WHITE PAPER

DELL EMC POWERSCALE

OneFS BEST PRACTICES

Abstract

This paper describes best practices for installing, configuring and managing a Dell EMC PowerScale cluster.

February 2021

Revisions

Version	Date	Comment
1.0	November 2017	Updated for OneFS 8.1.1
2.0	February 2019	Updated for OneFS 8.1.3
3.0	April 2019	Updated for OneFS 8.2
4.0	August 2019	Updated for OneFS 8.2.1
5.0	December 2019	Updated for OneFS 8.2.2
6.0	June 2020	Updated for OneFS 9.0
7.0	September 2020	Updated for OneFS 9.1

Acknowledgements

This paper was produced by the following:

Author: Nick Trimbee

TABLE OF CONTENTS

Contents

Intended Audience

This paper presents best practices for deploying and managing a Dell EMC PowerScale cluster. It also offers configuration and tuning recommendations to help achieve optimal performance for different workloads. This paper does not intend to provide a comprehensive background to the OneFS architecture.

📖 Please refer to the OneFS Technical Overview white paper for further details on the OneFS architecture.

The target audience for this white paper is anyone designing and deploying a Dell EMC PowerScale clustered storage environment. It is assumed that the reader has an understanding and working knowledge of the OneFS components, architecture, commands and features.

📖 More information on OneFS commands and feature configuration is available in the OneFS Administration Guide.

Data Layout Recommendations

Directory Structure and Layout

In general, it is more efficient to create a deep directory hierarchy that consolidates files in balanced subdirectories than it is to spread files out over a shallow subdirectory structure. Although the recommended maximum file limit per directory is one million, a best practice is to constrain the number of files in any one directory to one hundred thousand. A maximum of 100,000 directories per directory is also recommended. OneFS dynamically allocates new inodes from free file system blocks

ⓘ The key for file and directory layout always revolves around balance. The goal should be for a directory tree structure and is file contents to be as uniform as possible.

- Storing large numbers of files in a directory may affect enumeration and performance, but whether performance is affected depends on workload, workflow, applications, tolerance for latency, and other factors. To better handle storing a large number of files in a directory, use nodes that contain solid state drives (SSDs).
- Directory tree depth is limited to 509 directories and is determined by a maximum path length of 1,023 characters. However, depths greater than 275 directories may affect system performance.
- The maximum number of open files is 315,000 per node.
- Hard links are limited to a maximum of 65,535 per cluster. However, setting the number of per-file hard links to higher than 1,000 can slow down snapshot operations and file deletions. This per-file value can be configured via the efs.ifm.max_links syscontrol.

ⓘ The OneFS protocol daemons, such as the input-output daemon (lwio), may impose additional constraints on the number of files that a node can have open. The protocol daemons typically impose such constraints because the kernel places limits on per-process memory consumption.

File Count Limits

OneFS dynamically allocates new inodes from free file system blocks. The maximum number of possible inodes runs into the billions and depends on the number and density of nodes in the cluster, as expressed by the following formulas:

4Kn drives: ((number of nodes in the cluster) * (node raw TB) * 1000^4 * .99) / (8192 * (number of inode mirrors))

512n drives: ((number of nodes in the cluster) * (node raw TB) * 1000^4 * .73) / (512 * (number of inode mirrors))

File Size Limits

The largest file size that OneFS currently supports is increased to 16TB in OneFS 8.2.2, up from a maximum of 4TB in prior releases.

In order to support files larger than 4TB, adequate space is required in all of a cluster's disk pools in order to avoid a potential performance impact. As such, the following hard requirements apply:

Large File Support Requirement	Description
Version	A cluster must be running OneFS 8.2.2 in order to enable large file support.
Disk Pool	A maximum sized file (16TB) plus protection can consume **no more than 10% of any disk pool**. This translates to a minimum disk pool size of 160TB plus protection.
SyncIQ Policy	All SyncIQ remote clusters must be running OneFS 8.2.2 and also satisfy the restrictions for minimum disk pool size and SyncIQ policies.

After installing OneFS 8.2.2 on a cluster intended for large file support, the following CLI utility will verify that the cluster's disk pools and existing SyncIQ policies meet the above requirements:

```
# isi_large_file -c
```

Once the validation confirms that the cluster meets the requirements, the following CLI command is then run to enable large file support:

```
# isi_large_file -e
```

Upon successfully enabling large file support, the 'cluster full' alert threshold is automatically lowered to 85% from the OneFS default of 95%. This is to ensure that adequate space is available for large file creation, repair, and restriping. Additionally, any SyncIQ replication partners must also be running OneFS 8.2.2, adhere to the above minimum disk pool size, and have the large file feature enabled.

Any disk pool management commands that violate the large file support requirements are not allowed. Once enabled, disk pools are periodically checked for compliance and OneFS will alert if a disk pool fails to meet the minimum size requirement.

ⓘ Be aware that, once enabled, large file support cannot be disabled on a cluster. This may impact future expansion planning for the cluster and all of its SyncIQ replication partners.

🕮 Further information on OneFS limits and guidelines is available in the OneFS Technical Specifications guide.

OneFS Storage Efficiency

A typical data set consists of a mix of large and small files stored in a file system comprising a hierarchical directory structure. Usually, around 30 percent of the data is active; 70 percent is inactive. Snapshots typically back up the data for short-term retention combined with a long-term DR strategy, which frequently includes replication to a secondary cluster, and disk-to-disk or disk to tape NDMP backups.

In this document, large files are considered as those which are 128KB or greater and small files are those less than 128KB. This is significant because at 128KB and above, OneFS uses erasure coding (FEC) to parity protect a file, which results in high levels of storage efficiency. Conversely, files less than 128KB in size are essentially mirrored, so have a larger on-disk footprint. Large file efficiency via erasure coding offsets the penalty of mirroring of small files.

OneFS also provides additional storage efficiency via its native, post-process deduplication engine, SmartDedupe. Consider running deduplication primarily on archive or DR clusters. If system resources allow, deduplication can also be run during off-hours against lower-performance storage tiers or nodepools on primary storage.

Node Hardware Recommendations

Another key decision for cluster performance in an environment is the type and quantity of nodes deployed. Heterogeneous clusters can be architected with a wide variety of node styles and capacities, in order to meet the needs of a varied data set and wide spectrum of workloads. These node styles encompass several hardware generations and fall loosely into four main categories or tiers.

- Extreme performance (all-flash)

- Performance
- Hybrid/Utility
- Archive

The following table illustrates these tiers, and the associated Gen 6 hardware models:

Tier	I/O Profile	Drive Media	Node Type
Extreme Performance	High Perf, Low Latency	All-flash	F800 F810 F600 F200
Performance	Transactional I/O	SAS & SSD	H600 H5600
Hybrid / Utility	Concurrency & Streaming Throughput	SATA/SAS & SSD	H500 H400
Archive	Nearline & Deep Archive	SATA	A200 A2000

Figure 1: PowerScale and Isilon Node Hardware Tiers

Prior to OneFS 8.0, the recommendation was for a maximum cluster size of around 64 nodes based on balancing customer experience with the manageability of extra-large clusters, the risk profile associated with the size of the fault domain that represents for their business, and the ease and simplicity of a single cluster. However, since then, OneFS 8 and later releases have seen considerable backend network infrastructure enhancements removing this 64-node max recommendation and providing cluster stability up to the current supported maximum of 252 nodes per cluster in OneFS 8.2 and later.

Cluster Pool Size and Limits

OneFS SmartPools allows you to define the value of the data within your workflows based on policies, and automatically aligns data to the appropriate price/performance tier over time. Data movement is seamless, and with file-level granularity and control via automated policies, manual control, or API interface, performance and layout, storage tier alignment, and protection settings can be tuned and optimized with minimal impact to end-users.

OneFS Data Protection

A OneFS powered cluster eliminates much of the overhead that traditional storage systems consume. By not having RAID groups, OneFS evenly distributes, or stripes, data among a cluster's nodes with layout algorithms that maximize storage efficiency and performance. The system continuously reallocates data across the cluster, further maximizing space efficiency. At the same time, OneFS protects data with forward error correction, or FEC—a highly efficient method of reliably protecting data.

- With respect to Gen6 hardware in particular, the best practice is to use the recommended ‘hybrid’ protection level, typically 2d:1n, for cluster protection.
- The recommended protection level for a particular node pool is indicated as ‘Suggested’ in the list of requested protection levels.

This can be viewed from the WebUI by navigating to Data Management > Storage Pools > SmartPools and selecting the desired nodepool or tier. For example:

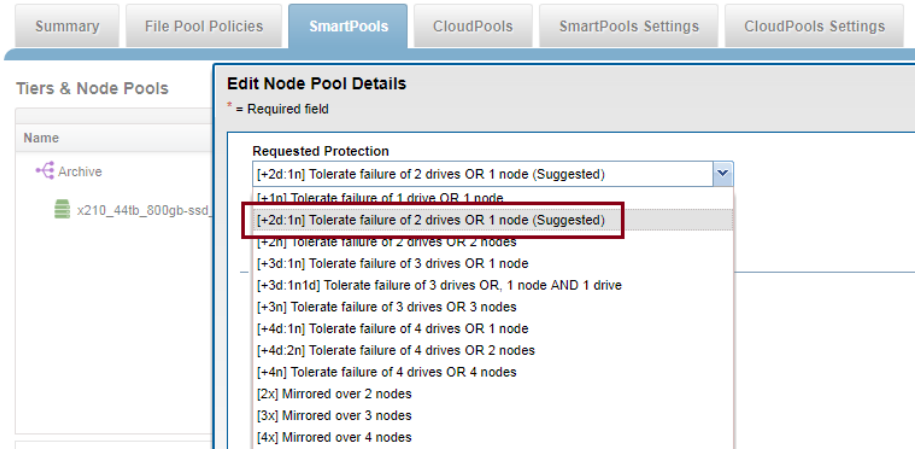

Figure 2: OneFS Suggested Protection Level

The hybrid protection schemes are particularly useful for Isilon Gen6 chassis and other high-density node configurations, where the probability of multiple drives failing far surpasses that of an entire node failure.

ⓘ For all current Gen6 hardware configurations, the recommended protection levels are '+2d:1n' or '+3d:1n1d'.

In the unlikely event that multiple devices have simultaneously failed, such that the file is "beyond its protection level", OneFS will re-protect everything possible and report errors on the individual files affected to the cluster's logs.

OneFS also provides a variety of mirroring options ranging from 2x to 8x, allowing from two to eight mirrors of the specified content. Metadata, for example, is mirrored at one level above FEC by default. For example, if a file is protected at +2n, its associated metadata object will be 4x mirrored.

The full range of OneFS protection levels are summarized in the following table:

Protection Level	Description
+1n	Tolerate failure of 1 drive OR 1 node (Not Recommended)
+2d:1n	Tolerate failure of 2 drives OR 1 node
+2n	Tolerate failure of 2 drives OR 2 nodes
+3d:1n	Tolerate failure of 3 drives OR 1 node
+3d:1n1d	Tolerate failure of 3 drives OR 1 node AND 1 drive
+3n	Tolerate failure of 3 drives or 3 nodes
+4d:1n	Tolerate failure of 4 drives or 1 node

+4d:2n	Tolerate failure of 4 drives or 2 nodes
+4n	Tolerate failure of 4 nodes
2x to 8x	Mirrored over 2 to 8 nodes, depending on configuration

Figure 3: OneFS protection levels

📖 Please refer to the OneFS Technical Overview white paper for further details on OneFS data protection levels.

OneFS enables the protection policy to be modified in real time, while clients are attached and reading and writing data. Be aware, however, that increasing a cluster's protection level may increase the amount of space consumed by the data on the cluster.

ⓘ OneFS also provides under-protection alerting for new cluster installations. If the cluster is under-protected, the cluster event logging system (CELOG) will generate alerts, warning the administrator of the protection deficiency and recommending a change to the appropriate protection level for that particular cluster's configuration.

Small File Considerations

In practice, a OneFS powered cluster typically delivers between 75 and 90 percent space efficiency for a typical dataset. Given a dataset with a wide range of file sizes, it is the large files that dominate utilization, saving as much as 20 to 30 percent of capacity over traditional storage systems. Even when small files make up more than 90 percent of a dataset by file count, they consume only 10 percent or less of the capacity. As such, any inefficiencies in storing small files are overshadowed by the efficiencies in storing large files. And as a data set increases in size, a cluster moves closer to 80 percent efficiency.

Data Tiering and Layout Recommendations

SmartPools Data Tiering

SmartPools enables a multi-tier architecture to be created using high performance nodes with SSD for performance tiers and high-capacity SATA-only nodes for the high-capacity archive tier. For example, a file pool policy could move files from the performance tier to a more cost-effective capacity-biased tier after the desired period of inactivity.

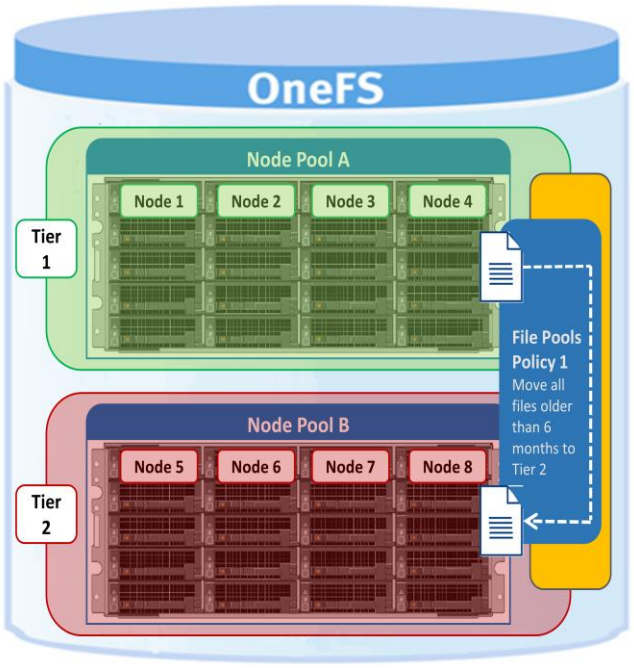

Figure 4: SmartPools tiering.

The following screenshot shows the creation of an 'archive' file pool policy for colder data, which moves files that have not been accessed for more than 30 days to a lower storage tier.

