PURESTORAGE®

TECHNICAL GUIDE

Protecting Pure FlashArray™ File Services with Commvault®

Solution Architecture and Best Practices

Contents

Introduction

Pure Storage® FlashArray™ File Services brings the reliability, data reduction, and simplicity of Purity//FA to network-attached storage (NAS). Where traditional NAS is often complicated to deploy, difficult to manage, and painful to refresh, FlashArray™ unified all-flash storage delivers the same experience that's been delighting customers for years. Thanks to Evergreen Storage™, FlashArray stays forever young, so you'll never need to plan another disruptive migration. FlashArray forms the foundation of a truly Modern Data Experience™.

Commvault® Backup & Recovery delivers scalable, high performance protection of FlashArray file data. Using a parallel backup model that provides easy performance scaling, plus a wealth of data management features, Commvault provides everything you need to back up and restore FlashArray file data.

How to Use This Guide

We wrote this solution overview and best practices guide for backup administrators, storage administrators, and others to help you understand and then implement Commvault to protect and recover data on FlashArray File Services. The guide covers the solution architecture, key performance factors for protecting FlashArray File Services, basic configuration for simple environments and POCs, and best practices for achieving optimal performance and efficiency.

This guide assumes a working knowledge of the concepts and interfaces used with Commvault and FlashArray File Services. You can learn more about Commvault interfaces from Commvault documentation. Refer to the FlashArray user guide in the FlashArray web console for more information about managing FlashArray File Services.

Solution Architecture

Commvault software uses a parallel model for protecting data stored in network shares. Multiple threads scan the file system for data changes, and multiple data access nodes read the file content. This parallelism increases backup speeds and reduces configuration compared to single-threaded, single-reader solutions.

Scan times are one of the biggest pain points when it comes to protecting NAS shares. With file counts in millions or even billions and deep directory structures, scanning can add hours or even days to backup times.

Commvault automatically distributes the backup workload between data access nodes to maximize efficient resource utilization, and scaling is as simple as adding access nodes. The storage has to be able to support the load, however, and many NAS platforms can’t handle parallel backups without significant performance degradation.

The Commvault solution for FlashArray File Services uses pools of resources, as shown in Figure 1. Pools of data access nodes running Commvault file system agents read data from the server message block (SMB) shares and network file system (NFS) exports on the FlashArray. These access nodes compress and deduplicate the data from the FlashArray and send it to one or more MediaAgent data movers, which in turn write the data to backup storage (such as Pure Storage FlashBlade). With Amazon S3-based object storage, such as FlashBlade object storage, the access nodes can leverage Commvault Cloud Accelerator and send data directly to the backup storage, bypassing the MediaAgents. Commvault can throttle the data traffic from the access nodes to limit the load amount put on the FlashArray.

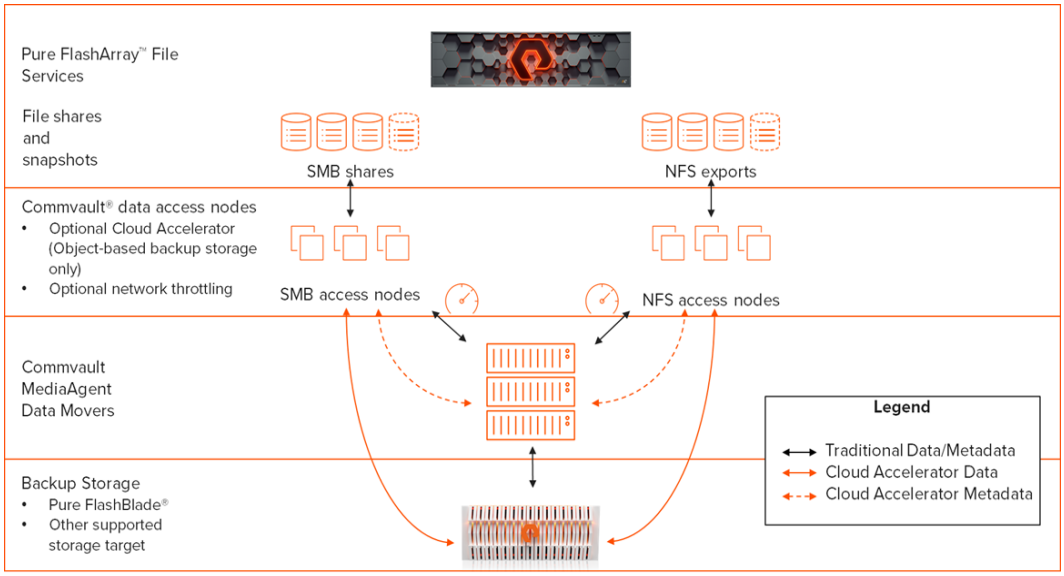

Figure 1. Reference Architecture

Components

Pure Storage FlashArray File Services

FlashArray File Services is the primary data tier, hosting unstructured data for access by users and applications. FlashArray File Services arranges the data in file systems and managed directories, which have policies applied. Users and applications access their data using the SMB or NFS protocol. Commvault groups file system content into subclients, logical links between data sets, and the policies that control backups.

Data Access Nodes

Data access nodes act as the readers and writers for backup and recovery operations. They run the Commvault file system agent for Windows or Linux. Windows access nodes process data on SMB shares, and Linux access nodes process NFS data. Commvault also supports UNIX for access nodes; however, Linux is the preferred operating system. Access nodes can be physical or virtual machines.

MediaAgents

Commvault MediaAgents serve several purposes. They are the data movers, receiving backup data from data access nodes and other backup clients, and writing it to backup storage. During recovery, they read data from backup storage and send it to recovery clients. MediaAgents also manage index data for the content they have protected or recovered, enabling users to browse and search protected data to find files to restore. They also track unique data in deduplication databases and provide lookup services to clients for source-side deduplication.

You can deploy both data access nodes and MediaAgent components to the same physical or virtual operating system instance if sufficient resources exist to meet the combined system requirements.

Cloud Accelerator

Commvault Cloud Accelerator is an optional component that you can install and enable on the data access nodes and other backup clients. It allows the nodes to send backup data directly to backup storage, bypassing the MediaAgent for backup and restore. Metadata is still sent to the MediaAgent for indexing purposes. Cloud Accelerator is only compatible with object storage platforms such as Pure FlashBlade.

Backup Storage

Commvault writes protected data to block, file, tape, and/or object-based storage for recovery and archive purposes. Commvault supports a broad list of storage products, including Pure FlashBlade.

System Recommendations for Commvault Components

Table 1 lists the Pure Storage recommended system configurations for deploying Commvault with FlashArray File Services. Please refer to System Requirements in Commvault documentation for up to date minimum requirements.

Component	Operating System	CPU	RAM
Data Access Nodes (SMB)	Windows Server 2016 or later	Minimum 8 cores Additional 2 cores for Cloud Accelerator	1GB per data reader (8GB or more recommended) Additional 2GB for Cloud Accelerator
Data Access Nodes (NFS)	Supported Linux or UNIX version	Minimum 8 cores Additional 2 cores for Cloud Accelerator	1GB per data reader Additional 2GB for Cloud Accelerator
MediaAgent	Windows Server 2016 or later Supported Linux version	See Hardware Specifications for Deduplication Mode in Commvault documentation for MediaAgent sizing	See Hardware Specifications for Deduplication Mode in Commvault documentation for MediaAgent sizing

Table 1: System specifications

Scaling Considerations

You can scale the solution vertically and horizontally at the data access nodes and MediaAgents. Larger data access nodes can support more data readers on fewer systems. Larger MediaAgents can service more backup and recovery streams.

Data stream scaling is nonlinear. Each additional stream you add will produce a smaller improvement until you reach a maximum throughput. Each additional stream will then decrease throughput. The actual maximum will vary based on several factors. See the Configuration Best Practices section for more detailed guidance on scaling for different data profiles.

Understanding Backup and Recovery Performance with FlashArray File Services

When setting recovery service level agreements (SLAs) for file data, you need to consider several factors; file access protocol, data profile, data streams, and data access node resources all have measurable effects on backup and restore throughput. Understanding these factors and their impacts will help you ensure can achieve your recovery point objective (RPO) and recovery time objective (RTO).

File Access Protocol

When reading or writing files on network storage, client systems use a protocol such as NFS or SMB. The protocol defines certain sequences of operations the client must perform to accomplish its specific task in a mixture of data and non-data operations. Protocol overheads are the non-data operations that the protocol forces. While protocol overhead exists with file systems on local or storage area network (SAN) block storage, it is generally more pronounced and visible with network file storage.

You can expect SMB to be slower than NFS by 10% or more, and the gap will grow as average file size goes down. The next section explains how file size affects backup and recovery.

You should use the same protocol for backup and recovery as you use for primary client access. While you can export the same file system over NFS and SMB, and use different protocols for client access and backup and recovery, this is not recommended. File system permissions are only approximated when you use a different protocol, and when permissions are not restored to the original state loss of access can result.

