PURESTORAGE® rubrik

TECHNICAL WHITE PAPER

Unlocking the Power of Unstructured Data with Rubrik and Pure Storage

Harness the power of unstructured data with performance, resilience, and security.

Contents

Introduction

Global data volumes are set to explode, with UBS projecting a 10-fold increase between 2020 and 2030, reaching an astonishing 660 zettabytes—equivalent to 129GB per person on Earth. Meanwhile, IDC reports that 90% of organizational data is unstructured, a category rich in untapped insights.

Unstructured data is the lifeblood of modern business, encompassing everything from documents and spreadsheets to social media posts, emails, medical images, and genomic sequences. This data drives AI/ML-powered decision-making and underpins innovations like generative AI, revolutionizing healthcare, finance, retail, and manufacturing industries.

The future belongs to those who can unlock the value hidden in their unstructured data. As unstructured data scales into the petabyte range and beyond, traditional storage systems can no longer keep pace. To harness its potential, organizations need modern storage and backup solutions designed for cyber resilience, massive scale, lightning-fast performance, airtight security, and unparalleled efficiency.

Breaking Free of Legacy Limitations

Modern applications are evolving at an unprecedented speed, leaving traditional IT infrastructure struggling to keep up. For IT teams, simplicity and flexibility are no longer optional—they are essential to adapt to ever-changing business needs. This applies to all core infrastructure functions, including file services, which must shift from rigid, pre-planned systems to ones that are dynamic with real-time adaptability. Unfortunately, traditional file storage systems remain mired in complexity. Over decades, legacy file architectures have accumulated layers of protocols and management tools, creating cumbersome, overloaded, and inflexible systems.

This complexity poses serious risks. With their patchwork of disparate technologies, legacy systems require constant monitoring, updates, and maintenance, increasing the likelihood of user error and amplifying exposure to external threats. Their rigidity hinders businesses from responding to modern demands, leaving them vulnerable in a landscape where agility is key. Simply put, outdated file services are a liability in today's dynamic IT environment, where resilience, speed, scalability, and security are non-negotiable.

Modern applications demand real-time file services tightly integrated with backup environments. Real-time Enterprise File, delivered via the Pure Storage® platform, provides the flexibility to adapt and reconfigure dynamically in response to changing needs. Unlike legacy systems, which struggle with block and file use cases, modern solutions eliminate the burdensome layers of complexity, delivering simplicity and scalability. These systems empower businesses to meet the demands of AI-driven workloads and other advanced applications without compromising performance or increasing risks.

When unstructured data reaches petabyte scale, legacy backup solutions falter. Traditional backup methods, such as network data management protocol (NDMP), were not designed for today's unstructured datasets' massive volume and throughput requirements. They lack the scalability and application awareness to manage these challenges, leading to incomplete protection, performance bottlenecks, and backup failures. Organizations require backup solutions built for massive scale to secure data effectively, delivering speed, efficiency, and comprehensive protection.

Beyond performance, legacy solutions expose businesses to increased security risks. Fragmented management layers create vulnerabilities that cyber threats, like ransomware, can exploit. Each additional layer of complexity expands the potential attack surface and increases the risk of vulnerabilities and human error. Modernizing file services and backup infrastructure is not just a matter of improving efficiency—it is a critical step toward safeguarding unstructured data in an era where cyber threats are becoming increasingly sophisticated and unstructured data is a primary target for malicious payloads.

Closing the Gaps

A scalable, flexible storage architecture is key to getting unstructured data under control, yielding the performance, resilience, security, and availability that organizations demand. The first step is eliminating data silos. The Pure Storage platform offers a range of unified block and file systems in the FlashArray™ family to meet your specific requirements as you consolidate multiple disparate workloads. Optimized for performance or capacity, FlashArray systems can scale to petabytes of data in a single storage pool in an ultra-dense footprint. Advanced data reduction through intelligent deduplication and compression gives you effective capacity that can be many times larger—replacing entire racks of legacy disk-based storage. Non-disruptive hardware upgrades let you add storage or even transition in place to a more powerful FlashArray system, with no data migration or downtime required, to keep up with growth and new workloads.

