen on a delete or if I copy over the file, or I were simply to change the file data. But notice that the space change only happens on the member volume where the file lives. This is reflected by the FlexGroup itself, but it doesn't incorporate the actual snapshot used space, because it's below the reserved capacity. In the example below, deleting the file uses a full 4.4GB of space in the snapshot and reduces the space used in the volume by the same amount:

```
cluster::*> vol show -vserver DEMO -volume FG_SM_400G* -fields size,used,size-used-by-
snapshots,snapshot-reserve-available
vserver volume     size    used    size-used-by-snapshots snapshot-reserve-available
------- ---------- ------- ------- ---------------------- --------------------------
DEMO    FG_SM_400G 420.9GB 772.7MB 4.41GB                 16.63GB
DEMO    FG_SM_400G__0001
                   105.2GB 596.7MB 4.41GB                 874.4MB
DEMO    FG_SM_400G__0002
                   105.2GB 58.65MB 624KB                  5.26GB
DEMO    FG_SM_400G__0003
                   105.2GB 58.69MB 1012KB                 5.26GB
DEMO    FG_SM_400G__0004
                   105.2GB 58.71MB 1.29MB                 5.26GB
```

If we start to overrun the snapshot reserve available, used snapshot space will start to spill into the AFS. For example, if I reduce the snap reserve to 1%, we can see that happen with our existing volume.

```
cluster::*> vol show -vserver DEMO -volume FG_SM_400G* -fields size,used,size-used-by-
snapshots,snapshot-reserve-available
```



```
vserver volume     size    used   size-used-by-snapshots snapshot-reserve-available
------- ---------- ------- ------ ---------------------- --------------------------
DEMO    FG_SM_400G 420.9GB 3.86GB 4.64GB                 3.15GB
DEMO    FG_SM_400G__0001
                   105.2GB 3.69GB 4.64GB                 0B
DEMO    FG_SM_400G__0002
                   105.2GB 58.72MB
                                  624KB                  1.05GB
DEMO    FG_SM_400G__0003
                   105.2GB 58.84MB
                                  1012KB                 1.05GB
DEMO    FG_SM_400G__0004
                   105.2GB 58.71MB
                                  1.29MB                 1.05GB
```

Now we see nearly all of the snapshot space reflected in used and we have 0 snapshot reserve available. This also affects the snap reserve used percentage we see in the df output:

```
cluster::*> df -g FG_SM_400G
Filesystem                total        used      avail capacity  Mounted on                 Vserver
/vol/FG_SM_400G/          416GB         3GB      412GB       0%  /FG_SM_400G                DEMO
/vol/FG_SM_400G/.snapshot   4GB         4GB        0GB     110%  /FG_SM_400G/.snapshot      DEMO
```

Now we are using 110% of our snapshot reserve. That space has to go somewhere, so it goes into the AFS and now we're using 3GB, when we were only using ~597MB before the snapshot reserve was adjusted. This is known as snapshot spill, and it can negatively affect capacity reporting and potentially cause a FlexGroup volume to run out of space, even if we are currently reporting free space available.

You can see the snapshot spill amount with the `volume show-space` command:

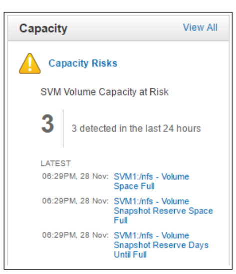
```
cluster::> volume show-space -vserver DEMO -volume FG_SM_400G* -fields snapshot-spill
vserver volume           snapshot-spill
------- ---------------- --------------
DEMO    FG_SM_400G__0001 3.58GB
DEMO    FG_SM_400G__0002 -
DEMO    FG_SM_400G__0003 -
DEMO    FG_SM_400G__0004 -
```

Capacity monitoring and alerting examples in Active IQ Unified Manager

Active IQ Unified Manager provides methods to monitor and alert on various storage system functionalities, including used and free capacities.

On the main Health page, Active IQ Unified Manager displays active warnings and errors about capacity.

Figure 113) Capacity view in Active IQ Unified Manager

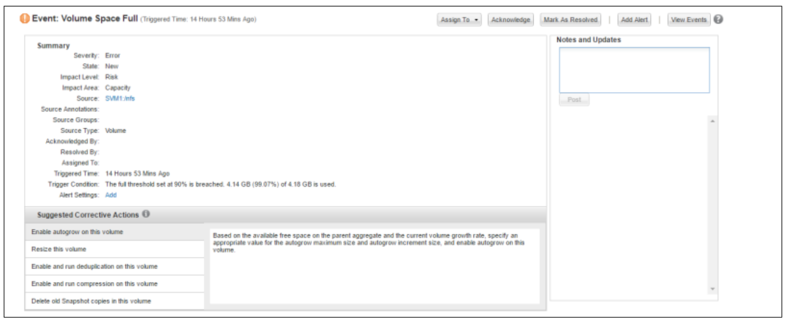


Also, Active IQ Unified Manager has a more detailed view of capacity-related events.

Figure 114) Capacity related events in Active IQ Unified Manager

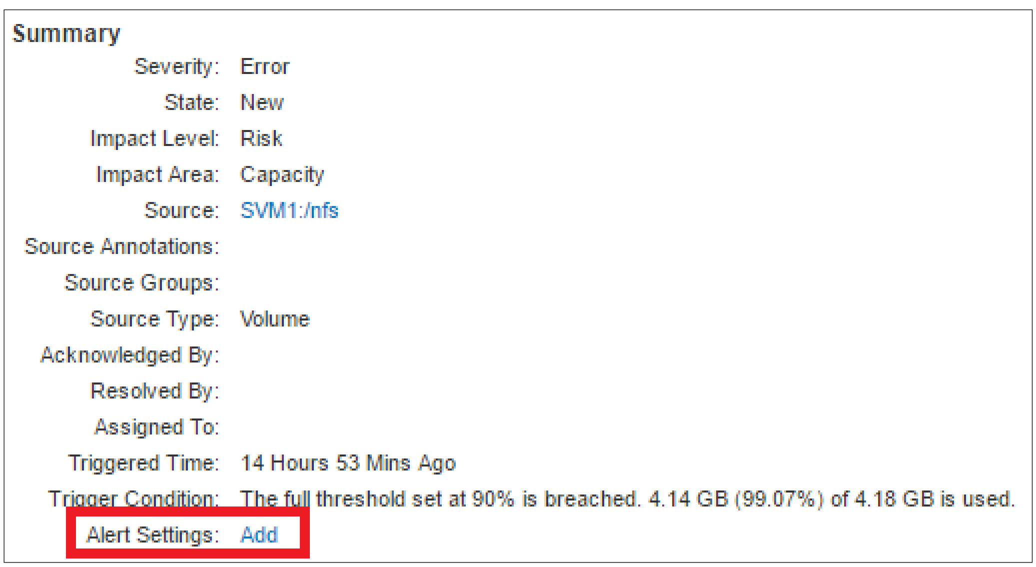
When you click one of the events, a full report of the issue is shown.

Figure 115) Event details, capacity — Active IQ Unified Manager

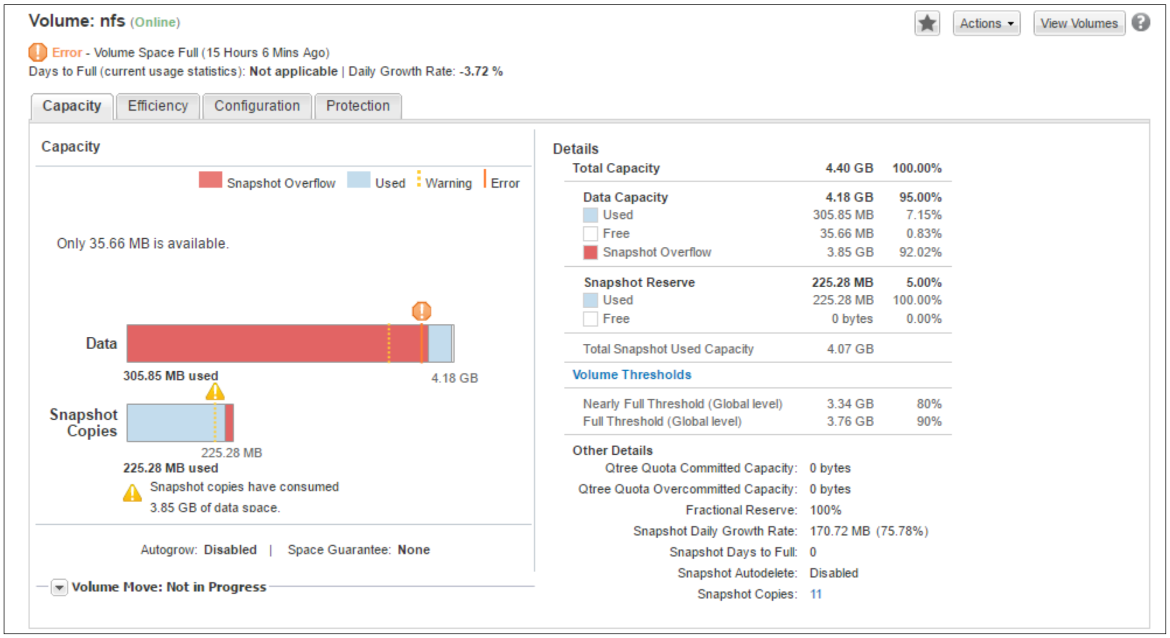


In this detailed view, you can also configure alerts.

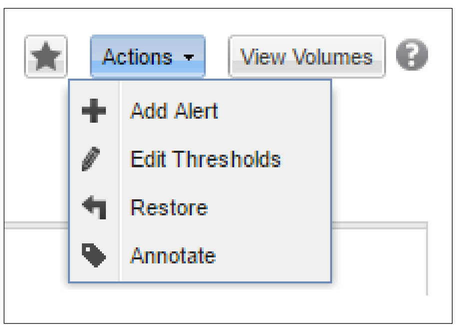
1. To configure alerts, beside Alert Settings, click Add.



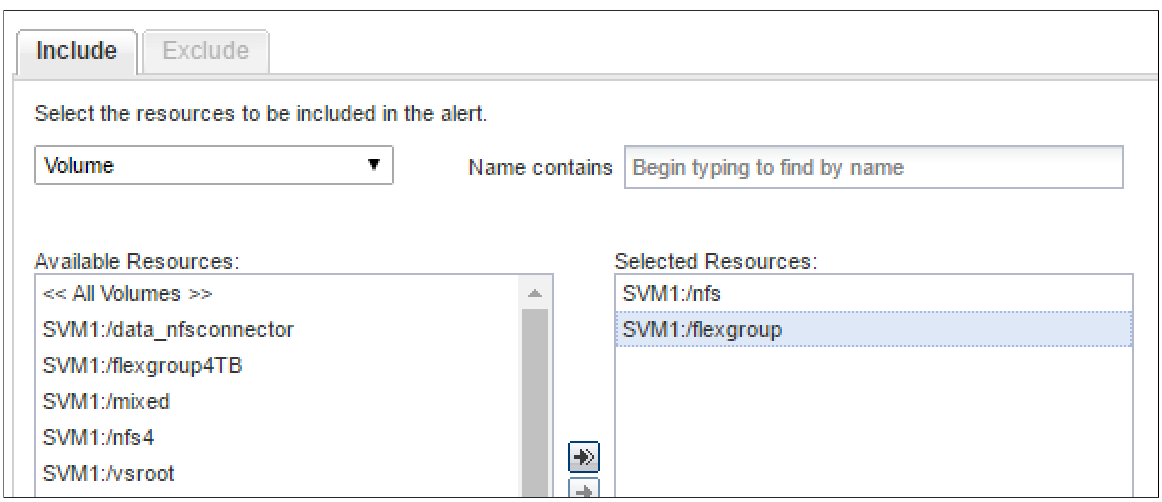
2. You can also view volume capacities from the Volume screen. To do so, click Storage → Volumes and select a volume:



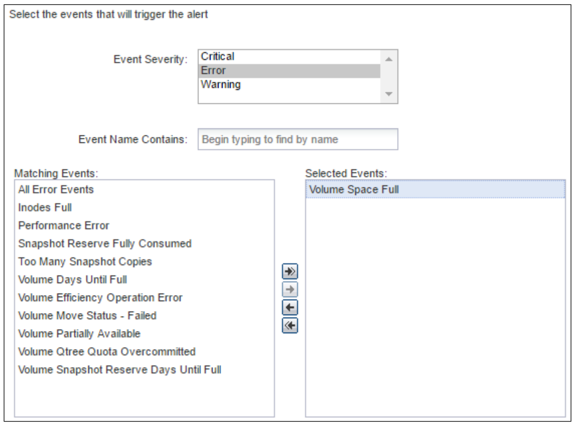
3. To create alerts that are specific to the volume, click Actions → Add Alert.