Data Profile

Both NFS and SMB use a combination of data and non-data operations, such as file locking and queries, to transfer file data and metadata. The ratio of data to non-data operations affects how fast the storage and client or backup agent can exchange data. The ratio varies with file size because both NFS and SMB can transfer data in blocks of 1MiB or more. With small files under 1MiB in size, it is possible to send the entire file content in a single data request, but the same transaction requires multiple non-data requests; protocol overhead is relatively high as a result. With large files, protocol overhead is lower since the same file may need many data requests, with larger data blocks, but the same number of non-data requests per transaction. While performance may vary somewhat between SMB and NFS for the same data set due to protocol differences, both will see better performance with large files than with small ones.

The data profile, therefore, has a direct impact on the performance you can expect during backup and recovery. You can back up and restore data sets with large average file sizes faster than data sets with small average file sizes, and you can back up a handful of large files much faster than a large number of small files.

File count plays a large part in file scan performance. At scales of millions of files, scan times can vary by minutes or even hours and have a significant impact on overall backup times. You will need to estimate your scan time based on the number of files. You will also need to review scan times periodically as your file systems grow.

File system structure is the other important factor in scan times. Directory width, which is the number of directories at the same level of the file system, affects how fast Commvault can scan the file system for changes and begin backing up file data. Multiple readers will look in parallel for files and directories altered since the last backup. Commvault divides the workload at the directory level, so the more directories there are, the more parallelism it can attain. In lab tests, scan rates

ranged from 6,000 files/second for a narrow but deep directory structure with six top-level directories to over 20,000 files/second for a very wide, deep structure with over 1,200 top-level directories.

Commvault runs all file scanning for a backup job on the same data access node. If file scan times are high due to large file counts, breaking up the data set into multiple backup jobs will let you distribute the load, but it brings more management complexity.

Make sure your RPO and RTO calculations factor in average file sizes, file counts, and directory structure in addition to total data size. If you have multiple managed directories or file systems with large file counts, consider creating separate subclients for them so you can back them up in separate jobs and distribute the file scan processing.

Data Streams

Backup software breaks data sets into multiple streams. Each stream represents a connection from the data source to the backup storage through a data mover. Each data access node can manage multiple parallel data streams for backup and restore, with more streams generally increasing throughput. Data streams have a 1:1 relationship with data readers, which perform the actual read activity on the data source.

Data profile has a significant effect on the impact of increasing streams. File systems with small average file sizes will reach a lower peak throughput with a smaller number of streams compared to file systems with large average file sizes. In lab tests, a full backup of a data set on SMB with an average file size under 1MiB scaled to 6 streams, while a data set with a 2MiB average file size scaled to 54 streams and over 7x the peak throughput. (For multiple concurrent jobs, the peak stream count for the latter data set was 18.) An NFS data set with an average file size of 100MiB scaled to 12 streams but with almost 15% higher throughput than either of the other data sets.

Commvault defaults to two data readers per data access node, which should give acceptable throughput for most data sets and SLAs. If you are seeing lower than expected throughput, you can increase the number of streams for the subclient so Commvault runs more streams per node, or you can assign more nodes but continue to run two streams per node.

Data Access Node Resources

Data access nodes perform several functions. In addition to reading and writing data on the FlashArray, they handle data compression and deduplication, and they transmit data to and from MediaAgents or storage. All these operations consume CPU and memory resources. The more parallel data streams an access node manages, the more CPU and memory it needs, so larger nodes can manage more streams and therefore process more data than smaller nodes. Adding nodes can also improve throughput by letting you run more parallel backup streams.

Scaling Recommendations for Data Access Nodes

When deciding how many data access nodes to deploy, bear in mind that:

- You should start with at least two data access nodes per protocol for any environment size to provide redundancy. Two nodes should support over 3 TiB/hour, depending on the data profile. Add nodes based on capacity and throughput needs.

- For the same number of data streams, spreading the streams across more nodes generally increases throughput compared to adding more streams to the same nodes. Consider adding more nodes before you increase the number of streams per node.
- Overall concurrency is easily overlooked, especially when jobs will share the same data access nodes. While a single job with 12 streams may achieve a high level of throughput, multiple jobs running in parallel, with 12 streams each, may see lower overall throughput. You should start with a single subclient each for SMB and NFS and break out data sets into more subclients only as necessary.
- Data profiles will affect how many streams and data access nodes can effectively process a given data set. With very small average file sizes, additional streams or nodes may not be able to increase throughput.
- Virtual data access nodes provide easy elasticity. You can quickly add resources or deploy new nodes if you need more throughput, and you can remove idle nodes. With virtual nodes, you will need to ensure the hypervisor hosts have sufficient available resources to support the number of nodes.

Basic Configuration

The basic configuration should help you implement the solution as quickly and simply as possible. To add FlashArray File Services into your existing Commvault environment, there are three steps you must perform. You must deploy data access nodes, create a network share client that represents the FlashArray, and configure one or more subclients.

NOTE: This document assumes you have a FlashArray joined to Active Directory and hosting data, one or more MediaAgents deployed, and a Commvault server plan created.

Deploying Data Access Nodes

Data access nodes are simply client systems with a Commvault file system agent installed. You must deploy the agent on the appropriate operating system. SMB shares require the Windows file system agent, while NFS exports require the Linux or UNIX file system agent.

See the System Recommendations for Commvault Components section for recommended system resources. See the Data Access Nodes section for best practices on deploying data access nodes.

You can deploy data access nodes using Command Center. In the navigation pane, expand **Manage**, then click **Servers**. Click the **Add server** link, then select **Windows/UNIX file server** from the dropdown to open the **Add server** form, shown in Figure 2. Complete the form as follows:

- In the **Host name** field, enter the DNS fully qualified domain name (FQDN) for the data access node, then click the **plus (+)** button. Enter all nodes.
- In the **User name**, **Password**, and **Confirm password** fields, enter the credentials for a user with full administrative rights to the access node.
- In the **OS Type** field, select the appropriate operating system family.
- In the **Select package(s)** dropdown, select **File System**.
- Leave the **Plan** field in its default unselected state.
- Click the **Install** button to go ahead with the deployment.

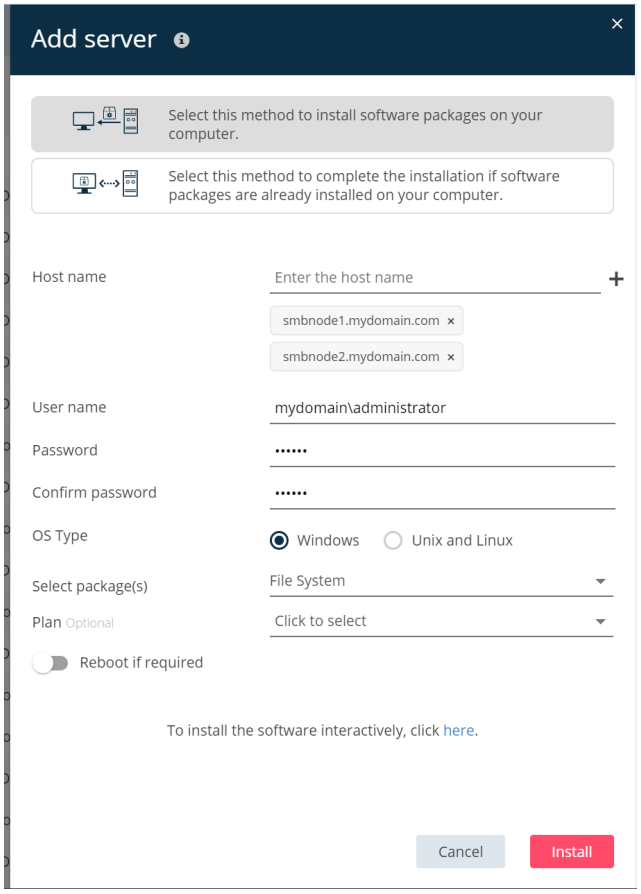

Figure 2. Add server form in Command Center

Creating the Network Share Client

You should manage all file data for the FlashArray through a single network share client. You can create the client using Command Center. In the navigation pane, expand **Protect**, then click **File Servers**. Click the **Add server** link, then select **NAS server** from the dropdown to open the **Add NAS server** form, shown in Figure 3. Complete the form as follows.

- In the **Server name** field, enter a display name for the new client.
- In the **Host name** field, enter the DNS fully FQDN for the FlashArray computer account. If you have segregated production and backup networks, enter the FQDN associated with the backup interface.
- Enable the **CIFS** and/or **NFS** checkboxes depending on the data you plan to protect.
- Click the **Save** button to create the client.

