Abstract

This document highlights best practices and implementation tips for using NetApp® ONTAP® with Electronic Design Automation (EDA) workloads. It also calls attention to NetApp ONTAP FlexGroup volumes, which are ideal for handling the high metadata overhead in EDA environments.
# TABLE OF CONTENTS

1 Overview.................................................................................................................. 5  
   1.1 NetApp ONTAP........................................................................................................... 5  
   1.2 NetApp ONTAP FlexGroup Volumes........................................................................ 5  
   1.3 Electronic Design Automation (EDA)...................................................................... 7  

2 Performance ............................................................................................................... 8  
   2.1 Git Workload Tests ................................................................................................. 8  
   2.2 FlexGroup Versus Scale-Out NAS Competitor: Do More with Less..................... 10  
   2.3 ONTAP 9.4 and 9.5 Performance Testing.............................................................. 11  
   2.4 Automatic Workload Adaptation............................................................................ 15  
   2.5 Preventing “Bully” Workloads with Quality of Service (QoS)............................... 16  

3 ONTAP Best Practices for EDA ................................................................................. 18  
   3.1 Hardware Considerations....................................................................................... 18  
   3.2 Aggregate Layout Considerations.......................................................................... 18  
   3.3 Networking Considerations................................................................................... 19  
   3.4 Volume Considerations......................................................................................... 22  
   3.5 High File Count/Inode Count Considerations....................................................... 32  
   3.6 Project Tiering Considerations................................................................................ 48  
   3.7 Security and ACL Style Considerations............................................................... 52  
   3.8 NFS Considerations............................................................................................... 53  
   3.9 NetApp Volume Encryption Considerations....................................................... 56  
   3.10 FlexCache Volume Considerations.................................................................... 56  

4 Interoperability .......................................................................................................... 57  

5 Migrating to ONTAP FlexGroup Volumes ................................................................. 57  
   5.1 NDMP ................................................................................................................... 58  
   5.2 FlexVol to FlexGroup Conversion......................................................................... 59  
   5.3 Migrating from Third-Party Storage to FlexGroup............................................... 60  
   5.4 Migrating from Data ONTAP Operating in 7-Mode to FlexGroup....................... 60  
   5.5 Migrating from SAN LUNs or Infinite Volume in ONTAP to FlexGroup.............. 60  
   5.6 XCP Migration Tool ............................................................................................. 61  

6 Additional Resources ................................................................................................ 61  

7 Acknowledgments ...................................................................................................... 62  
   7.1 Disclaimer............................................................................................................... 62
LIST OF BEST PRACTICES

Best Practice 1: Aggregate Usage with FlexGroup and Multiple FlexVol Volumes .................................................. 19
Best Practice 2: Network Design with FlexGroup ........................................................................................................... 19
Best Practice 3: Member Volume Count Recommendations .......................................................................................... 29
Best Practice 4: Storage Efficiency Recommendation in EDA Workloads ................................................................. 32
Best Practice 5: 64-Bit File Identifiers .......................................................................................................................... 37
Best Practice 6: Recommended ONTAP Version for High-File-Count Environments ................................................... 40
Best Practice 7: Avoiding maxdirsize Issues .................................................................................................................... 41
Best Practice 8: Directory Structure Recommendation ................................................................................................ 42
Best Practice 9: Maxdirsize Maximums .......................................................................................................................... 44
Best Practice 10: Special Character Handling in FlexGroup Volumes .............................................................................. 45
Best Practice 11: UTF-8 or utf8mb4? ............................................................................................................................... 46
Best Practice 12: SMB Change Notification Recommendation ........................................................................................ 46
Best Practice 13: Volume Security Style ........................................................................................................................ 53
Best Practice 14: NFS Version Considerations for EDA Environments ............................................................................. 53
1 Overview

1.1 NetApp ONTAP

NetApp ONTAP is a data-management software solution that offers the following advantages:

- **Performance.** Scale your cluster up by adding larger, stronger nodes. Scale your cluster out by providing more compute and capacity to electronic design automation (EDA) workloads for which you can rapidly increase the number of nodes. ONTAP allows you to grow your performance needs as your application grows, while providing a single namespace that can deliver millions of IOPS for your workloads.

- **Flexibility.** Provision unified storage with SAN and NAS connectivity with the ability to standardize data management across flash, disk, and the cloud. Deploy ONTAP on NetApp hardware with software-defined solutions such as NetApp ONTAP Select, or deploy in the cloud with NetApp Cloud Volumes ONTAP. ONTAP provides data access anywhere and anytime.

- **Resiliency and high availability.** Use patented RAID technologies, including triple erasure coding (NetApp RAID-TEC™), for extra protection against drive failures, particularly with larger drives that have longer rebuild times. Also, high-availability controller failover in ONTAP allows minimal downtime if planned or unplanned outages occur. Storage stacks are also connected by using the latest in multipath I/O technology for both redundancy and performance.

- **Scalability.** High-availability (HA) pairs can be clustered together to form a single NAS namespace or SAN target. Scale up by adding larger controllers and more disk to storage stacks to provide better capacity and performance. Scale out by adding more heads and disk to existing clusters nondisruptively. Deploy massive storage containers with ONTAP FlexGroup volume technology. Automate load and performance balancing with the latest ONTAP releases and their feature sets.

- **Efficiency.** Deliver multiple efficiency features to allow administrators to squeeze the most out of their existing storage. Inline deduplication, compression, and data compaction provide methods to shrink your data footprint as data is ingested. Inline aggregate deduplication allows deduplication across multiple volumes in the same aggregate. Thin provisioning and NetApp FlexClone® technology offers administrators flexible storage utilization without taking up valuable capacity.

- **Security.** Implement up-to-date security enhancements, including the following technologies:
  - FIPS 140-2-compliant data encryption technologies
  - AES-256 encryption for Kerberos in SMB and NFS
  - Transport Layer Security (TLS) 1.2 for NetApp SnapMirror® and NetApp FlexCache™
  - NetApp Volume Encryption (NVE)
  - NetApp Aggregate Encryption (NAE)
  - Self-encrypting drives (SEDs)
  - NetApp Storage Encryption (NSE) for encryption at rest

1.2 NetApp ONTAP FlexGroup Volumes

NetApp FlexVol® volumes have traditionally been a good fit for EDA workloads. However, as hard-drive costs decrease and flash hard-drive capacity grows exponentially, file systems are following suit. The days of file systems that number in the tens of gigabytes are over. Storage administrators face increasing demands from application owners for large buckets of capacity and millions to billions of files delivered with enterprise-level performance.

With this massive data growth, in which storage needs for a single namespace can extend into the petabyte range, an evolution of NAS file systems is overdue. ONTAP FlexGroup technology is the ideal solution for these architectures.

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 the
physical limits of hardware or the total volume limits of ONTAP. Limits are determined by the overall number of constituent member volumes that collaborate to dynamically balance load and space allocation evenly across all members. There is no required maintenance or management overhead with a FlexGroup volume. You simply create the FlexGroup volume and share it with your NAS clients. ONTAP does the rest.

Figure 1) Evolution of NAS file systems in ONTAP.

Advantages of FlexGroup Volumes

ONTAP FlexGroup volumes provide 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 creation events 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 lets you manage FlexGroup volumes like FlexVol volumes. ONTAP handles underlying member volume creation and balance across the cluster nodes and provides a single access point for NAS shares.

Superior Density for EDA Workloads

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 capabilities:

- 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.

1.3 Electronic Design Automation

EDA workloads present a unique set of challenges to storage systems, mainly due to the massive capacity, high-file-count, heavy metadata operations, and high performance requirements for manufacturers that need to ship products to stay competitive. Simplicity and usability in these environments are a must, because administrators need to focus on supporting the application and its users, rather than managing complex storage architectures. ONTAP can help address the challenges EDA workloads have with a multifaceted approach.

Figure 2) EDA workload types.

- **Capacity.** FlexVol volumes provide up to 100TB of space in a single container. However, EDA workloads might need to surpass that amount. FlexGroup volumes provide a multipetabyte container for EDA workloads over multiple NAS protocols. These volumes can scale up or scale out nondisruptively as the dataset grows; there is no need for maintenance windows.

- **High-file-count environments.** FlexVol volumes support up to two billion files in a single container. Sometimes that amount might not be enough. EDA file-system layouts can contain thousands of files per directory, with deep directory structures. FlexGroup volumes can scale across multiple nodes in a cluster to offer file counts in the hundreds of billions.

- **Performance.** FlexGroup volumes provide multithreaded parallel operations for high-file-count, metadata-heavy workloads, such as EDA. By spreading the ingest load across multiple cluster resources, FlexGroup volumes can deliver high throughput and IOPS at predictable, low latencies with near-linear scale. If you need to scale out your performance or capacity, you simply add more nodes to the cluster, nondisruptively. In addition, using ONTAP with NetApp AFF systems can provide performance and density for EDA workloads. For more information about FlexGroup volume performance, see the “Performance” section. For a performance validation of EDA workloads on AFF systems, see TR-4324.

- **Simplicity.** A FlexGroup volume blends capacity, high file count handling, and performance with a simple, easy-to-deploy container under a single NAS namespace. Ingest of data and balancing of
load is handled automatically through the ONTAP subsystems used by FlexGroup volumes, so you don’t have to worry about whether data is being placed locally or remotely. For more information, see TR-4557: NetApp ONTAP FlexGroup Volumes—A Technical Overview.

2 Performance

2.1 Git Workload Tests

NetApp ONTAP FlexGroup volumes can also handle Git workloads well, as evidenced by the following graphs.

The following configuration was used:

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

The workload was as follows:

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

Figure 3 compares the maximum achieved throughput (read and write) on Git clone operations on a single FlexVol volume versus a single FlexGroup spanning two nodes. The graph shows how the maximum throughput reaches nearly five times the amount of the FlexVol volume without the same degradation that FlexVol sees as the workload reaches 64 threads.

Figure 3) FlexVol versus FlexGroup: maximum throughput trends under increasing workload.
Figure 4 compares FlexVol and FlexGroup volumes in the same configurations. This time, we break down the maximum read and write throughput individually, as well as comparing that against the average throughput for the FlexVol and FlexGroup volumes.

Figure 4) FlexVol versus FlexGroup: maximum throughput trends under increasing workload—detailed.

Figure 5 shows the maximum total average IOPS for a FlexGroup volume versus a FlexVol volume on the AFF A700 system. Again, note the drastic increase of IOPS for the FlexGroup volume versus the degradation of IOPS at 64 threads for the FlexVol volume.
2.2 FlexGroup Versus Scale-Out NAS Competitor: Do More with Less

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

In the test, we saw that a single FlexGroup volume with eight member constituents was able to ingest nearly the same number of operations per second at essentially the same latency curve as the competitor’s 14-node cluster.
2.3 ONTAP 9.4 and 9.5 Performance Testing

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

The ONTAP 9.4 and 9.5 tests featured the following configurations:

- A NetApp AFF A700s cluster
- A FlexGroup volume spanning a single node and two nodes
- 14 NFSv3 clients
- 32 10GB logical interfaces, or 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 7) Standard NAS benchmark (EDA)—ONTAP 9.5: one node versus two nodes (operations/sec).

