International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Rubrik NAS Cloud Direct: A Technical Overview of Architecture and Workflow

Venkata Raman Immidisetti

Infrastructure Architect, Raleigh, North Carolina

vimmidisetti@gmail.com

Abstract:

Rubrik NAS Cloud Direct (NAS CD) is a modern data protection solution designed to efficiently back up and recover petabyte-scale unstructured data across on-premises and cloud environments. This paper provides a technical overview of the NAS CD architecture and its backup workflow. We describe how NAS CD combines a cloud-based control plane with stateless connector virtual machines to deliver high-throughput, scalable backups for network-attached storage (NAS) and object data. Key mechanisms such as advanced parallel scanning, an incremental-forever backup model, data chunking, and inline compression are detailed. We also discuss how the system ensures data integrity, security (through immutability and zero-trust access controls), and rapid recovery capabilities. Through an in-depth examination of the architecture and backup process, the paper illustrates how Rubrik NAS Cloud Direct addresses the challenges of protecting massive unstructured datasets with high performance and reliability.

Keywords:

Rubrik NAS Cloud Direct, Unstructured Data Backup, Cloud Data Protection, Incremental Forever Backup, Data Chunking, Parallel File Indexing, Zero-Trust Security, NAS Backup Architecture

I. INTRODUCTION

Unstructured data (such as files, documents, images, and videos) is growing at an unprecedented rate in modern organizations, reaching into the petabytes and billions of files. Traditional NAS backup approaches, like NDMP-based solutions, struggle to keep pace with this scale due to performance limitations and rigid architectures. Legacy backup systems often perform full backups infrequently and rely on single-threaded or serialized data capture, leading to prolonged backup windows and heavy impact on production storage performance. Additionally, ensuring the security and integrity of backups has become paramount as cyber threats like ransomware target backup data; backups must be stored in an immutable, air-gapped manner to prevent tampering or deletion.

Rubrik NAS Cloud Direct is a software-as-a-service (SaaS) solution engineered to tackle these challenges for enterprise NAS and object storage data. It introduces a high-performance, scalable architecture that can scan and index billions of files rapidly and transfer data to cloud or on-premises storage with minimal disruption. By leveraging parallelism and intelligent data management, NAS Cloud Direct achieves backup speeds often cited as up to an order of magnitude faster than legacy NDMP backups. NAS CD provides an incremental forever backup model—after an initial full backup, it efficiently captures only changes (new, modified, or deleted files) on subsequent runs. Combined with features like inline compression, small-file bundling, and zero-trust security principles, the system ensures that large datasets can be protected within tight windows while maintaining data safety. This paper explores the technical

architecture of Rubrik NAS Cloud Direct and explains in detail how its backup process works, from initial data capture to storing backup data and metadata, highlighting the components and techniques that enable its performance and reliability.

II. ARCHITECTURE

Rubrik NAS Cloud Direct employs a hybrid-cloud architecture composed of three primary components that work together to protect unstructured data:

Rubrik Security Cloud (RSC): This is the overarching SaaS platform which provides a secure front-end and global services for NAS Cloud Direct. RSC acts as the single pane of glass for authentication, policy management, and integration with other Rubrik services (such as anomaly detection and sensitive data monitoring). It enforces zero-trust security with strong authentication (SSO, MFA) and role-based access control. The NAS CD user interface is accessed through RSC, and it ensures only authorized users can manage or view backup data. RSC also facilitates SSO via SAML to integrate with enterprise identity providers, streamlining access to the NAS Cloud Direct UI.

Stateless Connector VM: The data-plane workhorse of NAS CD is a lightweight, stateless virtual machine deployed within the customer's environment. This VM is downloaded from the Rubrik service and can run on a variety of platforms (e.g. VMware vSphere, Nutanix AHV, Hyper-V, KVM, or as a cloud instance in AWS/Azure). The stateless VM is positioned "close" to the data sources to maximize throughput — for example, deployed in the same data center or VPC as the NAS storage. Its role is to perform file system scans, read data from the source, and stream backup data to the chosen storage target. The VM maintains no long-term state; it does not store backup data or indexes locally. Instead, it communicates outbound (over HTTPS) to the cloud control plane for instructions and to upload metadata. Notably, no inbound connections to the VM are required, which simplifies firewall configurations and enhances security. An environment can deploy multiple such VMs (connectors) to protect multiple NAS sources or to scale out performance for large datasets in different sites. Each connector VM is stateless and can be freely replaced or horizontally scaled as needed without losing any backup information (since all state is in the cloud control plane).