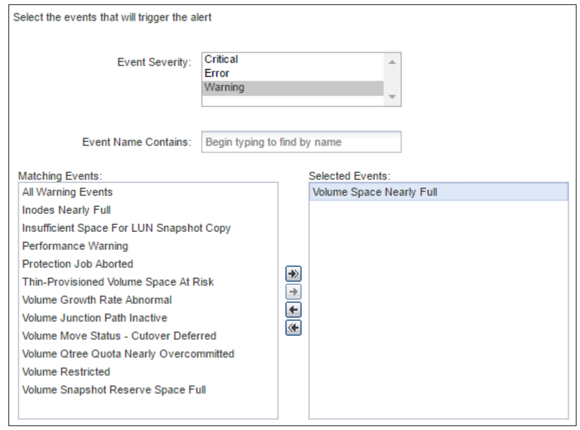


4. With the alert, add one or many volumes to various events (or exclude them).



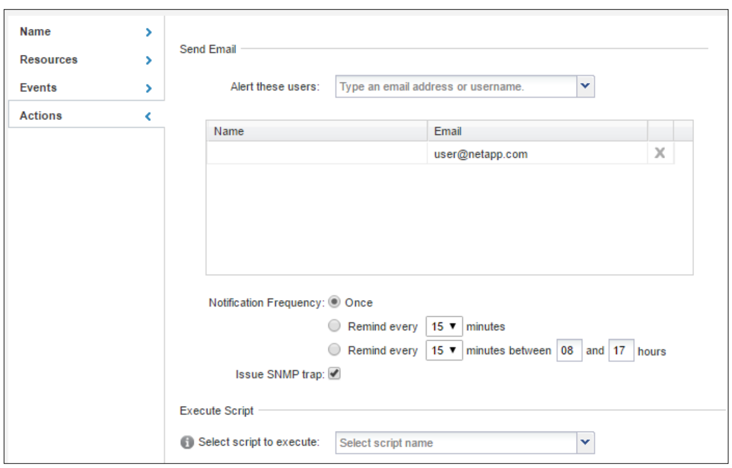
Events are organized by severity and include the Critical, Error, and Warning levels. Volume Space Full is included under the Error level, and Volume Space Nearly Full is under the Warning level.





When an event is triggered, the alert mechanism in Active IQ Unified Manager can do the following:

- Send an email to a user, a list of users, and a distribution list
- Trigger an SNMP trap
- Send reminders
- Execute scripts (such as an automated volume grow script)

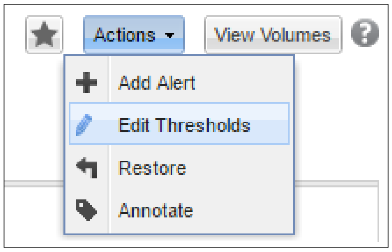


Editing volume thresholds in Active IQ Unified Manager

Thresholds for Volume Nearly Full and Volume Full control when an event management system event is triggered by the cluster. This control helps storage administrators stay on top of the volume capacities to prevent volumes from running out of space. In FlexGroup, this approach also involves remote allocation of files and folders, because ingest remoteness increases as a volume gets closer to full. As mentioned earlier, the Volume Nearly Full and Volume Full thresholds should be modified for a FlexGroup volume so that storage administrators are notified about potential capacity issues earlier than the defaults provide. For more information, see the section “Best practice 25: Volume space threshold recommendations for FlexGroup volumes.”

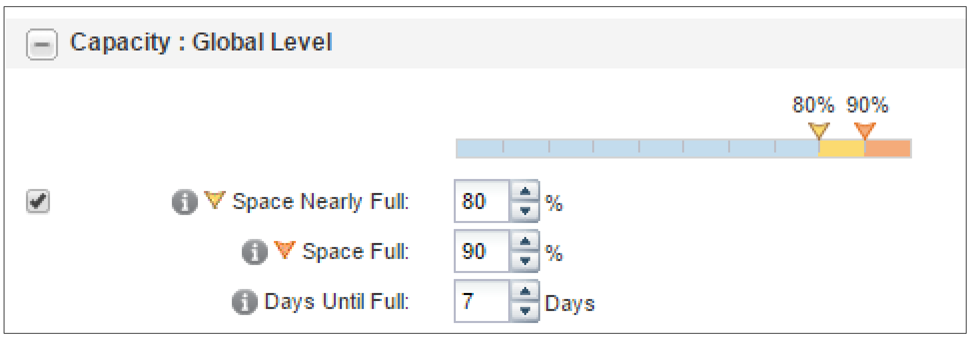
The command line provides a method to modify the thresholds, as does ONTAP System Manager. Under the Actions button of the volume detail, select Edit Thresholds to modify the volume threshold on a per-volume basis. With a FlexGroup volume, the setting is applied to the whole FlexGroup volume, and thresholds are set on each member volume individually.

Figure 116) Editing volume thresholds.



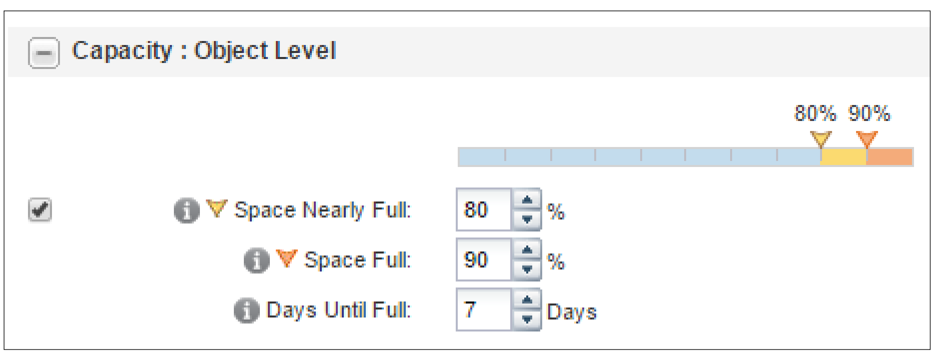
When you initially select the checkbox under Capacity: Global Level, the defaults display as shown following. These defaults are unaffiliated with the ONTAP event management system volume thresholds. Rather, they are specific to Active IQ Unified Manager.

Figure 117) Capacity: Global Level default settings.



Changing the values modifies the threshold to be an Object Level.

Figure 118) Capacity: Object Level.



On the cluster, the volume-level threshold options are unchanged.

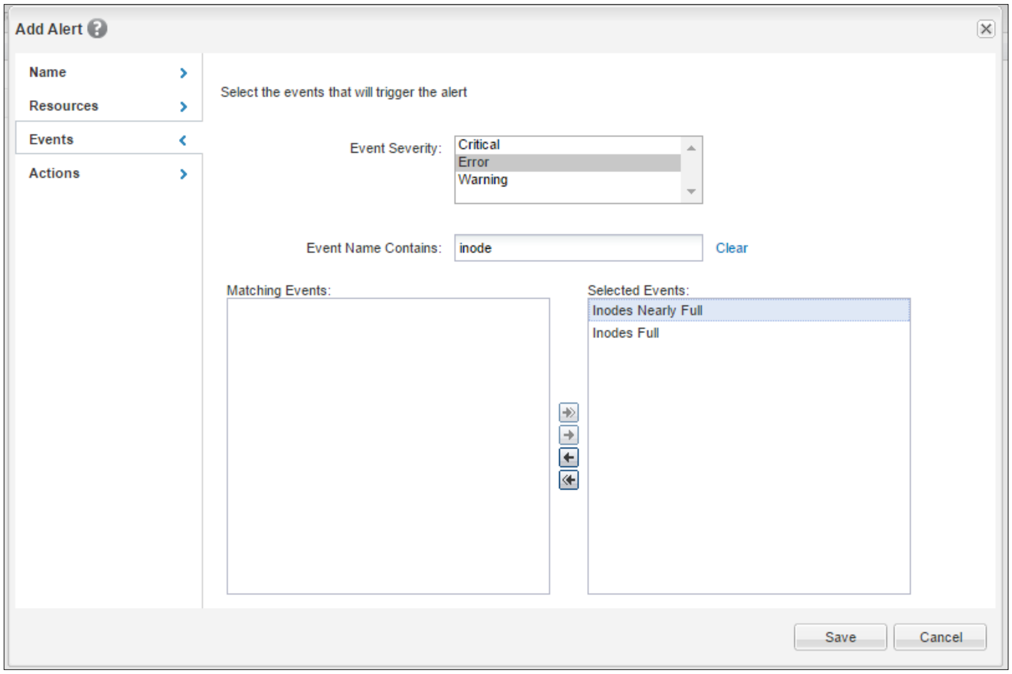
```
cluster::*> vol show -fields space-nearly-full-threshold-percent,space-full-threshold-percent -
sort-by space-nearly-full-threshold-percent -volume flexgroup
vserver volume    space-nearly-full-threshold-percent space-full-threshold-percent
------- --------- ----------------------------------- ----------------------------
SVM     flexgroup 95%                                 98%
```

You can use Active IQ Unified Manager alerting along with the cluster's event management system alerting and event destination logic, or independently of this logic.

Inode monitoring

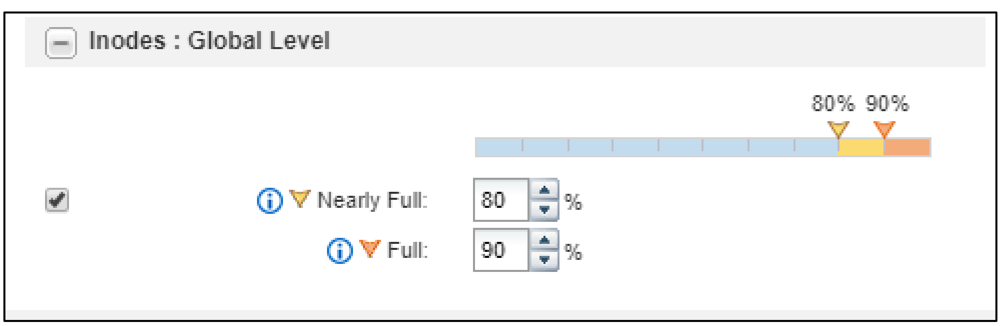
Active IQ Unified Manager also enables you to alert on inode count in FlexGroup volumes with the Inodes Nearly Full (Warning) and Inodes Full (Error) events. Alerts for inodes are configured similarly to the alerts for capacity.

Figure 119) Add Alert: Inode monitoring



You can also edit inode thresholds from the Edit Thresholds window for more granular control over alerting.

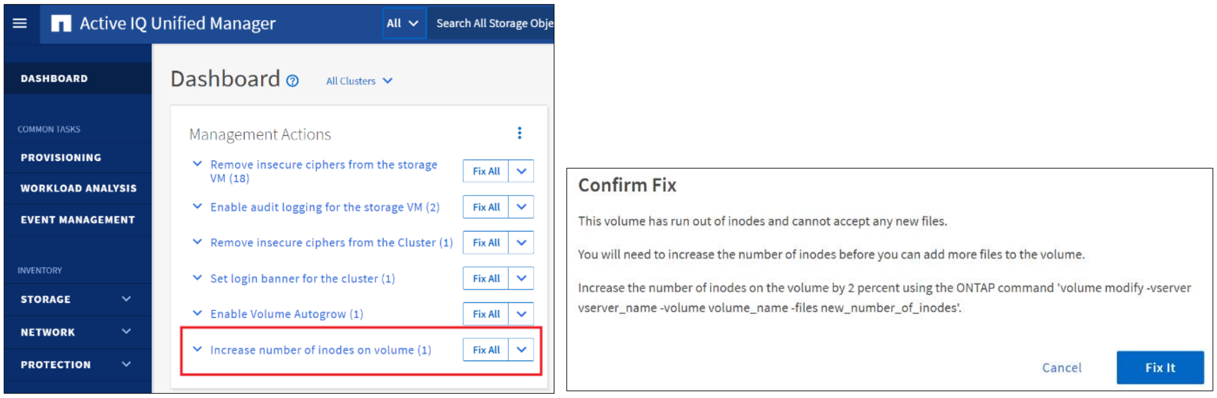
Figure 120) Edit Thresholds: Inodes.



Active IQ "Fix It"

Active IQ Unified Manager 9.8 and later introduces "fix it" functionality that allows storage admins to use a single click to resolve a set list of issues. One of those issues is when a volume reaches an inode threshold. These management actions are present on the dashboard when you log into Active IQ Unified Manager. In the example below, we have a volume that has exceeded the threshold for used inodes.

Figure 121) Active IQ Unified Manager — fix out of inodes.



When you click Fix, Active IQ Unified Manager increases the total files by 2%.

Sample FlexVol to FlexGroup volume conversion

In this sample conversion, before we convert a volume, we add around 300,000 files to help determine how long the process might take with many files present.

```
cluster::*> df -i lotsafiles
Filesystem iused ifree %iused Mounted on Vserver
/vol/lotsafiles/ 330197 20920929 1% /lotsafiles DEMO

cluster::*> volume show lots*
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
DEMO      lotsafiles   aggr1_node1  online     RW         10TB     7.33TB    0%
```

First, try out the validation.

```
cluster::*> volume conversion start -vserver DEMO -volume lotsafiles -foreground true -check-only
true
Error: command failed: Cannot convert volume "lotsafiles" in Vserver "DEMO" to a FlexGroup.
Correct the following issues and retry the command:
* SMB1 is enabled on Vserver "DEMO". Use the 'vserver cifs options modify -smb1-enabled false -
vserver DEMO' command to disable SMB1.
* The volume contains LUNs. Use the "lun delete -vserver DEMO -volume lotsafiles -lun *" command
to remove the LUNs, or use the "lun move start" command to relocate the LUNs to other
FlexVols.
* NFSv3 MS-DOS client support is enabled on Vserver "DEMO". Use the "vserver nfs modify -vserver
DEMO -v3-ms-dos-client disabled" command to disable NFSv3 MS-DOS client support on the
Vserver. Note that disabling this support will disable access for all NFSv3 MS-DOS clients
connected to Vserver "DEMO".
```

As you can see, there are some blockers, such as SMB1 and the LUN we created (to intentionally break conversion). So, we clear them with the recommendations and run the validation again. We see some caveats:

```
cluster::*> volume conversion start -vserver DEMO -volume lotsafiles -foreground true -check-only
true
Conversion of volume "lotsafiles" in Vserver "DEMO" to a FlexGroup can proceed with the following
warnings:
* After the volume is converted to a FlexGroup, it will not be possible to change it back to a
flexible volume.
* Converting flexible volume "lotsafiles" in Vserver "DEMO" to a FlexGroup will cause the state
of all Snapshot copies from the volume to be set to "pre-conversion". Pre-conversion Snapshot
copies cannot be restored.
```


Now, let us convert. First, we start a script that takes a while to complete while also using NetApp Active IQ Performance Manager to monitor performance during the conversion.