Figure 8) Standard NAS benchmark (EDA)—ONTAP 9.4 versus ONTAP 9.5 (operations/sec).
Figure 9) Standard NAS benchmark (EDA)—ONTAP 9.5: one node versus two nodes (MBps).

![Standard NAS Benchmark (EDA)
ONTAP 9.5: One node vs. Two nodes](image)

Figure 10) Standard NAS benchmark (EDA)—ONTAP 9.4 versus ONTAP 9.5 (MBps).

![Standard NAS Benchmark (EDA)
ONTAP 9.4 vs. ONTAP 9.5](image)

Figure 11 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 11) Standard NAS benchmark (software builds)—ONTAP 9.5 (operations/sec).

Figure 12) Standard NAS benchmark (software builds)—ONTAP 9.4 versus ONTAP 9.5 (operations/sec).
2.4 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 reach 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 single 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 storage virtual machine (SVM).

ONTAP 9.7 introduces ingest algorithm enhancements to recognize workload types (high metadata ingest versus streaming I/O). 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 enable even space allocation. This allows a wider variety of workloads to perform optimally on FlexGroup volumes, preventing space imbalance scenarios and reducing the need for administrator intervention.

**Ingest Algorithm Improvements**

Every ONTAP release further improves the ingest algorithms for FlexGroup volumes. These algorithms 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.

Therefore, NetApp strongly recommends that you run the latest patched ONTAP version when using FlexGroup volumes. You can download the latest release at [NetApp Support for ONTAP 9](https://www.netapp.com/support/ontap/).  

### 2.5 Preventing Bully Workloads with Quality of Service (QoS)

Starting in ONTAP 9.3, you can apply maximum storage QoS policies to help prevent a FlexGroup volume from acting as a bully workload in ONTAP. Storage QoS can help you manage risks around meeting your performance objectives. You use storage QoS to limit the throughput to workloads and to monitor workload performance. You can reactively limit workloads to address performance problems, and you can proactively limit workloads to prevent performance problems. For more information about storage QoS, see [TR-4211: Storage Performance Primer](https://www.netapp.com/support/library/tr-4211).

**How Storage QoS Maximums Work with FlexGroup**

With FlexGroup, storage QoS policies are applied to the entire FlexGroup volume. 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 15 and Figure 16 show how storage QoS gets applied to a FlexGroup volume.
Storage QoS Considerations with FlexGroup Volumes

Storage QoS is applied at the FlexGroup volume level. File-level QoS and nested policies are currently not supported with FlexGroup volumes.

QoS Minimums

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. This feature is supported for AFF systems only. For details on the feature, see TR-4211: Storage Performance Primer.

Adaptive QoS

ONTAP 9.4 also 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. This feature is covered in detail in TR-4211: Storage Performance Primer.
3 ONTAP Best Practices for EDA

This section covers NetApp ONTAP best practices for EDA environments. Although ONTAP FlexGroup volumes are a more natural fit for the type of workloads EDA throws at a storage system, this section also covers NetApp FlexVol volumes, because FlexGroup volumes might be missing features or functionality needed for specific EDA environments.

3.1 Hardware Considerations

EDA workloads perform best when the following storage hardware conditions exist:

- Large memory and RAM footprint
- Greater number of CPU cores for concurrent processing
- Large capacities

Therefore, NetApp highly recommends considering higher-end, all-flash storage systems, such as the NetApp AFF A-Series or AFF8xxx, to maximize the available RAM and CPU in each node. For project tiering, hot data workloads should reside on AFF systems. See the section “Project Tiering Considerations.” Cool and cold data workloads can reside on any platform and media type. The following table shows the CPU and RAM for AFF systems intended for hot data workloads. For more information, see the official hardware specifications on NetApp.com.

Table 1) NetApp all-flash system CPU and RAM per HA pair.

<table>
<thead>
<tr>
<th>System</th>
<th>CPU</th>
<th>RAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFF A700</td>
<td>Four 18-core Broadwell</td>
<td>1024GB</td>
</tr>
<tr>
<td>AFF A300</td>
<td>Two 16-core Broadwell</td>
<td>256GB</td>
</tr>
<tr>
<td>FAS9000</td>
<td>Four 18-core Broadwell</td>
<td>1024GB</td>
</tr>
<tr>
<td>FAS8200</td>
<td>Two 16-core Broadwell</td>
<td>256GB</td>
</tr>
</tbody>
</table>

Being able to throw more memory and CPU at EDA workloads can have a positive effect on the completion times of these workloads, which can present a greater return on investment, because the money saved in build times can offset the costs of more expensive nodes.

**Note:** NetApp highly recommends engaging with the appropriate NetApp sales account team to evaluate your business requirements before architecting the cluster scale-out setup in your environment.

Other Hardware Considerations

For the most consistent level of performance when spinning media is in place, use NetApp Flash Cache™ (formerly PAM) cards or NetApp Flash Pool™ aggregates in a cluster for EDA workloads. Flash Cache cards are expected to provide the same performance benefits for FlexGroup volumes that they provide for FlexVol volumes. Using Flash Cache or Flash Pool with AFF systems will provide little to no benefit, because these systems are already running the most optimal flash configurations.

3.2 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.

Adding more drives or spindles to a RAID group or aggregate can help improve performance, whether it's flash or spinning media.
Best Practice 1: Aggregate Usage with FlexGroup and Multiple FlexVol Volumes

For consistent performance when using FlexGroup volumes or multiple FlexVol volumes, make sure that the FlexGroup-volume or multiple-FlexVol-volume design 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 of disk types or aggregates is not necessary.

Table 2 shows NetApp best-practice recommendations for aggregate layout when you use FlexGroup volumes or multiple FlexVol volume layouts.

Table 2) Best practices for aggregate layout with FlexGroup volumes or multiple FlexVol volumes.

<table>
<thead>
<tr>
<th>Spinning Disk or Hybrid Aggregates</th>
<th>AFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two aggregates per node</td>
<td>One aggregate per node</td>
</tr>
</tbody>
</table>

Note: Aggregates should have the same number of drives and RAID groups.

For aggregate creation, follow the recommended aggregate guidelines available in ONTAP 9.4 and later by using ONTAP System Manager. For more information, see Storage Recommendations for Creating Aggregates.

3.3 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 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 other 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 this situation, 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.

FlexGroup volumes are not immune to this line of thinking. Although FlexVol volumes can be distributed across multiple nodes in a cluster just like a FlexGroup volume, the network connection can still be a bottleneck.

Best Practice 2: Network Design with FlexGroup

When you design a FlexGroup solution, consider the following networking best practices:

- Create at least one data LIF per node per SVM to provide a data path to each node.
- When possible, use LACP ports to host data LIFs for throughput and failover considerations.
- When you mount clients, spread the TCP connections across cluster nodes evenly.
- For clients that do frequent mounts and unmounts, consider using on-box DNS to help balance the load.
- Follow the latest general networking best practices that are listed in TR-4191.

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 Open Systems Interconnection (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 for throughput and resiliency, but you should consider the complexity of maintaining LACP environments when you are making the decision.

**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 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. However, it might be prudent to create multiple data LIFs per node per SVM and 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.

**Network Connection Concurrency: NFSv3**

In addition to the preceding considerations, it’s 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. You can configure clients 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 will send 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 can manifest as performance issues—for example, extra latency and slower job completion times—that aren’t explained by 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. The counter helps you determine whether the number of exec contexts blocked by the overrun connection is incrementing.

To gather those statistics, run the following command:

```
statistics start -object cid -instance cid
```

Then, review the statistics over a period to see if they are incrementing.

```
statistics show -object cid -instance cid -counter execs_blocked_on_cid
```

**Example of RPC Slot Table Impact on Performance**

In the following example, we ran a script to create 18 million files across 180,000 subdirectories. This load generation was created from three clients to the same NFS mount. The goal was to generate enough
NFS operations with clients that had the default RPC slot table settings to cause ONTAP to enter a flow-control scenario. We then ran the same scripts again on the same clients—but with the RPC slot tables set to 128.

The result was that the default slot tables (65,536) generated 18 million `execs_blocked_on_cid` events and added 3ms of latency to the workload versus the run with the lower RPC slot table setting (128).

**Figure 17** Effect of RPC slot tables on NFSv3 performance.

![Diagram showing performance comparison between 65536 slots and 128 slots](image)

Although 3ms might not seem like a lot of latency, it can add up over millions of operations, considerably slowing down job completion.

**Resolving Issues with RPC Slot Tables**

ONTAP cannot control the number of slot tables a client sends per TCP connection for NFSv3 operations. Therefore, clients must be configured to limit the maximum slot tables sent through NFS to 128. This setting varies depending on the client OS version. For example, Red Hat Enterprise Linux uses the following command to change the values:

```
# echo "options sunrpc udp_slot_table_entries=64 tcp_slot_table_entries=16
tcp_max_slot_table_entries=16" >> /etc/modprobe.d/sunrpc.conf
```

These values generally can be changed on the fly, but might not take effect until a client remounts the NFS export.

**Note:** To learn how to properly change these values, contact the client vendor.

It is possible to get more performance out of a client’s NFS connectivity by connecting more mount points to different IP addresses in the cluster on the same client, but that approach can create complexity. For example, rather than mounting a volume at SVM:/volumename, you could create multiple mount points on the same client across different folders and IP addresses in the volume.

LIF1:/volumename/folder1
LIF2:/volumename/folder2
LIF3:/volumename/folder3

Another possible option is to use the `nconnect` option available for some Linux distributions that can perform multiplexing of NFSv3 over the same TCP connection. This option provides more available concurrent sessions and better overall performance.
Does the RPC Slot Table Limit Affect Other NAS Protocols?

RPC slot table limits only affect NFSv3 traffic.

- SMB clients use different connection methodologies for concurrency, such as SMB multichannel, SMB multiplex, and SMB credits. The SMB connection methodology depends on client-server configuration and protocol version. For example, SMB 1.0 uses SMB multiplex (mpx), whereas SMB2.x uses SMB credits.
- NFSv4.x clients do not use RPC slot tables—instead, they use state IDs and session tables to control the flow of concurrent traffic from clients.

Border Gateway Protocol: ONTAP 9.5 and Later

Starting in version 9.5, ONTAP supports the Border Gateway Protocol (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 for more efficient port utilization.

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

3.4 Volume Considerations

FlexVol or FlexGroup Volumes?

When you design an EDA storage solution, you must consider the type of volume to use. A FlexGroup volume can provide exceptional performance for high metadata workloads, such as those seen in EDA environments. In the past, FlexGroup volumes didn’t necessarily support all the features that FlexVol volumes did, but, as of ONTAP 9.7, FlexGroup volumes have nearly complete feature parity with FlexVol volumes. Therefore, for EDA workloads, choosing a FlexGroup volume usually makes the most sense.