NAS Cloud Direct Control Plane: The control plane is hosted in a Rubrik-managed cloud environment (an isolated tenant in Rubrik's multi-tenant cloud, hosted on AWS for the SaaS offering). This control plane is the brains of NAS CD, coordinating all backup operations and storing the critical metadata. Internally, the control plane consists of several sub-components:

- **Cloud Slab:** The persistent storage layer within the control plane responsible for storing metadata, indices, and the "blob" data store that tracks backup content. This can be seen as the durable repository for job information and backup catalogs. It records details of each backup job (timestamps, success/failure status, retry info) and ensures consistency and durability of the backup metadata. The Cloud Slab underpins features like global file search and analytics by keeping a centralized record of all files protected.
- **Index Database (Catalog):** NAS CD maintains a highly scalable index catalog of all files and objects that have been backed up. This index database stores file metadata (attributes, sizes, timestamps such as last-modified time, permissions, etc.) and is updated incrementally with each backup. By cataloging which files have been added, changed, or removed between backups, this database enables rapid search across billions of entries and forms the basis of the incremental backup capability (the system can quickly determine differences since the last snapshot from this index). The index is also crucial during restores to locate which backup snapshot and object contains the data needed.

- **Cloud Apps:** This is the orchestration and management service layer in the control plane. Cloud Apps include the policy engine, job scheduler, and workflow orchestrators for NAS CD. They provide the user interface and API endpoints through RSC for administrators to define protection policies (such as backup frequency and retention), monitor jobs, and initiate restores. The Data Lifecycle Manager within Cloud Apps ensures that retention policies are applied (e.g., expiring old snapshots according to retention rules), and it manages replication or archival rules if configured. The job scheduler and other services coordinate activity between the connector VMs and the cloud components, ensuring tasks execute at the right times and resources are allocated optimally. Essentially, Cloud Apps act as the brains that translate user-defined policies into actions performed by the connector VMs and track the state of each dataset's backups.

All communication between these components is secure. The connector VM only initiates outbound connections to the control plane (for sending metadata or receiving commands) and to the backup storage target, using encrypted channels (HTTPS or secure storage protocols). No data is permanently stored on the VM; it streams data through to the target, which minimizes any risk of data exposure on intermediate systems. The architecture supports a variety of data sources: generic NFS and SMB shares from any NAS vendor, plus direct API integration with specific NAS platforms (such as Dell EMC Isilon/PowerScale, NetApp ONTAP, Pure Storage FlashBlade, Qumulo, as well as Amazon EFS and FSx for NetApp in cloud). These API integrations allow NAS CD to automatically discover shares/exports and set up least-privilege access, simplifying the onboarding of sources. On the backup target side, NAS CD can write to almost any S3-compatible object storage (public cloud or on-premises object stores) or to NFS storage. For cloud targets, customers can use their own cloud buckets (e.g. writing directly into an AWS S3 bucket in their account, including support for Glacier tiers for archive) or opt for Rubrik-managed storage (such as Rubrik Cloud Vault). The architecture is flexible: organizations can choose to send backups to cloud, keep copies on-prem for fast local restores, or both (to avoid vendor lock-in and support hybrid workflows). Each dataset protected by NAS CD is associated with a designated backup target and retains a one-to-one relationship (ensuring an incremental chain is maintained per source-target pair).

Overall, this architecture decouples the performance-critical data movement (handled by the stateless VM close to the data) from the control logic and metadata management (handled centrally in the cloud). By doing so, NAS Cloud Direct achieves both speed and scale: the heavy-lifting of reading and transferring data is distributed and parallelized via one or more connector VMs, while the cloud control plane centralizes indexing, search, and policy control. This design also inherently provides resiliency (the control plane data is safely stored in the cloud, and the stateless VMs can be redeployed if needed without losing knowledge of prior backups) and security (by isolating backup data and metadata in a separate environment with strict access controls, and by allowing backups to be stored as immutable objects in object storage).