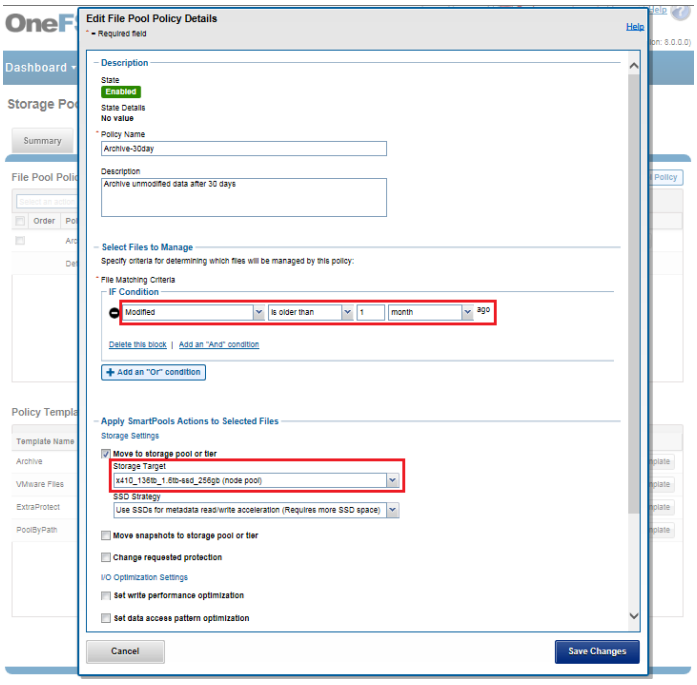

Figure 5: Creating a file pool policy

For optimal cluster performance, Dell EMC recommends observing the following OneFS SmartPools best practices:

- It is not recommended to tier based on modify time (-mtime). Access time is the preferred tiering criteria, with an –atime value of 1 day.
- Ensure that cluster capacity utilization (HDD and SSD) remains below 90% on each pool.
- If the cluster consists of more than one node type, direct the default file pool policy to write to the higher performing node pool. Data can then be classified and down-tiered as necessary.
- A file pool policy can have three 'OR' disjunctions and each term joined by an 'OR' can contain at most five 'AND's.
- The number of file pool policies should not exceed thirty. More than thirty policies may affect system performance.
- Define a performance and protection profile for each tier and configure it accordingly.
- File pool policy order precedence matters, as the policies are applied on first match basis (i.e., the first file pool policy to match the expression will be the applied policy).
- When employing a deep archiving strategy, ensure that the performance pool is optimized for all directories and metadata and the archive tier is just for cold file storage as they age out. This can be configured by adding a 'TYPE=FILE' statement to the aging file pool policy rule(s) to only move files to the archive tier.
- By default, the SmartPools job runs only once per day. If you create a file pool policy to be run at a higher frequency, ensure the SmartPools job is configured to run multiple times per day.
- Enable SmartPools Virtual Hot Spares with a minimum of 10% space allocation. This ensures that there's space available for data reconstruction and re-protection in the event of a drive or node failure, and generally helps guard against file system full issues.
- Avoid creating hardlinks to files which will cause the file to match different file pool policies
- If node pools are combined into tiers, the file pool rules should target the tiers rather than specific node pools within the tiers.
- Avoid creating tiers that combine node pools both with and without SSDs.
- The number of SmartPools tiers should not exceed 5. Although you can exceed the guideline of 5 tiers, doing so is not recommended because it might affect system performance.
- Where possible, ensure that all nodes in a cluster have at least one SSD, including nearline and high-density nodes.
- For performance workloads, SSD metadata read-write acceleration is recommended. The metadata read acceleration helps with getattr, access, and lookup operations while the write acceleration helps reduce latencies on create, delete, setattr, mkdir operations. Ensure that sufficient SSD capacity (6-10%) is available before turning on metadata-write acceleration.
- Determine if metadata operations for a particular workload are biased towards reads, writes, or an even mix, and select the optimal SmartPools metadata strategy.
- Avoid using OneFS Filesystem Explorer or the 'isi set' command to change file attributes, such as protection level, for a group of data. Instead use SmartPools file pool policies.
- If SmartPools takes more than a day to run on OneFS 8.2 or later, or the cluster is already running the FSAnalyze job, consider scheduling the FilePolicy (and corresponding IndexUpdate job) to run daily and reducing the frequency of the SmartPools job to monthly. The following table provides a suggested job schedule when deploying FilePolicy:

Job	Schedule	Impact	Priority
FilePolicy	Every day at 22:00	LOW	6
IndexUpdate	Every six hours, every day	LOW	5
SmartPools	Monthly – Sunday at 23:00	LOW	6

- If planning on using *atime*, be sure to enable *Access Time Tracking* as early as possible. The use of a 24-hour precision is recommended to prevent performance problems.

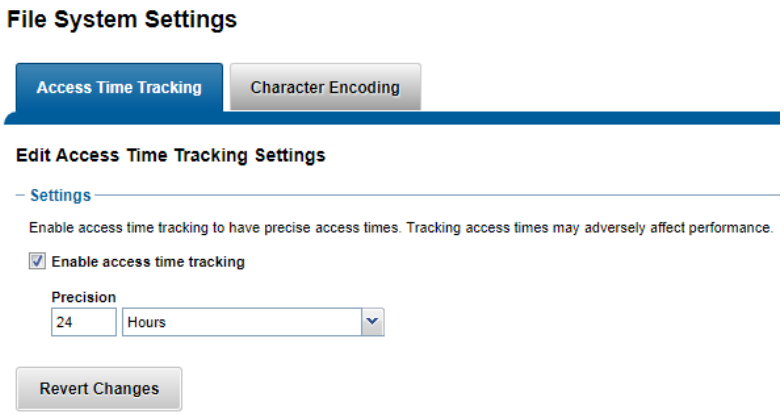

Figure 6: Access time tracking configuration

📖 More information on OneFS data tiering and file pool policies is available in the SmartPools white paper.

Data Access and On-disk Layout

Data Access Settings can be configured at the pool (or even the single file) level to optimize data access for the type of application accessing it. Data can be optimized for Concurrent, Streaming or Random access. Each one of these settings changes how data is laid out on disk and how it is cached.

Data Access Setting	Description	On-disk Layout	Caching
Concurrency	Optimizes for current load on the cluster, featuring many simultaneous clients. This setting provides the best behavior for mixed workloads.	Stripes data across the minimum number of drives required to achieve the data protection setting configured for the file.	Moderate prefetching
Streaming	Optimizes for high-speed streaming of a single file, for example to enable very fast reading with a single client.	Stripes data across a larger number of devices.	Aggressive prefetching
Random	Optimizes for unpredictable access to the file by performing almost no cache prefetching.	Stripes data across the minimum number of drives required to achieve the data protection setting configured for the file.	Little to no prefetching

Figure 7: OneFS data access settings

As the settings indicate, the 'Random' access setting performs little to no read-cache prefetching to avoid wasted disk access. This works best for workload with only small files (< 128KB) and large files with random small block accesses.

Streaming access works best for sequentially read medium to large files. This access pattern uses aggressive prefetching to improve overall read throughput, and on disk layout spreads the file across a large number of disks to optimize access.

Concurrency (the default setting for all file data) access is the middle ground with moderate prefetching.

- Concurrency is the preferred access setting for mixed workloads.

Attribute Optimization of Files and Directories

The attributes of a particular directory or file can be viewed by running the following command and replacing data in the example with the name of a directory or file. The command's output below, which shows the properties of a directory named 'data', has been truncated to aid readability:

```
# isi get -D data  ❶
POLICY   W   LEVEL PERFORMANCE COAL  ENCODING      FILE              IADDRS
default       4x/2 concurrency on  ❷  N/A           ./
<1,36,268734976:512>, <1,37,67406848:512>, <2,37,269256704:512>,
<3,37,336369152:512> ct: 1459203780 rt: 0
************************************************
* IFS inode: [ 1,36,268734976:512, 1,37,67406848:512, 2,37,269256704:512,
3,37,336369152:512 ]
************************************************
*  Inode Version:      6
*  Dir Version:        2
*  Inode Revision:     6
*  Inode Mirror Count: 4
*  Recovered Flag:     0
*  Restripe State:     0
*  Link Count:         3
*  Size:               54
*  Mode:               040777
*  Flags:              0xe0
*  Stubbed:            False
*  Physical Blocks:    0
*  LIN:                1:0000:0004
*  Logical Size:       None
*  Shadow refs:        0
*  Do not dedupe:      0
*  Last Modified:      1461091982.785802190
*  Last Inode Change:  1461091982.785802190
*  Create Time:        1459203780.720209076
*  Rename Time:        0
*  Write Caching:      Enabled  ❸
*  Parent Lin          2
*  Parent Hash:        763857
*  Snapshot IDs:       None
*  Last Paint ID:      47
```

```
*  Domain IDs:          None
*  LIN needs repair:    False
*  Manually Manage:
*       Access          False
*       Protection      True
*  Protection Policy:   default
*  Target Protection:   4x
*  Disk pools:          policy any pool group ID -> data target x410_136tb_1.6tb-
ssd_256gb:32(32), metadata target x410_136tb_1.6tb-ssd_256gb:32(32)
*  SSD Strategy:        metadata  ❹
*  SSD Status:          complete
*  Layout drive count: 0
*  Access pattern: 0
*  Data Width Device List:
*  Meta Width Device List:
*
*  File Data (78 bytes):
*    Metatree Depth: 1
*  Dynamic Attributes (40 bytes):
         ATTRIBUTE                OFFSET SIZE
         New file attribute       0      23
         Isilon flags v2          23     3
         Disk pool policy ID      26     5
         Last snapshot paint time 31     9
************************************************

*  NEW FILE ATTRIBUTES ❺
*  Access attributes:  active
*  Write Cache:  on
*  Access Pattern:  concurrency
*  At_r: 0
*  Protection attributes:  active
*  Protection Policy:  default
*  Disk pools:          policy any pool group ID
*  SSD Strategy:        metadata-write
*
************************************************
```

Figure 8: File and directory attributes

Here is what some of these lines mean:

❶ OneFS command to display the file system properties of a directory or file.

❷ The directory's data access pattern is set to concurrency

❸ Write caching (SmartCache) is turned on.

❹ The SSD strategy is set to metadata-read.

❺ Files that are added to the directory are governed by these settings, most of which can be changed by applying a file pool policy to the directory.

Optimal usage of SSD space

SSD Strategies

In addition to traditional hard disk drives (HDDs), OneFS nodes can also contain a smaller quantity of flash memory-based solid-state drives (SSDs), right up to all-flash nodes. There are a number of ways that SSDs can be utilized within a cluster.

OneFS SSD Strategies are configured on a per file pool basis. These strategies include:

- Metadata read acceleration: Creates a preferred mirror of file metadata on SSD and writes the rest of the metadata, plus all the actual file data, to HDDs.
- Metadata read & write acceleration: Creates all the mirrors of a file's metadata on SSD. Actual file data goes to HDDs.
- Avoid SSDs: Never uses SSDs; writes all associated file data and metadata to HDDs only. This strategy is used when there is insufficient SSD storage and you wish to prioritize its utilization.
- Data on SSDs: All of a node pool's data and metadata resides on SSD.

ⓘ Any node pools comprised off all-flash F-series nodes will automatically store all data and metadata on SSD, since those nodes do not contain any traditional hard disk drives.

The following SSD strategy decision tree explains the options available:

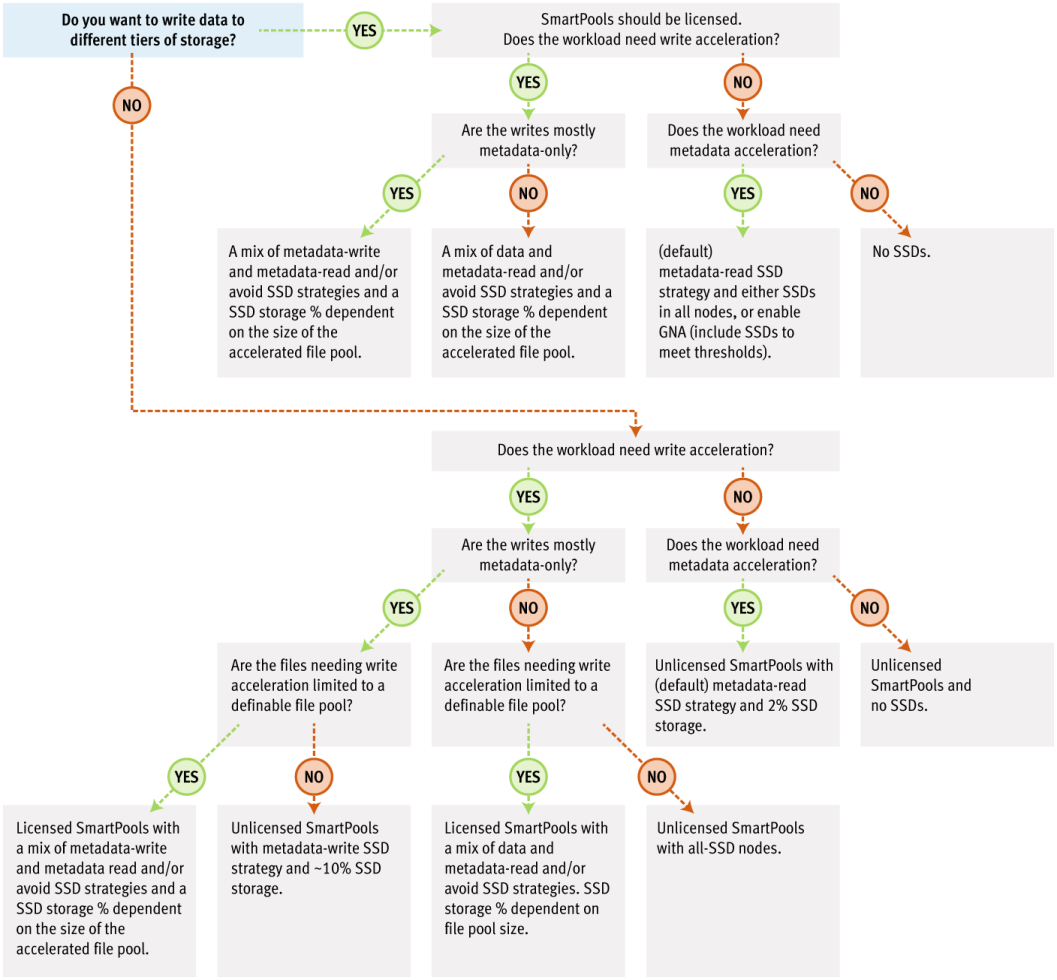

Figure 9: SSD usage decision tree

In all these cases, ensure that SSD capacity utilization remains below 90%.

If snapshots are enabled on a cluster, use the SSD Strategy “Use SSDs for metadata read/write acceleration” to enable faster snapshots deletes. The SSD metadata write strategy will require 6-10% of a pool’s capacity on SSD to accommodate all the metadata mirrors.

ⓘ In order to configure a particular tier to be the default for both data and metadata, the default file pool policy requires the SSD strategy to be set to “Use SSDs for data & metadata”.

🕮 More information on data tiering and management in OneFS is available in the SmartPools white paper.

OneFS Caching Recommendations

OneFS uses up to three levels of read cache, plus an NVRAM-backed write cache, or coalescer. These, and their high-level interaction, are illustrated in the following diagram.