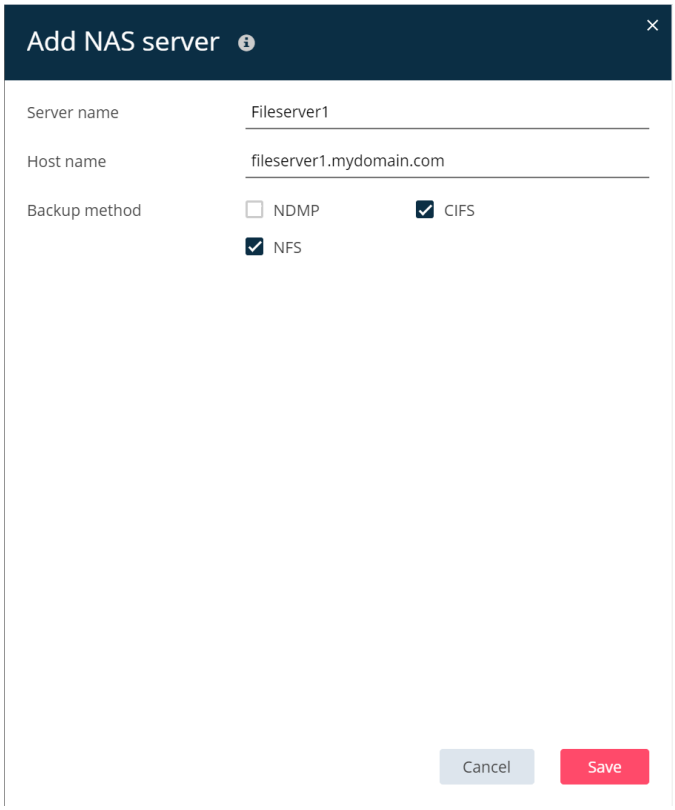

Figure 3. Add NAS server form in Command Center

Creating Subclients

You can create subclients using Command Center. In the navigation pane, expand **Protect**, then click **File Servers**. Locate and click the name of the network share client you created to open the client details view. In the **Agents** list, click either **Linux File System** or **Windows File System** based on the data you will protect. Click the **Add subclient** link to open the **Add subclient** form.

As shown in Figure 4, complete the **Add subclient** form as follows:

- in the **Subclient name** field, enter a display name for the subclient.
- In the **Backup method**, select either **CIFS** (for SMB shares) or **NFS** as appropriate.
- In the **Proxy** field, select any data access node you deployed earlier.
- In the **Plan** field, select the server plan you want to use to manage backup scheduling and retention.
- In the **Content** area, enter the UNC or NFS paths to the data on FlashArray. To add multiple paths, use the **+** button. You can also click the **Browse** link and select content directly from the FlashArray. See the Align Subclients and File Systems section for best practices on defining subclient content.
- Click the **Impersonate user** link and select saved credentials, or enter credentials in the **User name** and **Password** fields, then click the **OK** button to save the impersonation details. See the Managing Backup Access section for more information on saved credentials.

- If you need to manage exclusions, you can click the **Exclusions** and **Exceptions** links. Refer to Customizing the Backup Content for a Server in Commvault documentation for details on managing exclusions.
- Click the **Save** button to finish creating the subclient.

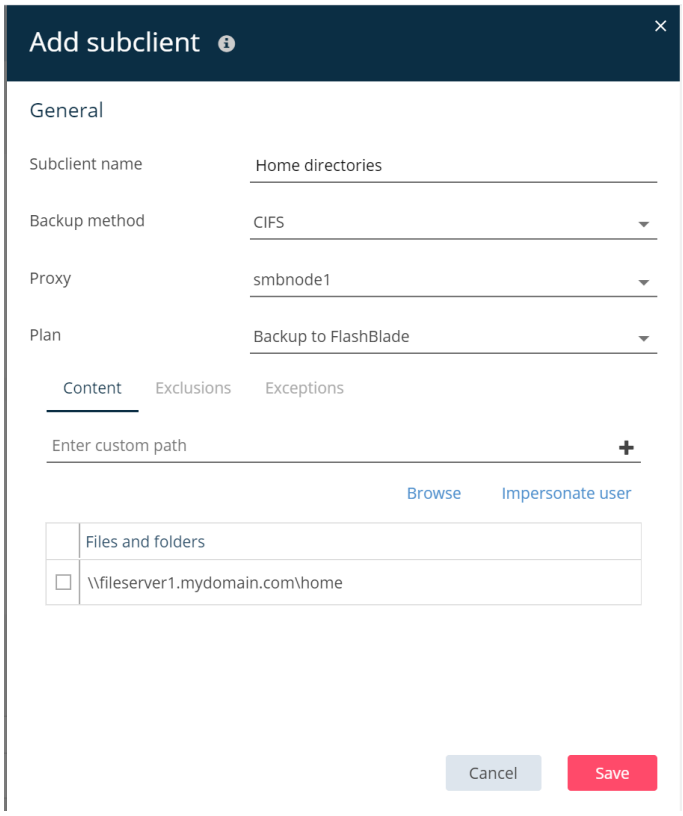

Figure 4. Add subclient form in Command Center

You have now successfully configured Commvault to protect FlashArray File Services data.

NOTE: Many of the best practices contained in the following sections require advanced options that aren't included in the basic configuration. Continue reading to learn more about optimizing backups for FlashArray File Services.

Configuration Best Practices

File Systems and Managed Directories

File systems act as top-level containers for FlashArray File Services data. Within a file system, FlashArray allows two types of directories: normal directories and managed directories. From the client perspective both appear identical, and you can apply permissions and other properties to them. However, there are several key differences between them.

Create normal directories through the file system from a client system as you would create directories on a local file system. The FlashArray administrative interfaces do not list normal directories and cannot manipulate them.

Managed directories are special directories that allow you to apply policies, such as export and protection policies. You can only create managed directories through the FlashArray administrative interfaces, either with GUI, CLI or API. When you create a managed directory, it adds a directory into the client view, with the path you specify. You can only take snapshots

on managed directories. Managed directories can't nest within normal directories, although they can nest in other managed directories to a certain depth. The file system root is also a managed directory.

There are several limitations on managed directories. You can't convert a normal directory to a managed directory, and vice versa. You also can't delete managed directories from client systems, and administrators cannot delete them unless they are empty.

For more information on managed directories and how to use them, see the Pure Storage white paper Introduction to Pure FlashArray™ File Services.

Group Associated Data Within File Systems

You should group data sets that you want to use, protect and recover together into the same file system. For example, create user home directories as managed directories in a single file system. If different teams need shared SMB directories, create managed directories for each team in a single file system. Grouping directories will make protection and recovery simpler, especially as relates to snapshots and open file protection.

Protect Managed Directory and Policy Configurations

Managed directories are a key part of a FlashArray File Services environment. It is therefore important to protect their configurations so you can recover them if there's a catastrophic loss.

You can create your own export script to run regularly to capture the key configuration details into a recovery script, which you can then execute prior to full recovery to recreate the key structures. You will need the following commands, in this order:

1. `purefs list –cli:` Outputs commands to recreate all file systems.
2. `puredir list –cli:` Outputs commands to recreate all managed directories.
3. `purepolicy nfs list --cli`: Outputs commands to recreate all NFS policies.
4. `purepolicy nfs rule list --cli`: Outputs commands to recreate the rules in each NFS policy.
5. `purepolicy smb list --cli`: Outputs commands to recreate all SMB policies.
6. `purepolicy smb rule list --cli`: Outputs commands to recreate the rules in each SMB policy.
7. `puredir export list --cli:` Outputs commands to recreate all managed directory exports.
8. `purepolicy snapshot list --cli`: Outputs commands to recreate all protection policies.
9. `purepolicy snapshot rule list --cli`: Outputs commands to recreate the rules for each protection policy.
10. `purepolicy snapshot list --member --cli`: Outputs commands to recreate the memberships for each protection policy, effectively enabling snapshot schedules.

You must protect the scripts external to the FlashArray to ensure you can recover from a catastrophic loss. If you save them to a directory on a FlashArray file system, you can use Commvault to back up the generated configuration scripts for extra protection.

These commands together will generate a script that will recreate the entire set of file systems, managed directories, export policies, and protection policies. You can skip any part of the configuration that already exists when performing the restore. This is ideal for most major recoveries. If you want to capture configurations for only a subset, such as a single file system or managed directory, you can apply filters (which all support wildcards) to the commands as follows:

- To output only a specific file system, use `purefs list <file system name> --cli`.
- To output only the directories in that file system, use `puredir list --file-system <file system name> --cli`.
- To output all exports for that file system, use `puredir export list --dir "<file system name>:*" --cli`.
- To output all snapshot policy memberships for that file system, use `purepolicy snapshot list --member --filter "member.name='<file system name>:*'" -cli`.

We don't recommend capturing a subset of policies and rules, as it's not possible to filter the list of policies or rules based on members, only by policy name and type. It is simpler to capture all policies and use the `puredir export list` or `purepolicy snapshot list --member` command to capture the specific directories' memberships.

IMPORTANT: Running the output of the `puredir export list --cli` command will make data accessible to end users. If you prefer to recover data before creating exports, save the command's output to a separate file that you can run independently of the other script. You may also wish to run a filtered capture of a subset of exports, such as all file system roots, by inserting a name or wildcard into the command. For example, to output the commands to recreate exports of file system roots, you would use the command `puredir export list --dir "*:root" --cli`.

Creation of Managed Directories Before Restoring Data

When recovering an entire file system or managed directory after a catastrophic loss, it is critical that you recreate managed directories before restoring data. Backup software has no way of knowing whether a directory is normal or managed. On recovery, the backup software will create any required top-level directories as normal directories if they do not already exist. The software can't convert these to managed directories without a significant effort, which may include repeating data recovery.