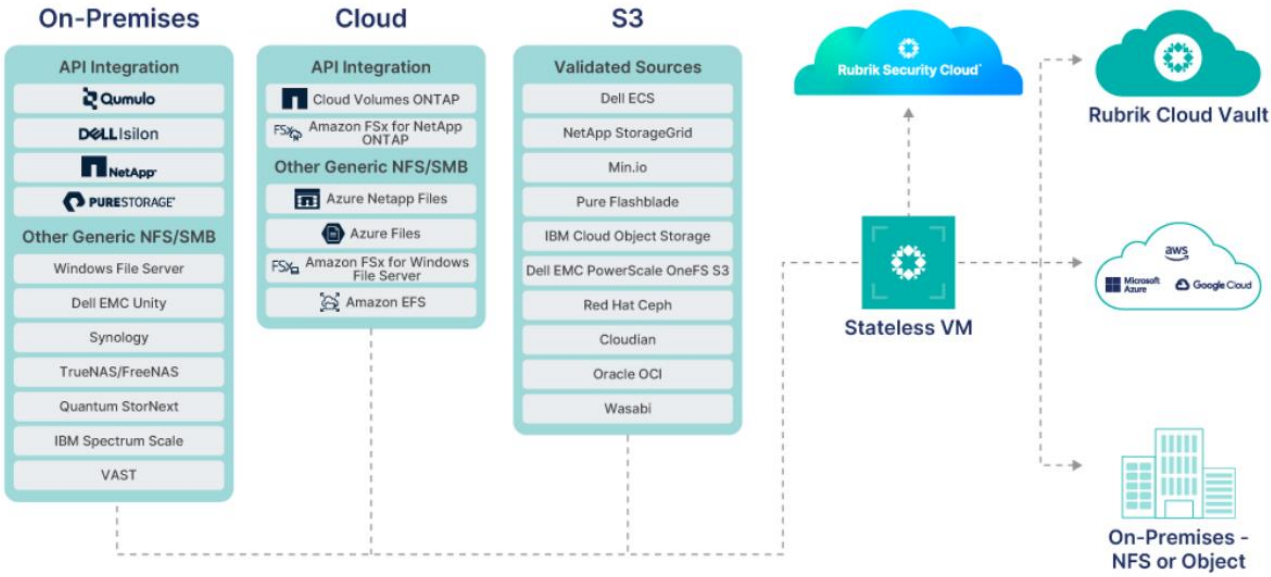

Figure 1: Rubrik Security Cloud

III. HOW IT WORKS

After a NAS or object source is added to Rubrik NAS CD and an initial full backup is performed, all subsequent backups run in an incremental forever fashion. An administrator defines a protection policy in the RSC interface, which includes: backup frequency (schedule), retention period for backups, whether the job is a regular backup or an archival job, and the backup destination (target storage). Once configured, the NAS Cloud Direct system orchestrates the backups automatically according to schedule. The backup process can be understood in a series of phases and steps:

Phase 1: Scan and Index Preparation. The first phase identifies what data needs to be backed up by comparing the current state of the file system to what was previously backed up.

1. Connect and Read Source: At the scheduled backup time, the NAS CD stateless VM connects to the source file system. It uses a custom-built NFS/SMB client (for file shares) or S3 client (for object stores) to efficiently enumerate files and metadata directly from the source NAS, without mounting the share like a regular NAS client. This direct, low-overhead scanning method allows the system to read directory trees and file metadata at very high speed. The connector VM uses multi-threaded parallel scanning – it spawns numerous threads to walk through the filesystem in parallel, which is especially important for large NAS volumes containing millions of files across many directories. Work is dynamically balanced across threads so that if one directory branch is slower or has fewer files, other threads can pick up more work, ensuring all scanner threads remain busy. Thanks to this parallelism and intelligent scheduling, NAS CD can discover and read hundreds of thousands of files per second during this phase. As the VM scans, it gathers the list of all files (and their metadata attributes such as size and last-modified timestamp).