Data security is a priority on FlashArray. Features such as always-on encryption, multifactor authentication (MFA), granular role-based access control (RBAC), and rule-based access policies help prevent unauthorized access. File-level auditing lets you monitor who's looking at your sensitive data. Immutable, indelible SafeMode™ Snapshots ensure your data is safe from accidental or malicious deletion. It's all managed through a simple, intuitive interface.

While a modern storage platform addresses your unstructured data needs, it amplifies the need for a modern approach to data protection and recovery. Rubrik® NAS Cloud Direct is a data management solution designed to secure unstructured data at petabyte scale, and it marries perfectly with FlashArray. It delivers rapid, efficient backup that keeps up with FlashArray performance and capacity. It accelerates and simplifies data recovery, including after cyber attacks. It also offers data visibility and anomaly detection when used with Rubrik Cloud Vault. With security as its foundation, NAS CD is designed with a zero-trust approach that keeps data immutable and protected from attackers. Its multi-layered defenses include air-gapped storage, encryption, object locking, and strict access controls. These measures ensure that attackers cannot discover, modify, or encrypt backups.

For an even better recovery experience, you can deploy Pure Storage FlashBlade® as a backup target. The Pure Storage platform provides an ideal object storage target for NAS CD backups. Like NAS CD, FlashBlade scales seamlessly to keep up with your data growth. Its performance ensures you can recover data quickly, whether someone just needs a few older versions of files or you need to mass recover after a cyber attack–and object locking and SafeMode Retention Lock ensure that an attacker can't destroy your backups. Configuring FlashBlade with NAS CD is easy, requiring little ongoing administration. Both software and hardware upgrades are nondisruptive.

Solution Overview

Combining the Pure Storage platform with NAS CD solves both primary and secondary storage challenges. FlashArray provides resilient, high performance, scalable, and secure unstructured data services that grow to petabyte scale, and NAS CD removes the pitfalls of legacy solutions.

Figure 1 illustrates the high-level solution architecture.

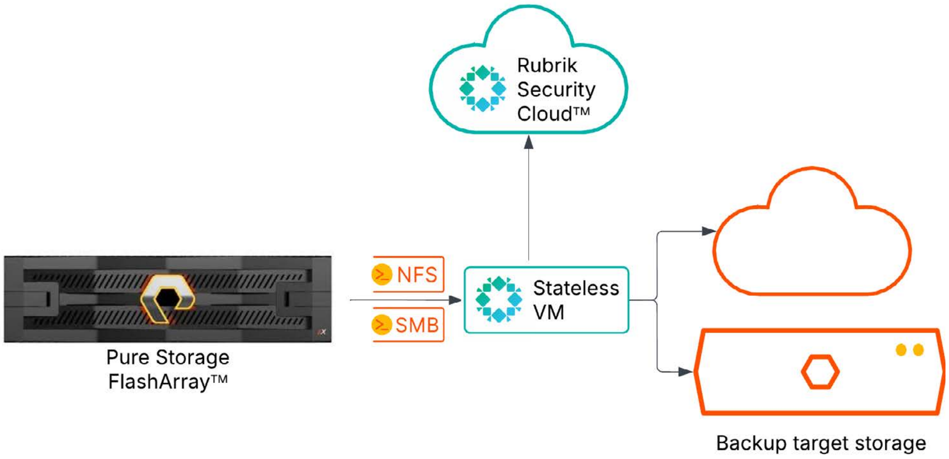


FIGURE 1 NAS Cloud Direct with Pure Storage FlashArray

FlashArray File Services

Real-time Enterprise File on the Pure Storage platform can meet demanding unstructured data needs with FlashArray File Services. FlashArray supports billions of files, thousands of concurrent users, and individual file sizes up to 4PB. Data services such as immutable snapshots and replication provide resilience against malicious or accidental data destruction and simplify disaster recovery.

In addition to performance and security, Real-time Enterprise File simplifies deployment and management. Data storage dynamically consumes all resources at a global level without the restrictions of creating reservations or allocations, so you can quickly onboard new use cases. Always-on features, including QoS, Auditing, and Multiprotocol, allow admins to take advantage of capabilities without pre-planning or extensive configuration and testing.

Pure Fusion™ provides automated fleet-wide operations, including workload placement and rebalancing of any combination of arrays across the entire Pure Storage platform, enabling you to easily and rapidly transform your environment from a single array up to the entire fleet under management with ease.