The following table lists the features that might be pertinent to EDA workloads, and tells whether the features are present in FlexVol volumes, FlexGroup volumes, or both. If a feature is not listed here, review the FlexGroup technical reports listed in the section “Additional Resources” at the end of this document or email us at flexgroups-info@netapp.com.

Table 3) Feature comparison of FlexVol and FlexGroup volumes.

<table>
<thead>
<tr>
<th>Feature</th>
<th>FlexVol Support?</th>
<th>FlexGroup Support?</th>
<th>ONTAP Release with First FlexGroup Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>All storage efficiencies (thin provision, inline and postprocess efficiencies)</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.1 (Aggregate level/cross-volume deduplication available in ONTAP 9.2)</td>
</tr>
<tr>
<td>AFF</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>Triple erasure coding</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>SAN (iSCSI/FCP)</td>
<td>Yes</td>
<td>No</td>
<td>n/a</td>
</tr>
<tr>
<td>SMB/CIFS</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.1 (no SMB1; see TR-4571)</td>
</tr>
<tr>
<td>NFS</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.1 (NFSv3) ONTAP 9.7 (NFSv4.x/pNFS)</td>
</tr>
<tr>
<td>FlexClone</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.7</td>
</tr>
<tr>
<td>Feature</td>
<td>FlexVol Support?</td>
<td>FlexGroup Support?</td>
<td>ONTAP Release with First FlexGroup Support</td>
</tr>
<tr>
<td>------------------------------</td>
<td>------------------</td>
<td>--------------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>NetApp Snapshot™</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>NetApp SnapMirror</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>NetApp SnapVault®</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>NetApp SnapLock®</td>
<td>Yes</td>
<td>No</td>
<td>n/a</td>
</tr>
<tr>
<td>NDMP</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.7</td>
</tr>
<tr>
<td>NetApp SnapCenter®</td>
<td>Yes</td>
<td>No</td>
<td>n/a</td>
</tr>
<tr>
<td>SnapDiff</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.4</td>
</tr>
<tr>
<td>QoS (minimum and maximum)</td>
<td>Yes</td>
<td>Yes</td>
<td>Maximums in ONTAP 9.3, Minimums in ONTAP 9.5</td>
</tr>
<tr>
<td>Qtrees</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>Quota reporting</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.1 (user), ONTAP 9.3 (tree and user)</td>
</tr>
<tr>
<td>Quota enforcement</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.5</td>
</tr>
<tr>
<td>Antivirus</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.4</td>
</tr>
<tr>
<td>FPolicy</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.4</td>
</tr>
<tr>
<td>Volume move</td>
<td>Yes</td>
<td>Yes (member volume level)</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>NetApp Volume Encryption</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.2</td>
</tr>
<tr>
<td>Volume autogrow</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.3</td>
</tr>
<tr>
<td>Increasing volume size</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>Shrinking volume size</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.6</td>
</tr>
<tr>
<td>NetApp MetroCluster™</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.6</td>
</tr>
<tr>
<td>ONTAP Select</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.1</td>
</tr>
<tr>
<td>Cloud Volumes ONTAP</td>
<td>Yes</td>
<td>Yes</td>
<td>ONTAP 9.5</td>
</tr>
</tbody>
</table>

**Cluster Considerations**

An ONTAP cluster that uses only NAS functionality (CIFS/SMB and NFS) can expand up to 24 nodes (12 HA pairs). Each HA pair is a homogeneous system (that is, two 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 and six hybrid nodes for storage-tiering functionality.
Cluster Considerations: FlexVol Volumes

A FlexVol volume is the standard container used in ONTAP to serve data to clients. It spans a single node and aggregate. Metadata operations are performed serially for NAS environments, which means a single CPU is being used for FlexVol volumes for metadata. In EDA environments, this can affect system performance and reduce the amount of hardware in a cluster that can be effectively used for the EDA workload.

To maximize the effectiveness of a FlexVol volume in an EDA workload, use the following tips.

- **Create multiple FlexVol volumes per node, across multiple nodes.**
  
  An ONTAP cluster can be designed in a way to maximize the hardware available. By creating multiple FlexVol volumes on a node, you can take advantage of more available CPU threads for a workload. Extending those FlexVol volumes across the cluster’s nodes gives the workload even more hardware resources to work with. When this is done, the FlexVol volumes appear as folders to the clients that mount the global namespace.

  Figure 18) Example of junctioned FlexVol volumes.

If EDA workloads can be designed to direct data into individual folders, this can be a viable design to get the most performance out of ONTAP for EDA workloads. If the application requires a namespace that can’t point to specific folders, consider a FlexGroup volume.

- **When using multiple FlexVol volumes, create them on homogeneous hardware.**
  
  Using the same types of disk, same size of RAID groups, same number of spindles, same node types, and so on for multiple FlexVol volumes helps ONTAP maintain consistency of performance. To avoid surprises, be sure to span nodes and aggregates that are identical. The exception would be if the EDA workload is designed to use some form of project tiering. For example, active workloads could reside on flash storage, whereas inactive workloads could be tiered off to spinning disk through nondisruptive volume moves or to disaster recovery sites through SnapMirror. For more information, see “Project Tiering Considerations.”

  Additionally, Snapshot copies, disaster recovery destination data, or inactive data could be tiered to Amazon Simple Storage Service (Amazon S3) through FabricPool. For more information about FabricPool, see TR-4598: FabricPool Best Practices.

- **Create a local data LIF per node and make sure that clients mount volumes local to the owning node.**
  
  FlexVol volumes are owned by aggregates, which in turn are owned by nodes in a cluster. Clients can access any volume in a cluster by using any of the SVM’s data LIFs. If an SVM has a single data LIF, but volumes that live on multiple nodes, then some of the cluster traffic ends up being remote.
Although this scenario is generally fine, it can introduce latency into NAS requests. EDA workloads can be sensitive to latency, so it’s best to avoid remote access to FlexVol volumes when possible. Having a data LIF per node, per SVM allows clients to mount to the local path and receive the performance benefits of accessing a volume locally. For more information, see the section “Networking Considerations.”

Figure 19) Local traffic versus remote traffic in a cluster.

With NAS protocols, ONTAP supports features such as CIFS autolocation, NFSv4.x referrals, and pNFS to help promote data locality in a clustered file system. For more information about those features, refer to the product documentation for your version of ONTAP.

- **Enable QoS for performance monitoring and throttling.**
  ONTAP enables you to collect statistics through QoS and to limit workloads at a volume, qtree, or file level to prevent bully workloads from affecting other workloads. For more information about QoS, see TR-4211: Storage Performance Primer.

**Cluster Considerations: FlexGroup Volumes**

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

- **FlexGroup volumes should span only hardware systems that are identical.**
  Because hardware systems in the same cluster can vary greatly in terms of CPU, RAM, and overall performance capabilities, the use of only homogeneous systems helps promote predictable performance across a FlexGroup volume.

- **FlexGroup volumes should span only disk types that are identical.**
  Like hardware systems, disk-type performance can vary greatly. For best results, make sure that the aggregates that are used are either all SSD, all spinning, or all hybrid.

- **FlexGroup volumes should use as many nodes in a cluster as possible.**
  A FlexGroup volume can be configured to span any node in the cluster, from a single node to all 24 nodes. You do not have to configure the FlexGroup volume to span the entire cluster, but doing so can take greater advantage of the hardware resources that are available, because more hardware resources provide better workload balance and resource utilization.
• **FlexGroup volumes should have multiple data network access points.**
  Although FlexGroup volumes always use the back-end cluster network, the parallelism of workloads that a FlexGroup volume offers outweigh the effect of remote cluster traffic. However, data network resources are still finite, so it’s important to spread network connections across all the participating nodes in a FlexGroup volume. When possible, to offer more even distribution of resources, create network interfaces on each node that has portions of the FlexGroup present.

### FlexVol Volume Layout Considerations
If you’re using multiple FlexVol volumes in a cluster for EDA workloads, consider the following recommendations:

• Balance the FlexVol volumes across nodes evenly. ONTAP 9.2 and later versions offer automatically balanced provisioning to place new volumes on nodes to help balance workloads.

• Create multiple FlexVol volumes per node to take advantage of volume affinities available in ONTAP to maximize CPU usage. See the section “Volume Affinity and CPU Saturation.” A general recommendation is four per aggregate, eight per node for spinning disk, and eight per aggregate for AFF.

• Mount volumes in the namespace to form a folder layout to replicate what the applications use for data layout, if possible. If a single namespace is needed, consider using a FlexGroup volume.

• If you have FlexVol volumes participating in the same application workload, avoid putting them on different aggregate or disk types, unless you’re using volumes to tier inactive data for archive purposes.

• Place volumes on nodes with local data LIFs and use local data LIFs when mounting from clients. This can help performance to those volumes.

### FlexGroup Member Volume Layout Considerations
FlexVol volumes are the building blocks of a FlexGroup volume. A FlexVol volume is 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 single aggregate can contain many FlexVol volumes. A FlexVol volume is not tied to any specific set of disks in the aggregate and is striped across all the disks in the aggregate. However, files themselves are not striped; they are allocated to individual FlexVol member volumes.

Each FlexGroup volume contains multiple member FlexVol volumes to provide concurrent performance and to expand the capacity of the volume past the usual 100TB limits. A FlexGroup volume also provides file counts well beyond the two-billion limit for FlexVol volumes.

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), `flexgroup deploy`, or ONTAP System Manager, the default number of member FlexVol volumes in a FlexGroup volume depends on the storage system and ONTAP release. It can range from eight per node (four per aggregate) to 16 per node (eight per aggregate).

  Here is an example of `volume create with the -auto-provision-as flexgroup option:

  ```bash
  volume create -vserver vs0 -volume fg -auto-provision-as flexgroup -size 200TB
  ```

  **Note:** For almost all use cases, NetApp recommends that you let ONTAP determine the member volume count per node. The ONTAP member count provides optimal performance through volume affinities and CPU slots. If needed, the number of member volume counts can be fewer or greater than eight, depending on the desired configuration. For more information, see [TR-4571](https://www.netapp.com/tr-4571).
Currently, when two aggregates of spinning disks are on a node, the automated FlexGroup creation methods create four member volumes per aggregate.

When a single SSD aggregate is on a node, the automated FlexGroup creation methods create eight member volumes per aggregate.

If a node with spinning disk does not contain two aggregates, the automated FlexGroup creation method might fail in some earlier ONTAP versions. If this happens, continue with manual creation. See the section “When Would I Need to Manually Create a FlexGroup Volume?”

Automated FlexGroup creation methods currently do not consider CPU, RAM, or other factors when deploying a FlexGroup volume. Instead, they follow a hard-coded methodology.

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.

When growing or shrinking (shrink is supported starting in ONTAP 9.6) a FlexGroup volume, use the volume size command at the FlexGroup level; avoid resizing individual FlexVol member volumes.

Consider disabling the space guarantee (thin provisioning) on the FlexGroup volume to allow the member volumes to be overprovisioned. For details on the effects of overprovisioning a FlexGroup volume, see TR-4571.