2. Fetch Last Backup Snapshot Metadata: In parallel with scanning the source, the connector VM queries the Index Database in the cloud control plane to retrieve the metadata from the last successful backup snapshot of this source. This is essentially a catalog of what files (and their versions) were previously backed up and their state (sizes,

timestamps, etc.) at that time. By retrieving this “last known index” and holding it in memory, the VM now has two datasets to compare: the current state from the live source vs. the last backup state from the index.

3. Calculate Differences: The VM compares the current file list against the last snapshot’s list to determine which files are new, which have changed (for example, a file whose last-modified time or size is different, indicating it was modified since last backup), and which files or directories have been deleted. This difference computation yields a set of change candidates that need to be acted upon. Only the new or modified files will need to be backed up (and the index will later be updated to reflect deletions as well). By performing this intelligent diff, NAS CD avoids re-copying unchanged files, dramatically reducing the amount of data that must be transferred on incremental backups. The output of this step is essentially a "to-do list" of files (and chunks) that require backup.

Phase 2: Data Transfer (Move Phase). Once the changed files are identified, the system proceeds to efficiently transfer the necessary data to the backup storage target.

4. **Read Changed Data:** The connector VM now begins reading the actual file data for each item identified in the diff list. Using the high-speed NFS/SMB or S3 client interfaces, it pulls the contents of new or modified files from the source. Importantly, NAS Cloud Direct pipelines these operations – even as the scanning/diff (Phase 1) might still be processing parts of the filesystem, the system can start reading and transferring data for portions it has already identified. This parallel pipeline (scan, read, and send concurrently) ensures maximum utilization of time and bandwidth. The VM also handles deleted files by marking them in the metadata (so the index knows those files were removed, ensuring the ability to purge or reflect deletions as per retention settings).

5. **Group and Chunk Files:** Before sending the data to storage, NAS CD performs an optimization step based on file sizes. Extremely small files, which are numerous in typical NAS environments, can incur a lot of overhead if each were stored separately in an object store (each would be a tiny object causing high API call counts and inefficiency). To mitigate this, NAS CD groups multiple small files together into larger composite blocks or "blobs". For example, files smaller than a certain threshold (e.g. 20 MB) may be bundled together into a single object or stream to optimize throughput. On the other end of the spectrum, very large files are split into manageable chunks (for instance, files larger than 100 MB might be broken into 100 MB segments). Medium-sized files in between might be sent as-is. This grouping and chunking strategy improves network utilization and storage efficiency: it reduces the total number of upload transactions and avoids the performance penalties of handling too many tiny pieces or extremely large singular streams. By chunking large files, NAS CD can also parallelize the transfer of different parts of a large file simultaneously across multiple threads, further speeding up the process.

6. **Inline Compression:** As data is being prepared for transfer, the system compresses it in-memory using an efficient algorithm (Rubrik uses the LZ4 compression for NAS CD). LZ4 is chosen for its high throughput and low CPU overhead, capable of significantly reducing data size on the fly with minimal latency. Compressing the data before sending reduces the amount of network bandwidth consumed and ultimately lowers the required capacity on the target storage. For example, large text-based files or redundancies can be shrunk, making the transfer faster and storage consumption smaller. This compression is transparent and inline, meaning the VM does it as part of the streaming process without needing extra staging.

7. **Transfer to Backup Target:** The connector VM then streams the compressed, chunked data to the configured backup repository. If the target is an object store (cloud or on-premises), the VM uses the storage provider’s native APIs (for instance, AWS S3 SDK for an S3 bucket, or a S3-compatible API for other object stores) to upload the data objects. If the target is a NAS (NFS) share used as an archive destination, the VM writes the data via NFS protocol.

NAS Cloud Direct is designed to be agnostic in this stage – it can write to any S3-compatible storage or NFSv3 share. Thanks to the earlier grouping of small files into larger blobs, the number of put/get operations on the object store is reduced, which can lower costs and improve speed. The transfer step also employs multi-threading and asynchronous operations where possible: multiple files or chunks can be in transit concurrently, maximizing utilization of available network bandwidth. The system does not need to constantly check what's already in the bucket because it relies on its internal metadata to avoid duplicates, so it can stream new objects directly, further avoiding latency. If network disruptions occur, the process can retry transfers without starting over, thanks to the job tracking in the control plane. Throughout the transfer, data integrity checksums are used to ensure what is written to the target matches the source data.