Data Security with Rubrik NAS Cloud Direct

Rubrik Security Cloud™ (RSC) delivers an intelligent, SaaS-based management platform for NAS Cloud Direct, transforming how organizations streamline policy management, reporting, and recovery workflows. Designed for seamless scalability, RSC not only orchestrates the deployment of data mover virtual machines but also performs advanced threat analysis on protected data, ensuring rapid and reliable recovery after ransomware incidents.

In the data center, one or more stateless VMs work in perfect sync with RSC, coordinating backup, restore, and copy operations as dedicated data movers. During backups, these VMs identify the content requiring protection, read files from FlashArray, index the data, and send this metadata to RSC. They then compress and process the data into NAS CD backup format before writing it to the target storage.

The process is reversed for recovery—data is retrieved from backup storage and restored to the FlashArray with precision. Leveraging NAS CD for efficient data migration? These VMs also execute high-performance copy operations to transfer your existing NAS data to FlashArray effortlessly, providing unmatched flexibility and efficiency in managing your data infrastructure.

For a more detailed description of NAS CD operations, please refer to How It Works: NAS Cloud Direct.

Backup Target Storage

NAS CD supports a range of NFS and object-based backup storage targets, including NFS and S3 protocol on Pure Storage FlashBlade and other compatible storage systems and public cloud options such as AWS, Azure, and Rubrik Cloud Vault.

Pure Storage FlashBlade

Pure Storage FlashBlade is an ideal object storage target for NAS CD backups. With high performance, FlashBlade ensures you can back up and restore your data within SLAs. High density and power efficiency minimize your data footprint and improve sustainability. Pure Storage simplicity makes deployment and management easy. FlashBlade will scale along with your unstructured data to tens of petabytes in a single cluster. Integration with NAS CD S3 object lock support, coupled with SafeMode Retention Lock, protects your backups from unwanted changes and deletions.

Rubrik Cloud Vault

Rubrik Cloud Vault is the ultimate cloud-based archival solution for organizations seeking secure, offsite data storage. Built on a robust zero-trust architecture, Rubrik Cloud Vault delivers unparalleled benefits, including unified data management through Rubrik Security Cloud, centralized reporting across your entire data security ecosystem, and predictable costs with a straightforward fixed pricing model.

Rubrik Cloud Vault goes beyond standard archival with seamless access to Rubrik's advanced security operators, Anomaly Detection and Sensitive Data Monitoring. These powerful tools empower customers to swiftly pinpoint the scope and sensitivity of critical unstructured data, enabling rapid response without the complexity or expense of additional security solutions. Rubrik Cloud Vault ensures peace of mind by combining comprehensive security, operational efficiency, and cost-effectiveness in one cohesive platform.

Validation

We validated NAS CD with FlashArray//C™ as a data source and FlashBlade//E™ as a backup target. The NAS CD stateless VMs were deployed on a VMware vSphere cluster consisting of five hosts, on an iSCSI datastore hosted on FlashArray//X70R3. The testing consisted of full and incremental NFS and SMB file share backups, data recovery, and protection of open files using FlashArray snapshots. Figure 2 illustrates the logical architecture of the lab environment.