### When Would I Need to Manually Create a FlexGroup Volume?

Usually, using the automated commands or GUI operations to let ONTAP choose the member volumes is a good option. In other words, don’t worry about CPU count or volume affinity best practices for FlexGroup creation—let ONTAP do that for you.

However, in some use cases (see the following examples), manual creation might be needed.

### Concern Regarding Overprovisioning Volume Counts

In ONTAP, FlexVol volumes are currently limited to a maximum of 1,000 per node. Because a FlexGroup volume is composed of FlexVol volumes, those limits also apply to FlexGroup volumes. If you have FlexGroup volumes with many member volumes or you want to create many FlexGroup volumes in a cluster, 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, Limited Capacity

If you have a workload with larger files and you want to comply with best practices for large files (see TR-4571), you might need to adjust the member volume count to create fewer members at larger individual capacities—particularly if your cluster is already limited in the amount of capacity that’s available. For example, if you want to create a 16TB FlexGroup volume across four nodes, that would create a minimum of 32 member volumes that are 500GB each in size. If your average file size is 250GB, then 500GB member volumes won’t be 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 let ONTAP create a FlexGroup volume, you will sometimes get at most 16 member volumes in a single FlexGroup volume, because it is code-limited to the best practice of eight per node. In the following example, the two-node cluster can create a FlexGroup volume with a maximum of only 1.56PB of capacity (eight members per node; 16 members total; 100TB per member volume).
Figure 20) Error when you create a FlexGroup volume beyond the allowed maximum in System Manager.

If you want a larger FlexGroup volume, you would need to create it manually to allow a higher number of member volumes by using the `-aggr-list-multiplier` option. For a 20PB FlexGroup volume, you would need 200 member volumes.

Make similar considerations if the file count needs to exceed the maximum files allowed. In the 16-member FlexGroup example, a maximum of 32 billion files would be allowed. If more files are needed, add more member volumes.

**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 have to 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_VBN 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
```

This example NetApp FAS8080 EX node is reporting that it can support fully concurrent operations on eight separate volumes simultaneously. It also says that to reach that maximum potential, it would work best with at least two separate aggregates hosting four constituents each. Therefore, when you’re
building a new FlexGroup volume that is served by this node, ideally that new FlexGroup volume would include eight constituents on this node, evenly distributed across two local aggregates. If two such nodes are in the cluster, then a well-formed FlexGroup volume would consist of four aggregates (two per node) and 16 constituents (four per aggregate).

Starting in ONTAP 9.4, the number of available affinities increased to eight per aggregate (two per node) for high-end systems 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 don’t 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 previous section, “When Would I Need to Manually Create a FlexGroup Volume?”

### Best Practice 3: Member Volume Count Recommendations

When you use automated deployment methods (CLI or GUI), ONTAP provisions the appropriate number of member volumes in a FlexGroup volume.

The best practice for member volume counts depends on storage system, aggregate count, and ONTAP version, because the member volume counts are tied to volume affinity availability. Generally, let ONTAP decide on member volume counts for most use cases.

ONTAP 9.4 and later versions provide up to 16 affinities per node (eight per aggregate) for higher-end systems. The best practice for volume count per node in a FlexGroup volume is to either use the automated methods or to match the member volume count per node to the number of VOL affinities listed in the `waffinity_stats` command shown earlier.

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.

### Initial Volume Size Considerations: FlexVol Volumes

When you set an initial FlexVol size, the most important consideration for EDA workloads is the default file count. EDA workloads can contain millions of files, so initially setting `maxfiles` to an appropriate value helps avoid future “out of space” warnings when the `maxfiles` value is exceeded. Table 5 shows a sample of FlexVol sizes, inode defaults (see Table 5), and maximums. If the initial FlexVol `maxfiles` value is not appropriate for the EDA workload, review the section in this document called “High File or Inode Count Considerations” for more information. Also keep in mind the maximum volume size and the maximum file count available to an individual FlexVol volume.
### Initial Volume Size Considerations: FlexGroup Volumes

One common deployment issue that customers run into is undersizing their FlexGroup volume capacity. FlexGroup volumes can be created at almost any capacity, but it’s important to remember that below the overall large container provided by the FlexGroup volume are several FlexVol member volumes that make up the total size of the FlexGroup. Generally, each node has eight member volumes by default, so the FlexGroup capacity is broken up into smaller FlexVol chunks in the form of (total FlexGroup size / number of member volumes in the FlexGroup volume). For example, in a 160TB FlexGroup volume with 16 member volumes, each member volume is 10TB in size.

**Why Is This Important?**

Available member volume size affects how often files are ingested locally or remotely in a FlexGroup volume, which in turn can affect performance and capacity distribution in the FlexGroup volume. File sizes are also important to consider when you design an initial FlexGroup volume, because large files can fill up individual members faster, causing more remote allocation, or even causing member volumes to run out of space prematurely (see the section “Large Files, Limited Capacity”). For example, suppose your FlexGroup volume has member volumes that are 100GB in size, and there are files of 10GB mixed with many smaller files. You might run into issues because the larger files create an imbalance that affects the smaller files, or the larger files might cause volumes to fill up faster. If possible, size your FlexGroup volumes to larger capacities (and use thin provisioning), or use fewer member volumes to allow larger capacities per member volume. As a general rule, aim for member volume sizes that have no more than 5% impact from a single file creation, and avoid creating FlexGroup volumes with fewer than two members per node. Running the latest patch release of ONTAP provides the most optimal ingest and balance functionality for FlexGroup volumes as well as useful safeguards against capacity errors, such as elastic sizing (see the section “Elastic Sizing”).

### Capacity Considerations: FlexGroup Volumes

Although FlexGroup allows massive capacity and file count possibilities, the FlexGroup volume itself is mostly limited to the physical maximums of the underlying hardware. The current maximums 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 need more member volumes, 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 9.2 and 800TB aggregates after ONTAP 9.2. Therefore, a maximum of four 100TB volumes per aggregate or eight 100TB volumes per aggregate are allowed, 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. Instead, aim to leave a cushion of 10% to 20% of the total maximum FlexVol member space to enable emergency space allocation.

These numbers are raw capacities, before features such as NetApp Snapshot reserve, NetApp WAFL® reserve, and storage efficiencies are factored in. For more information, see the [storage limits on the NetApp Support site](https://support.netapp.com/). To correctly size your FlexGroup solution, use the proper sizing tools, such as the [NetApp System Performance Modeler (SPM)](https://www.netapp.com/en-us/software/system-performance-modeler.html) (requires a NetApp login).
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, 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, your volume is composed of 4,000 files, but you then compress some files into a single giant file.

One solution is to grow volumes (either manually or by using volume autogrow) or delete data. However, administrators often don’t see the issue until it’s too late and “out of space” errors have occurred.

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), then 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. Therefore, if 20TB are available in a FlexGroup volume, then 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. For further details about elastic sizing, see TR-4571.

Storage Efficiency Considerations

ONTAP provides enterprise-class storage efficiency technologies, including the following capabilities:

- Thin provisioning
- Data compaction
- Inline data compression
- Inline deduplication, including aggregate inline deduplication starting in ONTAP 9.2

Many EDA workloads are mostly made up of small files. One benefit of the storage efficiencies in ONTAP is that files smaller than 128KB are never copied twice because of the architecture of the WAFL file system. This gives ONTAP an innate storage efficiency advantage over some competitors in the EDA space.

In environments containing data that is highly compressible with many duplicate blocks, space savings with ONTAP storage efficiencies can be tremendous, with a low performance overhead—especially on AFF systems. In fact, official NetApp SPEC SFS performance submissions are performed with storage efficiencies enabled to illustrate not only the performance capabilities of ONTAP but also the potential space savings, depending on the dataset.
Best Practice 4: Storage Efficiency Recommendation in EDA Workloads

To get the most out of your ONTAP storage system, NetApp recommends enabling all available storage efficiency options (inline and postprocess, as well as thin provisioning). Enabling these features has a very small performance impact that is outweighed by the cost savings that ONTAP storage efficiencies provide.

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

Thin Provisioning

Thin provisioning allows storage administrators to allocate more space to workloads than is physically available. For instance, if a storage system has 100TB available, it’s possible to create four 100TB volumes with space guarantees enabled. The benefit of this approach in EDA workloads is to free up available storage and to allow the applications to drive the capacity usage rather than the storage. When you use complementary features such as Snapshot autodelete, volume autogrow, and efficiencies such as compaction, deduplication, and compression, thin provisioning EDA workloads can prove beneficial. In addition, nondisruptive volume moves can be incorporated to move volumes automatically (through NetApp OnCommand® Workflow Automation) as capacity is exhausted on a physical node.

FlexClone Volumes

NetApp FlexClone volumes are Snapshot-backed copies of active FlexVol volumes. They take up nominal amounts of space while enabling administrators to provide proven productivity and efficiency to their end users.

Some use cases for EDA workloads include the following:

- Better developer productivity through quick workspace creation and faster builds
- Improved performance by offloading code checkouts and providing faster deletes
- Reduced license and storage costs through efficiencies
- Better DevOps lifecycle management with Snapshot copies for work in progress and tiering of workloads

FlexClone volumes, although not necessarily a best practice requirement for EDA, do offer value in EDA workflows that make them worth considering.

Backup Considerations

EDA workloads are CPU intensive. As such, it is a best practice to avoid running backups (either NDMP/tape or CIFS/NFS) on the primary storage. Instead, use SnapMirror to replicate the source volumes to a destination cluster and run the backups from the secondary storage system.

3.5 High File or Inode Count Considerations

An inode in ONTAP is a pointer to any file or folder within the file system. Each FlexVol volume has a finite number of inodes and has an absolute maximum of 2,040,109,451. The default or maximum number of inodes on a FlexVol volume depends on the FlexVol size and has a ratio of one inode to 32Kb of capacity. Inodes can be increased after a FlexVol volume has been created and can be reduced starting in Data ONTAP 8.0.

When the default volume inode count reaches 21,251,126, it remains at that value, regardless of the size of the FlexVol volume. This feature mitigates potential performance issues, but should be considered when you design a new FlexGroup volume. The FlexGroup volume can currently handle up to 400 billion files and 200 FlexVol member volumes, but the default inode count for 200 FlexVol members in a FlexGroup volume would be:
If the FlexGroup volume needs more inodes than the default value, you must increase the number of inodes with the `volume modify -files` command. Table 5 shows a sample of FlexVol sizes, inode defaults, and maximums.

Table 5) Inode defaults and maximums according to FlexVol volume size.

<table>
<thead>
<tr>
<th>FlexVol Size</th>
<th>Default Inode Count</th>
<th>Maximum Inode Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>20MB</td>
<td>566</td>
<td>4,855</td>
</tr>
<tr>
<td>1GB</td>
<td>31,122</td>
<td>249,030</td>
</tr>
<tr>
<td>100GB</td>
<td>3,112,959</td>
<td>24,903,679</td>
</tr>
<tr>
<td>1TB</td>
<td>21,251,126</td>
<td>255,013,682</td>
</tr>
<tr>
<td>10TB</td>
<td>21,251,126</td>
<td>2,040,109,451</td>
</tr>
<tr>
<td>100TB</td>
<td>21,251,126</td>
<td>2,040,109,451</td>
</tr>
</tbody>
</table>

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 6 shows various examples of FlexGroup configurations and the resulting default inode counts.

Table 6) Inode defaults resulting from FlexGroup member sizes and member volume counts.

<table>
<thead>
<tr>
<th>Member Volume Size</th>
<th>Member Volume Count</th>
<th>Default Inode Count (FlexGroup)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100GB</td>
<td>8</td>
<td>24,903,672</td>
</tr>
<tr>
<td>100GB</td>
<td>16</td>
<td>49,807,344</td>
</tr>
<tr>
<td>1TB</td>
<td>8</td>
<td>170,009,008</td>
</tr>
<tr>
<td>1TB</td>
<td>16</td>
<td>340,018,016</td>
</tr>
<tr>
<td>100TB</td>
<td>8</td>
<td>170,009,008</td>
</tr>
<tr>
<td>100TB</td>
<td>16</td>
<td>340,018,016</td>
</tr>
</tbody>
</table>

Note: FlexGroup members should not be smaller than 100GB.

High File Counts, Low Capacity Needs

As mentioned, ONTAP allocates a default inode and maximum inode count based on volume capacity. In Table 7, member volumes smaller than 10TB cannot use the two billion inodes available to a FlexVol volume. To get two billion inodes per member volume, the member volume capacity must be 10TB or greater. In a FlexGroup volume with eight member volumes, that would support 16 billion files, but would also provision 80TB of storage.

This can present a challenge to high-file-count environments, because file sizes might be small and won’t require that much total capacity. For example, if all files in a workload are 288 bytes each in size, 16 billion files would use up only about 4.6TB. To achieve a maximum of 16 billion files, you’d need at least eight member volumes that are 10TB each, which would take up 80TB. 4.6TB is around 6% of the total capacity needed to contain 16 billion files in ONTAP in an 80TB FlexGroup volume. In these cases, you would need to disable space guarantees to avoid reserving about 75TB of unused space.
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 member volumes of 10TB or larger with thin provisioning.** Thin-provisioning a volume simply means that you are telling ONTAP a volume will be 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, so it's important to monitor available aggregate space when you use thin provisioning.

- **Manually create the FlexGroup volume with more member volumes than the default.** If you want to keep space guarantees for volumes, another option for high-file-count/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 small 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 TR-4571: NetApp ONTAP FlexGroup Volumes.

<table>
<thead>
<tr>
<th>Total FlexGroup Size</th>
<th>Member Volume Count (Size)</th>
<th>Maximum Inode Count (Entire FlexGroup Volume)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80TB (no space guarantee)</td>
<td>8 (10TB)</td>
<td>16,320,875,608</td>
</tr>
<tr>
<td>64TB (space guarantee enabled)</td>
<td>32 (2TB)</td>
<td>16,320,875,608</td>
</tr>
<tr>
<td>64TB (space guarantee enabled)</td>
<td>64 (1TB)</td>
<td>16,320,875,608</td>
</tr>
</tbody>
</table>

**Planning for Inode Counts in ONTAP**

With utilities such as the NetApp XCP Migration Tool (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 Inodes and Available Inodes**

In ONTAP, you can view inode counts per volume by using the following command with advanced privileges:

```
cluster::*> volume show -volume flexgroup -fields files,files-used
tserver volume    files     files-used
-------         ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------- ------
Tips for Managing Slow Directory Listings in High-File-Count Environments

Some workflows in high-file-count environments include running `find`, `ls`, or some other read-metadata-heavy operation on an existing dataset. Generally, this process is inefficient and can take a long time to complete. If it’s 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 are client, protocol, or network related. The storage rarely is the bottleneck for slow reading of metadata. 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, that there might be millions of network requests to the storage. Each network request incurs n milliseconds of latency, which adds up over time. So, there are a few ways to speed this process up:

- **Send more `getattr` requests at a time.** By itself, `ls` can’t send requests in parallel. But, with utilities such as the XCP Migration Tool, 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 large time savings for workloads.

- **Run `ls` without unnecessary options, such as highlighting or 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/storage failover is required.

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.
Effect of Being Out of Inodes

When a volume runs out of inodes, no more files can be created in that volume until the inodes are increased or existing inodes are freed.

When a volume runs out of inodes, the cluster triggers an event management system event (callhome.no.inodes), and a NetApp AutoSupport® message is triggered.

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.

Note: If any member volume runs out of inodes, the entire FlexGroup volume reports being out of inodes, even if other members have available inodes.

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.

In addition to the call-home message, the following event management system messages are available:

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.

These messages can be used for monitoring or for triggering scripts that automatically increase inode counts to help avoid space errors.
64-Bit File Identifiers

By default, NFS in ONTAP uses 32-bit file IDs. 32-bit file IDs are limited to 2,147,483,647 maximum unsigned integers. With the two-billion inode limit in FlexVol, this value fits nicely into the architecture.

However, because FlexGroup volumes can officially support up to 400 billion files in a single container (and theoretically, many more), NetApp implemented 64-bit file IDs. These IDs support up to 9,223,372,036,854,775,807 unsigned integers.

Best Practice 5: 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 might exceed the two billion inode threshold.

The 64-bit file identifier option is set to "off/disabled" by default. This default prevents the ONTAP change from affecting legacy applications and operating systems that require 32-bit file identifiers before administrators can properly evaluate their environments. Check with your application and OS vendor for their support for 64-bit file IDs before enabling them. Alternatively, 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 currently be enabled only with advanced privileges on the command line.

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

After enabling or disabling this option, you must remount all clients. Otherwise, because the file system IDs change, the clients might receive stale file handle messages when attempting NFS operations. For more information, see “How FSIDs Operate with SVMs in High-File-Count Environments,” later in this document.

If a FlexGroup volume does not 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.

```bash
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.

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/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 throws an error.

The options are:

- `-v3-64bit-identifiers [enabled/disabled]`
- `-v4-64bit-identifiers [enabled/disabled]`
Using Quota Enforcement to Limit File Count

Starting with ONTAP 9.5, it's 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 policies don't apply to files created below the parent volume, create a qtree inside the FlexGroup volume. Then create a default quota rule with two billion files as the limit to help reduce the risk of users overrunning the 32-bit file ID limitations.

```
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

<table>
<thead>
<tr>
<th>Volume</th>
<th>Tree</th>
<th>Type</th>
<th>ID</th>
<th>Used</th>
<th>Limit</th>
<th>Used</th>
<th>Limit</th>
<th>Specifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>FG4</td>
<td>twobillionfiles</td>
<td>tree</td>
<td>1</td>
<td>0B</td>
<td>-</td>
<td>1</td>
<td>20000000000 twobillionfiles</td>
<td></td>
</tr>
<tr>
<td>FG4</td>
<td>tree</td>
<td>*</td>
<td>0B</td>
<td>-</td>
<td>0</td>
<td>2000000000 *</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2 entries were displayed.
```

After that is done, use file permissions to limit access, preventing users from creating files at the volume level. Apply SMB shares to the qtree rather than the volume, and 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

<table>
<thead>
<tr>
<th>Volume</th>
<th>Tree</th>
<th>Type</th>
<th>ID</th>
<th>Used</th>
<th>Limit</th>
<th>Used</th>
<th>Limit</th>
<th>Specifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>FG4</td>
<td>twobillionfiles</td>
<td>tree</td>
<td>1</td>
<td>0B</td>
<td>-</td>
<td>4</td>
<td>20000000000 twobillionfiles</td>
<td></td>
</tr>
<tr>
<td>FG4</td>
<td>tree</td>
<td>*</td>
<td>0B</td>
<td>-</td>
<td>0</td>
<td>2000000000 *</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Impact 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 impact can range from a "stale file handle" error on the client, to failure of directory and file listings, to application failure. In most cases, you should enable the 64-bit file ID option when you use FlexGroup volumes.

You can check a file's ID from the client by using the `stat` command. When an inode or file ID collision occurs, it might look like the following example text. Note that the inode is 3509598283 for both files.
This can manifest as issues such as “circular directory structure” errors on the Linux client and an inability to remove files.

```
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:** The 64-bit File ID option does not affect SMB operations and is unnecessary with volumes that use only SMB.

**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](https://bugzilla.netapp.com/show_bug.cgi?id=671319). If you disable the FSID change with NFSv3, be sure to enable the `-v3-64bit-identifiers` option in ONTAP 9 (see the section “64-Bit File Identifiers”). Keep in mind that this option could affect legacy applications that require 32-bit file IDs.

**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 will cause the 32-bit or 64-bit file identifiers to apply to the SVM itself, meaning that the two-billion-file limit with 32-bit would apply to all volumes. Therefore, the SVM would be limited to two billion files, rather than the FlexVol or FlexGroup volume. Leaving the FSID change option enabled allows volumes to operate as independent file systems with their own dedicated file counts.

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 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.

The -v4-fsid-change option does not apply to FlexGroup volumes earlier than ONTAP 9.7, because NFSv4 is unsupported with FlexGroup volumes in those releases.

**Directory Size Considerations**

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 storage system. 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.

<table>
<thead>
<tr>
<th>Best Practice 6: Recommended ONTAP Version for High-File-Count Environments</th>
</tr>
</thead>
<tbody>
<tr>
<td>For high-file-count environments, use the latest release of ONTAP available.</td>
</tr>
</tbody>
</table>

**What Directory Structures Can Affect maxdirsize?**

The `maxdirsize` value can be a concern when you’re 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 small effect 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 and files spread across directories.
Do FlexGroup Volumes Avoid `maxdirsize` Limitations?

In FlexGroup volumes, each member volume has the same `maxdirsize` setting. Even though a directory could potentially span multiple FlexVol member volumes and nodes, the `maxdirsize` performance effect can still come into play, because directory size is the key component, not individual FlexVol volumes. Directory size is tied to the parent volume and does not divide up across other member volumes. Therefore, the overall size of a directory is still an issue. Thus, FlexGroup volumes do not provide relief for environments facing `maxdirsize` limitations.

Best Practice 7: Avoiding `maxdirsize` Issues

Newer storage systems offer more memory and CPU capacity, and AFF systems provide performance benefits. However, the best way to reduce the performance effect in directories with large numbers of files is to spread files across more directories in a file system.

How Flat Directory Structures Can Affect Volumes

Flat directory structures (many files in one or a few directories) have a negative effect on a wide array of file systems, whether they’re NetApp systems or not. Areas of effect can include, but are not limited to:

- Memory pressure
- Network performance and latency (particularly during mass queries of files, `GETATTR` operations, `READDIR` operations, and so on)
- CPU utilization

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 nearly twice as fast as a FlexVol volume would.