8. **Commit Metadata (Index Update):** After all the new/modified data is successfully written to the backup target, the final phase is to update the central index and record the backup completion. The connector VM sends the updated metadata (the list of files, their versions, and the pointers to the objects or storage segments where those file data now reside) up to the Index Database in the cloud control plane. Essentially, it commits a new snapshot record. This updates the catalog to include the latest state of the protected dataset, which will be used as the "last snapshot" reference for the next backup iteration. The control plane also records the job's status (success or any failures) and other stats in the Cloud Slab (for monitoring and auditing). Once this metadata commit is done, the backup cycle for that snapshot is complete.

It is important to note that steps 4–8 are executed in parallel streams while the backup job runs. The architecture pipelines the scanning, reading, and writing so that there isn't a strict sequential wait for one step to finish entirely before the next begins. For example, as soon as some files are identified in step 3, the VM might start reading and sending them (steps 4–7) while continuing to scan further directories. Likewise, metadata commit (step 8) can happen for subsets of data that have completed transfer even as other files are still in transit. This overlapping of phases is a major factor in achieving high performance. By the time the backup job finishes, the system has effectively performed a coordinated set of tasks that result in a new backup version with minimal wasted time. The very first backup of a given source will operate slightly differently: since no previous index exists, the system will essentially mark all files as "new" and back up everything (a full backup). That establishes the baseline snapshot in the index. All future backups will then be incremental, sending only changes. NAS Cloud Direct maintains the one-to-one relationship between a source and its backup target path, which means if the target or source mapping is changed (say, you choose a new bucket or share to send data), that would constitute a new full backup path. As long as the pair is consistent, incremental forever is maintained. The result of this approach is significantly reduced backup windows after the initial run, even as data grows, because each cycle only transfers a fraction of the total data (the changes). The system is designed to handle very large numbers of files and high change rates by virtue of its scalable index and parallel data mover architecture. Throughout the backup process, data security is enforced. All data in transit from the VM to cloud is encrypted via TLS. Optionally, users can enable object-lock (Write-Once-Read-Many immutability) on the target storage (if using supported object stores like AWS S3 or Rubrik's Cloud Vault), so that once the backup data lands, it cannot be altered or deleted until a set retention period expires. Credentials used by the NAS CD VM to access sources or targets are stored securely (with the VM often creating a temporary least-privileged account on the NAS for reading, if using API integration for certain vendors). Backups are logically air-gapped since the control plane and storage can be in a different security domain than the source environment, and the backup data is not directly accessible via standard protocols by the source systems.

Restore Process

Although the focus of NAS Cloud Direct is on backup, an equally important capability is fast restore of data when needed. Restores in NAS CD leverage the global index to locate data quickly and only bring back what is necessary. In a restore operation, a user selects a specific snapshot (point in time) and either the original location or an alternate destination to recover data. The high-level restore workflow is as follows:

1. **Restore Initiation:** An administrator initiates a restore via the NAS CD interface or API, specifying the dataset, snapshot timestamp, and restore destination (which could be the original NAS path or a different NAS/filesystem). The system identifies the connector VM that will perform the restore (usually the same one that did backups for that dataset, or a suitable one with access to the destination).
2. **Listing of Required Files/Objects:** The control plane, in conjunction with the connector VM, consults the Index Database for that snapshot to generate a detailed list of all files and directories that existed in the source at that point in time. Essentially, this step recreates the file tree metadata for the restore set. The system then determines which backup objects (in the target storage) contain the needed data for those files. Because NAS CD had possibly chunked or grouped files, the restore logic figures out which composite objects or segments need to be retrieved to get all the requested files.
3. **Read Plan Creation:** A restore plan is formulated, which might involve retrieving a set of objects from the backup store. For example, if a directory with 1,000 small files is being restored, and those were stored in a few grouped blobs in the backup, the plan will note to fetch those specific blob files. If some large files were present, the plan will fetch the specific chunks for those files. This plan ensures that only the necessary chunks/blobs are read from the backup repository, not the entire backup set. If the backup was stored in an archive tier (like AWS Glacier), NAS CD will initiate a rehydration of those objects to a hotter tier (for instance, moving them temporarily to S3 Standard or a staging location) before proceeding, since archive storage has retrieval latency. This happens automatically when needed.
4. **Data Restoration:** The connector VM then pulls the required data from the backup storage and writes it to the restore destination. The restore process also uses parallel threads – multiple files can be restored simultaneously, and large files can be restored chunk by chunk in parallel. For efficiency, NAS CD applies similar logic as backups: it may restore small files in batches (combining multiple small file retrievals into one stream if they were stored together) to expedite the process. The original directory structure is recreated at the destination, and file metadata (timestamps, permissions, ownership) is preserved. NAS CD ensures that access control permissions (ACLs) on the files and directories are restored as recorded, maintaining security context. If restoring to an alternate location, the admin may choose to apply different permissions or just recover the data for analysis. During restore, the system provides progress updates and ensures data integrity (with checksums) as files are written out.