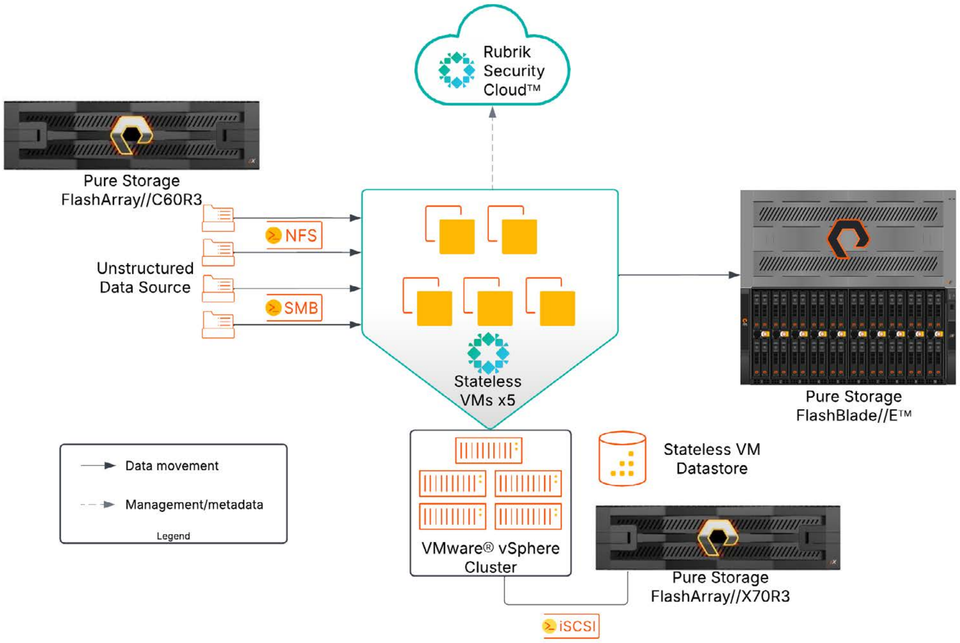


FIGURE 2 Solution validation architecture

System Details

Table 1 lists the system specifications for the storage arrays we used in testing.

Array Role	Array Model	Capacity (Usable)	Connectivity	Purity Release
Data source	FlashArray//C60R3	1.13PiB	25Gb Ethernet x2	Purity//FA 6.6.11
Backup target	FlashBlade//E	2.69PiB	100Gb Ethernet x4	Purity//FB 4.5.0
Virtual datastores	FlashArray//X70R3	26.87TiB	25Gb Ethernet x4	Purity//FA 6.6.11

TABLE 1 Storage specifications

Table 2 lists the system specifications for the test servers.

System	Processor	RAM	Network	Storage
VMware vSphere host x5 VMware ESXi 7.0.3	AMD EPYC 7713P 64-core Hyperthreading active	256GB	Broadcom BCM57414 NetXtreme-E 10Gb/25Gb RDMA Ethernet Controller x2	iSCSI connection to FlashArray//X70R3
NAS CD stateless VM x5	16 vCPU	64GB	vmxnet3	N/A

TABLE 2 Server specifications

Data Details

Testing was performed using several file sets with different profiles. The proprietary data generation tool we used creates a set of unique files that have a given average file size, total data size, and compressibility factor, distributed across a given directory depth. The tool adds some jitter to the file sizes to insert some variation in the data set. Table 3 lists the details of the various file sets.

File Set	Average File Sizes/Compressibility
Small files	98%: .5 MiB, 60% compressible 1%: 20 MiB, 10% compressible 1%: 1 GiB, incompressible
Mixed files	40%: .5 MiB, 50% compressible 40%: 20 MiB, 10% compressible 20%: 1 GiB, 10% compressible
Large files	20%: .5 MiB, 60% compressible 30%: 20 MiB, 10% compressible 50%: 1 GiB, incompressible

TABLE 3 Data set specifications

Backup and Recovery Tests

For each file set, we performed a sequence of full backup of the entire file share, incremental backup with 10% data change, and restore. We ran each test sequence using the Network File System (NFS) version 3 and Server Message Block (SMB) protocols.

NFS

With NFS (Figure 3), the average performance for full backups varied from approximately 6TiB/hour for the large file set to 8.5TiB/hour for the small file set. Small files are often the most challenging to protect and recover in a NAS environment, so it's impressive that the small file set outperformed the others. For incremental backups, the mixed file set achieved the best performance at around 7TiB/hour, while the other file sets delivered strong results at nearly 5TiB/hour.