Querying for Used `maxdirsize` Values

It is important to monitor and evaluate `maxdirsize` allocation in ONTAP. However, there are no commands to show `maxdirsize` values specific to ONTAP. Instead, `maxdirsize` allocation would need to be queried from the client.

The following command from an NFS client can 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.

```bash
# 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
781227871 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
360579347 328976 drwxr-xr-x  2 root root 335544320 May 29 21:23
/flexgroup/manyfiles/folder3/topdir_0/subdir_0
```
Using XCP to Check maxdirsize

The XCP Migration Tool is mostly considered a RapidData mover, but it also derives value in its robust file-scanning abilities. XCP is able to run find commands in parallel as well, so the previous examples can be run even faster on the storage system.

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}'
```

And this XCP command helps you find the directory size values:

```
# xcp -match "type == d" -fmt "{} {}" 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 "{} {}" localhost:/usr | 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.
```

**Best Practice 8: 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.
Best Practice 8: Directory Structure Recommendation

- NFS path lengths are defined by the client OS. Refer to the OS vendor for details.
- For information about SMB path lengths, see this Microsoft Dev Center link.

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 absolute maximum number of files in a single directory would be 4,341,760 for SMB or NFS. In FlexGroup volumes, because of hard links, the actual number of files can be closer to 2.6 million in a single directory. If at all possible, avoid putting millions of files in a single directory, and, instead, spread the workload across multiple folders.

NetApp strongly recommends that you keep the maxdirsize value at the defaults. If increases are necessary, contact NetApp Technical Support.

Event Management System Messages Sent When maxdirsize Is Exceeded

The following event management system 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. NetApp Active IQ® Unified Manager can be used to monitor maxdirsize, trigger alarms, and send a notification before the 90% threshold (see Figure 21). These messages also support SNMP traps.

- 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

- 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

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

Supports SNMP trap: true

Figure 21) ONTAP System Manager event screen with maxdirsize warning.

Effect of Increasing maxdirsize

When a single directory contains many files, the 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.

Best Practice 9: Maxdirsize Maximums

Values for maxdirsize are hard coded to not exceed 4GB. To avoid performance issues, NetApp recommends setting maxdirsize values no higher than 1GB if possible.

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. To remediate this problem, 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 “TR-4571.”

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 emoji, some UTF-8 character sizes grow 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:
When a special character is written to a FlexGroup volume, the following behavior occurs:

```bash
# 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 prior to ONTAP 9.5 (see “Support for utf8mb4 Volume Language”), as per 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 10: Special Character Handling in FlexGroup Volumes**

For special character handling with FlexGroup volumes, use ONTAP 9.5 or later.

Additionally, ONTAP has an event management system message for issues with bagofbits handling.

```bash
cluster::*> event route show -message-name wafl.bag* -instance
```

**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:

```bash
# touch "$(echo -e "file\xC")"
```

In ONTAP 9.1, the following command fails:

```bash
# touch "$(echo -e "file\xC")"
touch: cannot touch `file\374': Permission denied
```

In ONTAP 9.2 and later, the following command succeeds:

```bash
# touch "$(echo -e "file\xC")"
# 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 three-byte 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 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, use the corresponding volume language, which is likely to be UTF-8.

NFSv4 requires the 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, NetApp recommends using utf8mb4, because files with Unicode names above the basic plane are not converted properly on non-utf8mb4 volumes.

You can set language (using the `language` option) only when you create a volume. You cannot convert a volume’s language. To use files with a new volume language, create the volume and migrate the files by using a utility such as the XCP Migration Tool. For details about using XCP, see TR-4571.

**Best Practice 11: UTF-8 or utf8mb4?**

If you are running ONTAP 9.5 or later, use the utf8mb4 volume language to help prevent issues with file name translation unless clients are unable to support the language.

**Use of Change Notifications with SMB**

SMB change notifications are how SMB clients are informed of a file’s existence in a share without needing to close a session or refresh a window. The SMB clients are in constant contact 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 used, for example, for applications that must write files and then be able to immediately read the files in SMB shares. This functionality is 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.

NetApp added support for SMB change notifications for FlexGroup volumes in ONTAP 9.2. However, in rare cases, this share property can create performance issues. This is because a FlexGroup volume spans multiple nodes in a cluster, and latency can occur because of the number of change notification requests, particularly in high-file-count environments.

ONTAP 9.5 introduced inherited change notifications. This feature adds an improved algorithm that divides change notifications better and expends less CPU for operations. It should help prevent performance issues when you use SMB change notifications.

**Best Practice 12: SMB Change Notification Recommendation**

- **In ONTAP 9.4 and earlier**, disable SMB change notifications unless applications require them. You can do this by using System Manager or the `cifs share properties remove` command in the CLI.
- **If change notifications are not needed**, disable them on FlexGroup volumes to avoid potential performance impact.
- **If you’re using SMB change notifications**, use ONTAP 9.5 or later.

**File Deletions and FlexGroup Member Volume Balancing**

In general, a FlexGroup volume spreads data across multiple member volumes evenly when data is ingested. This makes file deletions operate a bit more efficiently on a FlexGroup volume than on a FlexVol volume. This is because the system is able to use more hardware and WAFL affinities to spread the delete load more efficiently and use less CPU per node for these operations. However, overall performance of file deletions might be slower because of remote access across the FlexGroup volume as compared with 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 command at the diag privilege level:

```
cluster::*> set diag
class::*> 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)

**Rebalancing Data Within a FlexGroup Volume**

It is not possible to rebalance the workload in a FlexGroup volume to even out capacities, but ONTAP generally does a good job of balancing the ingest load, so a rebalance is not necessary. In the rare case in which a member volume becomes a hot spot, you should analyze the workload. You can use the XCP Migration Tool to scan folders and files to identify file sizes and anomalies. Examples can be found in TR-4571.

After the files are identified, either delete them, move them, add space to the member volumes, or add more member volumes to help balance the ingest load in a FlexGroup volume. These options provide more affinities until the less-full members catch up to other member volumes or until you increase the size of the existing FlexGroup member volumes to provide some relief. In most cases, imbalances in data capacity won’t be noticeable to the end user.

**Why Doesn’t a FlexGroup Volume Rebalance 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 carefully and accurately place each new file where it could be used.
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.

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.
- 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.

However, this performance effect isn’t as serious as the performance effect of ONTAP moving data in a
FlexGroup volume to rebalance the workloads. Therefore, the current approach to FlexGroup volume
imbalances is to enable volume autogrow and set thresholds (~80% full) to help keep the system’s free
space in check.

**Performance Issues When Member Volumes Reach 80%**

In versions earlier than ONTAP 9.5P4, there might be 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.

### 3.6 Project Tiering Considerations

An ONTAP cluster can consist of up to 24 nodes in NAS-only environments and 12 nodes when SAN is
present. These nodes can be a mixture of high-performance nodes, such as AFF, and less-expensive
nodes with spinning disk for capacity needs. Clusters can be architected to tier workloads according to
SLAs, performance needs, capacity requirements, and various other considerations. For example, a
project can be provisioned on capacity nodes to start. As performance needs increase, the project can
then be moved nondisruptively to the AFF nodes in the cluster for high-throughput and low-latency
performance. When the project lifecycle is complete, use volume move to relocate the data to the less-
expensive nodes, or replicate to a disaster recovery site by using SnapMirror technology. Starting with
ONTAP 9.2, you can take your project tiering needs to the cloud by using FabricPool. You can also tier
cold data from Snapshot copies, SnapMirror destinations, or active datasets to the cloud or on-premises
environments by using Amazon S3. For more information regarding FabricPool, see TR-4598: FabricPool
Best Practices.

**Data Lifecycle Management**

When dealing with data lifecycle, you can think in terms of hot, cool, and cold.
Figure 23) Project lifecycle.

- **Hot data.** Consists of the latest project builds or the active jobs that are running. These workloads require the best possible performance, which is delivered with flash storage in ONTAP and, ideally, in parallel with FlexGroup volumes. It is also possible to use SAS or hybrid SAS/SSD aggregates to achieve performance that is suitable for these workloads if you want a more even balance of cost and performance (dollars per IOPS or dollars per MBps).

- **Cool data.** Consists of a recent build or release that has just finished production. The data is still being actively accessed, but high performance isn’t necessarily a requirement anymore. These workloads can survive on SAS or even SATA. For a performance boost, Flash Cache can be used.

- **Cold data.** Consists of archived projects and datasets. This type of data is rarely accessed and can live on less expensive storage. You can move this data by using a nondisruptive volume move from flash storage to capacity tiers, or move it automatically to cloud or Amazon S3 by using FabricPool.

**Table 8) Storage tiers.**

<table>
<thead>
<tr>
<th>Value</th>
<th>Performance</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hot</td>
<td>Fastest</td>
<td>SSD</td>
</tr>
<tr>
<td></td>
<td>Faster</td>
<td>SAS + Flash Cache</td>
</tr>
<tr>
<td></td>
<td>Fast</td>
<td>SAS</td>
</tr>
<tr>
<td>Cool</td>
<td>Good</td>
<td>SATA or SAS + Flash Cache</td>
</tr>
<tr>
<td></td>
<td>Good enough</td>
<td>SATA or SAS</td>
</tr>
<tr>
<td>Cold</td>
<td>Archive</td>
<td>SATA/object (S3)</td>
</tr>
</tbody>
</table>

**Data Lifecycle Management Challenges**

Often, build releases are kept in individual directories. In ONTAP, these directories can be set up as qtrees, if desired, to take advantage of more granular export policy rules and quotas.
However, the downside is that data in directories and qtrees is hard to move around easily. Tools such as rsync and XCP are necessary to migrate these large directories with many files. When you're dealing with live data, that challenge is even greater, because projects can ill afford the migration downtime. This makes project tiering difficult and untenable. Backups become time based, rather than data driven, and IT teams end up backing up too much data, because of complications of the build tree structures. There are a few approaches that could be taken here, each with benefits and drawbacks.

**Data Lifecycle Management Solution: Volume-Based Project Storage**

ONTAP allows storage administrators to present storage containers as folders to applications and projects in the form of FlexVol or FlexGroup volumes.
This approach allows benefits that traditional qtrees and directories cannot provide, including the ability to nondisruptively migrate data for individual projects and folders according to performance or capacity needs. Storage administrators can also replicate individual projects through SnapMirror or SnapVault, rather than needing to replicate multiple projects (by using qtrees) to a destination site.

Additionally, you can create FlexClone volumes to use as test, development, or scratch space scenarios for volumes. With the introduction of aggregate inline deduplication in ONTAP 9.2, this setup also allows storage efficiencies to be seen across multiple volumes in an aggregate.

Table 9) Pros and cons for volumes versus qtrees for project storage.