The conversion of the volume takes less than 1 minute, and the only disruption is a slight drop in IOPS.

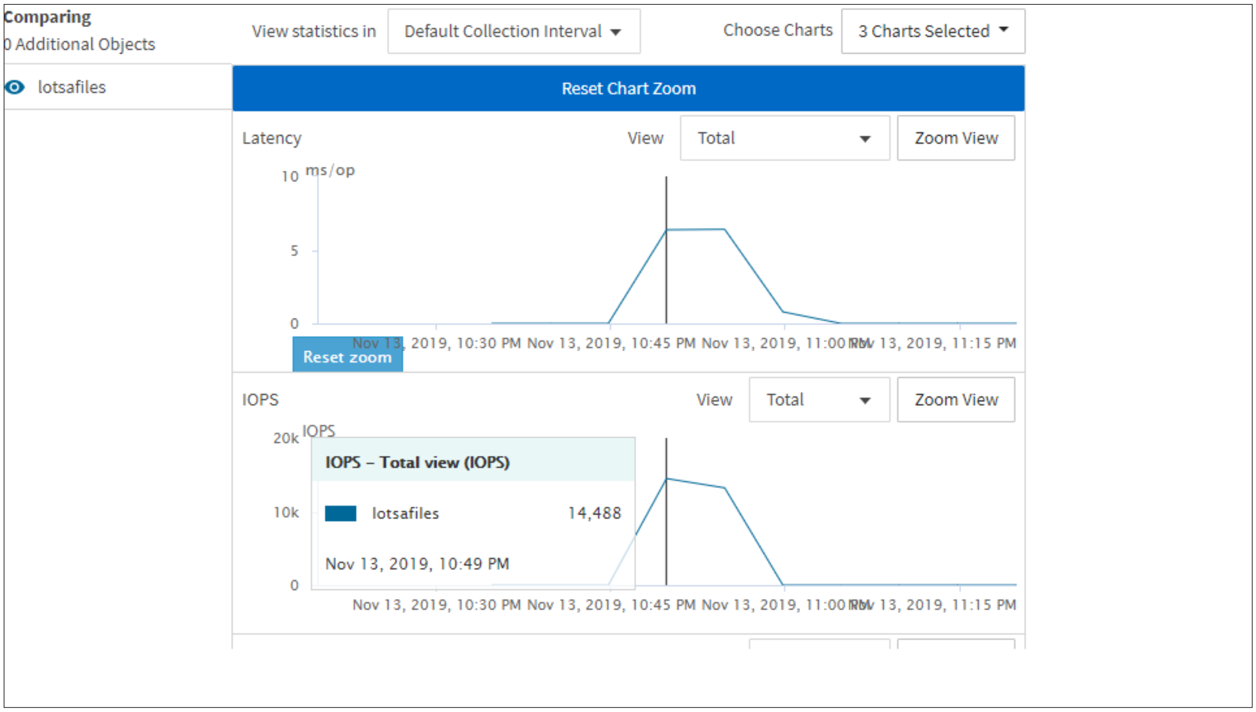
```
cluster::*> volume conversion start -vserver DEMO -volume lotsafiles -foreground true

Warning: After the volume is converted to a FlexGroup, it will not be possible to change it back
to a flexible volume.
Do you want to continue? {y|n}: y
Warning: Converting flexible volume "lotsafiles" in Vserver "DEMO" to a FlexGroup will cause the
state of all Snapshot copies from the volume to be set to "pre-conversion". Pre-conversion
Snapshot copies cannot be restored.
Do you want to continue? {y|n}: y
[Job 23671] Job succeeded: success
cluster::*> statistics show-periodic
cpu cpu total fcache total total data data data cluster cluster cluster disk disk pkts pkts
avg busy ops nfs-ops cifs-ops ops spin-ops recv sent busy recv sent busy recv sent read write
recv sent
---- ---- -------- -------- -------- -------- -------- -------- -------- ---- -------- -------- -
------ -------- -------- -------- -------- -------- --------
34% 44% 14978 14968 10 0 14978 14.7MB 15.4MB 0% 3.21MB 3.84MB 0% 11.5MB 11.6MB 4.43MB 1.50MB
49208 55026
40% 45% 14929 14929 0 0 14929 15.2MB 15.7MB 0% 3.21MB 3.84MB 0% 12.0MB 11.9MB 3.93MB 641KB 49983
55712
36% 44% 15020 15020 0 0 15019 14.8MB 15.4MB 0% 3.24MB 3.87MB 0% 11.5MB 11.5MB 3.91MB 23.9KB 49838
55806
30% 39% 15704 15694 10 0 15704 15.0MB 15.7MB 0% 3.29MB 3.95MB 0% 11.8MB 11.8MB 2.12MB 4.99MB
50936 57112
32% 43% 14352 14352 0 0 14352 14.7MB 15.3MB 0% 3.33MB 3.97MB 0% 11.3MB 11.3MB 4.19MB 27.3MB 49736
55707
37% 44% 14807 14797 10 0 14807 14.5MB 15.0MB 0% 3.09MB 3.68MB 0% 11.4MB 11.4MB 4.34MB 2.79MB
48352 53616
39% 43% 15075 15075 0 0 15076 14.9MB 15.6MB 0% 3.24MB 3.86MB 0% 11.7MB 11.7MB 3.48MB 696KB 50124
55971
32% 42% 14998 14998 0 0 14997 15.1MB 15.8MB 0% 3.23MB 3.87MB 0% 11.9MB 11.9MB 3.68MB 815KB 49606
55692
38% 43% 15038 15025 13 0 15036 14.7MB 15.2MB 0% 3.27MB 3.92MB 0% 11.4MB 11.3MB 3.46MB 15.8KB
50256 56150
43% 44% 15132 15132 0 0 15133 15.0MB 15.7MB 0% 3.22MB 3.87MB 0% 11.8MB 11.8MB 1.93MB 15.9KB 50030
55938
34% 42% 15828 15817 10 0 15827 15.8MB 16.5MB 0% 3.39MB 4.10MB 0% 12.4MB 12.3MB 4.02MB 21.6MB
52142 58771
28% 39% 11807 11807 0 0 11807 12.3MB 13.1MB 0% 2.55MB 3.07MB 0% 9.80MB 9.99MB 6.76MB 27.9MB 38752
43748
33% 42% 15108 15108 0 0 15107 15.1MB 15.5MB 0% 3.32MB 3.91MB 0% 11.7MB 11.6MB 3.50MB 1.17MB 50903
56143
32% 42% 16143 16133 10 0 16143 15.1MB 15.8MB 0% 3.28MB 3.95MB 0% 11.8MB 11.8MB 3.78MB 9.00MB
50922 57403
24% 34% 8843 8843 0 0 8861 14.2MB 14.9MB 0% 3.70MB 4.44MB 0% 10.5MB 10.5MB 8.46MB 10.7MB 46174
53157
27% 37% 10949 10949 0 0 11177 9.91MB 10.2MB 0% 2.45MB 2.84MB 0% 7.46MB 7.40MB 5.55MB 1.67MB 31764
35032
28% 38% 12580 12567 13 0 12579 13.3MB 13.8MB 0% 2.76MB 3.26MB 0% 10.5MB 10.6MB 3.92MB 19.9KB
44119 48488
30% 40% 14300 14300 0 0 14298 14.2MB 14.7MB 0% 3.09MB 3.68MB 0% 11.1MB 11.1MB 2.66MB 600KB 47282
52789
31% 41% 14514 14503 10 0 14514 14.3MB 14.9MB 0% 3.15MB 3.75MB 0% 11.2MB 11.2MB 3.65MB 728KB 48093
53532
31% 42% 14626 14626 0 0 14626 14.3MB 14.9MB 0% 3.16MB 3.77MB 0% 11.1MB 11.1MB 4.84MB 1.14MB 47936
53645
cluster: cluster.cluster: 11/13/2019 22:44:39
cpu cpu total fcache total total data data data cluster cluster cluster disk disk pkts pkts
avg busy ops nfs-ops cifs-ops ops spin-ops recv sent busy recv sent busy recv sent read write
recv sent
---- ---- -------- -------- -------- -------- -------- -------- -------- ---- -------- -------- -
------ -------- -------- -------- -------- -------- --------
30% 39% 15356 15349 7 0 15370 15.3MB 15.8MB 0% 3.29MB 3.94MB 0% 12.0MB 11.8MB 3.18MB 6.90MB 50493
56425
```



```
32% 42% 14156 14146 10 0 14156 14.6MB 15.3MB 0% 3.09MB 3.68MB 0% 11.5MB 11.7MB 5.49MB 16.3MB
48159 53678
```

This is what the performance looks like from Active IQ:



And now we have a single member FlexGroup volume.

```
cluster::*> volume show lots*
Vserver Volume Aggregate State Type Size Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
DEMO lotsafiles - online RW 10TB 7.33TB 0%
DEMO lotsafiles__0001 aggr1_node1 online RW 10TB 7.33TB 0%
2 entries were displayed.
```

The Snapshot copies are still there but are marked as `pre-conversion`.

```
cluster::> set diag
cluster::*> snapshot show -vserver DEMO -volume lotsafiles -fields is-convert-recovery,state
vserver volume snapshot state is-convert-recovery
------- ---------- -------- -------------- -------------------
DEMO lotsafiles base pre-conversion false
DEMO lotsafiles hourly.2019-11-13_1705
pre-conversion false
DEMO lotsafiles hourly.2019-11-13_1805
pre-conversion false
DEMO lotsafiles hourly.2019-11-13_1905
pre-conversion false
DEMO lotsafiles hourly.2019-11-13_2005
pre-conversion false
DEMO lotsafiles hourly.2019-11-13_2105
pre-conversion false
DEMO lotsafiles hourly.2019-11-13_2205
pre-conversion false
DEMO lotsafiles clone_clone.2019-11-13_223144.0
pre-conversion false
DEMO lotsafiles convert.2019-11-13_224411
pre-conversion true
9 entries were displayed.
```


When a Snapshot copy is in preconversion state, using it for a SnapRestore operation fails.

```
cluster::*> snapshot restore -vserver DEMO -volume lotsafiles -snapshot convert.2019-11-13_224411

Error: command failed: Promoting a pre-conversion Snapshot copy is not supported.
```

However, we can still obtain files from the client by using the Snapshot copies.

```
[root@centos7 scripts]# cd /lotsafiles/.snapshot/convert.2019-11-13_224411/pre-convert/
[root@centos7 pre-convert]# ls
topdir_0 topdir_14 topdir_2 topdir_25 topdir_30 topdir_36 topdir_41 topdir_47 topdir_52 topdir_58
topdir_63 topdir_69 topdir_74 topdir_8 topdir_85 topdir_90 topdir_96
topdir_1 topdir_15 topdir_20 topdir_26 topdir_31 topdir_37 topdir_42 topdir_48 topdir_53
topdir_59 topdir_64 topdir_7 topdir_75 topdir_80 topdir_86 topdir_91 topdir_97
topdir_10 topdir_16 topdir_21 topdir_27 topdir_32 topdir_38 topdir_43 topdir_49 topdir_54
topdir_6 topdir_65 topdir_70 topdir_76 topdir_81 topdir_87 topdir_92 topdir_98
topdir_11 topdir_17 topdir_22 topdir_28 topdir_33 topdir_39 topdir_44 topdir_5 topdir_55
topdir_60 topdir_66 topdir_71 topdir_77 topdir_82 topdir_88 topdir_93 topdir_99
topdir_12 topdir_18 topdir_23 topdir_29 topdir_34 topdir_4 topdir_45 topdir_50 topdir_56
topdir_61 topdir_67 topdir_72 topdir_78 topdir_83 topdir_89 topdir_94
topdir_13 topdir_19 topdir_24 topdir_3 topdir_35 topdir_40 topdir_46 topdir_51 topdir_57
topdir_62 topdir_68 topdir_73 topdir_79 topdir_84 topdir_9 topdir_95
```

Growing the newly converted FlexGroup volume is simple. We can add more member volumes by using `volume expand`.

```
cluster::*> volume expand -vserver DEMO -volume lotsafiles -aggr-list aggr1_node1,aggr1_node2 -
aggr-list-multiplier 2

Warning: The following number of constituents of size 10TB will be added to FlexGroup
"lotsafiles": 4. Expanding the FlexGroup will cause the state of all Snapshot copies to be set to
"partial".
Partial Snapshot copies cannot be restored.
Do you want to continue? {y|n}: y

Warning: FlexGroup "lotsafiles" is a converted flexible volume. If this volume is expanded, it
will no longer be able to be converted back to being a flexible volume.
Do you want to continue? {y|n}: y
[Job 23676] Job succeeded: Successful
```

But remember, the data doesn't redistribute. The original member volume keeps the files in place.

```
cluster::*> df -i lots*
Filesystem iused ifree %iused Mounted on Vserver
/vol/lotsafiles/ 3630682 102624948 3% /lotsafiles DEMO
/vol/lotsafiles__0001/ 3630298 17620828 17% /lotsafiles DEMO
/vol/lotsafiles__0002/ 96 21251030 0% --- DEMO
/vol/lotsafiles__0003/ 96 21251030 0% --- DEMO
/vol/lotsafiles__0004/ 96 21251030 0% --- DEMO
/vol/lotsafiles__0005/ 96 21251030 0% --- DEMO
6 entries were displayed.

cluster::*> df -h lots*
Filesystem total used avail capacity Mounted on Vserver
/vol/lotsafiles/ 47TB 2735MB 14TB 0% /lotsafiles DEMO
/vol/lotsafiles/.snapshot
2560GB 49MB 2559GB 0% /lotsafiles/.snapshot DEMO
/vol/lotsafiles__0001/ 9728GB 2505MB 7505GB 0% /lotsafiles DEMO
/vol/lotsafiles__0001/.snapshot
512GB 49MB 511GB 0% /lotsafiles/.snapshot DEMO
/vol/lotsafiles__0002/ 9728GB 57MB 7505GB 0% --- DEMO
/vol/lotsafiles__0002/.snapshot
512GB 0B 512GB 0% --- DEMO
/vol/lotsafiles__0003/ 9728GB 57MB 7766GB 0% --- DEMO
/vol/lotsafiles__0003/.snapshot
512GB 0B 512GB 0% --- DEMO
/vol/lotsafiles__0004/ 9728GB 57MB 7505GB 0% --- DEMO
/vol/lotsafiles__0004/.snapshot
512GB 0B 512GB 0% --- DEMO
```



```
/vol/lotsafiles__0005/ 9728GB 57MB 7766GB 0% --- DEMO
/vol/lotsafiles__0005/.snapshot
512GB 0B 512GB 0% --- DEMO
12 entries were displayed.
```

Converting FlexVol volumes in existing SnapMirror relationships: Example

Following is an example of how to convert a FlexVol volume that has an existing SnapMirror relationship.