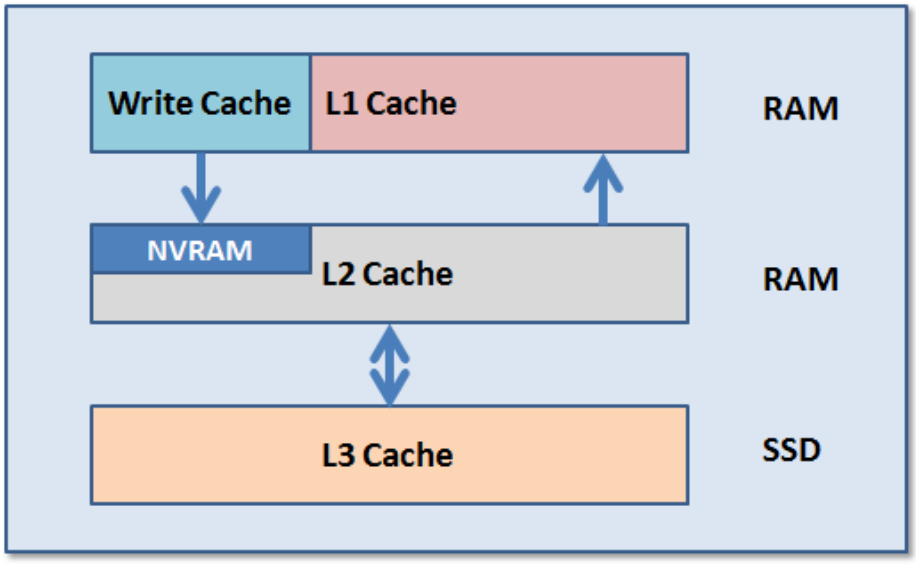

Figure 10: OneFS Caching Hierarchy

The first two types of read cache, level 1 (L1) and level 2 (L2), are memory (RAM) based, and analogous to the cache used in processors (CPUs). These two cache layers are present in all storage nodes.

An optional third tier of read cache, called SmartFlash or Level 3 cache (L3), is also configurable on nodes that contain solid state drives (SSDs). SmartFlash (L3 cache) is an eviction cache that is populated by L2 cache blocks as they are aged out from memory.

L3 Cache Best Practices

If using L3 cache, Dell EMC recommends the following best practices:

- Use a small number (ideally no more than two) of large capacity SSDs rather than multiple small SSDs.
- Use the appropriate capacity of SSD(s) that will fit your working data set. The isi_cache_stats utility can help to determine that on existing clusters. A useful general rule is to size L3 SSD capacity per node according to the following formula:

  L2 capacity + L3 capacity >= 150% of working set size.
- While L3 cache can potentially use up to a 2:1 HDD to SSD ratio per node, use at most 2-3 SSDs for L3 per node.
- Repeated random read workloads will typically benefit most from L3 cache via latency improvements.
- Although not recommended, both L3 cache and Global Namespace Acceleration (GNA) are supported within the same cluster.
- The same procedure is used for replacing failed L3 cache SSDs as for other storage drives. However, L3 cache SSDs do not require FlexProtect or AutoBalance to run post replacement, so it's typically a much faster process.
- For a legacy node pool using a SmartPools metadata-write strategy, don't convert to L3 cache unless:

  1. The SSDs are seriously underutilized.
  2. The SSDs in the pool are oversubscribed and spilling over to hard disk.
  3. Your primary concern is SSD longevity.

L3 Cache Considerations

When deploying L3 cache, the following considerations should be kept in mind:

- All the SSDs within a node pool can either be used for L3 cache, or for SmartPools data strategies (metadata-ro, metadata-rw, data) – but not mixed L3/SmartPools usage.
- L3 cache is not applicable for nodes containing 16 or more SSDs, and all SSD node pools are not eligible for L3 cache enablement.
- Enabling L3 cache on an existing nodepool with SSDs takes some time, since the data and metadata on the SSDs needs to be evacuated to other drives before the SSDs can be formatted for caching. Conversely, disabling L3 cache is a very fast operation, since no data needs to be moved and drive reformatting can begin right away.
- If you're concerned about metadata being evicted from L3, you can either deploy more SSDs per node to accommodate a large working set or disable L3 and stick with traditional SmartPools metadata acceleration (either metadata read-only or read-write) for the particular nodepool.
- It is possible to have GNA and L3 in the same cluster (different nodepools), although some manual setup is required including a SmartPools policy to avoid SSD storage on L3 nodepool. Note that L3 nodepool HDD space does count towards GNA limits
- All the SSDs in an L3 cache nodepool must be the same size.
- If an L3 cache SSD fails, OneFS does not need to run FlexProtect or AutoBalance jobs, like with a regular filesystem SSD. However, after the failed SSD is replaced, some period of time will be needed before the cache is repopulated.
- All new nodepools containing SSD will have L3 cache enabled by default.
- Existing nodepools with SSD will not be modified to use L3 cache on upgrade.
- SSDs displace HDDs. More SSDs and fewer HDD spindles can impact streaming and concurrency performance towards total capacity.
- The L3 cache is intentionally avoided for streaming reads during data prefetch operation. This keeps the streaming requests to the spinning disks (HDDs), while utilizing the SSDs for the random IO.
- L3 cache nodepool hard drive space DOES NOT count in GNA SSD percentage calculations.
- In L3 cache, metadata is preferentially cached over data blocks.
- When a node reboots, there's no automatic flushing of L2 blocks to L3 cache.
- Unlike HDDs and SSDs that are used for storage, when an SSD used for L3 cache fails, the drive state should immediately change to REPLACE without a FlexProtect job running. An SSD drive used for L3 cache contains only cache data that does not have to be protected by FlexProtect. After the drive state changes to REPLACE, you can pull and replace the failed SSD.
- Although there's no percentage completion reporting shown when converting nodepools to use L3 cache, this can be estimated by tracking SSD space usage throughout the job run. The Job impact policy of the FlexprotectPlus or SmartPools job, responsible for the L3 conversion, can also be reprioritized to run faster or slower.
- Current and historical L3 cache statistics are reported by InsightIQ.
- For L3 cache, the isi_cache_stats prefetch statistics will always read zero, since it's purely an eviction cache and does not utilize data or metadata prefetch.
- L3 cache has a metadata only mode (as opposed to data and metadata) to support high-density archive storage nodes.

📖 Further information is available in the OneFS SmartFlash white paper.

Network Recommendations

There are two separate network infrastructures associated with a Dell EMC PowerScale cluster:

External Front-end network

Clients connect to the cluster using Ethernet connections (1GbE, 10GbE or 40GbE) that are available on all nodes. Because each node provides its own Ethernet ports, the amount of network bandwidth available to the cluster scales linearly with performance and capacity. The cluster supports standard network communication protocols to a customer network, including NFS, SMB, HTTP, FTP, HDFS, and S3 object, plus full IPv4 and IPv6 support.

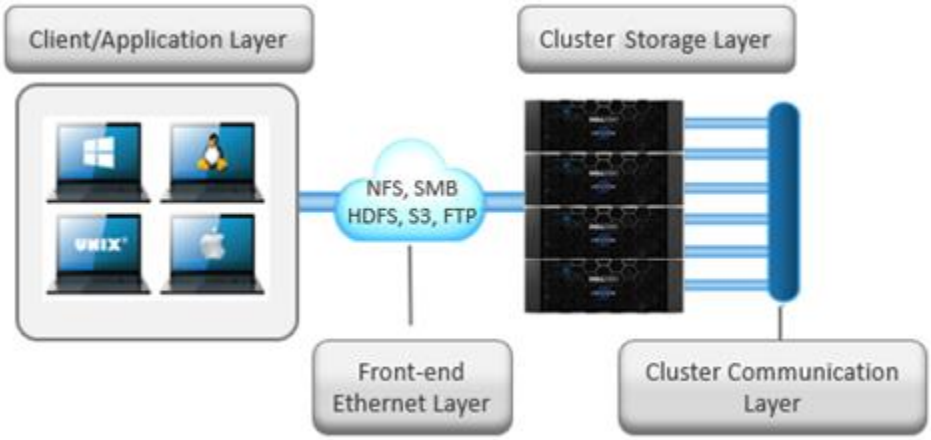

Figure 11: Cluster networking architectural overview

Front-end Connectivity Considerations

For most workflows, the recommendation is to configure at least one front-end 10 or 40 Gb Ethernet connection per node to support the high levels of network utilization that take place. Archive nodes and cold data workloads are often fine with 1Gb Ethernet connections per node.

A best practice is to bind multiple IP addresses to each node interface in a SmartConnect subnet pool. Generally, optimal balancing and failover is achieved when the number of addresses allocated to the subnet pool equals N * (N – 1), where N equals the number of node interfaces in the pool. For example, if a pool is configured with a total of five node interfaces, the optimal IP address allocation would total 20 IP addresses (5 * (5 – 1) = 20) to allocate four IP addresses to each node interface in the pool.

ⓘ For larger-scaled clusters, there is a practical number of IP addresses that is a good compromise between N * (N -1) approach and a single IP per node approach. Example: for a 35-node cluster, 34 IPs per node may not be necessary, depending on workflow.

Assigning each workload or data store to a unique IP address enables OneFS SmartConnect to move each workload to one of the other interfaces, minimizing the additional work that a remaining node in the SmartConnect pool must absorb and ensuring that the workload is evenly distributed across all the other nodes in the pool.

For a SmartConnect pool with four-node interfaces, using the N * (N – 1) model will result in three unique IP addresses being allocated to each node. A failure on one node interface will cause each of that interface's three IP addresses to fail over to a different node in the pool. This ensuring that each of the three active interfaces remaining in the pool receives one IP address from the failed node interface. If client connections to that node were evenly balanced across its three IP addresses, SmartConnect distributes the workloads to the remaining pool members evenly.

ⓘ The largest allocation per cluster that Dell EMC recommends is a /23 subnet, or 510 usable addresses. There are VERY few cases that would require such a large IP allocation.

Optimal Network Settings

Jumbo frames, where the maximum transmission unit (MTU) is set to 9000 bytes, yield slightly better throughput performance with slightly less CPU usage than standard frames, where the MTU is set to 1500 bytes. For example, with 10 Gb Ethernet connections, jumbo frames provide about 5 percent better throughput and about 1 percent less CPU usage.

📖 More information is available in the Advanced Networking Fundamentals guide.

Network Isolation

OneFS provides the ability to optimize storage performance by designating zones to support specific workloads or subsets of clients. Different network traffic types can be isolated on separate subnets using SmartConnect pools.

For large clusters, partitioning the cluster's networking resources and allocate bandwidth to each workload minimizes the likelihood that heavy traffic from one workload will affect network throughput for another. This is particularly true for SyncIQ replication and NDMP backup traffic, which can definitely benefit from its own set of interfaces, separate from user and client IO load.

Many customers as a best practice create separate SmartConnect subnets for the following traffic isolation:

- Workflow separation.
- SyncIQ Replication.
- NDMP backup on target cluster.
- Service Subnet for cluster administration and management traffic.
- Different node types and performance profiles.

ⓘ OneFS 8.0 and later include a new networking object as part of the support for multi-tenancy. Groupnets sit above subnets and pools and allow separate Access Zones to contain distinct DNS settings.

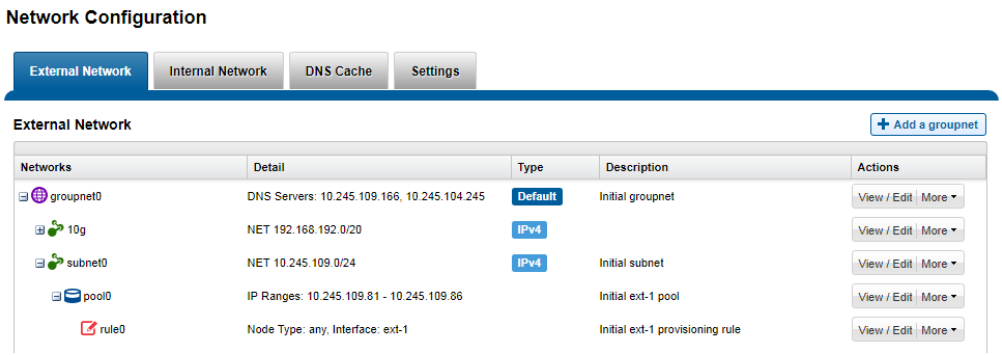

Figure 12: OneFS network object hierarchy

Connection-balancing and Failover Policies

By default, OneFS SmartConnect balances connections among nodes by using a round-robin policy and a separate IP pool for each subnet. A SmartConnect license adds advanced balancing policies to evenly distribute CPU usage, client connections, or throughput. It also lets you define IP address pools to support multiple DNS zones in a subnet.

Load-balancing Policy	General of Other	Few Clients with Extensive Usage	Many Persistent NFS & SMB Connections	Many Transitory Connections (HTTP, FTP, S3)	NFS Automounts or UNC Paths

Round Robin	✓	✓	✓	✓	✓
Connection Count *	✓	✓		✓	✓
CPU Utilization *					
Network Throughput *					

* Metrics are gathered every 5 seconds for CPU utilization and every 10 seconds for Connection Count and Network Throughput. In cases where many connections are created at the same time, these metrics may not be accurate, creating an imbalance across nodes.

Figure 13: Example usage scenarios and recommended balancing options

A 'round robin' load balancing strategy is the recommendation for both client connection balancing and IP failover.

Dynamic Failover

SmartConnect supports IP failover to provide continuous access to data when hardware or a network path fails. Dynamic failover is recommended for high availability workloads on SmartConnect subnets that handle traffic from NFS clients.

For optimal network performance, observe the following SmartConnect best practices:

- Do not mix interface types (40Gb / 10Gb / 1Gb) in the same SmartConnect Pool
- Do not mix node types with different performance profiles (for example, Isilon H600 and A200 interfaces).
- Use the 'round-robin' SmartConnect Client Connection Balancing and IP-failover policies.

SmartConnect Pool Sizing

To evenly distribute connections and optimize performance, the recommendation is to size SmartConnect for the expected number of connections and for the anticipated overall throughput likely to be generated. The sizing factors for a pool include:

- The total number of active client connections expected to use the pool's bandwidth at any time.
- Expected aggregate throughput that the pool needs to deliver.
- The minimum performance and throughput requirements in case an interface fails.

Since OneFS is a single volume, fully distributed file system, a client can access all the files and associated metadata that are stored on the cluster, regardless of the type of node a client connects to or the node pool on which the data resides. For example, data stored for performance reasons on a pool of all-flash nodes can be mounted and accessed by connecting to an archive node in the same cluster. The different types of platform nodes, however, deliver different levels of performance.

To avoid unnecessary network latency under most circumstances, the recommendation is to configure SmartConnect subnets such that client connections are to the same physical pool of nodes on which the data resides. In other words, if a workload's data lives on a pool of F-series nodes for performance reasons, the clients that work with that data should mount the cluster through a pool that includes the same F-series nodes that host the data.

SmartConnect Considerations

Keep in mind the following networking and name server considerations:

- Minimize disruption by suspending nodes in preparation for planned maintenance and resuming them after maintenance is complete

- If running OneFS 8.0 or later, leverage the groupnet feature to enhance multi-tenancy and DNS delegation, where desirable.
- Ensure traffic flows through the right interface by tracing routes. Leverage OneFS Source-Based Routing (SBR) feature to keep traffic on desired paths.
- If you have firewalls, ensure that the appropriate ports are open. For example, for the DNS service, if you open UDP port 53, ensure that TCP port 53 is also open.
- The client never sends a DNS request directly to the cluster. Instead, the site nameservers handle DNS requests from clients and route the requests appropriately.
- In order to successfully distribute IP addresses, the OneFS SmartConnect DNS delegation server answers DNS queries with a time-to-live (TTL) of 0 so that the answer is not cached. Certain DNS servers (particularly Windows DNS Servers) will fix the value to one second. If you have many clients requesting an address within the same second, this will cause all of them to receive the same address. If you encounter this problem, you may need to use a different DNS server, such as BIND.
- Certain clients perform DNS caching and might not connect to the node with the lowest load if they make multiple connections within the lifetime of the cached address.
- The site DNS servers must be able to communicate with the node that is currently hosting the SmartConnect service. This is the node with the lowest logical node number (LNN) with an active interface in the subnet that contains the SSIP address. This behavior cannot be modified.
- Connection policies other than round robin are sampled every 10 seconds. The CPU policy is sampled every 5 seconds. If multiple requests are received during the same sampling interval, SmartConnect will attempt to balance these connections by estimating or measuring the additional load.