<table>
<thead>
<tr>
<th>Option</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
</table>
| Using junctioned volumes for project-based storage | • Data mobility by way of volume moves.  
• Performance benefits in the form of multiple volume affinities/CPU threads.  
• Ability to apply export policies to CIFS if desired.  
• Ability to spread data volumes across multiple nodes.  
• Ability to create qtrees inside volumes and provide even more security granularity by using export policy rules.  
• Ability to create Snapshot copies of individual volumes/projects. | • Volume limits per node are much lower than qtree limits.  
• More storage management overhead (more export policies and rules, balancing across cluster resources). |
<table>
<thead>
<tr>
<th>Option</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Ability to create FlexClone volumes for development or testing scenarios.</td>
<td>• Volume moves migrate entire directory structure; no qtree-based moves.</td>
</tr>
<tr>
<td></td>
<td>• Granular data management at the volume level.</td>
<td>• Cannot spread data across nodes when using qtrees; node-limited.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No CIFS export policy support.</td>
</tr>
<tr>
<td>Using qtrees for project-based storage</td>
<td>• Ability to create many more qtrees than volumes in a cluster.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Ability to apply granular security at the qtree level.</td>
<td></td>
</tr>
</tbody>
</table>

### 3.7 Security and ACL Style Considerations

In ONTAP, you can access the same data through NFS and SMB/CIFS by using multiprotocol NAS access. The same general guidance that applies to a FlexVol volume applies to a FlexGroup volume. 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 by means of local files and/or servers such as LDAP or Network Information Service (NIS)
- Volume security style (NTFS, UNIX, or mixed)
- A default UNIX user (pcuser, created by default)

When you create a volume, you choose a security style. If you create the volume without specifying a security style, then the volume inherits the security style of the SVM root volume. The security style determines the style of access control lists (ACLs) that are 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 have the same security style settings.

#### Basic Volume Security Style Guidance

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

- For the UNIX security style, Windows users must map to valid UNIX users.
- For the NTFS security style, Windows users must map to a valid UNIX user, and UNIX users must map to valid Windows users to authenticate. Authorizations (permissions) are handled by the Windows client after the initial authentication.
- Neither the UNIX nor the NTFS security style allows users from the opposite protocol to change permissions.
- A mixed security style allows you to change permissions from any type of client, but the system has an underlying effective security style of NTFS or UNIX, based on the last client type to change ACLs.
- A mixed security style does not retain ACLs if the security style is changed. If the environment is not maintained properly and user mappings are not correct, this limitation can result in access issues.
Best Practice 13: Volume Security Style

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.

For more information about user mapping, name service best practices, and so on, see the product documentation. You can also find more information in TR-4073: Secure Unified Authentication, TR-4067: NFS Best Practice and Implementation Guide, and TR-4668: Name Services Best Practices Guide.

3.8 NFS Considerations

In most cases, EDA workloads run on NFS and predominantly on NFSv3 because of its statelessness, which works well in performance-driven workloads. The following section covers NFS best practices and considerations as they pertain to EDA workloads.

NFS Version Considerations

When a client using NFS attempts to mount a volume in ONTAP without specifying the NFS version (that is, `-o nfsvers=3`), a protocol version negotiation between the client and server takes place. The client asks for the latest 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.

When deciding which NFS version to use as your standard, you need to know the main differences between NFSv3 and NFSv4.x. Often, NFSv3 works fine for an environment and offers greater performance benefits than NFSv4.x, especially in high-file-count or high metadata environments.

Table 10) Basic differences between NFSv3 and NFSv4.x.

<table>
<thead>
<tr>
<th>NFSv3</th>
<th>NFSv4.x</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Stateless (nondisruptive in failovers)</td>
<td>• Stateful (disruptive in failovers)</td>
</tr>
<tr>
<td>• Multiple network ports for ancillary services</td>
<td>• Single network port</td>
</tr>
<tr>
<td>• External lock mechanisms through Network Lock Manager (NLM)</td>
<td>• Integrated locking (lease based)</td>
</tr>
<tr>
<td>• Kerberos (only for NFS; ancillary protocols like NLM, portmapper, and mount are not encrypted by Kerberos)</td>
<td>• Kerberos fully integrated and RFC standard</td>
</tr>
<tr>
<td>• Numeric user ID (UID) and group ID (GID)</td>
<td>• Name strings (user@domain)</td>
</tr>
<tr>
<td>• Standard UNIX mode bits (rwx/777)</td>
<td>• User and group granular ACLs</td>
</tr>
<tr>
<td>• Single packet per operation (portmap, mount, and so on)</td>
<td>• pNFS</td>
</tr>
<tr>
<td></td>
<td>• NFS mount referrals</td>
</tr>
<tr>
<td></td>
<td>• Delegations</td>
</tr>
<tr>
<td></td>
<td>• Compound NFS operations (multiple calls in same packet such as MOUNT/GETFH)</td>
</tr>
</tbody>
</table>

Best Practice 14: NFS Version Considerations for EDA Environments

The primary protocol for EDA simulations is NFSv3. If NFSv4.x and its features are required, avoid NFSv4.0 and use NFSv4.1 instead. If you are using FlexGroup volumes with NFSv4.x, use ONTAP 9.7 or later.

Note: Whenever you use NFSv4.x, plan to use the latest software releases for the NFS client and ONTAP.
Can NFSv3 and NFSv4.x Coexist in the Compute Farm?

Newer versions of Linux support both NFSv4.x and NFSv3. However, some EDA workloads might include a mix of older clients that don’t support both NFS versions and newer clients that do. Rather than taking maintenance windows to upgrade older clients to newer Linux releases, EDA workload environments can choose to use both NFSv3 and NFSv4.x in the same environment on the same datasets.

ONTAP supports NFSv3, NFSv4.0, and NFSv4.1/pNFS, and all three can be used in the same SVM concurrently for one or more exported file systems. Before using newer NFS versions, check with the application vendor for their statement of support, as well as their recommended client OS versions. Some applications might not handle the difference in locking mechanisms between NFSv3 and NFSv4.x very well.

The benefits of having both NFSv3 and NFSv4.1/pNFS protocols coexist in the compute farm are the following:

- No change is required to existing compute nodes that mount the file systems over NFSv3. There is no disruption to existing clients in the compute farm as more nodes on newer client versions are added to scale the number of jobs. The same file system can also be mounted over NFSv3 or NFSv4.1/pNFS from new pNFS-supported clients.
- NFSv4.1/pNFS can provide significant performance improvement in job completion times, as per the performance data in TR-4239. Critical chip designs can be isolated from the rest to provide faster job completion and better service-level objectives (SLO).

NFS Server Tuning

In ONTAP, most of server tuning, such as window sizing and NAS flow control, is performed dynamically, as is covered in TR-4067. The following section covers any NFS server-specific tuning recommended for EDA workloads.

Max TCP Transfer Size/Read and Write Size

Before ONTAP 9.0, NFS mounts negotiated `rsize` and `wsize` values according to the following options:

```
v3-tcp-max-read-size
v3-tcp-max-write-size
```

This value was 64K (65536) by default. ONTAP 9.0 and later versions deprecate those options and consolidate read and write sizes under the single option `tcp-max-xfer-size`. When a client mounts, it negotiates the read and write sizes to the value specified in `tcp-max-xfer-size` if no `rsize` and `wsize` are specified. The current recommended value for EDA workloads is 64K (65536), but if file sizes will exceed 64K, using a larger transfer size might help improve performance. ONTAP supports up to 1MB for `tcp-max-xfer-size`.

NFS/Compute Node Mount Considerations

When you’re mounting NFS to a client that is running EDA workloads, it’s important to understand how to get the best possible results. Use the following mount options in most cases, unless you need to deviate (for example, per application vendor recommendations):

```
vers=3,rw,bg,hard,rsize=65536,wsize=65536,proto=tcp,intr,timeo=600
```

Often, these mount options auto-negotiate or use default values from the client. Check your client version and the default mount options in the NFS client configuration files to verify. If desired, change the default mount options for your client to ensure that these mount options are always used, or use the `/etc/fstab` file to specify mount options.
• **Hard or soft mounts.** If the NFS server is unavailable, a hard mount causes the program using a file that is using NFS to stop and wait for the server to come back online. A soft mount, in contrast, causes the program to report an error.

If `hard` is specified, processes directed to an NFS mount that is unavailable cannot be terminated unless the `intr` option is also specified.

If `soft` is specified, the `timeo=<value>` option can be specified, where `<value>` is the number of seconds before an error is reported. This value should be no less than 60 seconds.

**Note:** For business-critical NFS exports such as EDA workloads, NetApp recommends using hard mounts. NetApp strongly discourages the use of soft mounts.

• **intr.** This option allows NFS processes to be interrupted when a mount is specified as a hard mount. This policy is deprecated in new clients such as Red Hat Enterprise Linux 6.4 and is hard-coded to `nointr`. Issuing `kill -9` is the only way to interrupt a process in newer kernels.

**Note:** For business-critical NFS exports, NetApp recommends using `intr` with hard mounts in clients that support it.

• **rsize=num and wsize=num.** The mount options `wsize` and `rsize` determine how much data is sent between the NFS client and server for each packet sent. This may help optimize performance for specific applications, but should be set according to application vendor best practices, because what is best for one application might not be best for other applications. Newer NFS clients autonegotiate the `wsize` and `rsize` values to what the `-tcp-max-xfer-size` value is set to on the ONTAP NFS server if the mount command does not explicitly set the values. ONTAP defaults `-tcp-max-xfer-size` to 64K and can be set to a maximum of 1MB.

**Note:** The general recommendation for `-tcp-max-xfer-size` is to increase that value in ONTAP to 262144 (256K) and then specify explicit mount options as the applications require it.

**RPC Slot Table Recommendations**

In versions earlier than Red Hat Enterprise Linux 6.3, the number of RPC slot requests per TCP connection was limited to a default of 16, with a maximum of 128 in-flight requests. In Red Hat Enterprise Linux 6.3, RPC slots were changed to dynamically allocate, allowing a much larger number of RPC slots (up to 65,535). As a result, clients running Red Hat Enterprise Linux 6.3 and later can potentially overload an ONTAP node’s NAS flow-control mechanisms, causing potential outages on the node. This is discussed in the section “Network Connection Concurrency: NFSv3.”

**10GbE (or Faster) Connections**

If compute nodes are using 10GbE or greater connections, NetApp recommends that you disable `irqbalance` on the nodes for better performance.

```
[root@ibmx3650-svl51 ~]# service irqbalance stop
Stopping irqbalance: [ OK ]
[root@ibmx3650-svl51 ~]# chkconfig irqbalance off
```

In addition, set the `sysctl` value `net.core.netdev_max_backlog = 300000` to avoid dropped packets on a 10GbE connection. For more information, see the ESnet page on Linux tuning.

**NFSv4.1/pNFS Client Tuning Recommendations**

• Turn off hyperthreading on the BIOS of each of the Linux nodes.

• Use the following mount options:

```
vers=4,rsize=65536,wsize=65536,hard,proto=tcp,timeo=600,minorversion=1
```

• Set up or configure NTP/time services on all compute nodes.
- **Set tuned-adm profile latency performance** for compute-intensive workloads. The following parameters are changed at the kernel level:
  
  - For `/sys/block/sdd/queue/scheduler`, set to `<deadline>; the default is `<cfq>`.
  - For `/etc/sysconfig/cpuspeed`, governor, set to `performance`; the default is `nothing`. This setting uses the performance governor for p-states through `cpuspeed`.