If you are exporting configuration scripts regularly, you can simply run the latest version to recreate the managed directories and policies before you restore file data.

Data Access Nodes

Placing Data Access Nodes

Commvault data access nodes need to read and write data as quickly as possible using the SMB and NFS protocols. If possible, locate nodes in the same site as the FlashArray and on the same virtual local area network (VLAN)..

Joining SMB Data Access Nodes to Active Directory

While data access nodes do not need to belong to Active Directory for backups to work, you can take advantage of Kerberos security if they do.

Managing Data Access Nodes Using Commvault Groups

Creating groups within Commvault will simplify managing data access nodes, both on initial configuration and as you scale. As data grows or performance needs change, you can add or remove nodes from the group, and later jobs will automatically use the correct nodes.

Create separate groups for SMB and NFS data to avoid extraneous errors. You can create separate groups for different data sets for various reasons, such as needing different levels of parallelism. The same access node can belong to multiple groups.

You can create a group using Commvault Command Center. In the navigation pane, click **Manage**, then click **Server groups**. Click the **Add** link in the upper right corner to open the **Add server group** form. Select the data access nodes to add to the group, then click the **Save** button, as shown in Figure 5.

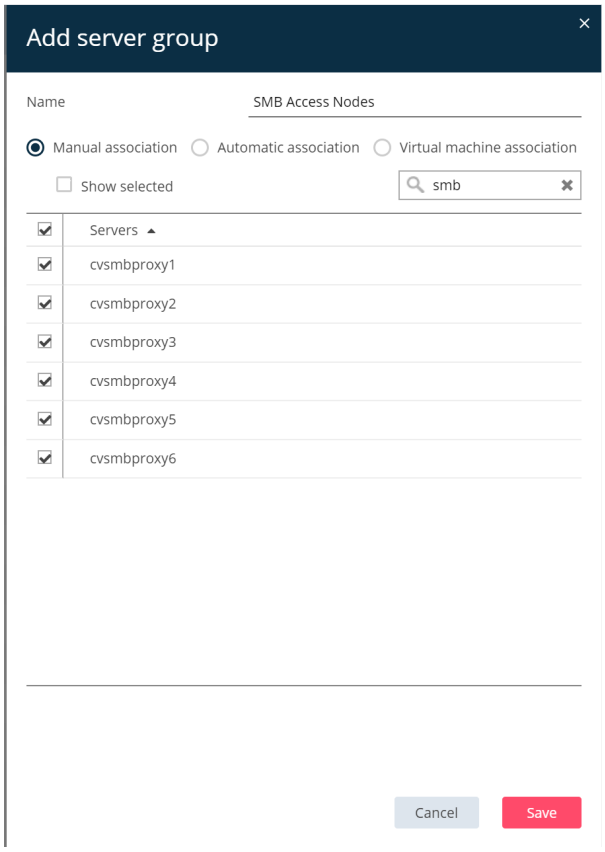

Figure 5. Add server group form in Commvault Command Center

NOTE: To associate a group to a data set for backup or restore, you must use the CommCell Console interface.

Using Cloud Accelerator with Fast Object Storage

Cloud Accelerator reduces the load on MediaAgents by allowing data streams to flow directly from clients to object-based backup storage. With fast object storage platforms such as Pure Storage FlashBlade, using Cloud Accelerator lets MediaAgents manage more data on fewer resources. We don't recommend using Cloud Accelerator with archive class object storage.

NOTE: There may be network implications from allowing traffic from data access nodes to backup storage. Using separate VLANs for user and backup traffic will mitigate these.

Refer to the Pure Storage guide Best Practices for Configuring Commvault with FlashBlade for more information about deploying and configuring Cloud Accelerator.

MediaAgents

Index Storage

DASH full backups rely heavily on the MediaAgent index. Performance on the underlying storage is critical to data processing speed. You should place the index on the highest performing storage available to the MediaAgent to minimize DASH full job times.

File data indexing collects more metadata and therefore requires more index space than other data types. Follow Commvault sizing recommendations to ensure you have enough capacity for the index.

You can set the index location during MediaAgent installation, and can change it at any time. You must use the CommCell Console to change the index location. In the CommCell Browser pane, expand **Storage Resources**, then **MediaAgents**. Right-click the MediaAgent, then select **Properties** to open the **MediaAgent Properties** dialog. Select the **Catalog** tab. Enter the new path in the **Index Directory** field, as shown in Figure 6. Click the **OK** button to commit the change. Commvault will move the index contents to the new location in the background.

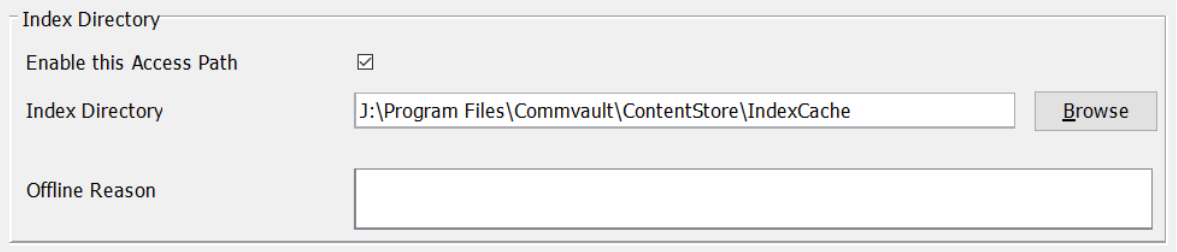

Figure 6. Setting the index directory location

Job Results Directory Storage

During recovery, the restore process builds a task list on one or more MediaAgents before starting data movement, or on the data access nodes if you are using Cloud Accelerator. You can find the list in the job results directory. The workload is almost entirely write I/O. The performance of the storage hosting the job results directory can have a significant impact on recovery times for large data sets, and it can even cause job timeouts and failures if it cannot process the I/O workload fast enough. The exact job size at which timeouts occur will vary based on the speed of the underlying storage. Full file system recovery is the most likely scenario where you'll encounter issues.

It is possible to mitigate the risk of timeouts. See the Limit File Count in Restore Jobs section for more details.

When recovering a large number of files, Commvault will generate a significant amount of data in the job results directory. The exact amount depends on the number of files and their path lengths. Ensure your MediaAgents have sufficient storage for the expected recovery scenarios and refer to Job Results Directory Disk Space Calculation in Commvault documentation for specific sizing guidelines.

You can set the job results location for a client when you install Commvault software if you install interactively or using the CommCell Console. You can also change the job results location at any time after installation. You must use the CommCell Console to change the job results location. In the CommCell Browser pane, expand **Client Computers**. Right-click the client, then select **Properties** to open the **Client Properties** dialog. Click the **Advanced** button to open the **Advanced Client Properties** dialog. Select the **Job Configuration** tab. Enter the new path in the **Job Results Directory** field, as shown in Figure 7. Click the **OK** button on both dialogs to commit the change. Commvault will copy the job results contents to the new location in the background.

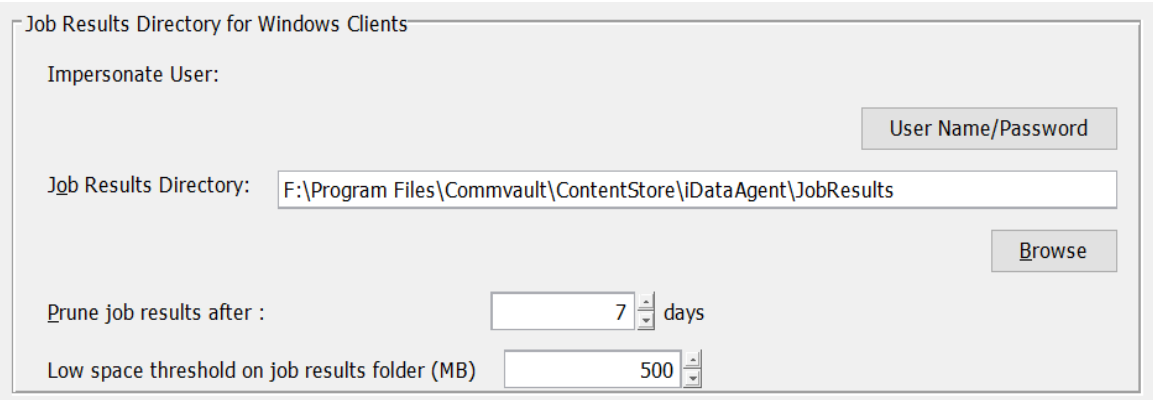

Figure 7. Setting job results directory location

Reducing Antimalware Impact

To prevent performance degradation due to antimalware software scanning, ensure that you apply all Commvault recommended exclusions to each MediaAgent. See the Antimalware Exclusions section for Pure Storage observations, and refer to Recommended Antivirus Exclusions for Windows and Recommended Antivirus Exclusions for UNIX and Macintosh in Commvault documentation for the full recommended exclusion list.

Performance

This section details the best practices for Commvault configuration elements when protecting FlashArray File Services data.