This is a volume in a SnapMirror relationship:

```
cluster::*> snapmirror show -destination-path data_dst -fields state
source-path destination-path state
----------- ---------------- ------------
DEMO:data   DEMO:data_dst    Snapmirrored
```

If you try to convert the source, you get an error:

```
cluster::*> vol conversion start -vserver DEMO -volume data -check-only true

Error: command failed: Cannot convert volume "data" in Vserver "DEMO" to a FlexGroup. Correct the
following issues and retry the command:
       * Cannot convert source volume "data" because destination volume "data_dst" of the
SnapMirror relationship with "data" as the source is not converted.  First check if the source
can be converted to a FlexGroup volume using "vol conversion start -volume data -convert-to
flexgroup -check-only true". If the conversion of the source can proceed then first convert the
destination and then convert the source.
```

So, you must convert the destination first.

1. To convert the destination, quiesce the SnapMirror relationship:

```
cluster::*> vol conversion start -vserver DEMO -volume data_dst -check-only true

Error: command failed: Cannot convert volume "data_dst" in Vserver "DEMO" to a FlexGroup. Correct
the following issues and retry the command:
* The relationship was not quiesced. Quiesce SnapMirror relationship using "snapmirror quiesce -
destination-path data_dst" and then try the conversion.
```

2. Next, convert the volume:

```
cluster::*> snapmirror quiesce -destination-path DEMO:data_dst
Operation succeeded: snapmirror quiesce for destination "DEMO:data_dst".

cluster::*> vol conversion start -vserver DEMO -volume data_dst -check-only true
Conversion of volume "data_dst" in Vserver "DEMO" to a FlexGroup can proceed with the following
warnings:
* After the volume is converted to a FlexGroup, it will not be possible to change it back to a
flexible volume.
* Converting flexible volume "data_dst" in Vserver "DEMO" to a FlexGroup will cause the state of
all Snapshot copies from the volume to be set to "pre-conversion". Pre-conversion Snapshot copies
cannot be restored.
```

When you convert the volume, the system lets you know your next steps.

```
cluster::*> vol conversion start -vserver DEMO -volume data_dst

Warning: After the volume is converted to a FlexGroup, it will not be possible to change it back
to a flexible volume.
Do you want to continue? {y|n}: y
Warning: Converting flexible volume "data_dst" in Vserver "DEMO" to a FlexGroup will cause the
state of all Snapshot copies from the volume to be set to "pre-conversion". Pre-conversion
Snapshot copies cannot be restored.
Do you want to continue? {y|n}: y
[Job 23710] Job succeeded: SnapMirror destination volume "data_dst" has been successfully
converted to a FlexGroup volume. You must now convert the relationship's source volume,
"DEMO:data", to a FlexGroup. Then, re-establish the SnapMirror relationship using the "snapmirror
resync" command.
```

3. Convert the source volume:


```
cluster::*> vol conversion start -vserver DEMO -volume data

Warning: After the volume is converted to a FlexGroup, it will not be possible to change it back
to a flexible volume.
Do you want to continue? {y|n}: y
Warning: Converting flexible volume "data" in Vserver "DEMO" to a FlexGroup will cause the state
of all Snapshot copies from the volume to be set to "pre-conversion". Pre-conversion Snapshot
copies cannot be restored.
Do you want to continue? {y|n}: y
[Job 23712] Job succeeded: success
```

4. Resync the mirror:

```
cluster::*> snapmirror resync -destination-path DEMO:data_dst
Operation is queued: snapmirror resync to destination "DEMO:data_dst".

cluster::*> snapmirror show -destination-path DEMO:data_dst -fields state
source-path destination-path state
----------- ---------------- ------------
DEMO:data DEMO:data_dst Snapmirrored
```

The conversion works, but the most important part of a SnapMirror relationship is the restore operation. Therefore, you must see if you can access files from the destination volume's Snapshot copy.

5. Mount the source and destination and compare `ls` output.

```
# mount -o nfsvers=3 DEMO:/data_dst /dst
# mount -o nfsvers=3 DEMO:/data /data
```

The following output shows the content of the source volume.

```
# ls -lah /data
total 14G
drwxrwxrwx 6 root root 4.0K Nov 14 11:57 .
dr-xr-xr-x. 54 root root 4.0K Nov 15 10:08 ..
drwxrwxrwx 2 root root 4.0K Sep 14 2018 cifslink
drwxr-xr-x 12 root root 4.0K Nov 16 2018 nas
-rwxrwxrwx 1 prof1 ProfGroup 0 Oct 3 14:32 newfile
drwxrwxrwx 5 root root 4.0K Nov 15 10:06 .snapshot
lrwxrwxrwx 1 root root 23 Sep 14 2018 symlink -> /shared/unix/linkedfile
drwxrwxrwx 2 root bin 4.0K Jan 31 2019 test
drwxrwxrwx 3 root root 4.0K Sep 14 2018 unix
-rwxrwxrwx 1 newuser1 ProfGroup 0 Jan 14 2019 userfile
-rwxrwxrwx 1 root root 6.7G Nov 14 11:58 Windows2.iso
-rwxrwxrwx 1 root root 6.7G Nov 14 11:37 Windows.iso
```

The destination volume matches exactly, as it should.

```
# ls -lah /dst
total 14G
drwxrwxrwx 6 root root 4.0K Nov 14 11:57 .
dr-xr-xr-x. 54 root root 4.0K Nov 15 10:08 ..
drwxrwxrwx 2 root root 4.0K Sep 14 2018 cifslink
dr-xr-xr-x 2 root root 0 Nov 15 2018 nas
-rwxrwxrwx 1 prof1 ProfGroup 0 Oct 3 14:32 newfile
drwxrwxrwx 4 root root 4.0K Nov 15 10:05 .snapshot
lrwxrwxrwx 1 root root 23 Sep 14 2018 symlink -> /shared/unix/linkedfile
drwxrwxrwx 2 root bin 4.0K Jan 31 2019 test
drwxrwxrwx 3 root root 4.0K Sep 14 2018 unix
-rwxrwxrwx 1 newuser1 ProfGroup 0 Jan 14 2019 userfile
-rwxrwxrwx 1 root root 6.7G Nov 14 11:58 Windows2.iso
-rwxrwxrwx 1 root root 6.7G Nov 14 11:37 Windows.iso
```

If you `ls` to the Snapshot copy in the destination volume, you see the expected files.

```
# ls -lah /dst/.snapshot/snapmirror.7e3cc08e-d9b3-11e6-85e2-00a0986b1210_2163227795.2019-11-
15_100555/
total 14G
drwxrwxrwx 6 root root 4.0K Nov 14 11:57 .
drwxrwxrwx 4 root root 4.0K Nov 15 10:05 ..
```



```
drwxrwxrwx 2 root root 4.0K Sep 14 2018 cifslink
dr-xr-xr-x 2 root root 0 Nov 15 2018 nas
-rwxrwxrwx 1 prof1 ProfGroup 0 Oct 3 14:32 newfile
lrwxrwxrwx 1 root root 23 Sep 14 2018 symlink -> /shared/unix/linkedfile
drwxrwxrwx 2 root bin 4.0K Jan 31 2019 test
drwxrwxrwx 3 root root 4.0K Sep 14 2018 unix
-rwxrwxrwx 1 newuser1 ProfGroup 0 Jan 14 2019 userfile
-rwxrwxrwx 1 root root 6.7G Nov 14 11:58 Windows2.iso
-rwxrwxrwx 1 root root 6.7G Nov 14 11:37 Windows.iso
```

6. Next, expand the FlexGroup source to provide more capacity:

```
cluster::*> volume expand -vserver DEMO -volume data -aggr-list aggr1_node1,aggr1_node2 -aggr-
list-multiplier

Warning: The following number of constituents of size 30TB will be added to FlexGroup "data": 4.
Expanding the FlexGroup will cause the state of all Snapshot copies to be set to "partial".
Partial Snapshot copies cannot be restored.
Do you want to continue? {y|n}: y
[Job 23720] Job succeeded: Successful
```

The source volume now has five member volumes. The destination volume has only one.

```
cluster::*> vol show -vserver DEMO -volume data*
Vserver Volume Aggregate State Type Size Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
DEMO data - online RW 150TB 14.89TB 0%
DEMO data__0001 aggr1_node2 online RW 30TB 7.57TB 0%
DEMO data__0002 aggr1_node1 online RW 30TB 7.32TB 0%
DEMO data__0003 aggr1_node2 online RW 30TB 7.57TB 0%
DEMO data__0004 aggr1_node1 online RW 30TB 7.32TB 0%
DEMO data__0005 aggr1_node2 online RW 30TB 7.57TB 0%
DEMO data_dst - online DP 30TB 7.32TB 0%
DEMO data_dst__0001
aggr1_node1 online DP 30TB 7.32TB 0%
8 entries were displayed.
```

7. Update the mirror, and ONTAP fixes it for you:

```
cluster::*> snapmirror update -destination-path DEMO:data_dst
Operation is queued: snapmirror update of destination "DEMO:data_dst".
```

The update initially fails with the following error message:

```
Last Transfer Error: A SnapMirror transfer for the relationship with destination FlexGroup
"DEMO:data_dst" was aborted because the source FlexGroup was expanded. A SnapMirror AutoExpand
job with id "23727" was created to expand the destination FlexGroup and to trigger a SnapMirror
transfer for the SnapMirror relationship. After the SnapMirror transfer is successful, the
"healthy" field of the SnapMirror relationship will be set to "true". The job can be monitored
using either the "job show -id 23727" or "job history show -id 23727" commands.
```

The job expands the volume, and then you can update again.

```
cluster::*> job show -id 23727
Owning
Job ID Name Vserver Node State
------ -------------------- ---------- -------------- ----------
23727 Snapmirror Expand cluster
node1
Success
Description: SnapMirror FG Expand data_dst


cluster::*> snapmirror show -destination-path DEMO:data_dst -fields state
source-path destination-path state
----------- ---------------- ------------
DEMO:data DEMO:data_dst Snapmirrored
```

Now both FlexGroup volumes have the same number of member volumes.

```
cluster::*> vol show -vserver DEMO -volume data*
Vserver Volume Aggregate State Type Size Available Used%
```



```
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
DEMO data - online RW 150TB 14.88TB 0%
DEMO data__0001 aggr1_node2 online RW 30TB 7.57TB 0%
DEMO data__0002 aggr1_node1 online RW 30TB 7.32TB 0%
DEMO data__0003 aggr1_node2 online RW 30TB 7.57TB 0%
DEMO data__0004 aggr1_node1 online RW 30TB 7.32TB 0%
DEMO data__0005 aggr1_node2 online RW 30TB 7.57TB 0%
DEMO data_dst - online DP 150TB 14.88TB 0%
DEMO data_dst__0001
aggr1_node1 online DP 30TB 7.32TB 0%
DEMO data_dst__0002
aggr1_node1 online DP 30TB 7.32TB 0%
DEMO data_dst__0003
aggr1_node2 online DP 30TB 7.57TB 0%
DEMO data_dst__0004
aggr1_node1 online DP 30TB 7.32TB 0%
DEMO data_dst__0005
aggr1_node2 online DP 30TB 7.57TB 0%
12 entries were displayed.
```

Sample FlexVol to FlexGroup volume conversion — 500 million files

In this example, a FlexVol volume with 500 million files is converted to a FlexGroup volume.

```
cluster::*> vol show -vserver DEMO -volume fvconvert -fields files,files-used,is-flexgroup
vserver volume    files      files-used is-flexgroup
------- --------- ---------- ---------- ------------
DEMO    fvconvert 2040109451 502631608  false
```

Because it takes so long to create that many files, we created a FlexClone volume of it and split it. This approach lets you keep the origin volume intact and test without risk.