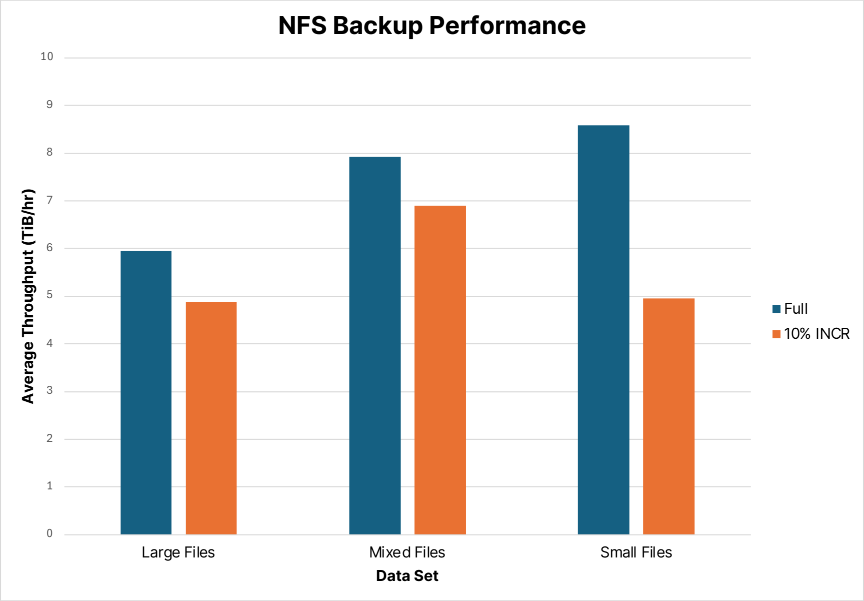


FIGURE 3 NFS backup performance

To test recovery performance, we performed out-of-place restores of each file set. We restored the entire file share to measure full-scale performance and 10% of the data to measure the performance of smaller-scale recovery. Figure 4 shows the results. The mixed file set performed the best during recovery, achieving 6TiB/hour. for the partial restore and almost 5 TiB/hour for the full restore. Small file recovery, one of the most challenging scenarios for any NAS, is performed at over 3.5 TiB/hour for the partial restore and over 2TiB/hour for the full restore. The large file set performed in the middle, above 3TiB/hour for both tests.

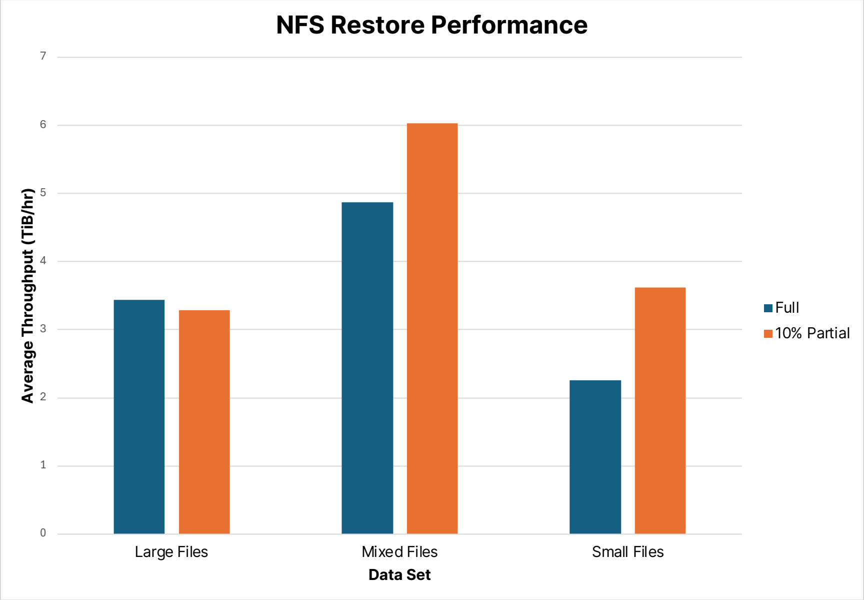


FIGURE 4 NFS restore performance

SMB

The SMB protocol has more overhead than NFS, so backup and restore are generally slower. This is most noticeable for file shares with many small files, where protocol operations comprise a high percentage of the total network traffic. Backup software can struggle to enumerate data changes. NAS platforms can have trouble keeping up with all the activity, leading to long scan times and long-running incremental backups. NAS CD takes a unique approach to file scans that eliminates much of the overhead. With NAS CD and the Pure Storage platform, we saw less than a 15% difference between the protocols–except for the small-file data set, where protocol overhead had a much more significant impact. Full backups ranged from almost 6TiB/hour to over 8TiB/hour, and incrementals averaged between 4.4TiB/hour for small files and around 6.5TiB/hour for mixed file sizes. Full backups for the mixed file set were faster over SMB than NFS. Figure 5 shows the breakdown.