Set Application Read Size

Application read size controls how much data Commvault tries to read at a time. Larger read sizes generally improve throughput but increase resource consumption on the data access nodes. A read size of 1024KB gives the best balance of performance and resource consumption.

You must configure application read size in the CommCell Console. In the CommCell Browser pane, expand **Client Computers**. Expand the network share client you created, then the appropriate file system agent. Click the backup set. In the right pane, right-click the subclient you want to modify, then select **Properties** to open the **Subclient Properties** dialog. Click the **Advanced** button to open the **Advanced Subclient Properties** dialog. Select the **Performance** tab. As shown in Figure 8, enable the **Application Read Size** checkbox, and set the spinner value to 1024. Click the **OK** button on both dialogs to commit the change.

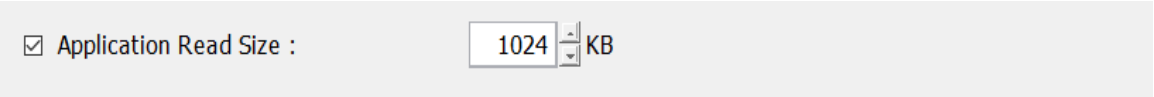

Figure 8. Setting application read size

Antimalware Exclusions

Portions of the backup and restore processes involve heavy write activity on the MediaAgents. Real-time malware protection can significantly reduce performance. In lab tests, excluding Commvault processes improved backup throughput by over 20%.

Lab tests identified the process and path exclusions as most impactful when using Microsoft Windows Defender, as seen in Tables 2 and 3.

Process name	Systems
cvd.exe	MediaAgent Access Node
cvfwd.exe	MediaAgent Access Node
ClMgrS.exe	MediaAgent Access Node
CVDistributor.exe	Access Node
IFind.exe	Access Node
CLBackup.exe	Access Node
CVODS.exe	MediaAgent
clRestore.exe	Access Node
SIDB2.exe	MediaAgent

Table 2. Recommended process exclusions

NOTE: To reduce the risk of spoofing, you should exclude processes using the full path or a file signature, if supported.

Path	Systems	Example
Index	MediaAgent	C:\Program Files\Commvault\ContentStore\IndexCache
Job Results	MediaAgent	C:\Program Files\Commvault\ContentStore\iDataAgent\JobResults

Table 3. Recommended path exclusions.

Refer to your antimalware product documentation for instructions on setting exclusions.

Limiting Network Throughput

You may want to limit backup traffic to reserve resources on the FlashArray. This will prevent backups from slowing down production workloads. There are several ways you can do this:

- Reducing the number of data readers. Each reader has an effective maximum processing rate, so fewer readers will result in fewer file operations and lower throughput. The exact amount will vary depending on factors such as processor type, CPU core count, and data profile. You can configure subclients to use as little as one reader. See Configure Additional Data Readers for Large File Systems for instructions on setting data readers.
- Using Commvault's network throttling capability to limit traffic between the data access nodes and MediaAgents. The throttled traffic is set by the post-deduplication size, so it doesn't give a hard cap on reads at the network shares. Unless low-bandwidth WAN links separate the nodes and MediaAgents, you should only throttle backup traffic from the nodes to the MediaAgent, which will reduce backup throughput while leaving restore traffic able to use all available bandwidth. Refer to Limiting Bandwidth Usage During Backups in Commvault documentation for instructions on enabling network throttling.
- Using quality of service (QoS) on your network to ensure production workloads take priority over backup.
- Using the hypervisor's traffic shaping capabilities to control bandwidth If you are using virtual data access nodes.
- Combining some or all of these options to minimize the chance of backups affecting production performance.

SMB-specific Best Practices

Managing SMB Backup Access

Since backups occur through network shares, you must configure a service account to access and back up the files. Create the service account in the Active Directory domain to the joined FlashArray File Services, and add it to the Backup Operators group on the domain. You should not grant the account local login access or elevated privileges on any Windows system.

IMPORTANT: You must set the uidNumber and gidNumber attributes on the service account in Active Directory. You should not use this account for any other purpose in your environment.

Adding the account to Credential Manager within Commvault allows you to easily reuse the account across multiple shares without having to reenter the password. You can also easily change the password through Credential Manager and avoid updating all the Commvault configurations. To add an account to Credential Manager:

1. In Command Center, navigate to **Manage**, then **Security**. Click the **Credential Manager** tile.
2. Click the **Add user** link.
3. Complete the **Add user** form as follows (Figure 9):
   - In the **Account type** dropdown, select **Windows Account**.
   - In the **Credential name** dropdown, enter a display name for the account.
   - In the **User name** field, enter the account's username in domain\username format.
   - In the **Password** field, enter the account's password.
   - (Optional) In the **Description** field, enter a description for the account.
   - (Optional) Set the **Owner** and **User/Group** field to the Commvault user and group that will manage the credentials.
   - Click the **Save** button to create the credential.

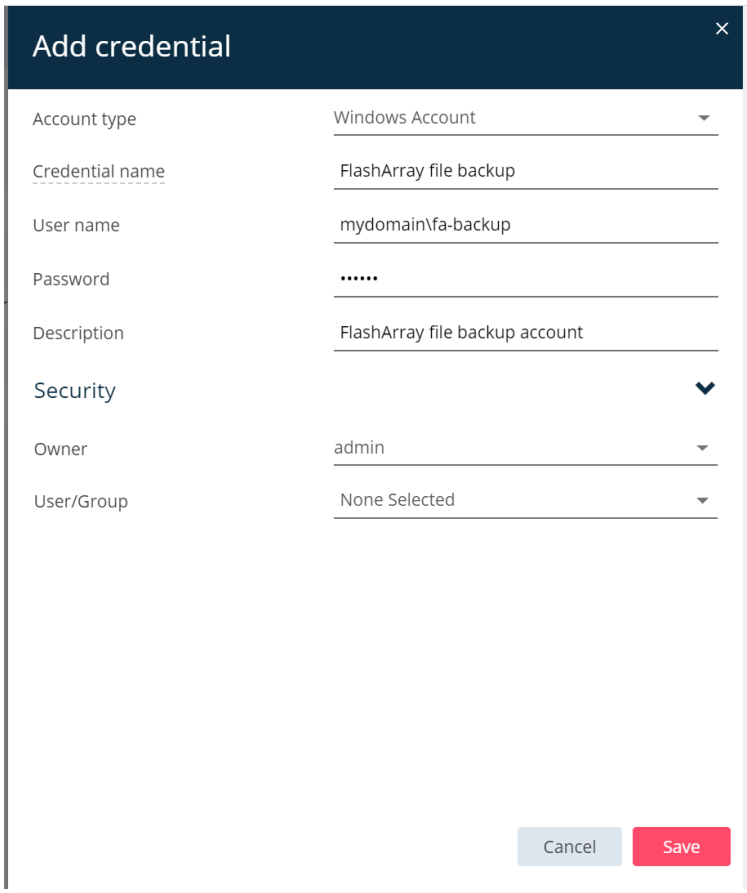

Figure 9. Add credential form

You should rotate the password regularly on the service account by changing it in both Active Directory and Credential Manager. Regular password changes reduce the risk of account compromise. To update the password in Commvault:

1. In the **Credential manager** view, click the display name for the account in the credential list.
2. In the **Edit credential** form, enter the new password in the **Password** field (Figure 10).
3. Click the **Save** button to commit the change.

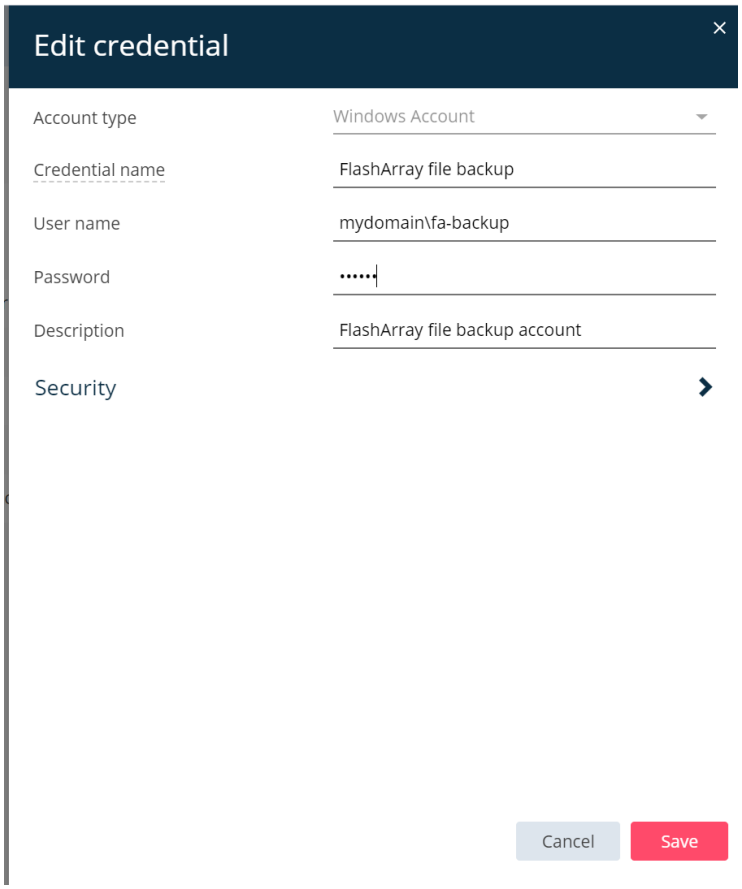

Figure 10. Edit credential form

NFS-Specific Best Practices

Use Automount Option with NFS Exports

For files where open file protection is not an issue, use Commvault's automount capability to simplify protection. Automount removes the need for matching static mounts on all NFS data access nodes. Simply specify the NFS path, in the format `<array name>:/<export>`. For example, if the array name is "fileserver1" and the export is "home," you would set the content to fileserver1:/home.