The restore process is optimized to minimize downtime: only needed data is fetched, and the ability to restore to alternate targets can be used to verify data or recover in isolation if needed (for example, in a ransomware scenario, one might restore affected data to a clean environment for validation before replacing the production data). Thanks to the indexed metadata, even locating a subset of files to restore (such as a specific directory or a range of files matching a criteria) is fast—administrators can search the catalog for specific files and initiate restores for just those, rather than retrieving entire volumes. The combination of the incremental forever backups and the efficient restore means NAS Cloud Direct not only reduces backup windows but also significantly improves Recovery Time Objectives (RTOs) for large NAS data sets.

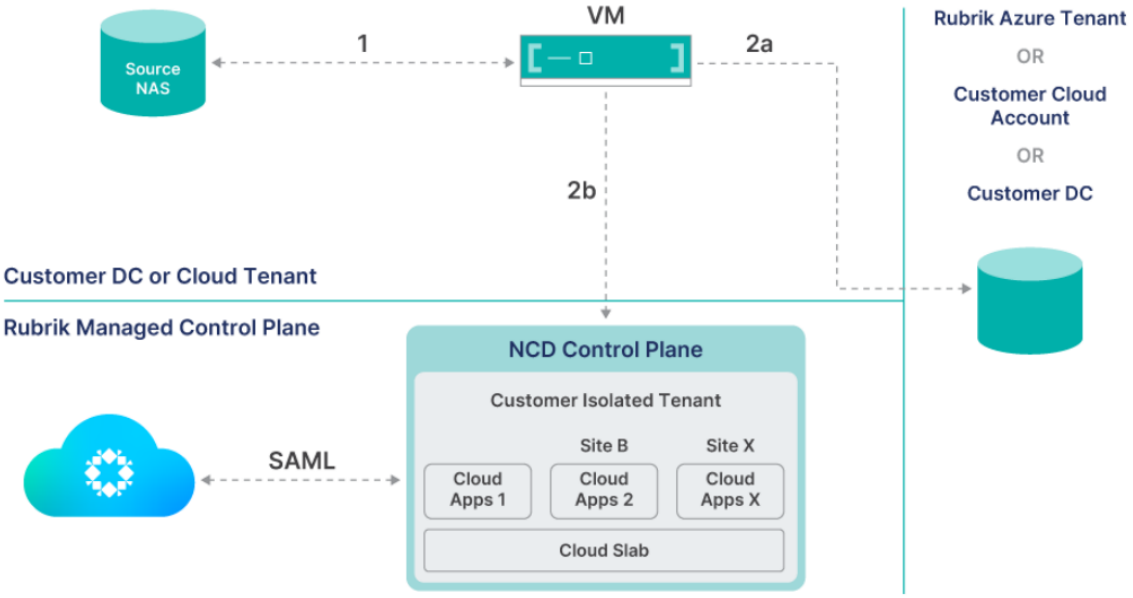

FIGURE 2: NAS CD initiates backup processes

IV. CONCLUSION

Rubrik NAS Cloud Direct introduces a highly technical and efficient approach to backing up NAS and unstructured data at massive scale. By separating the control plane (for coordination and indexing) from the data movers and leveraging a cloud-native architecture, it achieves a level of performance and scalability beyond traditional NAS backup methods. The architecture's stateless connector VMs near the data sources enable rapid scanning and reading of files with minimal impact on production systems, while the cloud control plane handles global index management and policy orchestration with secure, isolated multi-tenant principles. This design means that even as organizations face explosive growth in file data, the backup infrastructure can scale elastically and remain manageable centrally. Technically, NAS Cloud Direct's backup workflow is streamlined for speed: parallelized file discovery, intelligent differencing to send only changes, grouping of small files and chunking of large ones to optimize transfer, and on-the-fly compression all work in concert to maximize throughput while minimizing resource usage. These capabilities allow organizations to meet tight backup windows and reduce storage costs (by efficiently utilizing object storage and archive tiers for older data). At the same time, the solution maintains data integrity and security by incorporating features like immutability (protecting backups from deletion or modification) and strict access controls through the Rubrik Security Cloud interface. In summary, Rubrik NAS Cloud Direct provides a robust technical solution to the challenges of protecting vast unstructured datasets. Its architecture ensures high availability and centralized control, and its backup and restore processes ensure data is protected continuously with near-zero incremental impact and can be recovered swiftly when needed. For environments struggling with the limitations of legacy NAS backup tools, NAS Cloud Direct offers a modern, scalable alternative that is purpose-built for the era of petabyte-scale, distributed unstructured data. The system exemplifies how smart software design and cloud integration can overcome traditional performance bottlenecks, delivering reliable data protection even as data volumes and threats continue to grow.