In this example, the cloning process takes about 30 minutes:

```
cluster::*> vol clone split start -vserver DEMO -flexclone fvconvert -foreground true

Warning: Are you sure you want to split clone volume fvconvert in Vserver DEMO ? {y|n}: y
[Job 24230] 0% inodes processed.

cluster::*> job history show -id 24230 -fields starttime,endtime
node               record  vserver         endtime        starttime
------------------ ------- --------------- -------------- --------------
node1              2832338 cluster         12/09 10:27:08 12/09 09:58:16
```

After the clone split, we run the check. We must run `volume clone sharing-by-split undo` to get rid of shared FlexClone blocks, which takes some time, but then the check produces the following output:

```
cluster::*> volume conversion start -vserver DEMO -volume fvconvert -foreground true -check-only
true
Conversion of volume "fvconvert" in Vserver "DEMO" to a FlexGroup can proceed with the following
warnings:
* After the volume is converted to a FlexGroup, it will not be possible to change it back to a
flexible volume.
```

We then run the script that was run earlier to generate the load and watch the statistics on the cluster to see if we hit any outage. Again, the conversion takes seconds (with 500 million files) and there is just a small, barely noticeable delay.

```
cluster::*> volume conversion start -vserver DEMO -volume fvconvert -foreground true

Warning: After the volume is converted to a FlexGroup, it will not be possible to change it back
to a flexible volume.
Do you want to continue? {y|n}: y
[Job 24259] Job succeeded: success
```


Figure 122) Sample statistics from conversion process.

```
 cpu  cpu    total                            fcache      total      total data       data       data cluster   cluster   cluster       disk       disk       pkts       pkts
 avg busy      ops  nfs-ops cifs-ops       ops spin-ops       recv       sent busy       recv       sent    busy      recv      sent       read      write       recv       sent
---- ---- -------- -------- -------- -------- -------- -------- -------- ---- -------- -------- ------- -------- -------- -------- -------- -------- --------
 12%  21%      150      140        9        0      145   2.22MB   2.48MB   0%   50.9KB    315KB      0%   2.17MB   2.17MB   16.1MB   5.62MB      820      810
  7%  13%       46       46        0        0       46   1.36MB   1.46MB   0%   24.0KB    123KB      0%   1.34MB   1.34MB   11.0MB   19.9KB      452      446
 14%  24%     1721     1721        0        0     1721   2.25MB   2.48MB   0%    197KB    438KB      0%   2.06MB   2.05MB   24.3MB   9.55MB     2814     2917
 15%  25%     3576     3573        2        0     3575   5.49MB   5.77MB   0%    985KB   1.23MB      0%   4.53MB   4.53MB   18.1MB   1.22MB    14847    15681
 16%  21%     1211     1180       30        0     1209   2.41MB   2.68MB   0%    275KB    559KB      0%   2.14MB   2.14MB   16.9MB   2.35MB     4249     4751
 27%  34%     1979     1968       10        0     1978   39.4MB   22.2MB   1%   19.0MB   1.69MB      0%   20.4MB   20.5MB   14.3MB   1.14MB    21043     9869
 18%  23%     2666     2664        1        0     2665   3.43MB   4.54MB   0%    583KB   1.79MB      0%   2.86MB   2.75MB   14.6MB   19.9KB     7686     8755
 23%  34%     1917     1917        0        0     1917   2.88MB   4.22MB   0%    563KB   1.89MB      0%   2.33MB   2.33MB   19.7MB   19.1MB     7352     8323
 36%  58%     2264     2260        4        0     2264   3.25MB   4.40MB   0%    474KB   1.61MB      0%   2.79MB   2.79MB   34.2MB   19.0MB     5763     6342
 26%  45%     1351     1303       47        0     1350   7.98MB   5.64MB   0%   2.93MB    595KB      0%   5.05MB   5.05MB   34.4MB   19.7KB     8267     7036
 28%  45%     2032     2002       29        0     2031   24.9MB   13.9MB   0%   11.7MB    597KB      0%   13.2MB   13.3MB   33.5MB   1.66MB    15344     8798
 26%  49%     1813     1745       67        0     1812   28.6MB   16.5MB   1%   13.7MB    728KB      0%   14.9MB   15.8MB   36.7MB   19.8KB    17761     9963
 27%  50%     2438     2416       22        0     2437   18.6MB   10.3MB   0%   8.08MB    860KB      0%   10.5MB   9.48MB   37.8MB   11.9KB    13884     9831
 31%  58%     2043     2002       40        0     2043   18.8MB   12.6MB   0%   8.38MB    726KB      0%   10.4MB   11.9MB   64.3MB    150MB    13331     9469
 35%  67%     1475     1413       62        0     1474   22.8MB   13.1MB   0%   10.4MB    812KB      0%   12.3MB   12.3MB   86.6MB   53.7MB    14723     9296
 35%  66%     2028     1961       66        0     2022   29.3MB   15.9MB   1%   13.5MB    612KB      0%   15.8MB   15.3MB   86.3MB   2.84MB    16522     8716
 38%  71%     2446     2413       32        0     2444   13.6MB   9.10MB   0%   5.90MB    911KB      0%   7.70MB   8.21MB   78.9MB   19.8KB    13169    10819
 19%  34%     1771     1727       43        0     1770   34.0MB   17.9MB   1%   15.3MB    699KB      0%   18.7MB   17.2MB   11.0MB   11.9KB    19605    10707
 17%  30%      749      696       53        0      748   34.3MB   18.6MB   1%   17.5MB    419KB      0%   16.8MB   18.2MB   11.3MB   5.32MB    18226     7898
 19%  35%     1194     1137       56        0     1194   16.4MB   8.62MB   0%   6.54MB    261KB      0%   9.87MB   8.37MB   12.0MB   55.4MB     7586     3595
ontap9-tme-8040: cluster.cluster: 12/10/2019 11:03:06
 cpu  cpu    total                            fcache      total      total data       data       data cluster   cluster   cluster       disk       disk       pkts       pkts
 avg busy      ops  nfs-ops cifs-ops       ops spin-ops       recv       sent busy       recv       sent    busy      recv      sent       read      write       recv       sent
---- ---- -------- -------- -------- -------- -------- -------- -------- ---- -------- -------- ------- -------- -------- -------- -------- -------- --------
 25%  41%     2954     2915       38        0     2953   35.2MB   20.5MB   1%   16.9MB    814KB      0%   18.2MB   19.7MB   9.84MB    110MB    22564    13299
 23%  41%     2292     2250       41        0     2291   17.7MB   10.3MB   0%   6.73MB    866KB      0%   10.9MB   9.44MB   11.6MB    109MB    15241    12193
 25%  43%     2415     2369       46        0     2414   52.8MB   19.5MB   1%   16.8MB    909KB      1%   36.1MB   18.6MB   11.5MB   26.7MB    25076    13688
 29%  40%     2821     2792       29        0     2820   48.8MB   25.5MB   1%   22.6MB    852KB      1%   26.2MB   24.7MB   11.2MB   5.45MB    26766    13167
 26%  41%     2584     2550       33        0     2582   66.3MB   38.1MB   2%   35.1MB   1.13MB      1%   31.2MB   37.0MB   9.87MB   11.7KB    37901    17822
 34%  61%     3438     3397       40        0     3437   92.6MB   55.4MB   4%   51.7MB   1.65MB      2%   40.9MB   53.7MB   9.35MB   1.04MB    54703    25093
 25%  41%     5686     5664       22        0     5684   40.8MB   24.0MB   1%   18.9MB   1.63MB      0%   21.9MB   22.4MB   11.0MB   15.5MB    34334    25357
 19%  31%     4678     4650       28        0     4678   52.0MB   28.9MB   2%   25.1MB   1.51MB      1%   26.9MB   27.4MB   13.8MB    109MB    35020    21615
 18%  29%     3812     3794       18        0     3810   32.4MB   18.4MB   1%   14.1MB   1.13MB      0%   18.3MB   17.3MB   13.4MB   89.7MB    24517    17097
 23%  37%     4599     4564       35        0     4599   64.8MB   35.9MB   2%   31.4MB   1.53MB      1%   33.4MB   34.3MB   14.8MB   43.8MB    40398    22883
 19%  31%     3221     3180       41        0     3215   59.3MB   32.0MB   2%   28.6MB   1.31MB      1%   30.7MB   30.7MB   14.1MB   54.4MB    35120    18328
 12%  20%      661      633       27        0      661   42.3MB   22.9MB   1%   21.8MB    458KB      0%   20.5MB   22.5MB   19.3MB   63.7MB    18672     4920
 19%  29%     4844     4822       21        0     4842   47.7MB   25.1MB   1%   21.1MB   1.51MB      1%   26.6MB   23.6MB   6.85MB   19.9KB    32420    21523
 21%  34%     6205     6185       19        0     6205   44.2MB   26.5MB   1%   20.5MB   1.82MB      1%   23.7MB   24.7MB   8.10MB   2.64MB    37075    28031
 20%  33%     5652     5636       16        0     5651   34.5MB   20.0MB   1%   15.0MB   1.46MB      0%   19.5MB   18.5MB   5.43MB   19.8KB    27591    20579
 20%  29%     6400     6376       23        0     6399   40.7MB   24.7MB   1%   18.1MB   2.11MB      0%   22.6MB   22.6MB   5.79MB   1.10MB    37445    29159
 26%  41%     6493     6469       24        0     6492   49.6MB   28.3MB   1%   23.1MB   1.82MB      1%   26.5MB   26.5MB   9.06MB   61.3MB    39040    27688
 27%  40%     7860     7847       12        0     7860   30.3MB   19.9MB   1%   12.6MB   2.17MB      0%   17.7MB   17.7MB   8.96MB   94.1MB    35235    31193
 22%  34%     7093     7073       20        0     7092   41.0MB   24.9MB   1%   18.1MB   2.03MB      0%   22.9MB   22.9MB   7.22MB   1.10MB    38476    31338
```

Convert happened here:

Then, as the script is running, we add new member volumes to the FlexGroup volume. Again, there is no disruption.

```
cluster::*> volume expand -vserver DEMO -volume fvconvert -aggr-list aggr1_node1 -aggr-list-multiplier 3 -foreground true

Warning: The following number of constituents of size 40TB will be added to FlexGroup
"fvconvert": 3.
Do you want to continue? {y|n}: y
[Job 24261] Job succeeded: Successful
```

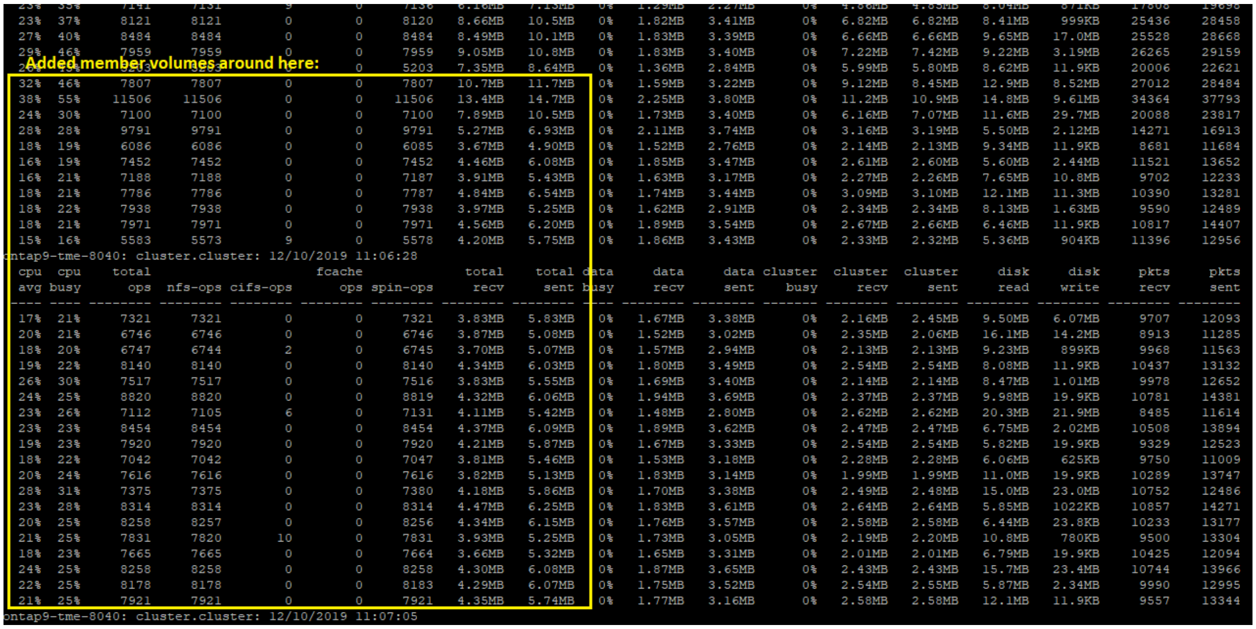
Then we add four more member volumes:

```
cluster::*> volume expand -vserver DEMO -volume fvconvert -aggr-list aggr1_node2 -aggr-list-multiplier 4

Warning: The following number of constituents of size 40TB will be added to FlexGroup
"fvconvert": 4.
Do you want to continue? {y|n}: y
[Job 24264] Job succeeded: Successful
```

As an added bonus, we start to see more total IOPS for the workload. The job itself takes much less time overall than when we ran it on a FlexVol volume, because the FlexGroup volume's parallel ingest starts to help the script run faster.

Figure 123) Sample statistics during conversion process — Adding member volumes.



You can view a [video of this capture](#).

We also capture the time of completion for each job.

```
This was the job on the FlexVol before it was converted:
# python file-create.py /fvconvert/files
Starting overall work: 2019-12-09 10:32:21.966337
End overall work: 2019-12-09 12:11:15.990707
total time: 5934.024611
```

Converting the FlexVol volume to a FlexGroup volume (with added member volumes) saves some time:

```
# python file-create.py /fvconvert/files2
Starting overall work: 2019-12-10 11:02:28.621532
End overall work: 2019-12-10 12:22:48.523772
total time: 4819.95753193
```

That is a savings of about 1100 seconds, or 18 minutes—which saves us around 20% of the total time of completion.