*IMPORTANT: To prevent failures during backup, you must specify the nolock mount option. You must set mount options using the CommCell Console. In the CommCell Browser pane, expand **Client Computers**. Expand the network share client you created, then the Linux file system agent. Click the backup set. In the right pane, right-click the subclient you want to modify, then select **Properties** to open the **Subclient Properties** dialog. Click the **Advanced** button to open the **Advanced Subclient Properties** dialog. Select the **Auto Mount Options** tab. As shown in Figure 11, in the **Mount Options** field, enter "nolock." If you wish to specify any other NFS mount options, enter them as a comma-separated list. Click the **OK** button in both dialogs to commit the change.*

Figure 11. Setting NFS automount options

Refer to Creating a Subclient with Auto-mount Contents for NFS Exports in Commvault documentation for more information on using automount.

Statically Mount File System Roots to Data Access Nodes for Restore

During recovery of NFS data, Commvault allows only a local target path and does not support NFS format target paths. This requires you to set a consistent NFS mount on each access node. If you create a static mount to the file system root, you can restore data into both normal and managed directories through that path. Adding the NFS mount to the /etc/fstab file or autofs will ensure that you can recover data without extra steps.

Keep the root mounts separate from any mounts used for backups, even if the backup mounts also point to the file system root. Separating backup and recovery paths will let you reconfigure backups as your needs change without affecting recovery processes.

Backup Best Practices

Use an Incremental Forever Backup Strategy

Commvault features DASH full backups, a deduplication-integrated synthetic full backup technology. The combination of incremental and DASH full backups is far more efficient than using traditional full backups that must read the entire file set. In lab tests, DASH full backups were over 25 times faster than traditional full backups.

You should schedule synthetic full backups in your server plan to run regularly. A weekly schedule is best, but the frequency should be at least monthly and no longer than the shortest retention period.

Configure Additional Data Readers for Large File Systems

By default, Commvault will assign two data streams to every access node assigned to the subclient, either directly or through inheritance, during backup jobs. Depending on the number of nodes and the size and profile of the data set, you may need more streams to achieve your desired backup time.

To increase the number of streams, you can either add data access nodes or increase the number of data readers for the subclient. The recommended access node size can handle more than two streams, so increasing the data readers will often be the better choice.

You must use the CommCell Console to increase the number of data readers. Open the **Subclient Properties** dialog for the subclient, then click the **Advanced** button to open the **Advanced Subclient Properties** dialog. Select the **Performance** tab. As shown in Figure 12, in the **Number of Data Readers** frame, select the **Use __ data readers** option, then set the desired number of streams. Enable the **Allow multiple data readers within a drive or mount point** check box. Click the **OK** button to commit the change.

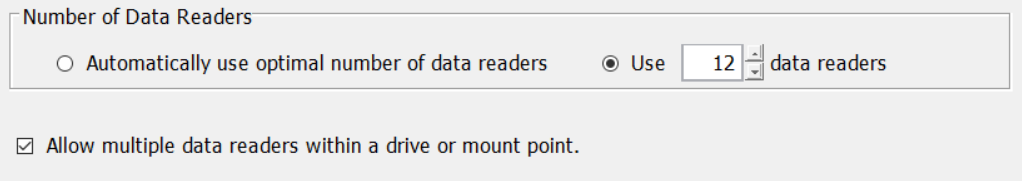

Figure 12. Setting the number of data readers

NOTE: During a backup job, Commvault will divide the data readers among the nodes based partly on available resources, which may or may not result in even load distribution.

Align Subclients and File Systems

Commvault supports any number of configurations using subclients to protect file system data. A single subclient can protect multiple file systems, and you can divide a file system up into multiple subclients. Each approach has its pros and cons.

Management complexity increases significantly with each subclient. Best practice is to start with a single subclient that encompasses all protected file systems. You can create more subclients for file systems or directories that need different retention or scheduling, or to manage stream counts separately. File systems you back up using snapshots should also have separate subclients. Adding multiple subclients for the same file system has implications for open file protection. See the Protecting Open Files section for further details.

IMPORTANT: While you can create subclients in both the Command Center and CommCell Console interfaces, only the CommCell Console allows you to use groups to assign data access nodes and exposes settings for data readers and scripts.

NOTE: Each subclient you create will add data streams during backups. Consider the total stream count across all your backup jobs and your access node size when configuring the number of data readers.

Use Inheritance with Data Access Node Assignments

Commvault clients use a hierarchy of components. For network share backups, the client has one or more Linux and/or Windows File System agents; these contain backup sets, which in turn contain the subclients that represent specific data sets. You can assign data access nodes at the agent and subclient. If you assign nodes to the agent, any new subclient will inherit the assignment, unless you associate a different set of nodes to the subclient. This will simplify administration since multiple subclients can automatically use the same pool of nodes.

To assign data access nodes to an agent, you use the CommCell Console. In the **CommCell Browser** view, expand the network share client you created for the FlashArray. Right-click either the **Linux File System** or **Windows File System** agent, then select **Properties** from the popup menu. Select the **Data Access Nodes** tab. As shown in Figure 13, in the **Available Access Nodes** list box, select the nodes and/or groups you want to assign, then click the **Add** button. Click the **OK** button to commit the assignment.

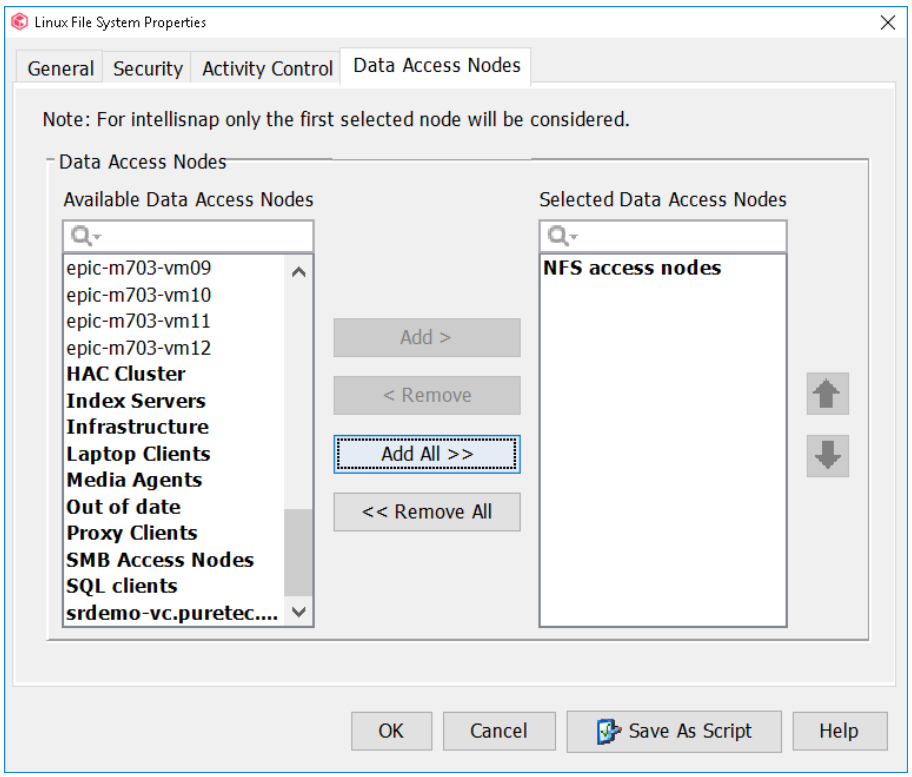

Figure 13. Assigning a group of data access nodes

*NOTE: The **Available Access Nodes** list box will show all clients of the appropriate OS type, e.g. the Windows File Systems agent will show the Windows clients. You can enter a full or partial client name in the search box above the list box to filter the available clients.*

Protecting Open Files

Ensuring that you back up critical files is an important part of an effective NAS protection strategy. There are several reasons why your backup software might not protect all of your NAS data. Your backup software can't back up locked files, files in use by users or applications, or files that users have modified or locked between the file scan and data read stages of a backup job.

Storage snapshots are a simple and effective way to provide that assurance. Integrating snapshots into Commvault network share backups requires several other configuration changes, which vary based on the file protocol. This section details the procedures and best practices for using Commvault with FlashArray File Services snapshots.