📖 Further information is available in the OneFS SmartConnect white paper.

Protocol Recommendations

NFS Considerations

NFSv3 is the ubiquitous protocol for clients accessing storage. This is due to the maturity of the protocol version, ease of implementation, and wide availability of client and server stacks.

There are some useful configuration settings to keep in mind when using a OneFS powered cluster with NFS clients in a performance-oriented environment:

Client NFS Mount Settings

For NFS3 and NFS4, the maximum read and write sizes (rsize and wsize) are 1 MB. When you mount NFS exports from a cluster, a larger read and write size for remote procedure calls can improve throughput. The default read size in OneFS is 128 KB. An NFS client uses the largest supported size by default. Setting the value too small on a client overrides the default value and can undermine performance.

For performance workloads, the recommendation is to avoid explicitly setting NFS rsize or wsize parameters on NFS clients when mounting a cluster's NFS exports directly, or via the automounter. Instead, for NFSv3 clients, use the following mount parameters:

```
mount -vers=3,rw,tcp,hard,intr,retry=2,retrans=5,timeo=600
```

ⓘ For NFS clients that support it, the READDIRPLUS call can improve performance by 'prefetching' file handle, attribute information, and directory entries – plus information to allow the client to request additional directory entries in a subsequent readdirplus transaction. This relieves the client from having to query the server for that information separately for each entry.

For an environment with a high file count, the readdirplus prefetch can be configured to a value higher than the default value of 10. For a low file count environment, you can experiment with setting it lower than the default.

Another recommendation for performance NFS workflows is to use asynchronous (async) mounts from the client. Conversely, using sync as a client mount option makes all write operations synchronous, usually resulting in poor write performance. Sync mounts should be used only when a client program relies on synchronous writes without specifying them.

Optimal Thread Count

The number of threads used by the OneFS NFS server is dynamically allocated and auto-tuning and is dependent on the amount of available RAM.

NFS Connection Count

As a conservative best practice, active NFS v3 or v4 connections should be kept under 1,000, where possible. Although no maximum limit for NFS connections has been established, the number of available TCP sockets can limit the number of NFS connections. The number of connections that a node can process depends on the ratio of active-to-idle connections as well as the resources available to process the sessions. Monitoring the number of NFS connections to each node helps prevent overloading a node with connections.

NFS Recommendations

The recommended limit for NFS exports per cluster is 40,000. To maximize performance, configure NFS exports for asynchronous commit.

For larger NFS environments consider the following:

- Use 10 or 40 Gb Ethernet whenever available
- Consider aggregating client interfaces using LACP.
- Where possible, use Jumbo frames (MTU 9000) to increase network payload.
- Use SmartConnect load-balancing, typically with a round-robin balancing policy.
- Optimize mount point organization.
- Consider using NFS netgroups for large, complex NFS environments

SMB Considerations

Server Message Block (SMB), also known as Common Internet File System, is Microsoft's application-layer network protocol for Windows file sharing. While SMB1 is rarely used these days, OneFS also provides support for SMB2 and SMB3, including features such as continuous availability (CA) for transparent failover, encryption, and multi-channel for increased application throughput.

Best practices for the SMB protocol on OneFS include:

- Static pools are recommended for connecting SMB workloads, including SMB CA.
- The recommendation is to use either SMB2 or SMB3 Windows clients. Where possible, avoid using SMB1.
- Create no more than 80,000 SMB shares per cluster and keep share names below 80 characters.
- For SMB 2 & 3, do not exceed 3,000 active sessions and 27,000 idle connections per node. For SMB1, the recommended limit is 1000 connections per node.
- SMB read and write performance improvements can often be achieved by setting the data-access pattern to Streaming.
- An access zone can authenticate users with only one Active Directory domain. Although you can add more than one of the other directory services to a zone, a best practice is to limit each zone to no more than one of each of the directory services. User mapping rules apply only in the access zone in which you created them.

- As a best practice, if you create access zones, make sure that the directory paths for each zone under /ifs do not overlap. Instead, you should designate separate directory trees for each zone.
- In general, a best practice is to use Microsoft Active Directory with Windows Services for UNIX and RFC 2307 attributes to manage Linux, UNIX, and Windows systems. In some versions of Microsoft Windows, Windows Services for UNIX is also known as Identity Management for Unix (IDMU). Integrating UNIX and Linux systems with Active Directory centralizes identity management and eases interoperability. Make sure your domain controllers are running Windows Server 2003 R2 or later.
- Where possible, a best practice is to authenticate all users with Kerberos because it is a highly secure protocol. If you are authenticating users with Kerberos, ensure that both the cluster and clients use either Active Directory or the same NTP server as their time source.
- In an environment with two or more identity management systems, the simplest configurations name users consistently so that each UNIX user corresponds to a similarly named Windows user. Before assigning a UID and GID, OneFS searches its other authentication providers, such as LDAP, for other identities with the same name. If OneFS finds a match, the mapping service by default selects the associated UID and group memberships. Naming users consistently also allows user mapping rules with wildcards to match names and map them without explicitly specifying each pair of accounts.
- The native identity option is likely to be the best for a network with UNIX and Windows systems. In native mode, OneFS favors setting the UID as the on-disk identity because doing so improves NFS performance. OneFS stores only one type of identifier—either a UID and a GID or a SID—on disk at a time. As a best practice, if you change the on-disk identity, you should run the repair permissions job; see the OneFS Administration Guide.

SMB3 Multi-channel

SMB3 Multichannel allows storage devices to use multiple network connections simultaneously and dramatically increase throughput to clients and is enabled by default on OneFS. The SMB client will automatically detect, negotiate, and use multiple network connections if a proper configuration is identified.

On the client side, use Windows 2012 or Windows 8 or later, with at least one of the following 10 or 40 Gb Ethernet network configurations:

- Multiple network adapters
- One or more network adapters that support RSS (Receive Side Scaling)
- One of more network adapters configured with NIC Teaming

Additionally:

- Use a high-performance non-blocking 10 or 40 Gb Ethernet switch.
- Avoid configuring LACP.
- Each cluster interface is given its own static IP address in the same subnet – do not bond the 10GbE connections using LACP.
- Configure the client's NIC driver as follows:
    - "Max Number of RSS Queues" set to the physical core count for client's CPU.
    - "Receive Side Scaling" set to "Enabled",
    - "Receive Buffers" set to 4096
    - "Transmit Buffers" set to 16384.
- If using Windows Server 2012, make sure the "Enable Virtual Machine Queues" setting under "Virtualization" is set to disabled.

- Windows 8 clients may need the “interrupt moderation rate” in the NIC driver set to “disabled” in order to achieve optimal performance.

New Cluster Best Practices

When initializing a new cluster, the OneFS installation process creates three directories under the clustered filesystem root, /ifs:

- /ifs/data
- /ifs/home
- /ifs/.ifsvar

Consider the following guidelines for directories under /ifs.

- Avoid modifying permissions on /ifs/.ifsvar (mode 755). Do not use directory for general storage.
- Avoid modifying /ifs/data/Isilon_Support, which is created the first time the isi_gather_info command is run to collect cluster logs.
- Create hidden shares for /ifs and set ACLs appropriately.

ⓘ Beginning with OneFS 9.0 CLI, /ifs/.ifsvar is hidden from view, utilities, scripts, and recursive tools. Anything that attempts to access .ifsvar by full or relative path will still succeed. This feature works by omitting /ifs/.ifsvar when listing this one directory. It only effects the ./ifsvar directory and doesn't prevent protocols from accessing /ifs/.ifsvar. Prior to OneFS 9.0, /ifs/.ifsvar is hidden from view, but not from tools, scripts, or protocols.

Proper directory layout is critical to a successful OneFS disaster recovery plan.

When planning a cluster’s initial directory structure, consider multi-tenancy, availability and disaster recovery requirements.

- During a failover event, NFS clients require their exported paths to remain the same to enable accessing the data. The mount entry for any NFS connection must have a consistent mount point so that during failover, you don’t have to manually edit the file system table (fstab) or automount entries on all connected clients.
- OneFS balances a cluster’s connections among all the nodes that service external (front-end) connections. Regularly monitor cluster connections (e.g. with InsightIQ). If the number of connections frequently approaches the maximum number the node can support, consider adding another node.
- SmartConnect balances incoming network connections across all the configured network interfaces in a SmartConnect Zone or pool with one of several load-balancing policies. The most common of these is round robin, which is effective for most workflows. However, it is important to understand whether your front-end connections are being evenly distributed, either in count or bandwidth. Monitor front-end connection distribution.
- Regularly monitor cluster usage with InsightIQ, the web administration interface, or the command-line interface (CLI). When disk space usage reaches 90 percent, we strongly recommend adding additional capacity.
- Many cluster configuration settings are global and have cluster-wide effects. Before changing cluster-wide configuration settings, ensure that you fully understand the global settings and their implications. For information about global cluster configuration settings, see the OneFS Web Administration Guide or the OneFS CLI Administration Guide.
- Confirm that remote support functions work correctly through EMC Secure Remote Support (ESRS) and/or internal email/SNMP notifications. Note that OneFS 9.1 and beyond will only support ESRSv3. Current ESRSv2 clusters will be automatically transitioned to v3 on upgrade to OneFS 9.1.

- OneFS provides a recommended protection level based on cluster configuration. This 'suggested' protection level strikes the best balance between data protection and storage efficiency. Follow its recommendations.
- Recommend disabling client DNS caching, where possible. To handle client requests properly, SmartConnect requires that clients use the latest DNS entries. If clients' cache SmartConnect DNS information, they might connect to incorrect SmartConnect zone names. In this situation, SmartConnect might not appear to be functioning properly.
- Use LACP on interfaces that carry clients when LACP can be configured across two separate switches to provide switch redundancy.
- Use consistent Ethernet MTU sizes across your network. For example, if using jumbo frames ensure MTU 9000 is enable uniformly across the network infrastructure to prevent packet fragmentation or dropped data.
- If round-robin is used as the SmartConnect load balancing strategy with only a small number of high-throughput clients (i.e. less than 10), the distribution of SMB/NFS connections could result in overloading a few nodes while starving others.

Data Availability and Protection Recommendations

Availability and recovery objectives

At the core of every effective data protection strategy lays a solid business continuance plan. An explicitly defined and routinely tested plan is essential to minimize the potential impact to the workflow when a failure occurs or in the event of a natural disaster.

Among the primary approaches to data protection are fault tolerance, redundancy, snapshots, replication (local and/or geographically separate), and backups to nearline storage, VTL, or tape.

Some of these methods are biased towards cost efficiency but have a higher risk associated with them, and others represent a higher cost but also offer an increased level of protection. Two ways to measure cost versus risk from a data protection point of view are:

- **Recovery Time Objective (RTO):** RTO is the allotted amount of time within a Service Level Agreement (SLA) to recover data. For example, an RTO of four hours means data must be restored and made available within four hours of an outage.
- **Recovery Point Objective (RPO):** RPO is the acceptable amount of data loss that can be tolerated per an SLA. With an RPO of 30-minutes, this is the maximum amount of time that can elapse since the last backup or snapshot was taken.

The availability and protection of data can be illustrated in terms of a continuum:

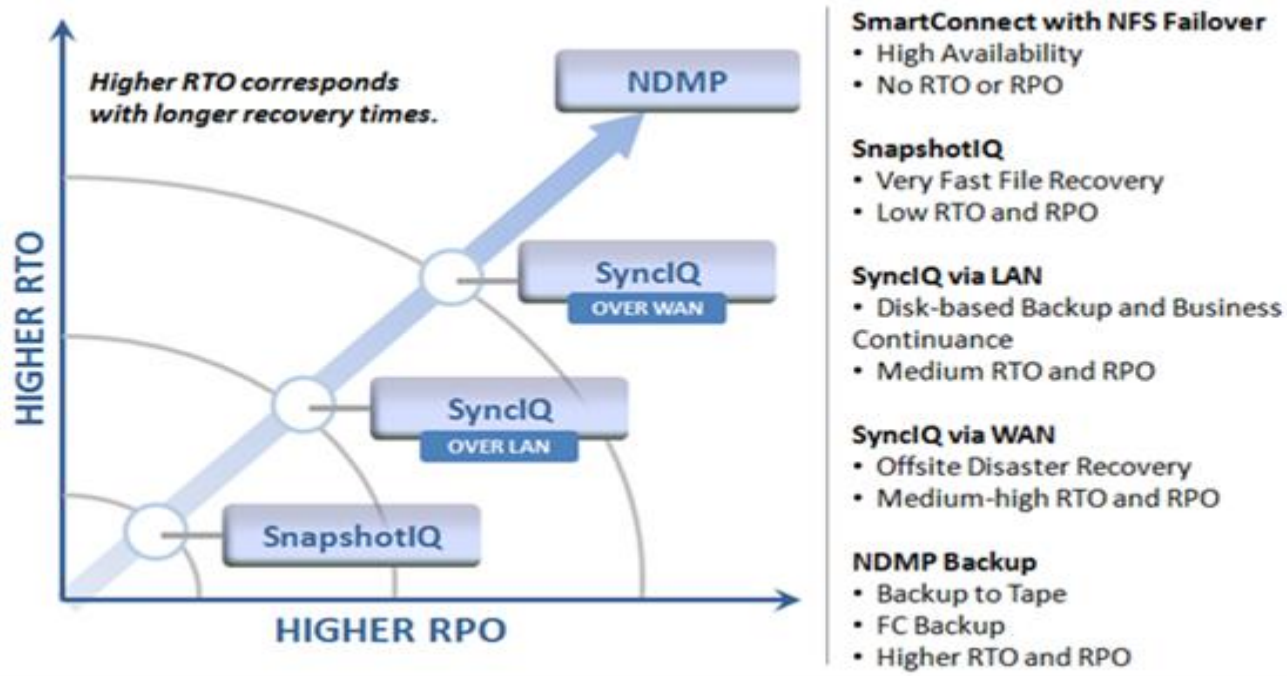

Figure 14: OneFS Data Protection technology alignment with protection continuum

At the beginning of the continuum sits high availability. This requirement is usually satisfied by redundancy and fault tolerant designs. The goal here is continuous availability and the avoidance of downtime by the use of redundant components and services. Further along the continuum lie the data recovery approaches in order of decreasing timeliness: SnapshotIQ for fast recovery, followed by SyncIQ, and finally traditional backup, providing insurance against large scale data loss, natural disasters and other catastrophic events.