- In Red Hat Enterprise Linux 6.5 and later, the profile requests a `cpu_dma_latency` value of 1.

- **Disable irqbalance**.

- **Set** `net.core.netdev_max_backlog = 300000`.

### NFSv4.1/pNFS: ONTAP Considerations

- If you are using pNFS, verify that a data LIF exists on each node to ensure a local path to data.
- pNFS provides data locality. A volume can be accessed over a direct path from anywhere in a cluster.
- If a volume is moved for capacity or workload balancing, there is no requirement to move or migrate the LIF around in the cluster namespace to provide local access to the volumes. pNFS handles the pathing.
- NFSv4.1 is a stateful protocol, unlike NFSv3. If there is ever a requirement to migrate a LIF, the I/O operations stall for up to 45 seconds to migrate the lock states over to the new location.

### READIRPLUS (READDIR+) with FlexGroup Volumes

If you are running a version of ONTAP earlier than 9.1P5 and use the READDIR+ functionality in NFS, you might experience some latency on rename operations in FlexGroup volumes. This is caused by [bug 1061496](https://bugzilla.netapp.com/show_bug.cgi?id=1061496), which is fixed in 9.1P5 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.

### 3.9 NetApp Volume Encryption Considerations

ONTAP 9.2 introduced support for NetApp Volume Encryption (NVE) for FlexGroup volumes. To implement this feature with FlexGroup volumes, follow the same recommendations and best practices as stated for FlexVol volumes. Rekeying an existing FlexGroup volume is possible in ONTAP 9.5 and later. See "TR-4571" 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 clusterwide 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 TR-4571, “NetApp Encryption Power Guide,” and “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, as long as all aggregates that contain member volumes belonging to the same FlexGroup volume are encrypted.

### 3.10 FlexCache Volume Considerations

ONTAP 9.5 included 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.

FlexCache volumes offer various benefits for EDA workloads, including the following:

- Improved performance by providing load distribution for local or remote ONTAP instances
- Reduced latency by locating data closer to the point of client access
- Enhanced availability by serving cached data in a network disconnection situation
- The ability to take advantage of the cloud for burst workloads to drive down infrastructure costs

For example, if a server farm of hundreds of clients all mount the same volume at the same time, that could add extra load to the system and affect other workloads. However, if those mounts are balanced across several FlexCache volumes, the load strain can be mitigated for that volume and there will be less workload effect.

Starting in ONTAP 9.7, FlexGroup volumes can be origin volumes for FlexCache. For more information about FlexCache, see TR-4743: FlexCache in NetApp ONTAP.

4 Interoperability

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 IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on your installation in accordance with published specifications.

5 Migrating to ONTAP FlexGroup Volumes

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 to FlexGroup can take three general forms:

- 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 logical unit numbers (LUNs), or Infinite Volume in ONTAP to FlexGroup

Data migrations to FlexGroup volumes are the best way to migrate, but they 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 (copy-based transitions [CBT] and copy-free transition [CFT])

The following sections cover the data migration use cases and tell how to approach them.
5.1 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 volume to a FlexGroup volume. For information about setting up `ndmpcopy`, see the article [How to Run ndmpcopy in Clustered Data ONTAP](#).

In the following example, `ndmpcopy` was used to migrate around five million folders and files from a FlexVol volume to a FlexGroup volume. The process took around 51 minutes:

```bash
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
```

The same dataset using `cp` over NFS took 316 minutes—six times as long as `ndmpcopy`:

```bash
# time cp -R /flexvol/nfs/* /flexgroup/nfsACP/
real  316m26.531s
user  0m35.327s
sys   14m8.927s
```
Using the XCP Migration Tool, that dataset took 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 performed on a virtual machine (VM) with a 1GB network and not much RAM or CPU; more robust servers perform even better.

## 5.2 FlexVol to FlexGroup Conversion

In ONTAP 9.7, it is now possible to convert a single FlexVol volume to a FlexGroup volume with a single member volume, in place, with less than 40 seconds of disruption. This is regardless of how much data capacity or how many 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 expand the capacity.

### Why to Convert a FlexVol Volume to a FlexGroup Volume

FlexGroup volumes offer a few advantages over FlexVol volumes, such as the following:

- The ability to expand beyond 100TB and two billion files in a single volume
- The ability to scale out capacity or performance nondisruptively
- Multithreaded performance for high-ingest workloads
- Simplification of volume management and deployment

For example, perhaps you have a workload that is growing rapidly and you don’t want to have to migrate the data, but you still want to provide more capacity. Or perhaps a workload’s performance isn’t good enough on a FlexVol volume, so you want to provide better performance handling with a FlexGroup volume. Converting can help with these cases.

### 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 aren't available in FlexGroup volumes, then you should hold off. For example, SVM-DR and cascading SnapMirror relationships aren't supported in ONTAP 9.7, so, if you need them, you should stay with FlexVol volumes.

Also, if you have a FlexVol volume that's already very large (80–100TB) and already very full (80–90%), you might want to copy the data rather than convert, because the converted FlexGroup volume would have a very large, very full member volume. This could create performance issues, and it doesn't fully resolve your capacity issues, particularly if that dataset contains files that grow over time.

For example, if you have a FlexVol volume that is 100TB in capacity with 90TB used, then it would look like this:

![Figure 26] FlexVol volume that is nearly full and at maximum capacity.
If you were to convert this 90% full volume to a FlexGroup volume, you would have a 90% full member volume. If you add new member volumes, they would be 100TB each and 0% full, so they would take on a majority of new workloads. The data would not rebalance and if the original files grew over time, you could still run out of space with nowhere to go, because 100TB is the maximum member volume size.

Figure 27) Converted FlexVol volume that now has a member volume that is at maximum capacity.

For more detailed information, examples, and caveats about FlexGroup conversion, see TR-4571.

### 5.3 Migrating from Third-Party Storage to FlexGroup

When you migrate from third-party storage, the migration path is a file-based copy. Various methods are available to perform this migration; some are free, and some are third-party solutions that require a fee.

- For NFSv3-only data, NetApp strongly recommends that you consider NetApp XCP software. 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. See the section “XCP Migration Tool.”
- For NFSv4 data, especially data with NFSv4 ACLs, you should use a tool that has ACL preservation and NFSv4 support.
- For CIFS/SMB data, Robocopy is a free tool, but the speed of transfer depends on using its multithreaded capabilities. Third-party providers, such as NetApp partner Peer Software, can also perform this type of data transfer.

### 5.4 Migrating from Data ONTAP Operating in 7-Mode to FlexGroup

Migrate data from Data ONTAP operating in 7-Mode to FlexGroup in one of two ways:

- Perform full migration of 7-Mode systems to clustered ONTAP systems by using the copy-based or copy-free transition methodology. If you use copy-free transition, the process is followed by copy-based migration of data in FlexVol volumes to FlexGroup volumes.
- Perform copy-based transition from a FlexVol volume or host-based copy from a LUN by using the tools mentioned earlier for migrating from third-party storage to FlexGroup.

**Note:** FlexVol to FlexGroup conversion currently cannot be performed on volumes transitioned from Data ONTAP operating in 7-Mode, so you should strongly consider using copy-based migration.

### 5.5 Migrating from SAN LUNs or Infinite Volume in ONTAP to FlexGroup

When you migrate from existing ONTAP objects such as SAN-based LUNs or Infinite Volume, the current migration path is copy-based. The previously mentioned tools for migrating from third-party storage to FlexGroup can also be used for migrating from ONTAP objects.

**Deprecation of Infinite Volume**

Starting in ONTAP 9.4, infinite volumes can no longer be created 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, infinite volumes can no longer be created or modified, 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.

5.6 XCP Migration Tool

The NetApp XCP Migration Tool is free and was designed specifically for scoping, migration, and management of large sets of unstructured NAS data. The initial version is NFSv3 only. To use the tool, download it and request a free license (for software tracking purposes only).

For more information, see the official XCP website. For further details about XCP with FlexGroup volumes, see TR-4571.

Additional Resources

**Technical Reports Specific to NFS and FlexGroup**

- TR-4063: pNFS Configuration and Best Practices for Clustered Data ONTAP 8.2 and Later
- TR-4067: NFS Best Practice and Implementation Guide
- TR-4379: Name Services Best Practice Guide (versions earlier than ONTAP 9.3)
- TR-4668: Name Services Best Practice Guide (ONTAP 9.3 and later)
- TR-4557: NetApp ONTAP FlexGroup Volumes—Technical Overview
- TR-4598: FabricPool Best Practices
- TR-4808: NetApp XCP Best Practices

**EDA Technical Reports**

- TR-4143: Optimizing Synopsys VCS Performance on NetApp Storage
- TR-4238: Optimizing Synopsys VCS Performance on NetApp Storage with Clustered Data ONTAP 8.2—Best Practices Guide
- TR-4239: Synopsys VCS Performance Validation with NetApp Clustered Data ONTAP 8.2 and NFSv4.1/pNFS
- TR-4270: Optimizing Standard Cell Library Characterization with Cadence Virtuoso Liberate and NetApp Clustered Data ONTAP 8.2
• TR-4324: Electronic Device Automation (EDA) Verification Workloads and All Flash FAS (AFF) Array—Performance Validation of Synopsys VCS with FAS8080EX and All-Flash Aggregates www.netapp.com/us/media/tr-4324.pdf

Miscellaneous Content
• Tech OnTap Podcast Episode 188: FlexGroup Update https://soundcloud.com/techontap_podcast/episode-188-flexgroup-update
• 7 Myths about NetApp ONTAP FlexGroup Volumes https://blog.netapp.com/blogs/seven-myths-about-netapp-ontap-flexgroup-volumes/
• FlexGroup lightboard video https://www.youtube.com/watch?v=Wp6jEd4VkgI&t=4s

6 Acknowledgments
Thanks to NetApp Technical Account Manager Charlie Bryant for providing some of the content in this document.

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

Version History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Document Version History</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version 1.0</td>
<td>August 2017</td>
<td>Justin Parisi: Initial commit</td>
</tr>
<tr>
<td>Version 2.0</td>
<td>April 2020</td>
<td>Justin Parisi: Updates for new FlexGroup functionality in ONTAP 9.7</td>
</tr>
</tbody>
</table>
7 Contact Us

Let us know how we can improve this technical report.
Contact us at docfeedback@netapp.com.
Include TECHNICAL REPORT 4617 in the subject line.
Refer to the **Interoperability Matrix Tool (IMT)** on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer’s installation in accordance with published specifications.

**Copyright Information**

Copyright © 2017–2020 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

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

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

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

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

Data contained herein pertains to a commercial item (as defined in FAR 2.101) and is proprietary to NetApp, Inc. The U.S. Government has a non-exclusive, non-transferrable, non-sublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used.

**Trademark Information**

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