FlashArray File Services Snapshot Behavior

FlashArray File Services creates snapshots of managed directories, including file system roots, capturing the state of the entire directory structure at the time of the capture. Snapshots will capture nested managed directories. You can manage snapshots manually, by script, or by policy, and access them through the .snapshot directory in the base of the managed directory where you took the snapshot.

Snapshot names have two parts. When creating a snapshot or a protection policy, the administrator supplies a client name, which the array appends to the managed directory name. The array also appends a numeric suffix, which increments automatically, to the client name. For example, for a FlashArray called "fafile," a file system called "homedirs" might have a

managed directory named “alan.” If the snapshot client name is “commvault” and the suffix has incremented to 3, the snapshot name would be “homedirs:alan.commvault.3.” Figure 14 shows the management view of the managed directory and snapshot.

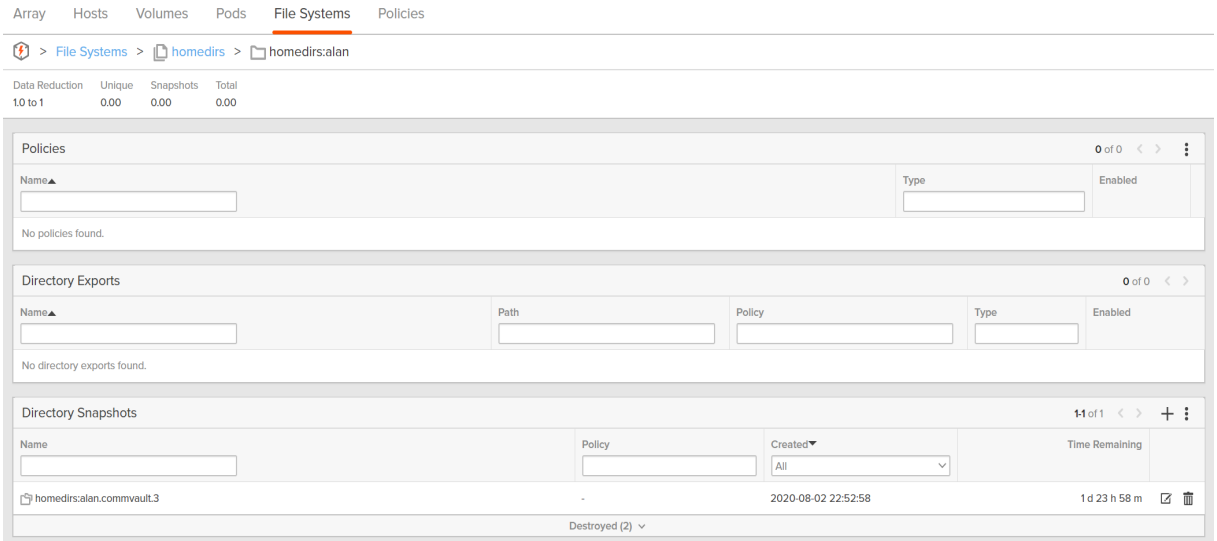

Figure 14. Managed directory snapshot, management view

Within the client view, you see only the client name and suffix of the snapshot, so for this example, an SMB client could access the snapshot at \\fafile\homedirs\alan\.snapshot\commvault.3. Figure 15 shows the client view.

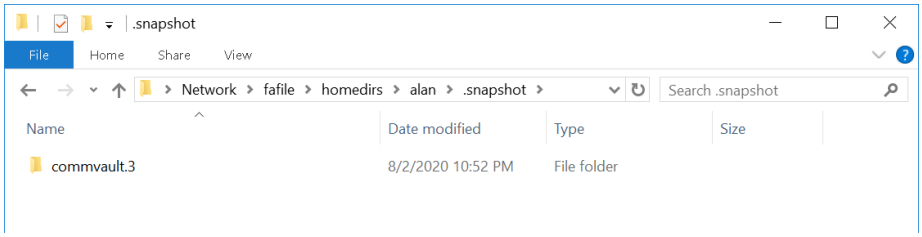

Figure 15. Managed directory snapshot, client view

It is important to note that there is no guarantee that the suffix will always increment by one between snapshots, only that each snapshot for a file system or managed directory will have a higher suffix than its earlier snapshot.

This naming scheme has two key implications for Commvault backups. First, it means snapshot names will be different for each snapshot, requiring configuration changes each backups to ensure Commvault sees a consistent content path. Second, controlling or predicting snapshot names is not possible. Processes to update the configuration will need to account for these restrictions. We cover recommendations for SMB and NFS exports in detail in the coming subsections.

General Snapshot Best Practices

Create Snapshots at the File System Root

Since snapshots of the file system root contain all managed directories for the file system, creating snapshots at the file system root is the simplest way to drive backups for the entire file system. For example, if you use managed directories for individual user home directories in a single file system, you can create a single snapshot of the file system and use that as

the backup source for all the home directories. You can still create snapshots of individual managed directories if you need to protect them on separate schedules.

NOTE: You can configure multiple Commvault subclients to use the same snapshot as the backup source. See the NFS and SMB subsections for specific details on how to best use snapshots with Commvault.

Align Backup Content to Snapshot Locations

Consider setting backup content to the same level where you create snapshots. If you create snapshots at managed directories, configure backups to protect the managed directories. If you create snapshots at file system roots, target the file system root for backup. With this configuration, a single Commvault subclient ties to a single snapshot or associated snapshots. This in turn avoids contention and more complexity, such as multiple jobs trying to map the same snapshot or backup jobs having to remap the same snapshot, and it makes schedule coordination between protection policies and Commvault simpler.

Use Policies to Create Snapshots for Backups

While you can create snapshots manually or through scripts, it is simpler to use a protection policy and let the array manage the scheduling. Regardless of the approach, your process will have to find the appropriate snapshot to use for backup.

Automate Configuration Updates for Backups

For each backup job, you will need to find the appropriate snapshot to use as the content source and to update the access path before the data backup begins. A scripted process can identify the latest snapshot through the .snapshot directory in the client file system view or through the array management interfaces, using either timestamps or suffixes to differentiate between snapshots.

Best Practices for Using Snapshots with NFS

There are several best practices related to using snapshots of NFS file systems with Commvault. Following these practices will help ensure smooth backup operations and avoid protection gaps.

Use Static Mounts Instead of Automount

While Commvault has an automount feature with NFS backups, detailed in the NFS-Specific Best Practices section, you cannot use automount together with snapshots due to the changing snapshot names. You must create static NFS mounts on the data access nodes and configure Commvault to use the local file system path to the mounts. For example, you might create a directory at /mnt/home for home directories and configure /mnt/home as the content in Commvault.

IMPORTANT: Mounts must match across all data access nodes. Mismatches will compromise data integrity or cause job errors and negative performance impact.

Mount Snapshot Directories to Data Access Nodes

Always mount the snapshot directories to the data access nodes rather than the .snapshot directory itself. Mounting the snapshots to local mount paths on the nodes ensures that each backup has properly refreshed file system content. Continuing the previous examples, to protect the “home” file system using a snapshot named commvault.1, mount fileserver1:/home/.snapshot/commvault.1 to /mnt/home and configure /mnt/home as the subclient content in Commvault. If

you create a new snapshot named commvault.2 before the next backup, change the mount for /mnt/home to fileserver1:/home/.snapshot/commvault.2, and repeat the process for each subsequent backup.

IMPORTANT: Snapshot names may not increment predictably. Always confirm snapshot names when updating NFS mounts.

Dismount Snapshots Between Backups

Dismounting the snapshot from the data access nodes after the backup completes will ensure there is no unexpected behavior when mounting the next snapshot.

Use Pre-scan and Post-backup Scripts in Commvault to Manage Mounts

For scripted mount and dismount processes, configure the scripts to run as part of the Commvault backup jobs. Mount scripts should be set to run as the pre-scan process, and dismount scripts should run as the post-backup process.

You must use the CommCell Console to assign scripts to subclients. In the CommCell Browser pane, expand **Client Computers**. Right-click the client, then select **Properties** to open the **Client Properties** dialog. Click the **Advanced** button to open the **Advanced Client Properties** dialog. Select the **Pre/Post Process** tab. As shown in Figure 16, enter or browse to the mount script path in the **PreScan Process** field. Enter or browse to the dismount script path in the **PostBackup Process** field. Click the **OK** button on both dialogs to commit the change. Commvault will copy the job results contents to the new location in the background.