- Snapshots are frequently used to back up the data for short-term retention and to satisfy low recovery objective SLAs.
- Replication of data from the primary cluster to a target DR cluster, ideally located at a geographically separate location, is strongly recommended.
- NDMP backup to tape or VTL (virtual tape library) typically satisfies longer term high recovery objective SLAs and any regulatory compliance requirements.

📖 Further information is available in the OneFS high availability and data protection white paper.

Snapshot Considerations

Snapshots always carry a trade-off between cluster resource consumption (CPU, memory, disk) and the benefit of increased data availability, protection, and recovery.

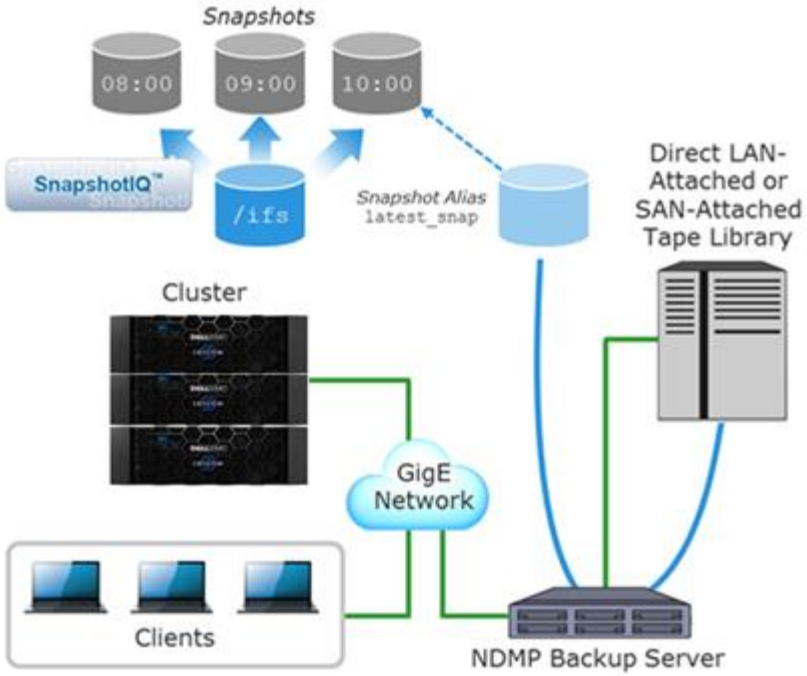

Figure 15: SnapshotIQ integration with NDMP backups.

OneFS SnapshotIQ creates snapshots at the directory-level instead of the volume-level, thereby providing improved granularity.

There is no requirement for reserved space for snapshots in OneFS. Snapshots can use as much or little of the available file system space as desirable.

Snapshots can either be manually taken on-demand or automated with a snapshot schedule.

Snapshot scheduling allows cluster administrators to automatically generate snapshots according to a pre-defined itinerary. OneFS snapshot schedules can be configured at daily, weekly, monthly or yearly intervals, with single or multiple job frequency per schedule, and down to a per-minute granularity. Similarly, automatic snapshot deletion can be configured per defined schedule at an hourly through yearly range.

- An ordered deletion schedule is simple to configure but retains a larger number of snapshots and is recommended for datasets with a lower rate of change.
- For more active data, an unordered deletion schedule can prove more effective. The configuration and monitoring overhead is slightly higher, but fewer snapshots are retained.

The following table provides a suggested snapshot schedule for both ordered and unordered deletion configurations.

Deletion Type	Snapshot Frequency	Snapshot Time	Snapshot Expiration	Max Retained Snapshots
Ordered deletion (for mostly static data)	Every four hours	Start at 12:00AM End at 11:59AM	1 month	180

Unordered deletion (for frequently modified data)	Every other hour	Start at 12:00AM End at 11:59AM	1 day	27
	Every day	At 12:00AM	1 week	
	Every week	Saturday at 12:00AM	1 month	
	Every month	First Saturday of month at 12:00AM	3 months	

Figure 16: Snapshot Schedule Recommendations

For optimal cluster performance, Dell EMC recommends observing the following SnapshotIQ best practices.

- Use an ordered snapshot deletion strategy where viable.
- Configure the cluster to take fewer snapshots, and for the snapshots to expire more quickly, so that less space will be consumed by old snapshots. Take only as many snapshots as you need and keep them active for only as long as you need them.
- Using SmartPools, snapshots can physically reside on a different disk tier than the original data. The recommendation, however, is to keep snapshots on the same tier on which they were taken.
- The default snapshot limit is 20,000 per cluster and recommend limiting snapshot creation to 1,024 per directory.
- Limit snapshot depth to a maximum of 275 directories.
- Avoid creating snapshots of directories that are already referenced by other snapshots.
- It is recommended that you do not create more than 1000 hard links per file in a snapshot to avoid performance degradation.
- Creating snapshots of directories higher on a directory tree will increase the amount of time it takes to modify the data referenced by the snapshot and require more cluster resources to manage the snapshot and the directory.
- Avoid taking snapshots at /ifs level. Taking snapshots at a parent dataset level is recommended, enabling faster snapshot deletions and avoiding management complexities. In particular, avoid taking nested snapshots, redundant snapshots, or overly scoped snapshots. For example, if you schedule snapshots of /ifs/data and /ifs/data/foo and/ifs/data/foo/bar, consider taking snapshots of only the intermediate or most granularly scoped part (/ifs/data/foo or /ifs/data/foo/bar).
- If you intend on reverting snapshots for a directory, it is recommended that you create SnapRevert domains for those directories while the directories are empty. Creating a domain for a directory that contains less data takes less time.
- Delete snapshots in order, beginning with the oldest. Where possible, avoid deleting snapshots from the middle of a time range. Newer snapshots are mostly pointers to older snapshots and deleting them will not free up much space. Deleting the oldest snapshot ensures you will actually free up the space. You can determine snapshot order (if not by name or date) by using the isi snapshot snapshots list command. The snapshot IDs (first column) are non-conserved, serial values.
- Configure the SSD Strategy to “Use SSDs for metadata read/write acceleration” for faster snapshots deletes.
- Quotas can be used to calculate a file and directory count that includes snapshot revisions, provided the quota is configured to include snaps in its accounting via the “--snaps=true” configuration option.
- SnapshotDelete will only run if the cluster is in a fully available state, i.e., no drives or nodes are down.
- A snapshot schedule cannot span multiple days: To generate snapshots from 5:00 PM Monday to 5:00 AM Tuesday, create one schedule that generates snapshots from 5:00 PM to 11:59 PM on Monday, and another schedule that generates snapshots from 12:00 AM to 5:00 AM on Tuesday.
- If a directory is moved, you cannot revert any snapshots of that directory which were taken prior to its move.

- Do not delete SyncIQ snapshots (snapshots with names that start with SIQ), unless the only remaining snapshots on the cluster are SyncIQ snapshots, and the only way to free up space is to delete those SyncIQ snapshots.

📖 Further information is available in the OneFS SnapshotIQ white paper.

Replication Considerations

OneFS SyncIQ delivers high-performance, asynchronous replication of unstructured data to address a broad range of recovery point objectives (RPO) and recovery time objectives (RTO). This enables customers to make an optimal tradeoff between infrastructure cost and potential for data loss if a disaster occurs. SyncIQ does not impose a hard limit on the size of a replicated file system so will scale linearly with an organization's data growth up into the multiple petabyte ranges.

SyncIQ is easily optimized for either LAN or WAN connectivity in order to replicate over short or long distances, thereby providing protection from both site-specific and regional disasters. Additionally, SyncIQ utilizes a highly-parallel, policy-based replication architecture designed to leverage the performance and efficiency of clustered storage. As such, aggregate throughput scales with capacity and allows a consistent RPO over expanding datasets.

A secondary cluster synchronized with the primary production cluster can afford a substantially improved RTO and RPO than tape backup and both implementations have their distinct advantages. And SyncIQ performance is easily tuned to optimize either for network bandwidth efficiency across a WAN or for LAN speed synchronization. Synchronization policies may be configured at the file-, directory- or entire file system-level and can either be scheduled to run at regular intervals or executed manually.

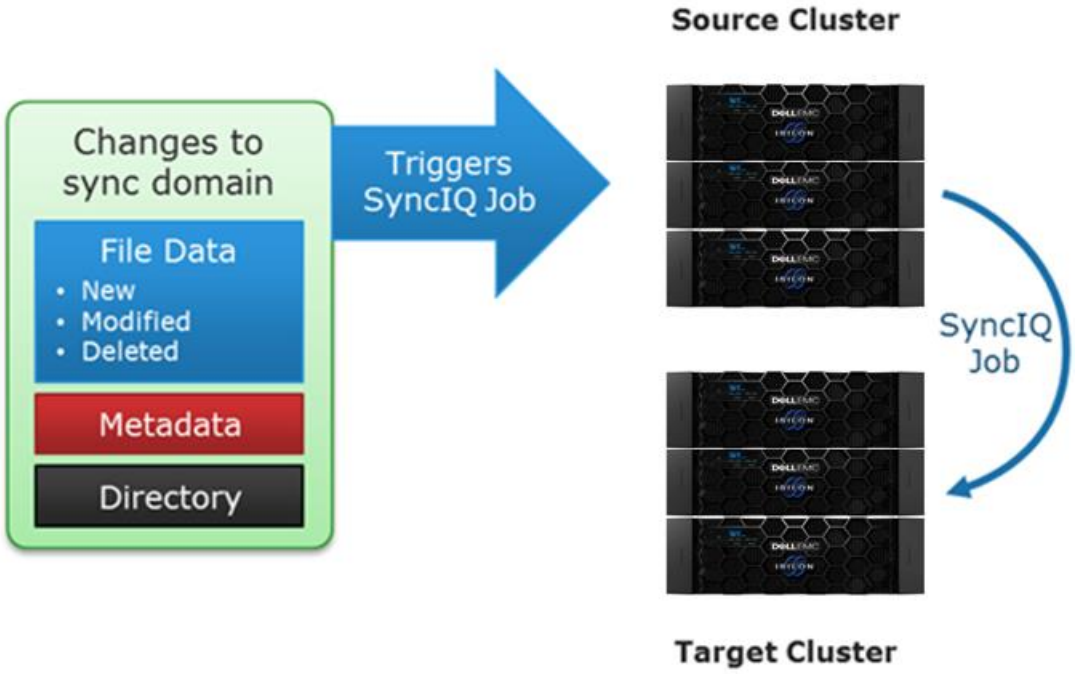

Figure 17: SyncIQ change-based replication.

By default, a SyncIQ source cluster can run up to fifty concurrent replication jobs.

ⓘ For OneFS versions prior to 8.0, this limit is five consecutive jobs.

OneFS queues any additional jobs until a job execution slot becomes available, and jobs that are queued can be easily cancelled. SyncIQ policies also have a priority setting to allow favored policies to preempt others. In addition to chronological scheduling, replication policies can also be configured to start whenever the source is modified (change based replication). If preferred, a delay period can be added to defer the start of a change-based policy.

Bear in mind the following SyncIQ recommendations:

- Highly recommend implementing Superna Eyeglass for failover/failback.
- The recommended limit of running SyncIQ policies is 1000 policies and 50 concurrent jobs per cluster (for a cluster with 4 or more nodes).
- While the maximum number of workers per node per policy is eight, the default and recommended number of workers per node is three.
- The recommended limit of workers per replication policy is 40.
- Recommend having the target cluster running the same or a later version of OneFS as the source cluster.
- After creating a policy and before running the policy for the first time, use the policy assessment option to see how long it takes to scan the source cluster dataset with default settings.
- Increase workers per node in cases where network utilization is low. This can help overcome network latency by having more workers generate I/O on the wire. If adding more workers per node does not improve network utilization, avoid adding more workers because of diminishing returns and worker scheduling overhead.
- Increase workers per node in datasets with many small files to push more files in parallel. Be aware that as more workers are employed, more CPU is consumed, due to other cluster operations.
- Consider using SmartConnect pools to constrain replication to a dedicated set of cluster network interfaces, and to avoid contention with other workflows accessing the cluster through these nodes.
- Use SyncIQ network throttling to control how much network bandwidth SyncIQ can consume.
- Avoid full dataset replications where possible. Changing any of the following parameters will trigger a full baseline sync of the policy:
    - Source path(s): root path, include and exclude paths
    - Source file selection criteria: type, time, and regular expressions
- With a policy of type ‘Sync’, modifying file attributes comparison options and values causes a re-sync and deletion of any non-matching files from the target next time the job runs. This does not apply to policies of type ‘Copy’.
- Specifying file criteria in a SyncIQ policy will slow down a copy of sync job.
- Full baseline replication takes much longer than incremental synchronizations, so to optimize performance, avoid triggering full synchronizations unless necessary. Changing any of the following parameters will trigger a baseline sync of the policy:
    - Source path(s): root path, include and exclude paths
    - Source file selection criteria: type, time, and regular expressions
    - Remember that “target aware synchronizations” are much more CPU-intensive than regular baseline replication. However, they potentially generate far less network traffic if both source and target datasets are already seeded with similar data.
    - Setting a target cluster password is useful for security and to verify that the source cluster is replicating to the correct target. The target cluster password is different from a cluster’s root password. Do not specify a target password unless you create the required password file on the target cluster.

- If a cluster is running OneFS 8.2 or later, use SyncIQ encryption to protect any replication sessions that traverse WAN or other insecure or untrusted network segments.

📖 Further information is available in the OneFS SyncIQ white paper.

Data Management Recommendations

Quota Best Practices

OneFS SmartQuotas tracks disk usage with reports and allows for the enforcement of storage limits with alerts and hard boundaries. SmartQuotas comprises two types of capacity quota:

- Accounting Quotas: monitor and report on the amount of storage consumed, but do not take any action.
- Enforcement Quotas: Restrict how much storage that a user, group, or directory can use and send notifications.

A OneFS SmartQuota can have one of four enforcement types:

- **Hard**: A limit that cannot be exceeded.
- **Soft**: A limit that can be exceeded until a grace period has expired.
- **Advisory**: An informal limit that can be exceeded.
- **None**: No enforcement. Quota is accounting only.

All three quota types have both a limit, or threshold, and a grace period. A hard quota has a zero-time grace period, an advisory quota has an infinite grace period, and a soft quota has a configurable grace period.

ⓘ Even when a hard quota limit is reached, there are certain instances where operations are not blocked. These include administrative control via root (UID 0), system maintenance activities, and the ability of a blocked user to free up space.