The following output shows the file distribution before the script run. Note that the first member volume has the most files, because it was previously the FlexVol volume.

```
cluster::*> volume show -vserver DEMO -volume fvconvert* -fields files,files-used
vserver volume          files      files-used
------- --------------  ---------- ----------
DEMO    fvconvert__0001 2040109451 502848737
DEMO    fvconvert__0002 2040109451 12747
DEMO    fvconvert__0003 2040109451 12749
DEMO    fvconvert__0004 2040109451 12751
```

At the end of the job, we can see that the files spread out fairly evenly.

```
cluster::*> volume show -vserver DEMO -volume fvconvert* -fields files,files-used
vserver volume          files      files-used
------- --------------  ---------- ----------
DEMO    fvconvert__0001 2040109451 506770209
DEMO    fvconvert__0002 2040109451 3345330
DEMO    fvconvert__0003 2040109451 3345330
DEMO    fvconvert__0004 2040109451 3345319
DEMO    fvconvert__0005 2040109451 3331657
```



```
DEMO     fvconvert__0006 2040109451 3331635
DEMO     fvconvert__0007 2040109451 3331657
DEMO     fvconvert__0008 2040109451 3331657
```

We run the script again on the newly converted FlexGroup volume. This time, we want to see how much faster the job runs and how the files distribute on the emptier FlexVol member volumes.

Remember, when we started out, the newer member volumes all had less than 1% of files used (3.3 million of two billion possible files). The member volume that was converted from a FlexVol volume was using 25% of the total files (500 million of two billion).

After the job runs, we see a file count delta of about 3.2 million on the original member volume and of about 3.58 million on all the other members. We're still balancing across all member volumes but favoring the less full ones for new file and folder creations.

```
cluster::*> volume show -vserver DEMO -volume fvconvert* -fields files,files-used
vserver volume          files      files-used
------- --------------  ---------- ----------
DEMO     fvconvert__0001 2040109451 509958440
DEMO     fvconvert__0002 2040109451 6808792
DEMO     fvconvert__0003 2040109451 6809225
DEMO     fvconvert__0004 2040109451 6806843
DEMO     fvconvert__0005 2040109451 6798959
DEMO     fvconvert__0006 2040109451 6800054
DEMO     fvconvert__0007 2040109451 6849375
DEMO     fvconvert__0008 2040109451 6801600
```

With the new FlexGroup volume, converted from a FlexVol volume, our job time drops from 5900 seconds to 4656 seconds. We are also able to push two times the amount of IOPS:

```
# python file-create.py /fvconvert/files3
Starting overall work: 2019-12-10 13:14:26.816860
End overall work: 2019-12-10 14:32:03.565705
total time: 4656.76723099
```

Figure 124) Sample statistics of conversion process — two times performance.

```
 cpu  cpu    total                    fcache             total    total data     data     data cluster  cluster  cluster     disk     disk     pkts     pkts
 avg busy      ops  nfs-ops cifs-ops      ops spin-ops     recv     sent busy     recv     sent    busy     recv     sent     read    write     recv     sent
---- ---- -------- -------- -------- -------- -------- -------- -------- ---- -------- -------- ------- -------- -------- -------- -------- -------- --------
 26%  29%    10403    10403        0        0    10403   7.55MB   8.31MB   0%   2.25MB   3.00MB      0%   5.30MB   5.31MB   4.20MB   21.6MB    17915    23123
 27%  31%     9262     9262        0        0     9262   6.93MB   7.82MB   0%   2.05MB   2.56MB      0%   4.87MB   5.26MB   5.67MB   35.3MB    16487    20524
 25%  29%     8773     8773        0        0     8773   7.47MB   7.68MB   0%   2.39MB   3.01MB      0%   5.08MB   4.67MB    851KB   7.92KB    18667    22978
 18%  22%     6592     6592        0        0     6591   4.21MB   4.57MB   0%   1021KB   1.34MB      0%   3.21MB   3.23MB   1.33MB   23.9KB     8963    10892
 20%  21%     9400     9400        0        0     9399   6.72MB   7.32MB   0%   2.26MB   2.87MB      0%   4.46MB   4.45MB   1.05MB   8.22MB    18350    21814
 25%  26%    12010    12010        0        0    12010   7.25MB   8.00MB   0%   2.18MB   2.93MB      0%   5.07MB   5.07MB   4.67MB   17.2MB    17918    22028
 22%  23%    11266    11266        0        0    11266   8.23MB   9.06MB   0%   2.49MB   3.31MB      0%   5.73MB   5.74MB   5.11MB   12.1MB    20029    25981
 25%  26%    12445    12445        0        0    12445   11.0MB   12.0MB   0%   3.82MB   4.84MB      0%   7.18MB   7.12MB    915KB   10.3MB    27291    35571
 25%  26%    12253    12253        0        0    12253   8.04MB   8.77MB   0%   2.53MB   3.26MB      0%   5.51MB   5.52MB    976KB   11.7KB    20328    25953
 29%  34%    12699    12699        0        0    12699   8.42MB   9.29MB   0%   2.65MB   3.52MB      0%   5.77MB   5.77MB   1.41MB   3.73MB    20937    27166
 28%  30%    12599    12599        0        0    12599   8.34MB   9.09MB   0%   2.62MB   3.38MB      0%   5.71MB   5.71MB   4.20MB   21.5MB    20958    26748
 30%  34%    13929    13919        9        0    13924   9.41MB   10.5MB   0%   3.00MB   4.11MB      0%   6.41MB   6.40MB   3.29MB   65.7KB    23395    30206
 26%  28%    14499    14499        0        0    14499   9.68MB   10.6MB   0%   3.08MB   4.00MB      0%   6.60MB   6.60MB   3.77MB   25.3MB    24627    31571
 29%  34%    13231    13231        0        0    13230   8.44MB   9.46MB   0%   2.75MB   3.78MB      0%   5.69MB   5.68MB   1.77MB   11.9KB    21565    27726
 26%  28%    13505    13502        2        0    13503   9.10MB   10.3MB   0%   3.01MB   4.19MB      0%   6.09MB   6.09MB   2.02MB   3.45MB    24130    30584
 25%  29%    13553    13553        0        0    13553   8.94MB   9.86MB   0%   2.82MB   3.73MB      0%   6.12MB   6.13MB   4.62MB   23.1MB    22491    28837
```

2x!

As you can see, there's an imbalance of files and data in these member volumes (much more in the original FlexVol volume), but performance is still much better than the previous FlexVol performance because work across multiple nodes is more efficient. That's the power of the FlexGroup volume.

Event management system examples

Inode-related EMS examples

```
Message Name: callhome.no.inodes
Severity: ERROR

Corrective Action: Modify the volume's maxfiles (maximum number of files) to increase the inodes
on the affected volume. If you need assistance, contact NetApp technical support.

Description: This message occurs when a volume is out of inodes, which refer to individual files,
other types of files, and directories. If your system is configured to do so, it generates and
```



```
transmits an AutoSupport (or 'call home') message to NetApp technical support and to the
configured destinations. Successful delivery of an AutoSupport message significantly improves
problem determination and resolution.
```

```
Message Name: fg.inodes.member.nearlyFull
Severity: ALERT

Corrective Action: Adding capacity to the FlexGroup by using the "volume modify -files +X"
command is the best way to solve this problem. Alternatively, deleting files from the FlexGroup
might work, although it can be difficult to determine which files have landed on which
constituent.

Description: This message occurs when a constituent within a FlexGroup is almost out of inodes.
This constituent will receive far fewer new create requests than average, which might impact the
FlexGroup's overall performance, because those requests are routed to constituents with more
inodes.
```

```
Message Name: fg.inodes.member.full
Severity: ALERT

Corrective Action: Adding capacity to the FlexGroup by using the "volume modify -files +X"
command is the best way to solve this problem. Alternatively, deleting files from the FlexGroup
may work, but it is difficult to determine which files have landed on which constituent.

Description: This message occurs when a constituent with a FlexGroup has run out of inodes. New
files cannot be created on this constituent. This might lead to an overall imbalanced
distribution of content across the FlexGroup.
```

```
Message Name: fg.inodes.member.allOK
Severity: NOTICE

Corrective Action: (NONE)

Description: This message occurs when conditions that led to previous
"fg.inodes.member.nearlyFull" and "fg.inodes.member.full" events no longer apply for any
constituent in this FlexGroup. All constituents within this FlexGroup have sufficient inodes for
normal operation.
```

Example of maxdirsize message

```
Message Name: wafl.dir.size.max
Severity: ERROR

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 after a directory has reached its maximum directory size
(maxdirsize) limit.
                               Supports SNMP trap: true
                                     Destinations: -
            Number of Drops Between Transmissions: 0
Dropping Interval (Seconds) Between Transmissions: 0


Message Name: wafl.dir.size.max.warning
Severity: ERROR

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 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.
                               Supports SNMP trap: true
                                     Destinations: -
            Number of Drops Between Transmissions: 0
Dropping Interval (Seconds) Between Transmissions: 0


Message Name: wafl.dir.size.warning
Severity: ERROR

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 mesaage occurs when a directory surpasses 90% of its current maximum directory
size (maxdirsize) limit.
                               Supports SNMP trap: true
                                     Destinations: -
            Number of Drops Between Transmissions: 0
Dropping Interval (Seconds) Between Transmissions: 0
```

Examples of capacity-related event management system messages

```
Message Name: monitor.volume.full
Severity: DEBUG
Corrective Action: (NONE)

Description: This message occurs when one or more file systems are full, typically indicating at
least 98% full. This event is accompanied by global health monitoring messages for the customer.
The space usage is computed based on the active file system size and is computed by subtracting
the value of the "Snapshot Reserve" field from the value of the "Used" field of the "volume show-
space" command. The volume/aggregate can be over 100% full due to space used or reserved by
metadata. A value greater than 100% might cause Snapshot(tm) copy space to become unavailable or
cause the volume to become logically overallocated. See the "vol.log.overalloc" EMS message for
more information.
                               Supports SNMP trap: true
                                     Destinations: -
            Number of Drops Between Transmissions: 0
Dropping Interval (Seconds) Between Transmissions: 0
```

```
Message Name: monitor.volume.nearlyFull
Severity: ALERT
Corrective Action: Create space by increasing the volume or aggregate sizes, or by deleting data
or deleting Snapshot(R) copies. To increase a volume's size, use the "volume size" command. To
delete a volume's Snapshot(R) copies, use the "volume snapshot delete" command. To increase an
aggregate's size, add disks by using the "storage aggregate add-disks" command. Aggregate
Snapshot(R) copies are deleted automatically when the aggregate is full.
                                      Description: This message occurs when one or more file
systems are nearly full, typically indicating at least 95% full. This event is accompanied by
global health monitoring messages for the customer. The space usage is computed based on the
active file system size and is computed by subtracting the value of the "Snapshot Reserve" field
from the value of the "Used" field of the "volume show-space" command.
                               Supports SNMP trap: true
                                     Destinations: -
            Number of Drops Between Transmissions: 0
Dropping Interval (Seconds) Between Transmissions: 0
```



```
Message Name: monitor.volume.ok
Severity: DEBUG
Corrective Action: (UNKNOWN)

Description: The previously-reported volume full condition is fixed. * We log this event, as well
as the other monitor.volume events, at LOG_DEBUG level to avoid spamming the messages file with
events which are already being reported as part of the global health messages.
                               Supports SNMP trap: true
                                     Destinations: -
            Number of Drops Between Transmissions: 0
Dropping Interval (Seconds) Between Transmissions: 0
```

```
Message Name: monitor.volumes.one.ok
Severity: DEBUG
Corrective Action: (NONE)

Description: This message occures when one file system that was nearly full (usually this means
>= 95% full) is now OK. This event and other "monitor.volume" events are logged at LOG_DEBUG
level to avoid spamming the messages file with events that are already being reported as part of
the global health messages. The space usage is computed based on the active file system size and
is computed by subtracting the value of the "Snapshot Reserve" field from the value of the "Used"
field of the "volume show-space" command.
                               Supports SNMP trap: true
                                     Destinations: -
            Number of Drops Between Transmissions: 0
Dropping Interval (Seconds) Between Transmissions: 0
```

```
Message Name: vol.log.overalloc
Severity: ALERT
Corrective Action: Create space by increasing the volume or aggregate size, deleting data,
deleting Snapshot(R) copies, or changing the provisioning from thick to thin. To increase a
volume's size, use the "volume size" command. To delete a volume's Snapshot(R) copies, use the
"volume snapshot delete" command. To change provisioning in a volume, reserved files can be
unreserved by using the "volume file reservation" command. To increase an aggregate's size, add
disks by using the "storage aggregate add-disks" command. Aggregate Snapshot(R) copies are
deleted automatically when the aggregate is full. To change provisioning of a volume in an
aggregate, change the volume guarantee from "volume" to "none" by using the "space-guarantee"
field of the "volume modify" command.

Description: This message occurs when the volume or aggregate allocates more space than it can
honor by way of reservations, or the aggregate has allocated more space than it can honor by way
of guarantees. If the reserved or guaranteed space is consumed, there is insufficient physical
space, which can cause the volume or aggregate to be taken offline.
                               Supports SNMP trap: true
                                     Destinations: -
            Number of Drops Between Transmissions: 0
Dropping Interval (Seconds) Between Transmissions: 0
```

```
Message Name: fg.member.elastic.sizing
Severity: NOTICE
Corrective Action: (NONE)
Description: This message occurs when a FlexGroup constituent undergoes elastic sizing, either to
restore balance among constituents or to resize constituents to accommodate space needs.
Supports SNMP trap: false
Destinations: -
Number of Drops Between Transmissions: 0
Dropping Interval (Seconds) Between Transmissions: 0
```