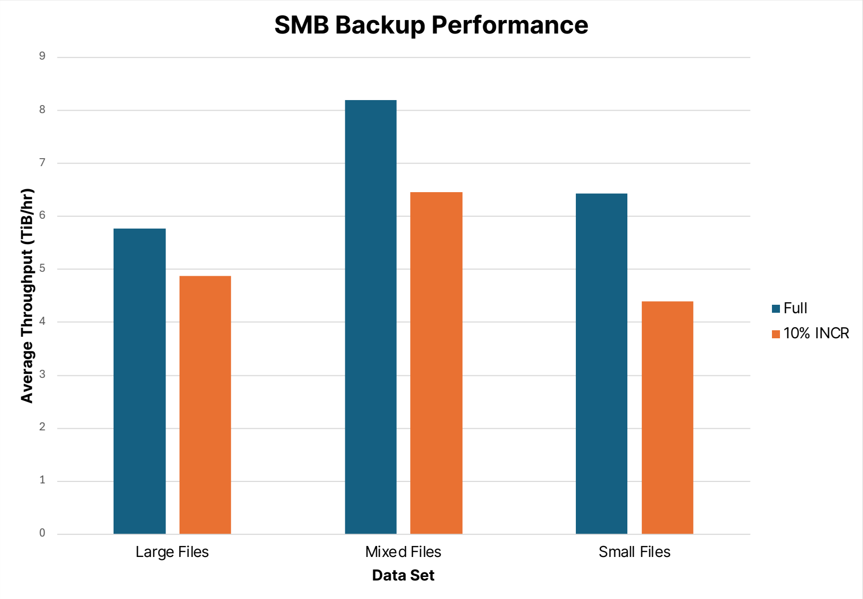


FIGURE 5 SMB backup performance

With SMB file recovery, file size had a pronounced effect on performance. With the highest protocol overhead, recovery of the small file set achieved just under 1.5TiB/hour. Large file recovery performed best, at over 5TiB/hour–outperforming the same file set over NFS. Figure 6 shows the breakdown.

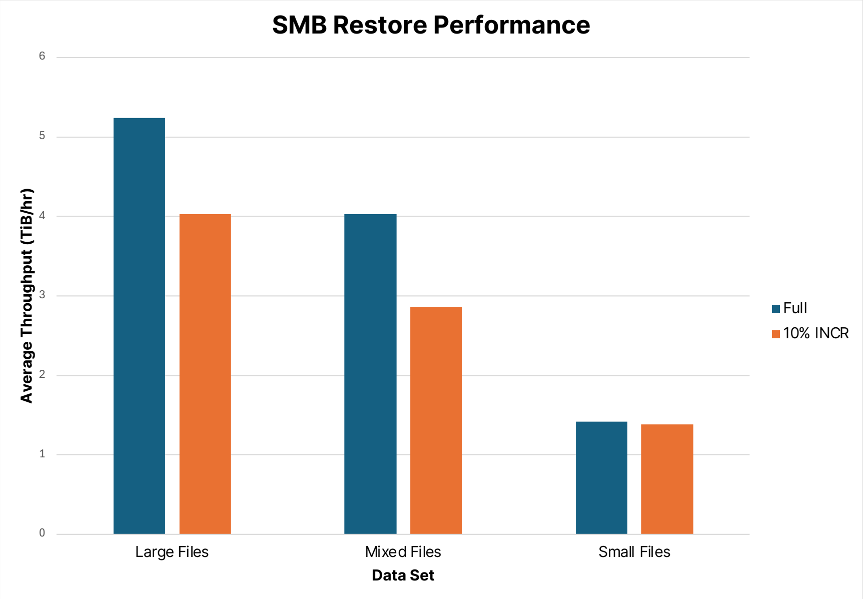


FIGURE 6 SMB restore performance

Parallelism and Scaling

All of the results so far have shown the performance of individual workloads executed separately. This isn't how typical NAS environments operate, though. Many organizations will have tens or even hundreds of file shares that need to be backed up within a window, with file recoveries needed from multiple shares simultaneously. With this in mind, we tested how NAS CD scales across multiple proxy VMs with multiple parallel backups. Figure 7 shows we saw more than 30% better average throughput when running a second backup job through a second data mover and an additional 20% gain with a third job. We had modest gains beyond the third proxy, but with five concurrent backups, NAS CD and the Pure Storage platform exceeded 14TiB/hour.