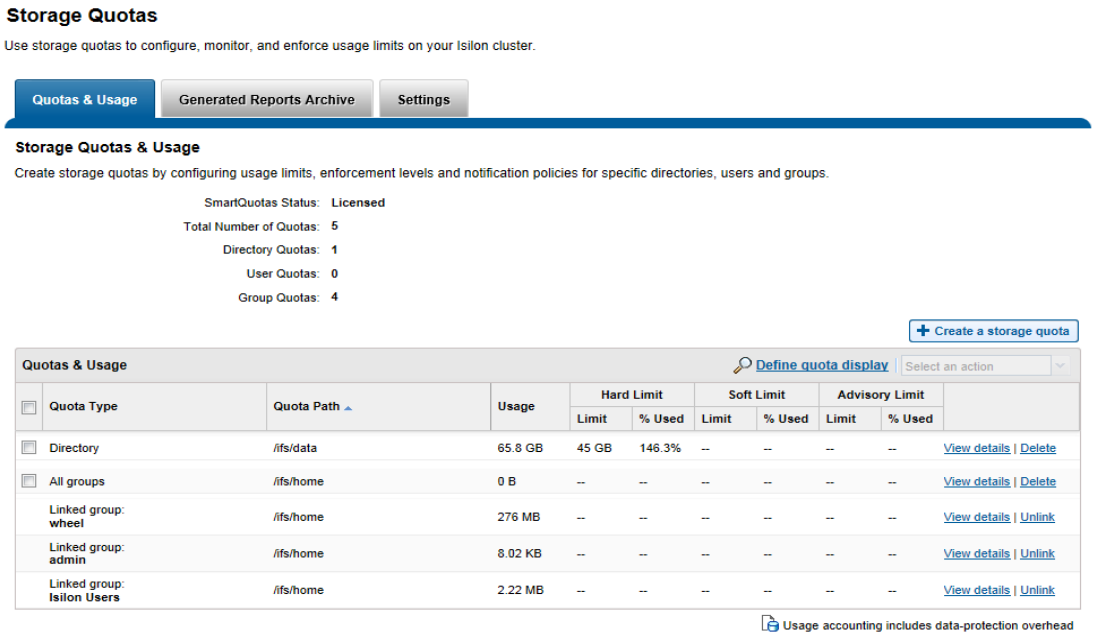

Figure 18: SmartQuotas usage overview.

SmartQuotas best practices include:

- Avoid creating quotas on the root directory of the default OneFS share (/ifs). A root-level quota may result in performance degradation.
- Where possible, observe the best practice of a maximum number of 500,000 quotas per cluster in OneFS 8.2 and later, and 20,000 quotas per cluster in prior releases.
- Limit quota depth to a maximum of 275 directories.
- Governing a single directory with overlapping quotas can also degrade performance.
- Directory quotas can also be used to alert of and constrain runaway jobs, preventing them from consuming massive amounts of storage space.
- Enforcement quotas are not recommended for snapshot-tracking quota domains.
- Before using quota data for analysis or other purposes, verify that no QuotaScan jobs are running.
- Rather than editing the quota email notification templates directly, copy them to another directory to edit and deploy them.
- If quota reports are not in the default directory, you can run the 'isi quota settings reports view' command to find the directory where they are stored.
- Use the 'isi quota quotas notifications disable' command to disable all notifications for a quota.

Quota Considerations

- The maximum tested quota limit is 400,000 (although the file system has no hard-coded limit on quotas). However, when listing a large number of quotas, only a partial list may be returned.
- With CloudPools data, the quota is calculated based on the size of the data local to the cluster. For example, for a 100MB file tiered to a cloud provider, SmartQuotas would calculate just the size of the local stub file (8K).
- SmartQuotas reports the logical capacity of the files, whether they are deduplicated or not.
- The QuotaScan job runs after the creation of a quota, but not after a change. However, it does run on a schedule and incorporates any changes then.
- If two quotas are created on the same directory – for example an accounting quota without Snapshots and a hard quota with Snapshots - the quota without Snapshot data overrules the limit from the quota with Snapshot data.
- SmartQuotas also provide a low impact way to provide directory file count reports.
- Configuration changes for linked quotas must be made on the parent quota that the linked quota is inheriting from. Changes to the parent quota are propagated to all children. To override configuration from the parent quota, you must unlink the quota first.
- If a quota type uses the accounting-only option, enforcement limits cannot be used for that quota.
- Cloned and deduplicated files are treated as ordinary files by quotas. If the quota includes data protection overhead, the data protection overhead for shared data is not included in the usage calculation.
- Moving quota directories across quota domains is not supported.
- You can edit or delete a quota report only when the quota is not linked to a default quota.

- A quota can only be unlinked when it's linked to a default quota. Configuration changes for linked quotas must be made on the parent (default) quota that the linked quota is inheriting from. Changes to the parent quota are propagated to all children. If you want to override configuration from the parent quota, you must first unlink the quota.
- Disabling all quota notifications also disables all system notification behavior. Use the --clear options to remove specific quota notification rules and fall back to the system default.
- Quota containers compartmentalize /ifs, so that a directory with a container will appear as its own separate 'file system slice'. For example, to configure a directory quota with a 4TB container on /ifs/data/container1, you could use the following CLI command:

```
# isi quota quotas create /ifs/data/container1 directory --hard-threshold 4T --
container true
```

Within OneFS, quota data is maintained in Quota Accounting Blocks (QABs). Each QAB contains a large number of Quota Accounting records, which need to be updated whenever a particular user adds or removes data from a filesystem on which quotas are enabled. If a large number of clients are accessing the filesystem simultaneously, these blocks can become highly contended and a potential bottleneck.

To address this, quota accounts have a mechanism to avoid hot spots on the nodes storing QABs. Quota Account Constituents (QACs) help parallelize the quota accounting by including additional QAB mirrors on other nodes.

The following sysctl increases the number of quota accounting constituents, which allows for better scalability and reduces latencies on create/delete flurries when quotas are used.

Using this parameter, the internally calculated QAC count for each quota is multiplied by the specified value. If a workflow experiences write performance issues, and it has many writes to files or directories governed by a single quota, then increasing the QAC ratio (efs.quota.reorganize.qac_ratio) may improve write performance.

The QAC ration can be changed to value 8 from the default value of 1 by running the following OneFS CLI command:

```
# isi_sysctl_cluster efs.quota.reorganize.qac_ratio=8
```

📖 Further information is available in the OneFS SmartQuotas white paper.

SmartDedupe Best Practices

OneFS SmartDedupe maximizes the storage efficiency of a cluster by decreasing the amount of physical storage required to house an organization's data. Efficiency is achieved by scanning the on-disk data for identical blocks and then eliminating the duplicates. This approach is commonly referred to as post-process, or asynchronous, deduplication.

After duplicate blocks are discovered, SmartDedupe moves a single copy of those blocks to a special set of files known as shadow stores. During this process, duplicate blocks are removed from the actual files and replaced with pointers to the shadow stores.

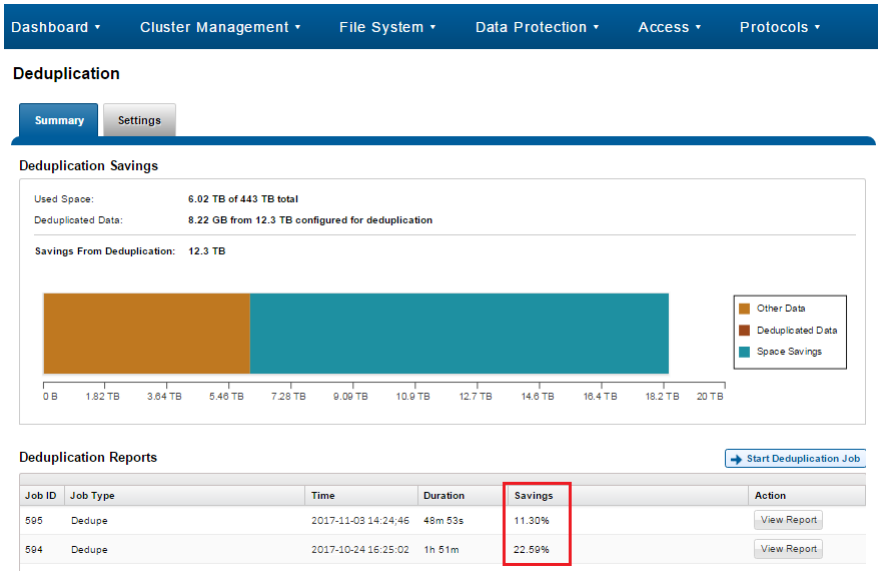

Figure 19: SmartDedupe capacity savings.

For optimal cluster performance, Dell EMC recommends observing the following SmartDedupe best practices. Please note that some of this information may be covered elsewhere in this paper.

- Deduplication is typically applied to data sets with a lower rate of change – for example, file shares, home directories, and archive data.
- Enable SmartDedupe to run at subdirectory level(s) below /ifs.
- Avoid adding more than ten subdirectory paths to the SmartDedupe configuration policy,
- SmartDedupe is ideal for home directories, departmental file shares and warm and cold archive data sets.
- Run SmartDedupe against a smaller sample data set first to evaluate performance impact versus space efficiency.
- Schedule deduplication to run during the cluster's low usage hours – i.e. overnight, weekends, etc. By default, the SmartDedupe job runs automatically.
- After the initial dedupe job has completed, schedule incremental dedupe jobs to run every two weeks or so, depending on the size and rate of change of the dataset.
- Always run SmartDedupe with the default 'low' impact Job Engine policy.
- Run the dedupe assessment job on a single root directory at a time. If multiple directory paths are assessed in the same job, you will not be able to determine which directory should be deduplicated.
- When replicating deduplicated data, to avoid running out of space on target, it is important to verify that the logical data size (i.e. the amount of storage space saved plus the actual storage space consumed) does not exceed the total available space on the target cluster.

ⓘ Bear in mind that deduplication isn't free. There's always trade-off between cluster resource consumption (CPU, memory, disk), and the benefit of increased space efficiency.

SmartDedupe Considerations

- SmartDedupe will not share blocks across files with different protection policies applied.

- OneFS metadata, including the deduplication index, is not deduplicated.
- SmartDedupe will not attempt to deduplicate files smaller than 32KB in size.
- Dedupe job performance will typically improve significantly on the second and subsequent job runs, once the initial index and the bulk of the shadow stores have already been created.
- SmartDedupe will not deduplicate the data stored in a snapshot. However, snapshots can certainly be created of deduplicated data.
- If deduplication is enabled on a cluster that already has a significant amount of data stored in snapshots, it will take time before the snapshot data is affected by deduplication. Newly created snapshots will contain deduplicated data, but older snapshots will not.
- From OneFS 8.0 onwards, SmartDedupe deduplicates common blocks within the same file, resulting in even better data efficiency.

📖 Further information is available in the OneFS SmartDedupe white paper.

In-line Data Reduction Best Practices

OneFS in-line data reduction is available exclusively on the PowerScale F600 & F200 nodes and Isilon F810 all-flash and H5600 deep-hybrid platforms. The OneFS architecture is comprised of the following principle components:

- Data Reduction Platform
- Compression Engine and Chunk Map
- Zero block removal phase
- Deduplication In-memory Index and Shadow Store Infrastructure
- Data Reduction Alerting and Reporting Framework
- Data Reduction Control Path

The in-line data reduction write path comprises three main phases:

- Zero Block Removal
- In-line Deduplication
- In-line Compression

If both in-line compression and deduplication are enabled on a cluster, zero block removal is performed first, followed by dedupe, and then compression. This order allows each phase to reduce the scope of work each subsequent phase.

Figure 20: In-line Data Reduction workflow.

The Isilon F810 platform includes a compression off-load capability, with each node in an F810 chassis containing a Mellanox Innova-2 Flex Adapter. This means that compression and decompression are transparently performed by the Mellanox adapter with minimal latency, thereby avoiding the need for consuming a node's expensive CPU and memory resources.

The OneFS hardware compression engine uses zlib, with a software implementation of igzip as fallback in the event of a compression hardware failure. OneFS employs a compression chunk size of 128KB, with each chunk comprising sixteen 8KB data blocks. This is optimal since it is also the same size that OneFS uses for its data protection stripe units, providing simplicity and efficiency, by avoiding the overhead of additional chunk packing.

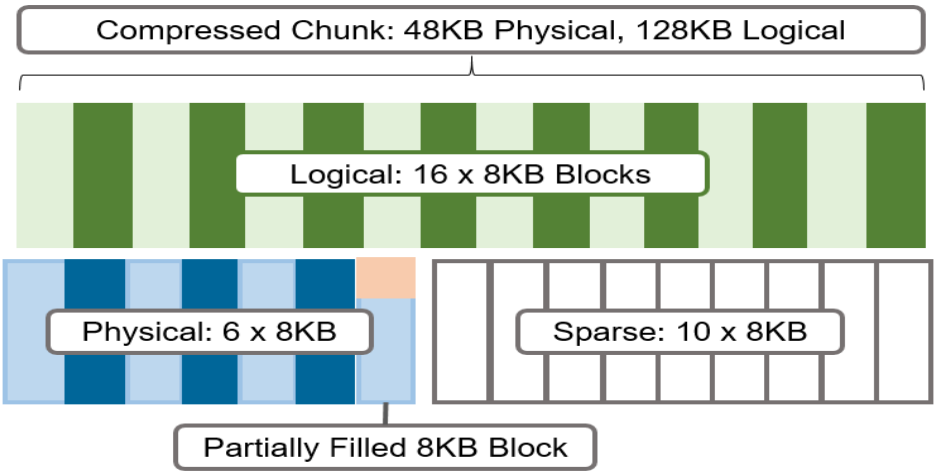

Figure 21: Compression chunks and OneFS transparent overlay.

Consider the diagram above. After compression, this chunk is reduced from sixteen to six 8KB blocks in size. This means that this chunk is now physically 48KB in size. OneFS provides a transparent logical overlay to the physical attributes. This overlay describes whether the backing data is compressed or not and which blocks in the chunk are physical or sparse, such that file system consumers are unaffected by compression. As such, the compressed chunk is logically represented as 128KB in size, regardless of its actual physical size.

Efficiency savings must be at least 8KB (one block) in order for compression to occur, otherwise that chunk or file will be passed over and remain in its original, uncompressed state. For example, a file of 16KB that yields 8KB (one block) of savings would be compressed. Once a file has been compressed, it is then FEC protected.

Compression chunks will never cross node pools. This avoids the need to de-compress or recompress data to change protection levels, perform recovered writes, or otherwise shift protection-group boundaries.

For optimal cluster performance, Dell EMC recommends observing the following in-line compression best practices. Please note that some of this information may be covered elsewhere in this paper.

- In-line data reduction is for Isilon F810 installations only. Legacy Isilon F800 nodes cannot be upgraded or converted to F810 nodes.
- Run the assessment tool on a subset of the data to be compressed/deduplicated.
- When replicating compressed and/or deduplicated data, to avoid running out of space on target, it is important to verify that the logical data size (i.e. the amount of storage space saved plus the actual storage space consumed) does not exceed the total available space on the target cluster.
- Avoid running SmartDedupe on any node pools with in-line deduplication enabled.
- Data reduction can be disabled on a cluster if the overhead of compression and deduplication is considered too high and/or performance is impacted.
- The software data reduction fall back option is less performant, more resource intensive, and less efficient (lower compression ratio) that hardware data reduction. Consider removing Isilon F810 nodes with failing offload hardware from the node pool.

- Run the dedupe assessment job on a single root directory at a time. If multiple directory paths are assessed in the same job, you will not be able to determine which directory should be deduplicated.