REFERENCES

[1] https://www.rubrik.com/content/dam/rubrik/en/resources/white-paper/hiw-rubrik-nas-cloud-direct.pdf

[2] https://www.rubrik.com/solutions/nas

[3] https://docs.rubrik.com/en-us/saas/saas/ncd_data_sets.html

[4] https://cloudian.com/solutions/data-protection/rubrik-nas-cloud-direct/

[5] Kumar, Deepak. "NAS Storage and Data Recovery." In *Proceedings of the International Conference on Embedded Systems, Cyber-physical Systems, and Applications (ESCS)*, p. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[6] D. S. Adji, G. Eduardus, Michael, Minawati and W. Budiharto, "Performance Analysis Between Cloud Storage and NAS to Improve Company's Performance: A Literature Review," 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), Jakarta, Indonesia, 2021, pp. 263-268, doi: 10.1109/ICCSAI53272.2021.9609792.

[7] Kumar, MG Ravi, Ayudh Nagaraj, Benjamin Paul, and Sharat P. Dixit. "Network-Attached Storage: Data Storage Applications." *Turkish Journal of Computer and Mathematics Education* 12, no. 12 (2021): 2385-2396.

[8] A. El kamel, "A Fast Failure Recovery Mechanism using On-Premise/Cloud-based NAS in SDN," 2023 IEEE Symposium on Computers and Communications (ISCC), Gammarth, Tunisia, 2023, pp. 1090-1093, doi: 10.1109/ISCC58397.2023.10217995.

[9] N. Davidović, M. Marković and M. Popović, "Backup of Cloud Data to On-Premises Locations," 2025 24th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 2025, pp. 1-6, doi: 10.1109/INFOTEH64129.2025.10959170.

[10] S. -H. Zou, N. -S. Fang and W. -J. Gao, "Research on online cloud storage technology," 2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), Xuzhou, China, 2020, pp. 62-65, doi: 10.1109/DCABES50732.2020.00025.

[11] B. I. Ismail, M. N. Mohd Mydin and M. F. Khalid, "Architecture of scalable backup service for private cloud," 2013 IEEE Conference on Open Systems (ICOS), Kuching, Malaysia, 2013, pp. 174-179, doi: 10.1109/ICOS.2013.6735069.

[12] Y. -H. Kuo, Y. -L. Jeng and J. -N. Chen, "A Hybrid Cloud Storage Architecture for Service Operational High Availability," 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops, Japan, 2013, pp. 487-492, doi: 10.1109/COMPSACW.2013.94.

[13] V. Oujezsky, P. Novak, T. Horvath, M. Holik and M. Jurcik, "Data Backup System with Integrated Active Protection Against Ransomware," 2023 46th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 2023, pp. 65-69, doi: 10.1109/TSP59544.2023.10197687.