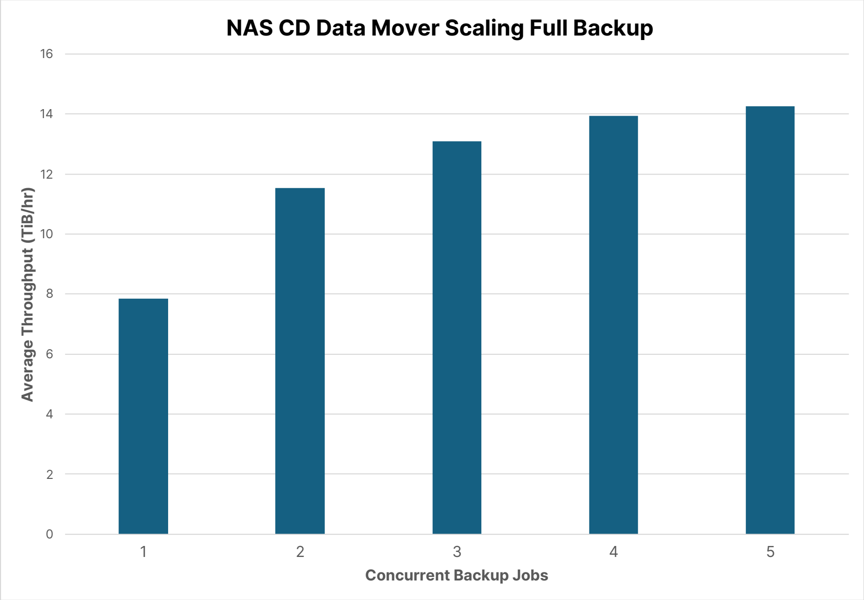


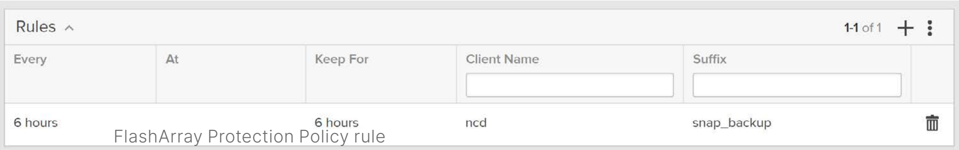
FIGURE 7 NAS CD data mover scaling

Open File Backups with FlashArray Snapshots

Two common problems with protecting unstructured data are files that are locked by users and point-in-time consistency. When user applications lock files, backup processes can't access them, leaving holes in backups that can lead to data loss. Similarly, since unstructured data backups can run for a long time, files and objects can change between when a backup starts and when it ends. Snapshots solve both problems by creating a read-only copy of all data and freezing all files in a consistent state. By leveraging FlashArray snapshots, you can ensure that all files within a file share are available and ready to be protected.

To use FlashArray snapshots with NAS CD, you need to take several actions on the array and in NAS CD:

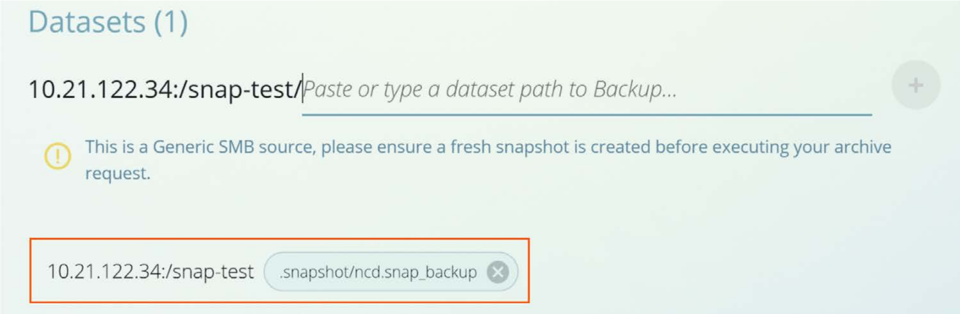
1. **Create a FlashArray protection policy**: Protection policies define how often to create snapshots, how long to keep them, and how to name them. To work properly with NAS CD, snapshot names need to remain consistent. This requires a protection policy rule with retention no longer than the snapshot frequency. Snapshot names follow the convention of <client name>.<suffix>, as defined in the policy rule. In the example shown in Figure 8, the FlashArray will create a snapshot called ncd.snap_backup, which will be refreshed every six hours. Typically, you would set the frequency and retention for one day, but the system is flexible enough to accommodate many