In-line Data Reduction Considerations

In-line data reduction is supported with the following caveats:

- OneFS 8.2.1 will support from 4 to 252 Isilon F810 nodes, or 36 chassis, per cluster. OneFS 8.2.2 will support from 4 to 252 Isilon H5600 or F810 nodes per cluster. OneFS 9.0 will support from 3 to 252 F600 or F200 PowerScale nodes, or 4 to 252 Isilon H5600 or F810 nodes per cluster.
- Data reduction savings depend heavily on factors like the data, cluster composition, and protection level.
- Compressed and deduplicated data does not exit the filesystem compressed or deduplicated in any shape or form.
- Decompression is substantially less expensive than compression.
- Inline data reduction is exclusive to the Isilon F810 and H5600 platforms and PowerScale F600 and F200 nodes and does not require a software license. In-line data reduction will be automatically disabled on non-F810, F600, F200, or H5600 node pools.
- In OneFS 8.2.1 and later, in-line compression is automatically enabled on supporting nodes, whereas in-line dedupe is disabled. The following command line syntax will activate in-line dedupe on a compression cluster: '`isi dedupe inline settings modify --mode enabled`'.
- There is no compatibility or equivalency between Isilon F800 and F810 nodes: They cannot share the same node pool and the F800 nodes will not be able to store compressed data.
- There is no OneFS WebUI support for data reduction. Configuration and management are via the CLI only.
- Partial writes to compression chunks may require reading the entire compression chunk first and decompressing it. This is true even if most of the compression chunk is being written.
- Modifications to compression chunks may require rewriting the entire compression chunk even if only a single logical block is changed.
- Some workloads will have data access patterns that exacerbate the above issues and have the potential to cause more writes than if compression was not used.
- Data integrity failures with compressed data will likely mean that corruption does not just affect a single block but instead the entire compression chunk.
- If SmartPools is used on a mixed cluster containing compression nodes, data will only be compressed and/or in-line deduplicated when it physically resides on the compression nodepool(s). If data is tiered to non-compression node pools it will be uncompressed on the compression nodes before it is moved, so full uncompressed capacity will be required on the compressed pool.
- Post-process SmartDedupe can run in concert with compression and in-line deduplication. It is supported but not widely used. The SmartDedupe job will have to decompress data first to perform deduplication, which is an addition resource expense.
- In-line compression on the Isilon F810 uses a dedicated backend network card, so F810 nodes will not support 2-way NDMP backend via a Gen6 backend Ethernet / fibre channel controller.
- Even though compressed files are unintelligible when stored on disk, this does not satisfy the encryption requirements for secure data at rest compliance. However, Isilon F810 and H5600 nodes are available with SED drives.
- InsightIQ is not yet fully integrated with in-line data reduction and will not report compression savings. This will be addressed in a future release.

- As discussed earlier, in-line compression isn't free. There's always trade-off between cluster resource consumption (CPU, memory, disk), the potential for data fragmentation and the benefit of increased space efficiency.
- Since compression extends the capacity of a cluster, it also has the effect of reducing the per-TB compute resource ratio (CPU, memory, I/O, etc).
- Depending on an application's I/O profile and the effect of In-line data reduction on the data layout, read and write performance and overall space savings can vary considerably.
- In-line data reduction is limited to PowerScale F600 & F200 and Isilon F810 & H5600 clusters or node pools.
- OneFS metadata structures (inodes, b-trees, etc) are not compressed.
- Since compression trades cluster performance for storage capacity savings, compression may not be ideally suited for heavily accessed data, or high-performance workloads.
- SmartFlash (L3) caching is not applicable to F-series nodes since they contain exclusively SSD flash media anyway.
- If a heterogeneous cluster contains PowerScale F600 or F200 nodes or Isilon F810 or H5600 chassis plus other F800, H-series or A-series nodes, data will be uncompressed on the fly when it moves between pools. A non-compression node on the cluster can be an initiator for compressed writes to a compression pool and will perform compression in software. However, this may generate significant overhead for lower powered Archive class nodes.
- In-line dedupe will not permit block sharing across different hardware types or node pools to reduce the risk of performance asymmetry.
- In-line dedupe will not share blocks across files with different protection policies applied.
- OneFS metadata is not deduplicated.
- In-line dedupe will not deduplicate the data stored in a snapshot.
- There is no in-line deduplication of CloudPools files.
- In-line dedupe can deduplicate common blocks within the same file and a sequence of consecutive blocks to a single block in a shadow store, resulting in even better data efficiency.

Data Immutability Recommendations

SmartLock offers advanced data immutability and security capabilities, such as the protection of directories and files from deletion in a WORM state, the ability to disable privileged delete, and overall security. It is available in two different modes: Compliance and Enterprise. The administrative restrictions of Compliance mode have the potential to affect both compliance data as well as enterprise data. In order to make an informed decision consider the following guidelines and suggested best practices:

- Use SmartLock in Compliance mode only if your organization is legally obligated to do so under SEC rule 17-a4(f). As the Compliance mode installation or upgrade involves careful planning and preparation, it is recommended to be done with the assistance of Dell EMC Support.
- Enterprise mode offers more than adequate security requirements for most users, in the majority of situations. Moreover, the 'superuser' account remains available in Enterprise mode. Therefore, it is more administrator friendly compared to Compliance mode. Following are the best practices that need to be performed prior to putting an existing cluster in Compliance mode.
- Test and validate all workflows using a proof-of-concept Compliance mode cluster.
- Verify that the cluster time is correct before putting the cluster in Compliance Mode.

- Avoid using 'run-as-root' on SMB shares. If you have previously configured SMB shares to 'run-as-root' then change the settings for those shares to specify access permissions to either 'Full-Control', 'Read-Write' or 'Read' before putting the cluster in compliance mode.
- Use Role based access control (RBAC) for cluster access to execute file management and administrative operations. Enable RBAC, grant appropriate privileges, and then log on through the RBAC-enabled account to the command line interface (CLI). 'compadmin' represents a regular data user in the context of the CLI.
- For any data migrations from a non-Compliance mode cluster to a cluster that you intend to put in compliance mode, first verify that current ownership and access permissions are valid and appropriate on both clusters so as to allow data migration.
- Review the permissions and ownership of any files that exclusively permit the root account to manage or write data to them. Once an upgrade to Compliance mode is complete, if the existing OneFS configuration limits relevant POSIX access permissions to specific directories or files in any way, writing data or changing ownership of these objects will be blocked.
- If any root-owned workflow or data files exist, all ownership or permission changes should be executed before upgrading to Compliance mode. You should not change the ownership of any system files. The Compliance mode conversion process automates all required ownership changes to system files. Avoid changing the ownership of any files outside of /ifs, as no user data should reside outside of /ifs. As a best practice, change the ownership of files under /ifs that are owned by 'root' to the 'compadmin' account before upgrading to Compliance mode.
- In Compliance mode, the default POSIX permissions permit the compadmin account to write data. However, the following directories should not be modified unless the default permissions for these directories have been changed: /ifs/.ifsvar and /ifs/.snapshot
- Verify the available disaster recovery options on compliance mode clusters in relation to SyncIQ.

ⓘ If NDMP is being used for backups, the NDMP backups of SmartLock Compliance data are not considered to be compliant with the SEC regulation 17-a4f.

📖 Further information is available in the SmartLock white paper.

Permissions, Authentication and Access Control Recommendations

Access Zones Best Practices

Access Zones provide secure, isolated storage pools for specific departments within an organization, allowing consolidation of storage resources without compromising security.

A minimum of two AD, LDAP or NIS servers provides redundancy and helps avoid access control lookups being a bottleneck. For larger environments, scaling the number of domain servers may be required.

The best practices for Access Zones include:

- The number of access zones should not exceed 50. The number of local users and groups per cluster should not exceed 25,000 for each. While possible, creating a larger number of local groups and/or users may affect system performance.
- Separate corporate tenants with Access Zones, up to a maximum of 50 zones.
- Use the system access zone for cluster management.
- Constrain different protocols (for example, NFS, SMB) to separate access zone

- If you create access zones, ensure that the directory paths for each zone under /ifs do not overlap. Instead, you should designate separate directory trees for each zone.
- As part of SmartConnect's support for multi-tenancy, Groupnet objects sit above subnets and pools and allow separate Access Zones to contain distinct DNS settings.

📖 For more information on identity management, authentication, and access control in combined NFS and SMB environments, please refer to the OneFS Multiprotocol Security Guide.

Job Engine Recommendations

In a clustered architecture, there are cluster jobs that are responsible for taking care of the health and maintenance of the cluster itself—these jobs are all managed by the OneFS job engine. The Job Engine runs across the entire cluster and reduces a task into smaller work items and then allocates these to multiple worker threads on each node. Jobs are typically executed as background tasks across the cluster, using spare or especially reserved capacity and resources. The jobs themselves can be categorized into three primary classes:

File System Maintenance Jobs

These jobs perform background file system maintenance, and typically require access to all nodes. These jobs are required to run in default configurations, and often in degraded cluster conditions. Examples include file system protection and drive rebuilds.

Feature Support Jobs

The feature support jobs perform work that facilitates some extended storage management function, and typically only run when the feature has been configured. Examples include deduplication and anti-virus scanning.

User Action Jobs

These jobs are run directly by the storage administrator to accomplish some data management goal. Examples include parallel tree deletes and permissions maintenance.

The Job Engine allows up to three jobs to be run simultaneously. This concurrent job execution is governed by the following criteria:

- Job Priority
- Exclusion Sets - jobs which cannot run together (i.e., FlexProtect and AutoBalance)
- Cluster health - most jobs cannot run when the cluster is in a degraded state.

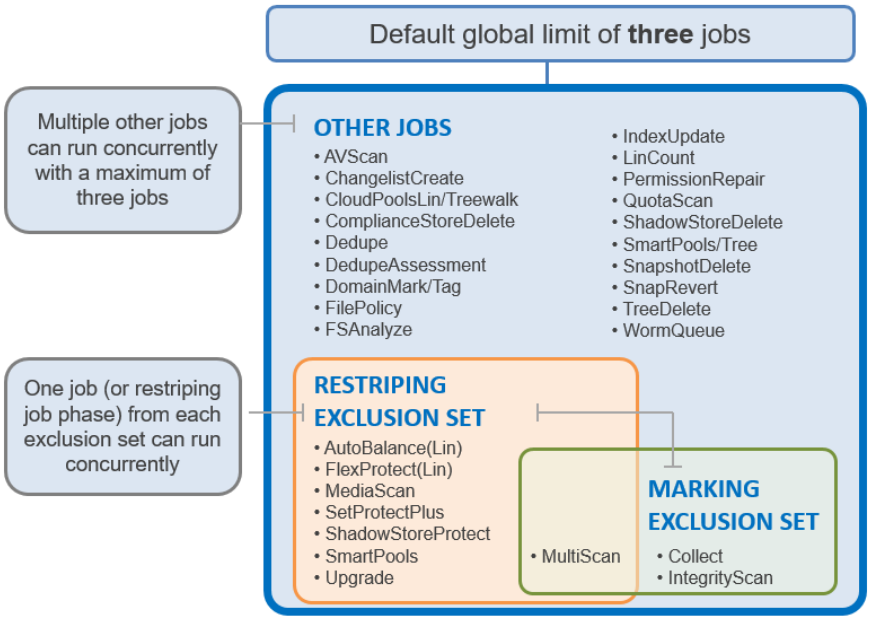

Figure 22: Job Engine exclusion sets.

The default, global limit of 3 jobs does not include jobs for striping or marking; one job from each of those categories can also run concurrently.

For optimal cluster performance, Dell EMC recommends observing the following Job Engine best practices.

- Schedule jobs to run during the cluster's low usage hours – overnight, weekends, etc.
- Where possible, use the default priority, impact and scheduling settings for each job.
- To complement the four default impact profiles, create additional profiles such as "daytime_medium", "after_hours_medium", "weekend_medium", etc, to fit specific environment needs.
- Ensure the cluster, including any individual node pools, is less than 90% full, so performance is not impacted and that there's always space to re-protect data in the event of drive failures. Also enable virtual hot spare (VHS) to reserve space in case you need to smartfail devices.
- Configure and pay attention to alerts. Set up event notification rules so that you will be notified when the cluster begins to reach capacity thresholds, etc. Make sure to enter a current email address to ensure you receive the notifications.
- It is recommended not to disable the snapshot delete job. In addition to preventing unused disk space from being freed, disabling the snapshot delete job can cause performance degradation.
- Delete snapshots in order, beginning with the oldest. Do not delete snapshots from the middle of a time range. Newer snapshots are mostly pointers to older snapshots, and they look larger than they really are.
- If you need to delete snapshots and there are down or smartfailed devices on the cluster, or the cluster is in an otherwise "degraded protection" state, contact Dell EMC Technical Support for assistance.
- Only run the FSAnalyze job if you are using InsightIQ and require filesystem analytics. FSAnalyze creates data for InsightIQ's file system analytics tools, providing details about data properties and space usage within /ifs. Unlike SmartQuotas, FSAnalyze only updates its views when the FSAnalyze job runs. Since FSAnalyze is a fairly low-priority job, it can sometimes be preempted by higher-priority jobs and therefore take a long time to gather all of the data.

- Schedule deduplication jobs to run every 10 days or so, depending on the size of the dataset.
- SmartDedupe will automatically run with a “low-impact” Job Engine policy. However, this can be manually reconfigured.
- In a heterogeneous cluster, tune job priorities and impact policies to the level of the lowest performance tier.
- Before running a major (non-rolling) OneFS upgrade, allow active jobs to complete, where possible, and cancel out any outstanding running jobs.
- Before running TreeDelete, ensure there are no quotas policies set on any directories under the root level of the data for deletion. TreeDelete cannot delete a directory if a quota has been applied to it.
- If FlexProtect is running, allow it to finish completely before powering down any node(s), or the entire cluster. While shutting down the cluster during restripe won't hurt anything directly, it does increase the risk of a second device failure before Flexprotect finishes re-protecting data.
- When using metadata read or metadata write acceleration, always run a job with the *LIN suffix where possible. For example, favor the FlexProtectLIN job, rather than the regular FlexProtect job.
- OneFS 8.1 and earlier, before running TreeDelete, ensure there are no quotas policies set on any directories under the root level of the data for deletion. In OneFS 8.2 and later, TreeDelete can delete a directory to which a quota has been applied using the ‘--delete-quota’ flag. For example: `#isi job start TreeDelete --paths=/ifs/quota –delete quota`
- Avoid data movement as much as possible during daily operations. SmartPools data placement requires resources which can contend with client IO. With a mixed node cluster where data tiering is required, the recommendation is to schedule the SmartPools job to run during off-hours (nights and/or weekends) when the client activity is at its lowest.

Job Engine Considerations

For optimal cluster performance, bear in mind the following OneFS Job Engine considerations:

- When reconfiguring the default priority, schedule and impact profile for a job, consider the following questions:
- What resources am I impacting?
- What would I be gaining or loosing if I re-prioritized this job?
- What are my impact options and their respective benefits and drawbacks?
- How long will the job run and what other jobs will it contend with?
- SyncIQ, the OneFS replication product, does not use job engine. However, it has both influenced, and been strongly influenced by, the job engine’s design. SyncIQ also terms its operations "jobs", and its processes and terminology bear some similarity to job engine. The job engine impact management framework is aware of the resources consumed by SyncIQ, in addition to client load, and will throttle jobs accordingly.
- A job with a name suffixed by ‘Lin’, for example FlexProtectLin, indicates that this job will scan an SSD-based copy of the LIN tree metadata, rather than access the hard drives themselves. This can significantly improve job performance, depending on the specific workflow.
- When more than three jobs with the same priority level and no exclusion set restrictions are scheduled to run simultaneously, the three jobs with the lowest job ID value will run, and the remainder will be paused.
- There is only one ‘mark cookie’ available for marking jobs. So, if marking job A is interrupted by another marking job B, job A will automatically be cancelled when it resumes after the completion of B.