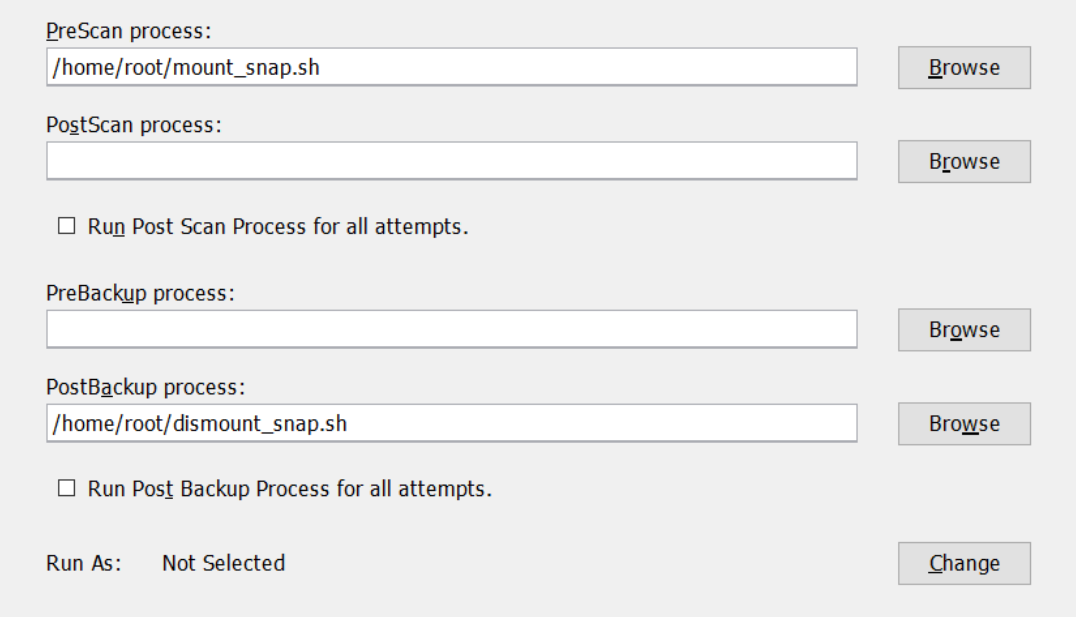

Figure 16. Assigning scripts to a subclient

*IMPORTANT: Do not set the **Run Post Backup Process for all attempts** option in Commvault. If the backup encounters a retriable error, having this option enabled would dismount the snapshot before the backup is complete, making the backup content unavailable and causing failures for all subsequent attempts.*

Scripts should update fstab file. Scripts managing mount and dismount operations should update the /etc/fstab file in addition to the actual mount operations. This will ensure that snapshots are still available in the event an access node

reboots unexpectedly. Mount scripts should add snapshots to fstab, while dismount scripts should remove them. You can also use autofs in place of fstab.

Best Practices Using Snapshots with SMB

There are several best practices related to using Commvault and SMB with snapshots of file systems. Following these practices will help ensure smooth backup operations and avoid protection gaps.

For SMB backups, Commvault uses UNC paths to content rather than local file system paths. Maintaining a consistent subclient configuration requires creating symbolic links, or symlinks, in the SMB file system.

You can create symlinks using the mklink command `mklink /d <link path> <target path>`. You can delete symlinks using the rmdir command `rmdir <link path>`.

NOTE: You cannot modify Symlinks after you create them. To redirect a symlink to a new target path, you must delete the symlink and create a new one.

Configure Windows Data Access Nodes to Resolve Remote Symbolic Links

By default, Windows will not create or resolve symlinks that point to locations on network shares to prevent a malicious user from redirecting clients to malware or other unwanted files. You must enable remote symbolic link evaluation on the data access nodes using the built-in fsutil command `fsutil behavior set SymlinkEvaluation R2R:1`.

NOTE: Setting the option L2R:1 is not required for the solution and is not recommended.

IMPORTANT: You must run the command on all Windows data access nodes to prevent backup failures.

Do Not Enable Remote Symbolic Link Resolution on Other Clients

To minimize risk, do not enable remote symlink evaluation on any systems that will not be backing up SMB shares from snapshots, and do not grant users login access to the data access nodes.

Use Relative Paths to Link Targets

You can configure Symlinks for either absolute paths, which use the entire path to the target, or relative paths, which contain the directory route from the symlink to the target. Absolute paths take the form C:\parent\child, while relative paths exclude the top level. For example, a link called C:\dir1\mylink that points to C:\dir2 could either specify the absolute path C:\dir2 or the relative path ..\dir2, since .. would represent C:\. Windows doesn't allow you to create a symlink within a UNC path that uses an absolute UNC path as its target, so you must create symlinks using relative paths to snapshots.

Create Symlinks in Subdirectories

To prevent users from accessing snapshots using symlinks, you should put links in a subdirectory beneath the file system root, parallel to the managed directories. You should restrict access for this directory to allow only administrators and the backup service account and use the same directory for all symlinks that point to the same file system.

For example, to create a link to a snapshot named "commvault.3" under the "home" file system on FlashArray fileserver1, you can create a directory called "links" under the file system root and create a link called "backup." The link target would be "..\.snapshot\commvault.3." The associated `mklink` command would be

```
mklink /d \\fileserver1\home\links\backup ..\.snapshot\commvault.3.
```

Use Pre-scan and Post-backup Scripts in Commvault to Manage Symlinks

You should configure scripts to manage symlinks as the pre-scan and post-backup processes in the Commvault subclient. The pre-scan process should create a symlink, and the post-backup process should remove it.

*IMPORTANT: Do not set the **Run Post Backup Process for all attempts** option in Commvault. If the backup encounters a retriable error, having this option enabled would dismount the snapshot before the backup is complete, making the backup content unavailable and causing failures for all subsequent attempts.*

Recovery Best Practices

Recreate Managed Directories Before Restoring Data

After a catastrophic loss that requires creating file system structures from scratch, you *must* recreate managed directories before you restore any data. File recovery will automatically create any missing directories, and since clients like backup software can only create normal directories, you would not be able to apply export or protection policies after restoring files. If you have protected your configuration, you can easily recover the managed directories before you begin the data recovery process.

For more information on protecting the managed directory configuration, see the Protect Managed Directory and Policy Configurations section.

Use Multiple Streams and Data Access Nodes for Large Restores

As with backup, each data stream has a functional performance limit. This can extend recovery times beyond SLA requirements for large amounts of data. You should use multiple streams across multiple data access nodes for large restores to maximize recovery performance.

Use the CommCell Console to enable multiple streams and data access nodes during restore. After you browse for data to recover, the **Restore Options** dialog appears. As shown in Figure 17, you can use the **Number of streams** field to specify how many data streams the restore will use.

In the same dialog, you can enable the **Use Multiple Nodes** checkbox to distribute the streams across multiple data access nodes. When you enable the checkbox, the **Advanced Restore Options** dialog appears, open to the **Data Access Nodes** tab. You can add or remove Commvault clients and groups in the **Selected Data Access Nodes** list. Click the **OK** button to commit the access node selection. This will distribute the streams as evenly as possible across the nodes.

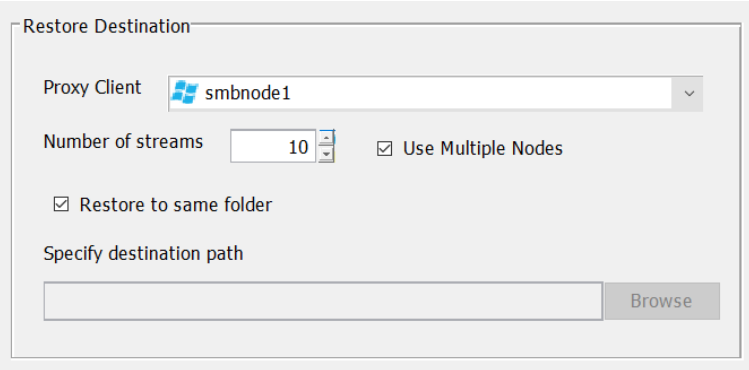

Figure 17. Setting stream options for restore

Limit File Count in Restore Jobs

As described in the Job Results Directory Storage section, restore jobs generate a task list before moving data. The amount of time this takes depends directly on the number of files to restore. At some point, the time required to create the task list will exceed the job timeout and the restore will fail. The exact number of files will vary between environments based on MediaAgent specifications. To avoid failures due to the timeout, you should limit recovery to under 15 million files in a single job. If you need to restore over 15 million files, create multiple restore jobs.

Conclusion

Commvault Backup & Recovery provides a powerful, scalable platform for data protection and recovery that you can use with confidence to ensure the availability of your FlashArray File Services data. Commvault can meet all your backup and recovery needs, from individual files to entire arrays.

Adding Pure FlashBlade and Commvault Cloud Accelerator further improves the performance and scalability of both FlashArray File Services backup and recovery and the broader Commvault environment by shifting data traffic away from MediaAgents.

To learn more about FlashArray File Services or Commvault Complete Backup & Recovery, contact your Pure and Commvault account teams, or your reseller of choice.

Additional Resources

- Commvault Backup & Recovery
- Commvault public documentation
- Best Practices for Configuring Commvault with FlashBlade
- Introduction to Pure FlashArray™ File Services

About the Author

Roy Child is a Senior Solution Architect with Pure Storage, responsible for defining data recovery solutions around various data protection applications. He defines Pure Storage solutions and reference architectures for protecting and recovering primary workloads such as Oracle, SQL, and VMware. Roy has worked in and with the data protection industry for 20 years, from end-user to IT architect with multiple backup and recovery products, followed by product management with Commvault. Roy joined Pure Storage in April 2019.

Pure Storage, Inc.
650 Castro Street, #400
Mountain View, CA 94041