Every	At	Keep For	Client Name	Suffix
6 hours		6 hours	ncd	snap_backup

FlashArray Protection Policy rule

2. **Attach the policy to FlashArray Directories**: For each managed directory on the FlashArray that you want to use snapshots to protect, you must add the directory to the policy as a member. Policies can have multiple member directories; all will follow the same schedule, retention, and snapshot naming. Snapshots will be accessible through their respective file shares under a directory called .snapshot. For example, attaching the policy shown in Figure 8 to a directory on a FlashArray named server1, exported as snap-test, the snapshot would be accessible to NAS CD at server1:/snap-test/.snapshot/ncd.snap_backup.
3. **Configure NAS CD to protect the snapshot directory**: By default, NAS CD will back up an entire file share and exclude snapshot directories. Using the Backup Directory option for the file share, you can configure individual directories to be backed up using policies. Entering the snapshot path (Figure 9) as a dataset will instruct NAS CD to back up the snapshot instead of the entire share.


Datasets (1)

10.21.122.34:/snap-test/ *Paste or type a dataset path to Backup...*

This is a Generic SMB source, please ensure a fresh snapshot is created before executing your archive request.

10.21.122.34:/snap-test .snapshot/ncd.snap_backup

FIGURE 9 Configuring NAS CD to back up from a snapshot

Restoring files from a snapshot is similar to restoring from a regular file share backup.

Best Practices

Below are some best practices for NAS CD connections, backup, and deployment.

- **Use LACP for FlashArray File interfaces**: Aggregating physical network connections through Link Aggregation Control Protocol (LACP) gives users, application workloads, and NAS CD more bandwidth to access file shares, adding a redundancy layer. You can work with your Pure Storage account team to plan for implementing LACP.
- **Consider multiple stateless VMs**: While many NAS CD customers use a single data mover, adding a second or third one can significantly reduce your backup window, especially if you have many file shares. Consider deploying two or three stateless VMs to increase your overall throughput.
- **Distribute stateless VMs across separate ESXi hosts**: If you decide to deploy multiple NAS CD data movers, separating them across ESXi hosts reduces resource contention and gives each one access to more bandwidth, translating to faster backup and restore. If you use VMware Distributed Resource Scheduler™ (DRS), you should consider using affinity rules to enforce separation.
- **Use Snapshots to protect open files**: If you have file shares where users and applications regularly maintain locks on open files, or if you have file systems where backups must represent a consistent view of the entire data set, use FlashArray snapshots to achieve these outcomes. Refer to the Open File Backups with FlashArray Snapshots section for more details.

Conclusion

The partnership between Rubrik and Pure Storage redefines how organizations manage and secure unstructured data at scale. Combining high-performance, cyber-resilient solutions from Pure Storage with Rubrik's advanced data protection and zero-trust security, businesses can operate with unmatched speed, agility, and peace of mind. Pure Storage FlashArray and FlashBlade systems deliver robust, scalable storage tailored for modern workloads, offering robust data encryption, real-time auditing, disaster recovery, and built-in safeguards like SafeMode to protect against administrative errors and malicious attacks. Together, these technologies enable seamless data access and retrieval, even in the most demanding environments, ensuring business continuity and efficiency.

Rubrik complements this with its NAS Cloud Direct and Cloud Vault solutions, which provide comprehensive protection for unstructured data, whether on-premises or in the cloud. With policy-driven automation, granular recovery options, immutable backups, and anomaly detection for ransomware, Rubrik ensures that organizations are prepared to counter cyber threats while maintaining operational continuity and simplicity. The seamless integration of Pure Storage and Rubrik enables enterprises to eliminate performance bottlenecks, enhance data visibility, and achieve unparalleled cyber resilience. Together, these solutions empower businesses to navigate exponential data growth confidently, turning unstructured data from a challenge into a strategic advantage in today’s dynamic digital world.

When you’re ready to see for yourself how Rubrik and Pure Storage can transform your unstructured data, reach out to your account teams or your favorite reseller.

Additional Resources

- Learn about protecting unstructured data with Rubrik.
- Discover the Pure Storage FlashArray family.
- Read the Rubrik How It Works: NAS Cloud Direct white paper.
- Learn more about Pure Storage FlashBlade//E.
- Read about Pure Storage FlashBlade//S.
- Explore Pure Storage Evergreen® subscriptions.