Command examples

FlexGroup capacity commands

```
cluster::*> aggr show-space -instance -aggregate aggr1_node1

                          Aggregate Name: aggr1_node1
                       Volume Footprints: 2.05TB
               Volume Footprints Percent: 26%
Total Space for Snapshot Copies in Bytes: 0B
      Space Reserved for Snapshot Copies: 0%
                      Aggregate Metadata: 15.20MB
              Aggregate Metadata Percent: 0%
                              Total Used: 2.05TB
                      Total Used Percent: 26%
                                    Size: 7.86TB
               Snapshot Reserve Unusable: -
       Snapshot Reserve Unusable Percent: -
                Total Physical Used Size: 143.7GB
                Physical Used Percentage: 2%

                          Aggregate Name: aggr1_node2
                       Volume Footprints: 2.02TB
               Volume Footprints Percent: 26%
Total Space for Snapshot Copies in Bytes: 0B
      Space Reserved for Snapshot Copies: 0%
                      Aggregate Metadata: 8.63MB
              Aggregate Metadata Percent: 0%
                              Total Used: 2.02TB
                      Total Used Percent: 26%
                                    Size: 7.86TB
               Snapshot Reserve Unusable: -
       Snapshot Reserve Unusable Percent: -
                Total Physical Used Size: 69.71GB
                Physical Used Percentage: 1%
2 entries were displayed.
```

```
cluster::*> volume show-space -vserver SVM -volume flexgroup__*
      Vserver : SVM
      Volume  : flexgroup__0001

      Feature                                   Used      Used%
      --------------------------------     ----------     ------
      User Data                               57.06MB         0%
      Filesystem Metadata                      3.51MB         0%
      Inodes                                  87.26MB         0%
      Snapshot Reserve                          512GB         5%
      Deduplication                              12KB         0%
      Performance Metadata                       48KB         0%

      Total Used                              512.1GB         5%
      Total Physical Used                     148.3MB         0%


      Vserver : SVM
      Volume  : flexgroup__0002

      Feature                                   Used      Used%
      --------------------------------     ----------     ------
      User Data                               57.03MB         0%
      Filesystem Metadata                      4.66MB         0%
      Inodes                                  83.66MB         0%
      Snapshot Reserve                          512GB         5%
      Deduplication                              20KB         0%
      Performance Metadata                       44KB         0%

      Total Used                              512.1GB         5%
```



```
Total Physical Used                 145.7MB         0%


Vserver : SVM
Volume  : flexgroup__0003

Feature                                Used      Used%
--------------------------------    ----------   ------
User Data                              57.02MB      0%
Filesystem Metadata                     3.66MB      0%
Inodes                                 84.55MB      0%
Snapshot Reserve                        512GB       5%
Deduplication                            12KB       0%
Performance Metadata                     44KB       0%

Total Used                             512.1GB      5%
Total Physical Used                    145.6MB      0%


Vserver : SVM
Volume  : flexgroup__0004

Feature                                Used      Used%
--------------------------------    ----------   ------
User Data                              57.19MB      0%
Filesystem Metadata                     8.93MB      0%
Inodes                                 82.09MB      0%
Snapshot Reserve                        512GB       5%
Deduplication                            12KB       0%
Performance Metadata                     44KB       0%

Total Used                             512.1GB      5%
Total Physical Used                    148.5MB      0%


Vserver : SVM
Volume  : flexgroup__0005

Feature                                Used      Used%
--------------------------------    ----------   ------
User Data                               3.99GB      0%
Filesystem Metadata                     4.88MB      0%
Inodes                                 83.54MB      0%
Snapshot Reserve                        512GB       5%
Deduplication                            12KB       0%
Performance Metadata                     52KB       0%

Total Used                             516.1GB      5%
Total Physical Used                     4.08GB      0%


Vserver : SVM
Volume  : flexgroup__0006

Feature                                Used      Used%
--------------------------------    ----------   ------
User Data                              57.04MB      0%
Filesystem Metadata                     3.50MB      0%
Inodes                                 87.26MB      0%
Snapshot Reserve                        512GB       5%
Deduplication                            12KB       0%
Performance Metadata                     44KB       0%

Total Used                             512.1GB      5%
Total Physical Used                    148.2MB      0%


Vserver : SVM
Volume  : flexgroup__0007

Feature                                Used      Used%
```



```
      --------------------------------     ----------     ------
      User Data                              57.02MB         0%
      Filesystem Metadata                     3.50MB         0%
      Inodes                                 85.03MB         0%
      Snapshot Reserve                         512GB         5%
      Deduplication                             12KB         0%
      Performance Metadata                      44KB         0%

      Total Used                             512.1GB         5%
      Total Physical Used                    145.9MB         0%


      Vserver : SVM
      Volume  : flexgroup__0008

      Feature                                   Used      Used%
      --------------------------------     ----------     ------
      User Data                              57.03MB         0%
      Filesystem Metadata                     3.52MB         0%
      Inodes                                 86.12MB         0%
      Snapshot Reserve                         512GB         5%
      Deduplication                             12KB         0%
      Performance Metadata                      44KB         0%

      Total Used                             512.1GB         5%
      Total Physical Used                    147.0MB         0%
```

```
cluster::> vol show -is-constituent true -volume flexgroup__*
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
SVM       flexgroup__0001
                       aggr1_node1  online     RW         10TB     5.05TB   49%
SVM       flexgroup__0002
                       aggr1_node2  online     RW         10TB     5.08TB   49%
SVM       flexgroup__0003
                       aggr1_node1  online     RW         10TB     5.05TB   49%
SVM       flexgroup__0004
                       aggr1_node2  online     RW         10TB     5.08TB   49%
SVM       flexgroup__0005
                       aggr1_node1  online     RW         10TB     5.05TB   49%
SVM       flexgroup__0006
                       aggr1_node2  online     RW         10TB     5.08TB   49%
SVM       flexgroup__0007
                       aggr1_node1  online     RW         10TB     5.05TB   49%
SVM       flexgroup__0008
                       aggr1_node2  online     RW         10TB     5.08TB   49%
8 entries were displayed.

cluster::*> storage aggregate show -aggregate aggr1* -fields usedsize,size,percent-used -sort-by
percent-used
aggregate   percent-used size   usedsize
----------- ------------ ------ --------
aggr1_node1 26%          7.86TB 2.05TB
aggr1_node2 26%          7.86TB 2.02TB
2 entries were displayed.
```

Example of statistics show-periodic command for entire cluster

```
cluster::*> statistics show-periodic
cluster: cluster.cluster: 11/30/2016 11:49:46
  cpu  cpu    total                      fcache            total    total data     data     data
cluster  cluster  cluster     disk     disk     pkts     pkts
  avg busy      ops  nfs-ops cifs-ops      ops spin-ops     recv     sent busy     recv     sent
busy     recv     sent     read    write     recv     sent
 ---- ---- -------- -------- -------- -------- -------- -------- -------- ---- -------- --------
------- -------- -------- -------- -------- -------- --------
   5%   5%        0        0        0        0        0   65.3KB   64.4KB   0%   2.22KB   1.13KB
0%   62.7KB   63.2KB    489KB    407KB       91       83
```



```
   5%   5%        0        0        0        0        0   62.5KB   61.6KB   0%   1.28KB     767B
0%   61.0KB   60.9KB   23.8KB   23.8KB       64       60
   4%   5%        0        0        0        0        0   62.3KB   61.3KB   0%   1.43KB     708B
0%   60.7KB   60.7KB   15.8KB   15.8KB       69       58
cluster: cluster.cluster: 11/30/2016 11:49:53
  cpu  cpu    total                          fcache            total    total data     data     data
cluster  cluster  cluster     disk     disk     pkts     pkts
  avg busy      ops  nfs-ops cifs-ops      ops spin-ops     recv     sent busy     recv     sent
busy     recv     sent     read    write     recv     sent
 ---- ---- -------- -------- -------- -------- -------- -------- -------- ---- -------- --------
------- -------- -------- -------- -------- -------- --------
Minimums:
   4%   5%        0        0        0        0        0   62.3KB   61.3KB   0%   1.28KB     708B
0%   60.7KB   60.7KB   15.8KB   15.8KB       64       58
Averages for 3 samples:
   4%   5%        0        0        0        0        0   63.4KB   62.4KB   0%   1.64KB     877B
0%   61.5KB   61.6KB    176KB    149KB       74       67
Maximums:
   5%   5%        0        0        0        0        0   65.3KB   64.4KB   0%   2.22KB   1.13KB
0%   62.7KB   63.2KB    489KB    407KB       91       83
```

Real-time SVM-level statistics show-periodic for NFSv3 read and write operations

```
cluster::*> statistics show-periodic -instance SVM -interval 2 -iterations 0 -summary true -
vserver SVM -object nfsv3 -counter nfsv3_ops|nfsv3_read_ops|nfsv3_write_ops
cluster: nfsv3.SVM: 11/30/2016 13:29:57
             nfsv3    nfsv3
    nfsv3     read    write    Complete    Number of
      ops      ops      ops Aggregation Constituents
 -------- -------- -------- ----------- ------------
     2360        0      697      Yes      16
     2245        0      652      Yes      16
     2126        0      629      Yes      16
cluster: nfsv3.SVM: 11/30/2016 13:30:04
             nfsv3    nfsv3
    nfsv3     read    write    Complete    Number of
      ops      ops      ops Aggregation Constituents
 -------- -------- -------- ----------- ------------
Minimums:
     2126        0      629       -       -
Averages for 3 samples:
     2243        0      659       -       -
Maximums:
     2360        0      697       -       -
```

Real-time FlexGroup local and remote statistics

```
cluster::*> statistics show-periodic -instance 0 -interval 2 -iterations 0 -summary true -object
flexgroup -counter
cat1_tld_local|cat1_tld_remote|cat2_hld_local|cat2_hld_remote|cat3_dir_local|cat3_dir_remote|cat4
_fil_local|cat4_fil_remote
cluster: flexgroup.0: 11/30/2016 13:34:55
     cat1     cat1     cat2     cat2     cat3     cat3     cat4     cat4
      tld      tld      hld      hld      dir      dir      fil      fil    Complete    Number of
    local   remote    local   remote    local   remote    local   remote Aggregation Constituents
 -------- -------- -------- -------- -------- -------- -------- -------- ----------- ------------
        1        0       17      113        0        0      619        0      n/a      n/a
        0        1       17      114        0        0      654        0      n/a      n/a
        0        2       17      112        0        0      647        0      n/a      n/a
cluster: flexgroup.0: 11/30/2016 13:35:02
     cat1     cat1     cat2     cat2     cat3     cat3     cat4     cat4
      tld      tld      hld      hld      dir      dir      fil      fil    Complete    Number of
    local   remote    local   remote    local   remote    local   remote Aggregation Constituents
 -------- -------- -------- -------- -------- -------- -------- -------- ----------- ------------
Minimums:
        0        0       17      112        0        0      619        0       -        -
Averages for 3 samples:
        0        1       17      113        0        0      640        0       -        -
Maximums:
```



```
        1          2         17       114          0          0       654          0         -         -
```

Example of creating a FlexGroup volume and specifying fewer member volumes than the default value

This command creates a 10TB FlexGroup volume with two 5TB member volumes across two nodes.

```
cluster::> volume create -vserver DEMO -volume flexgroup -aggr-list aggr1_node1,aggr1_node2 -
aggr-list-multiplier 1 -junction-path /flexgroup -size 10t

Warning: The FlexGroup "flexgroup" will be created with the following number of constituents of
size 5TB: 2.

Do you want to continue? {y|n}: y
```

Note: You must use the `-aggr-list` flag to make sure that the volume is a FlexGroup volume.