- For mixed node (heterogeneous) clusters that do not have a license for SmartPools data tiering module, the SetProtectPlus job will run instead, and apply the default file policy. SetProtectPlus will be automatically disabled in the event that a valid SmartPools license key is added to the cluster.
- By default, FlexProtect is the only job allowed to run if a cluster is in degraded mode. Other jobs will automatically be paused and will not resume until FlexProtect has completed and the cluster is healthy again.
- Restriping jobs only block each other when the current phase may perform restriping. This is most evident with MultiScan, whose final phase only sweeps rather than restripes. Similarly, MediaScan, which rarely ever restripes, is usually able to run to completion more without contending with other restriping jobs.
- MediaScan restripes in phases 3 and 5 of the job, and only if there are disk errors (ECCs) which require data re-protection. If MediaScan reaches its third phase with ECCs, it will pause until AutoBalanceLin is no longer running. If MediaScan's priority were in the range 1-3, it would cause AutoBalanceLin to pause instead.
- If two jobs happen to reach their restriping phases simultaneously and the jobs have different priorities, the higher priority job (ie. priority value closer to "1") will continue to run, and the other will pause. If the two jobs have the same priority, the one already in its restriping phase will continue to run, and the one newly entering its restriping phase will pause.
- During MediaScan's verify and repair phases, attempts to re-read bad sectors can occasionally cause drives to stall briefly while trying to correct the error. This is typically only a very brief and limited interruption.

📖 Further information is available in the OneFS Job Engine white paper.

Cluster Management Recommendations

There are three access methods for configuring and administering a OneFS powered cluster:

- Command line interface (CLI) - either via SSH or serial console.
- Web interface (WebUI)
- RESTful platform API (PAPI)

ⓘ OneFS 9.0 introduces IPMI (Intelligent Platform Management Interface) support for Isilon Gen6 and PowerScale platforms. This functionality allows out-of-band console access and remote node power control via a dedicated Ethernet management port. and is configured and managed via the 'isi ipmi' CLI command set.

While the Web Interface is the most intuitive, menu driven, and simple to use cluster administration method, it is also the most limited in terms of scope. The CLI has a more comprehensive set of administrative commands that the WebUI, making it a popular choice for OneFS power users.

However, where possible the recommendation is use scripts to automate management of the cluster via the platform API. This also avoids challenges with the CLI and WebUI in parsing large numbers of configuration policies – for example, tens of thousands of NFS exports.

Cluster Capacity Management

When a cluster, or any of its nodepools, becomes more than 95% full, OneFS can experience slower performance and possible workflow interruptions in degraded mode and high-transaction or write latency sensitive operations. Furthermore, when a large cluster approaches full capacity (over 98% full), the following issues can occur:

- Performance degradation in some cases
- Workflow disruptions - failed file operations and inability to write data.
- Inability to make configuration changes or run commands to delete data and free up space

- Increased disk and node rebuild times.

To ensure that a cluster or its constituent pools do not run out of space:

- Add new nodes to existing clusters or pools
- Replace smaller-capacity nodes with larger-capacity nodes
- Create more clusters.

When deciding to add new nodes to an existing large cluster or pool, contact your sales team to order the nodes well in advance of the cluster or pool running short on space. The recommendation is to start planning for additional capacity when the cluster or pool reaches 75% full. This will allow sufficient time to receive and install the new hardware, while still maintaining sufficient free space.

The following table presents a suggested timeline for large cluster capacity planning:

The following table presents a suggested timeline for cluster capacity planning:

Used Capacity	Action
75%	Plan additional node purchases.
80%	Receive delivery of the new hardware.
85%	Rack and install the new node(s).

Figure 23: Capacity Panning Timeline

If an organization's data availability and protection SLA varies across different data categories (for example, home directories, file services, etc), ensure that any snapshot, replication and backup schedules are configured accordingly to meet the required availability and recovery objectives, and fit within the overall capacity plan.

Consider configuring a separate accounting quota for /ifs/home and /ifs/data directories (or wherever data and home directories are provisioned) to monitor aggregate disk space usage and issue administrative alerts as necessary to avoid running low on overall capacity.

ⓘ For optimal performance in any size cluster, the recommendation is to maintain 10% free space in each pool of a cluster.

To better protect smaller clusters (containing 3 to 7 nodes) we recommend that you maintain 15 to 20% free space. A full smartfail of a node in smaller clusters may require more than one node's worth of space. Maintaining 15 to 20% free space can allow the cluster to continue to operate while EMC helps with recovery plans.

Plan for contingencies: having a fully updated backup of your data can limit the risk of data loss if a node fails.

Maintaining appropriate protection levels

Ensure your cluster and pools are protected at the appropriate level. Every time you add nodes, re-evaluate protection levels. OneFS includes a 'suggested protection' function that calculates a recommended protection level based on cluster configuration, and alerts you if the cluster falls below this suggested level

OneFS supports several protection schemes. These include the ubiquitous +2d:1n, which protects against two drive failures or one node failure.

ⓘ The best practice is to use the recommended protection level for a particular cluster configuration. This recommended level of protection is clearly marked as 'suggested' in the OneFS WebUI storage pools configuration pages and is typically configured by default. For all current Isilon Gen6 hardware configurations, the recommended protection level is '+2d:1n' or '+3d:1n1d'.

Monitoring cluster capacity

- **Configure alerts**. Set up event notification rules so that you will be notified when the cluster begins to reach capacity thresholds. Make sure to enter a current email address in order to receive the notifications.
- **Monitor alerts**. The cluster sends notifications when it has reached 95 percent and 99 percent capacity. On some clusters, 5 percent (or even 1 percent) capacity remaining might mean that a lot of space is still available, so you might be inclined to ignore these notifications. However, it is best to pay attention to the alerts, closely monitor the cluster, and have a plan in place to take action when necessary.
- **Monitor ingest rate.** It is very important to understand the rate at which data is coming in to the cluster or pool. Options to do this include:
    - SNMP
    - SmartQuotas
    - FSAnalyze
- **Use SmartQuotas** to monitor and enforce administrator-defined storage limits. SmartQuotas manages storage use, monitors disk storage, and issues alerts when disk storage limits are exceeded. Although it does not provide the same detail of the file system that FSAnalyze does, SmartQuotas maintains a real-time view of space utilization so that you can quickly obtain the information you need.
- **Run FSAnalyze jobs**. FSAnalyze is a job-engine job that the system runs to create data for InsightIQ's file system analytics tools. FSAnalyze provides details about data properties and space usage within the /ifs directory. Unlike SmartQuotas, FSAnalyze updates its views only when the FSAnalyze job runs. Since FSAnalyze is a fairly low-priority job, it can sometimes be preempted by higher-priority jobs and therefore take a long time to gather all of the data. An InsightIQ license is required to run an FSAnalyze job.

Managing data

Regularly archive data that is rarely accessed and delete any unused and unwanted data. Ensure that pools do not become too full by setting up file pool policies to move data to other tiers and pools.

Provisioning additional capacity

To ensure that your cluster or pools do not run out of space, you can create more clusters, replace smaller-capacity nodes with larger-capacity nodes, or add new nodes to existing clusters or pools. If you decide to add new nodes to an existing cluster or pool, contact your sales representative to order the nodes long before the cluster or pool runs out of space. EMC recommends that you begin the ordering process when the cluster or pool reaches 80% used capacity. This will allow enough time to receive and install the new equipment and still maintain enough free space.

Managing snapshots

Sometimes a cluster has many old snapshots that take up a lot of space. Reasons for this include inefficient deletion schedules, degraded cluster preventing job execution, expired SnapshotIQ license, etc.

Ensuring all nodes are supported and compatible

Each version of OneFS supports only certain nodes. Refer to the "OneFS and node compatibility" section of the Supportability and Compatibility Guide for a list of which nodes are compatible with each version of OneFS. When upgrading OneFS, make sure that the new version supports your existing nodes. If it does not, you might need to replace the nodes.

Space and performance are optimized when all nodes in a pool are compatible. When you add new nodes to a cluster, OneFS automatically provisions nodes into pools with other nodes of compatible type, hard drive capacity, SSD capacity, and RAM. Occasionally, however, the system might put a node into an unexpected location. If you believe that a node has been placed into a

pool incorrectly, contact Dell EMC Technical Support for assistance. Different versions of OneFS have different rules regarding what makes nodes compatible

Enabling Virtual Hot Spare

The purpose of Virtual Hot Spare (VHS) is to keep space in reserve in case you need to smartfail drives when the cluster gets close to capacity. Enabling VHS will not give you more free space, but it will help protect your data in the event that space becomes scarce. VHS is enabled by default. Dell EMC strongly recommends that you do not disable VHS unless directed by a Support engineer. If you disable VHS in order to free some space, the space you just freed will probably fill up again very quickly with new writes. At that point, if a drive were to fail, you might not have enough space to smartfail the drive and re-protect its data, potentially leading to data loss. If VHS is disabled and you upgrade OneFS, VHS will remain disabled. If VHS is disabled on your cluster, first check to make sure the cluster has enough free space to safely enable VHS, and then enable it.

Enabling Spillover

Spillover allows data that is being sent to a full pool to be diverted to an alternate pool. Spillover is enabled by default on clusters that have more than one pool. If you have a SmartPools license on the cluster, you can disable Spillover. However, it is recommended that you keep Spillover enabled. If a pool is full and Spillover is disabled, you might get a “no space available” error but still have a large amount of space left on the cluster.

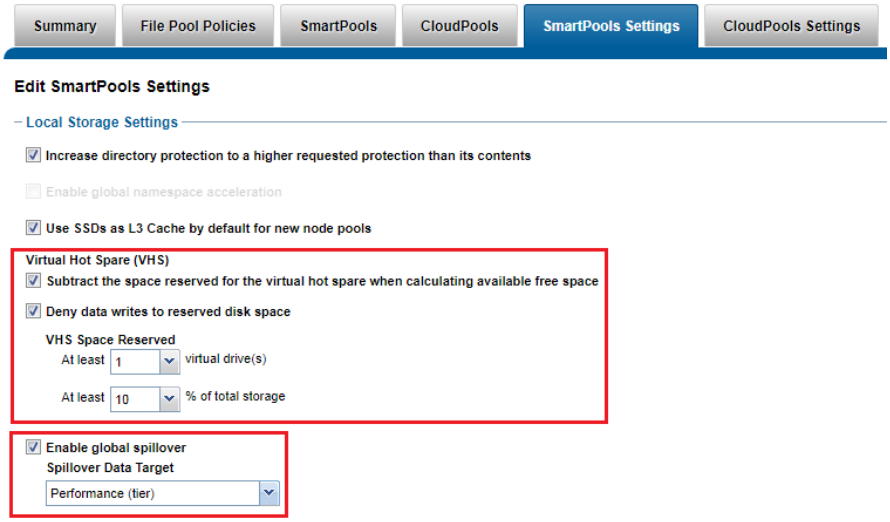

Figure 24: OneFS SmartPools VHS and Spillover configuration.

Confirm that:

- There are no cluster issues.
- OneFS configuration is as expected.

In summary, best practices on planning and managing capacity on a cluster include the following:

- Maintain sufficient free space.
- Plan for contingencies.
- Manage your data:
- Maintain appropriate protection levels.
- Monitor cluster capacity and data ingest rate.

- Consider configuring a separate accounting quota for /ifs/home and /ifs/data directories (or wherever data and home directories are provisioned) to monitor aggregate disk space usage and issue administrative alerts as necessary to avoid running low on overall capacity.
- Ensure that any snapshot, replication and backup schedules meet the required availability and recovery objectives and fit within the overall capacity.
- Manage snapshots.
- Use InsightIQ, ClusterIQ and DataIQ, usage forecasting, and verifying cluster health.

Best Practices Checklist

For optimal cluster performance, Dell EMC recommends observing the following best practices. Please note that this information will likely be covered elsewhere in this paper.

- ✓ Define, implement and regularly test a data protection strategy and business continuance plan.
- ✓ Maintain sufficient free space and pay attention to data ingest rate.
- ✓ Ensure that cluster capacity utilization (HDD and SSD) remains below 90%.
- ✓ If SmartPools is licensed, ensure that spillover is enabled (default setting).
- ✓ Manage your data: Archive infrequently accessed data and delete unused data.
- ✓ Maintain appropriate data protection levels.
- ✓ Periodically check the data protection level as the cluster grows.
- ✓ Recording your original settings before making any configuration changes to OneFS or its data services.
- ✓ Monitor cluster capacity and data ingest rate.
- ✓ Ensure that all desired data services are licensed and configured.
- ✓ Observe NFS and SMB connection limits.
- ✓ Many cluster configuration settings are global and have cluster-wide effects. If you consider changing cluster-wide configuration settings, be sure that you fully understand the global settings and their implications
- ✓ Manage snapshot creation and deletion schedules.
- ✓ Setup SmartConnect for load balancing and use Round Robin as the balancing policy.
- ✓ Recommend turning off client DNS caching, where possible. To handle client requests properly, SmartConnect requires that clients use the latest DNS entries.
- ✓ Ensure Virtual Hot Spare and SmartPools spillover both remain enabled (the default).
- ✓ Ensure the SmartPools job only runs during off-hours.
- ✓ If using SmartPools tiering, reconfigure the Storage Target field from “anywhere” to a specific tier or node pool to direct ingest to a performance node pool or tier.
- ✓ Add cluster to an InsightIQ monitoring instance
- ✓ Deploy a lab cluster or OneFS Simulator environment to test and validate any new cluster configurations before making changes that affect the production environment.
- ✓ Confirm that remote support functions work correctly through EMC Secure Remote Support (ESRS) and internal email/SNMP notifications.

- ✓ Upgrade OneFS to a newer release at least once a year via the non-disruptive 'rolling-upgrade' option.
- ✓ Configure and pay attention to cluster events and alerts.
- ✓ Regularly run and monitor OneFS Healthchecks.

Summary

Dell EMC PowerScale overcomes the problems that undermine traditional NAS systems by combining the three traditional layers of storage architecture—file system, volume manager, and data protection—into a scale-out NAS cluster with a distributed file system. The Dell EMC PowerScale architecture eliminates controller bottlenecks to reduce wall-clock runtimes for concurrent jobs, accelerates metadata operations, improves storage efficiency to lower capital expenditures, centralizes management to reduce operating expenses, and delivers strategic advantages to increase competitiveness in a global market.

TAKE THE NEXT STEP

Contact your Dell EMC sales representative or authorized reseller to learn more about how Dell EMC PowerScale scale NAS storage solutions can benefit your organization.

Visit Dell EMC PowerScale to compare features and get more information.

Learn more about Dell EMC PowerScale solutions

Contact a Dell EMC Expert

View more resources

Join the conversation with #DellEMCStorage