Sample REST API for creating a FlexGroup volume

The following REST API example creates a 2TB, eight-member thin-provisioned FlexGroup volume across a single aggregate.


```
{
  "aggregates": [
    {
      "name": "aggr1_node1"
    }
  ],
  "constituents_per_aggregate": 8,
  "efficiency": {
    "compaction": "inline",
    "compression": "inline",
    "cross_volume_dedupe": "inline",
    "dedupe": "inline"
  },
  "guarantee": {
    "type": "none"
  },
  "name": "RESTAPI_FG",
  "nas": {
    "export_policy": {
      "id": 42949672961,
      "name": "default"
    },
    "gid": 0,
    "path": "/RESTAPI_FG",
    "security_style": "unix",
    "uid": 0,
    "unix_permissions": 755
  },
  "size": "2T",
  "style": "flexgroup",
  "svm": {
    "name": "DEMO",
    "uuid": "7e3cc08e-d9b3-11e6-85e2-00a0986b1210"
  }
}
```


This is what the FlexGroup looks like after it is created:

```
cluster::*> vol show -vserver DEMO -volume REST*
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
DEMO      RESTAPI_FG   -            online     RW          2TB     1.90TB    0%
DEMO      RESTAPI_FG__0001
                       aggr1_node1  online     RW        256GB    243.1GB    0%
DEMO      RESTAPI_FG__0002
```



```
                         aggr1_node1  online     RW         256GB    243.1GB    0%
DEMO      RESTAPI_FG__0003
                         aggr1_node1  online     RW         256GB    243.1GB    0%
DEMO      RESTAPI_FG__0004
                         aggr1_node1  online     RW         256GB    243.1GB    0%
DEMO      RESTAPI_FG__0005
                         aggr1_node1  online     RW         256GB    243.1GB    0%
DEMO      RESTAPI_FG__0006
                         aggr1_node1  online     RW         256GB    243.1GB    0%
DEMO      RESTAPI_FG__0007
                         aggr1_node1  online     RW         256GB    243.1GB    0%
DEMO      RESTAPI_FG__0008
                         aggr1_node1  online     RW         256GB    243.1GB    0%
9 entries were displayed.
```

To include more than one aggregate in the list, use this REST API as an example:


```
{
  "aggregates": [
    { "name": "aggr1_node1" },{ "name": "aggr1_node2" }
  ],
  "efficiency": {
    "compaction": "inline",
    "compression": "inline",
    "cross_volume_dedupe": "inline",
    "dedupe": "inline"
  },
  "guarantee": {
    "type": "none"
  },
  "name": "RESTAPI_FG3",
  "nas": {
    "export_policy": {
      "id": 42949672961,
      "name": "default"
    },
    "gid": 0,
    "path": "/RESTAPI_FG3",
    "security_style": "unix",
    "uid": 0,
    "unix_permissions": 755
  },
  "size": "2T",
  "style": "flexgroup",
  "svm": {
    "name": "DEMO",
    "uuid": "7e3cc08e-d9b3-11e6-85e2-00a0986b1210"
  }
}
```


This is how it looks:

```
cluster::*> vol show -vserver DEMO -volume *FG3*
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
DEMO      RESTAPI_FG3  -            online     RW          2TB     1.90TB    0%
DEMO      RESTAPI_FG3__0001
                       aggr1_node1  online     RW        256GB    243.1GB    0%
DEMO      RESTAPI_FG3__0002
                       aggr1_node2  online     RW        256GB    243.1GB    0%
DEMO      RESTAPI_FG3__0003
                       aggr1_node1  online     RW        256GB    243.1GB    0%
DEMO      RESTAPI_FG3__0004
                       aggr1_node2  online     RW        256GB    243.1GB    0%
DEMO      RESTAPI_FG3__0005
                       aggr1_node1  online     RW        256GB    243.1GB    0%
DEMO      RESTAPI_FG3__0006
                       aggr1_node2  online     RW        256GB    243.1GB    0%
DEMO      RESTAPI_FG3__0007
                       aggr1_node1  online     RW        256GB    243.1GB    0%
```



```
DEMO      RESTAPI_FG3__0008
                       aggr1_node2  online     RW        256GB    243.1GB    0%
9 entries were displayed.
```

This REST API creates a four-member FlexGroup volume y using the `style` option and does not specify the `constituents_per_aggregate` option.


```
{
  "aggregates": [
    {
      "name": "aggr1_node1"
    }
  ],
  "efficiency": {
    "compaction": "inline",
    "compression": "inline",
    "cross_volume_dedupe": "inline",
    "dedupe": "inline"
  },
  "guarantee": {
    "type": "none"
  },
  "name": "RESTAPI_FG2",
  "nas": {
    "export_policy": {
      "id": 42949672961,
      "name": "default"
    },
    "gid": 0,
    "path": "/RESTAPI_FG2",
    "security_style": "unix",
    "uid": 0,
    "unix_permissions": 755
  },
  "size": "2T",
  "style": "flexgroup",
  "svm": {
    "name": "DEMO",
    "uuid": "7e3cc08e-d9b3-11e6-85e2-00a0986b1210"
  }
}
```


And this is the resulting FlexGroup volume:

```
cluster::*> vol show -vserver DEMO -volume RESTAPI_FG2*
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
DEMO      RESTAPI_FG2  -            online     RW          2TB     1.90TB    0%
DEMO      RESTAPI_FG2__0001
                       aggr1_node1  online     RW        512GB    486.3GB    0%
DEMO      RESTAPI_FG2__0002
                       aggr1_node1  online     RW        512GB    486.3GB    0%
DEMO      RESTAPI_FG2__0003
                       aggr1_node1  online     RW        512GB    486.3GB    0%
DEMO      RESTAPI_FG2__0004
                       aggr1_node1  online     RW        512GB    486.3GB    0%
5 entries were displayed.
```

Example of increasing a FlexGroup volume’s size

```
cluster::*> volume show -vserver SVM -volume flexgroup*
SVM       flexgroup    -            online     RW      70.20TB    10.14TB   85%
SVM       flexgroup__0001
                       aggr1_node1  online     RW         10TB     5.06TB   49%
SVM       flexgroup__0002
                       aggr1_node2  online     RW         10TB     5.08TB   49%
SVM       flexgroup__0003
                       aggr1_node1  online     RW         10TB     5.06TB   49%
SVM       flexgroup__0004
```



```
                      aggr1_node2  online     RW         10TB     5.08TB   49%
SVM       flexgroup__0005
                      aggr1_node1  online     RW         10TB     5.06TB   49%
SVM       flexgroup__0006
                      aggr1_node2  online     RW         10TB     5.08TB   49%
SVM       flexgroup__0007
                      aggr1_node1  online     RW         10TB     5.06TB   49%
SVM       flexgroup__0008
                      aggr1_node2  online     RW         10TB     5.08TB   49%

cluster::*> vol size -vserver SVM -volume flexgroup -new-size 100t
vol size: Volume "SVM:flexgroup" size set to 100t.

cluster::*> volume show -vserver SVM -volume flexgroup*
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
SVM       flexgroup    -            online     RW        100TB    10.14TB   89%

SVM       flexgroup__0001
                      aggr1_node1  online     RW      12.50TB     5.06TB   59%
SVM       flexgroup__0002
                      aggr1_node2  online     RW      12.50TB     5.08TB   59%
SVM       flexgroup__0003
                      aggr1_node1  online     RW      12.50TB     5.06TB   59%
SVM       flexgroup__0004
                      aggr1_node2  online     RW      12.50TB     5.08TB   59%
SVM       flexgroup__0005
                      aggr1_node1  online     RW      12.50TB     5.06TB   59%
SVM       flexgroup__0006
                      aggr1_node2  online     RW      12.50TB     5.08TB   59%
SVM       flexgroup__0007
                      aggr1_node1  online     RW      12.50TB     5.06TB   59%
SVM       flexgroup__0008
                      aggr1_node2  online     RW      12.50TB     5.08TB   59%
```

Example of expanding a FlexGroup volume

```
cluster::*> volume show -vserver SVM -volume flexgroup4*
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
SVM       flexgroup4TB -            online     RW          4TB     3.78TB    5%
SVM       flexgroup4TB__0001
                      aggr1_node1  online     RW        512GB    485.5GB    5%
SVM       flexgroup4TB__0002
                      aggr1_node2  online     RW        512GB    481.2GB    6%
SVM       flexgroup4TB__0003
                      aggr1_node1  online     RW        512GB    481.5GB    5%
SVM       flexgroup4TB__0004
                      aggr1_node2  online     RW        512GB    485.5GB    5%
SVM       flexgroup4TB__0005
                      aggr1_node1  online     RW        512GB    485.5GB    5%
SVM       flexgroup4TB__0006
                      aggr1_node2  online     RW        512GB    485.5GB    5%
SVM       flexgroup4TB__0007
                      aggr1_node1  online     RW        512GB    485.5GB    5%
SVM       flexgroup4TB__0008
                      aggr1_node2  online     RW        512GB    485.5GB    5%

cluster::*> volume expand -vserver SVM -volume flexgroup4TB -aggr-list aggr1_node1,aggr1_node2 -
aggr-list-multiplier 4


cluster::*> volume show -vserver SVM -volume flexgroup4*
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
SVM       flexgroup4TB -            online     RW          8TB     7.78TB    1%
SVM       flexgroup4TB__0001
                      aggr1_node1  online     RW        512GB    485.5GB    1%
SVM       flexgroup4TB__0002
                      aggr1_node2  online     RW        512GB    481.2GB    1%
```



```
SVM       flexgroup4TB__0003
                         aggr1_node1  online     RW        512GB    481.5GB    1%
SVM       flexgroup4TB__0004
                         aggr1_node2  online     RW        512GB    485.5GB    1%
SVM       flexgroup4TB__0005
                         aggr1_node1  online     RW        512GB    485.5GB    1%
SVM       flexgroup4TB__0006
                         aggr1_node2  online     RW        512GB    485.5GB    1%
SVM       flexgroup4TB__0007
                         aggr1_node1  online     RW        512GB    485.5GB    1%
SVM       flexgroup4TB__0008
                         aggr1_node2  online     RW        512GB    485.5GB    1%
SVM       flexgroup4TB__0009
                         aggr1_node1  online     RW        512GB    485.5GB    1%
SVM       flexgroup4TB__0010
                         aggr1_node2  online     RW        512GB    481.2GB    1%
SVM       flexgroup4TB__0011
                         aggr1_node1  online     RW        512GB    481.5GB    1%
SVM       flexgroup4TB__0012
                         aggr1_node2  online     RW        512GB    485.5GB    1%
SVM       flexgroup4TB__00013
                         aggr1_node1  online     RW        512GB    485.5GB    1%
SVM       flexgroup4TB__0014
                         aggr1_node2  online     RW        512GB    485.5GB    1%
SVM       flexgroup4TB__0015
                         aggr1_node1  online     RW        512GB    485.5GB    1%
SVM       flexgroup4TB__0016
                         aggr1_node2  online     RW        512GB    485.5GB    1%
```

Other command-line examples

Create a FlexGroup volume by using `flexgroup deploy`:

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

Use the ONTAP 9.2 `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
```

Create 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
```

Modify the FlexGroup Snapshot policy:

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

Apply storage QoS:

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

Apply 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.
```


Where to find additional information

To learn more about the information described in this document, refer to the following documents:

Technical reports

- TR-4067: NFS Best Practice and Implementation Guide
  http://www.netapp.com/us/media/tr-4067.pdf
- TR-4100: Nondisruptive Operations with SMB File Shares
  http://www.netapp.com/us/media/tr-4100.pdf
- TR-4476: NetApp Data Compression, Deduplication, and Data Compaction
  http://www.netapp.com/us/media/tr-4476.pdf
- TR-4523: DNS Load Balancing in ONTAP
  http://www.netapp.com/us/media/tr-4523.pdf
- TR-4568: FabricPool Best Practices
  http://www.netapp.com/us/media/tr-4568.pdf
- TR-4570: NetApp Storage Solutions for Apache Spark
  http://www.netapp.com/us/media/tr-4570.pdf
- TR-4571-a: NetApp ONTAP FlexGroup Volumes: Top Best Practices
  http://www.netapp.com/us/media/tr-4571-a.pdf
- TR-4597: VMware vSphere with ONTAP
  https://www.netapp.com/us/media/tr-4597.pdf
- TR-4616: NFS Kerberos in ONTAP
  http://www.netapp.com/us/media/tr-4616.pdf
- TR-4617: Electronic Design Automation Best Practices
  http://www.netapp.com/us/media/tr-4617.pdf
- TR-4668: Name Services Best Practices Guide (ONTAP 9.3 and later)
  http://www.netapp.com/us/media/tr-4668.pdf
- TR-4678: Data Protection Best Practices with FlexGroup Volumes
  https://www.netapp.com/us/media/tr-4678.pdf
- TR-4705: NetApp MetroCluster Solution Design and Architecture
  https://www.netapp.com/us/media/tr-4705.pdf
- TR-4740: SMB 3.0 Multichannel
  https://www.netapp.com/media/17136-tr4740.pdf
- TR-4743: NetApp FlexCache Volumes in ONTAP 9.7
  https://www.netapp.com/us/media/tr-4743.pdf
- TR-4808: NetApp XCP Best Practices
  https://www.netapp.com/us/media/tr-4808.pdf
- TR-4835: How to Configure LDAP in ONTAP
  https://www.netapp.com/us/media/tr-4835.pdf
- TR-4867: Best Practice Guide for File System Analytics
  https://www.netapp.com/pdf.html?item=/media/20707-tr-4867.pdf
- TR-4863: Best Practice Guidelines for XCP
  https://www.netapp.com/media/20029-tr-4863.pdf

Miscellaneous content

- Tech ONTAP Podcast Episode 46: FlexGroups
  https://soundcloud.com/techontap_podcast/episode-46-flexgroups-1

- Tech ONTAP Podcast Episode 188: FlexGroup Update
https://soundcloud.com/techontap_podcast/episode-188-flexgroup-update
- Tech ONTAP Podcast Episode 219: FlexVol to FlexGroup Conversion
https://soundcloud.com/techontap_podcast/episode-219-flexvol-to-flexgroup-conversion
- Tech ONTAP Podcast Episode 270: File System Analytics
https://soundcloud.com/techontap_podcast/episode-270-netapp-ontap-file-systems-analytics
- What's New For FlexGroup Volumes in ONTAP 9.3?
https://blog.netapp.com/whats-new-for-netapp-flexgroup-volumes-in-ontap-9-3/
- FlexGroup Volumes: An Evolution of NAS
https://newsroom.netapp.com/blogs/netapp-flexgroup-volumes-an-evolution-of-nas/
- 7 Myths about NetApp ONTAP 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

Version history

Version	Date	Document Version History
Version 1.0	January 2017	Initial release
Version 1.0.1	February 2017	Minor revisions
Version 1.1	May 2017	ONTAP 9.2RC1
Version 1.2	December 2017	ONTAP 9.3GA
Version 1.3	May 2018	ONTAP 9.4RC1
Version 1.4	November 2018	ONTAP 9.5
Version 1.5	June 2019	ONTAP 9.6
Version 1.6	January 2020	ONTAP 9.7
Version 1.7	January 2021	ONTAP 9.8
Version 1.8	May 2021	ONTAP 9.9.1
Version 1.9	October 2021	ONTAP 9.10.1

Acknowledgments

We would like to thank Chad Morgenstern, Richard Jernigan, Dan Tennant, Ken Cantrell, Sunitha Rao, Shriya Paramkusam, and Mrinal Devadas for their help preparing this report.

.

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.

TR-4571-1021

NetApp