COHESITY

White Paper

Cohesity SpanFS and SnapTree

Cohesity SpanFS™ and SnapTree™

Web-Scale File System, Designed for Secondary Storage in the Cloud Era

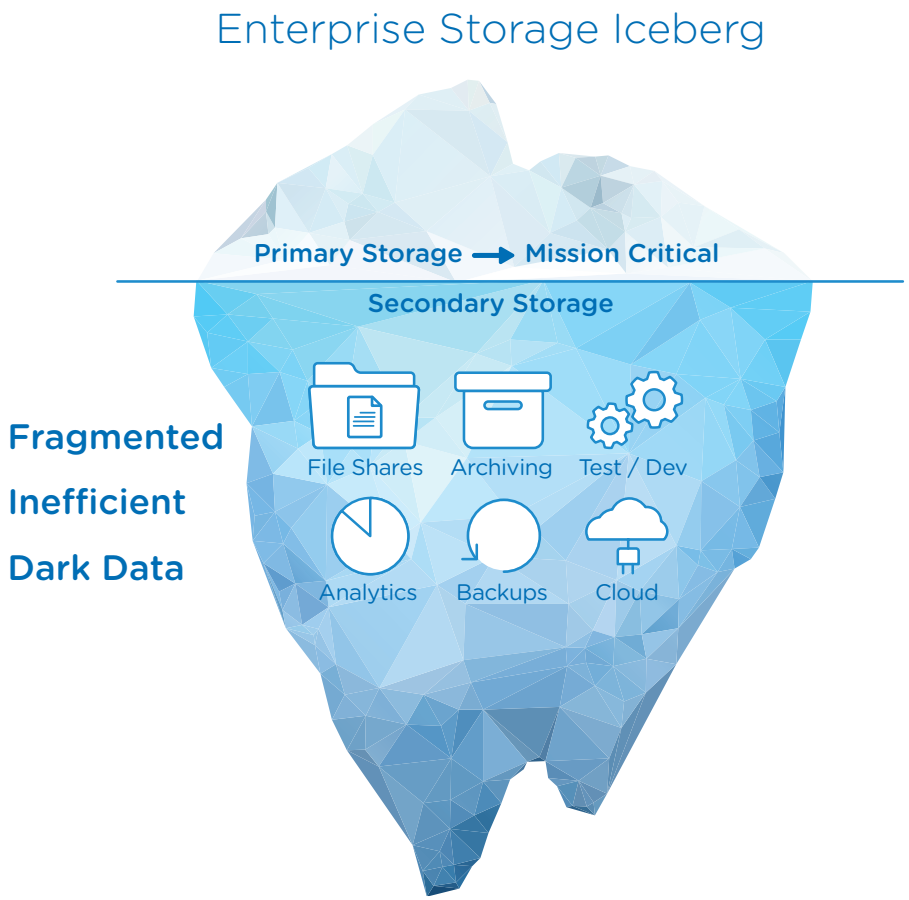
Why Do Enterprises Need a New File System?

The Enterprise Data Tsunami

Enterprises are struggling to manage ever-increasing volumes of unstructured data with antiquated secondary storage silos. According to IDC, enterprises managed a total of 6 Zettabytes of data in 2016, and that number is expected to grow to 93 Zettabytes by 2025. About 80% of that data is secondary data in the form of files, objects, and backups.

The Legacy Secondary Storage Iceberg

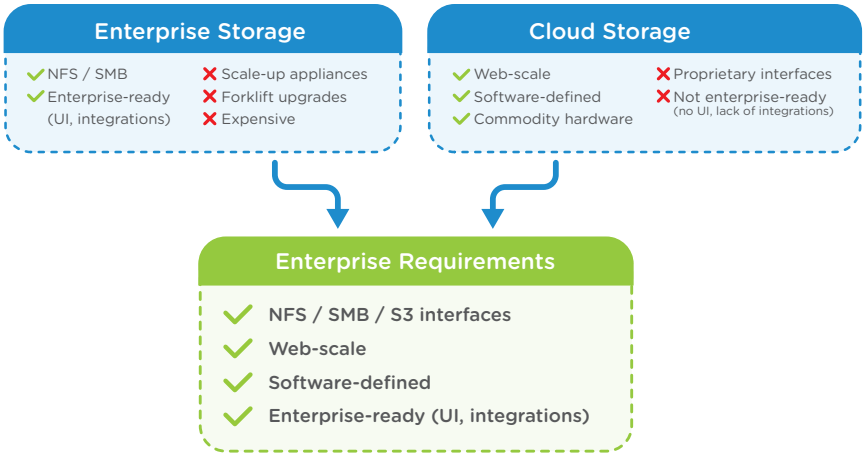
Enterprise storage can be compared to an iceberg. Above the surface of the water is highly-visible primary storage supporting mission-critical applications. But primary storage typically only comprises 20% of overall capacity. The remaining 80% below the surface consists of secondary storage that covers all your secondary use cases, including backup, file and object storage, test/dev, and analytics. Secondary storage infrastructure is fragmented across a patchwork of point appliances, including deduplication appliances, backup servers, cloud gateways, NAS, and data lakes. This approach is complex – each silo needs to be provisioned, configured, managed and updated through its own proprietary UI and processes. Data must be copied and stored across the silos, with enterprises keeping an average of 10 to 15 redundant copies of data across silos. Enterprises need a simpler, more efficient way to manage their ever-increasing volumes of data.



Tale of Two Storage Architectures – Enterprise and Cloud

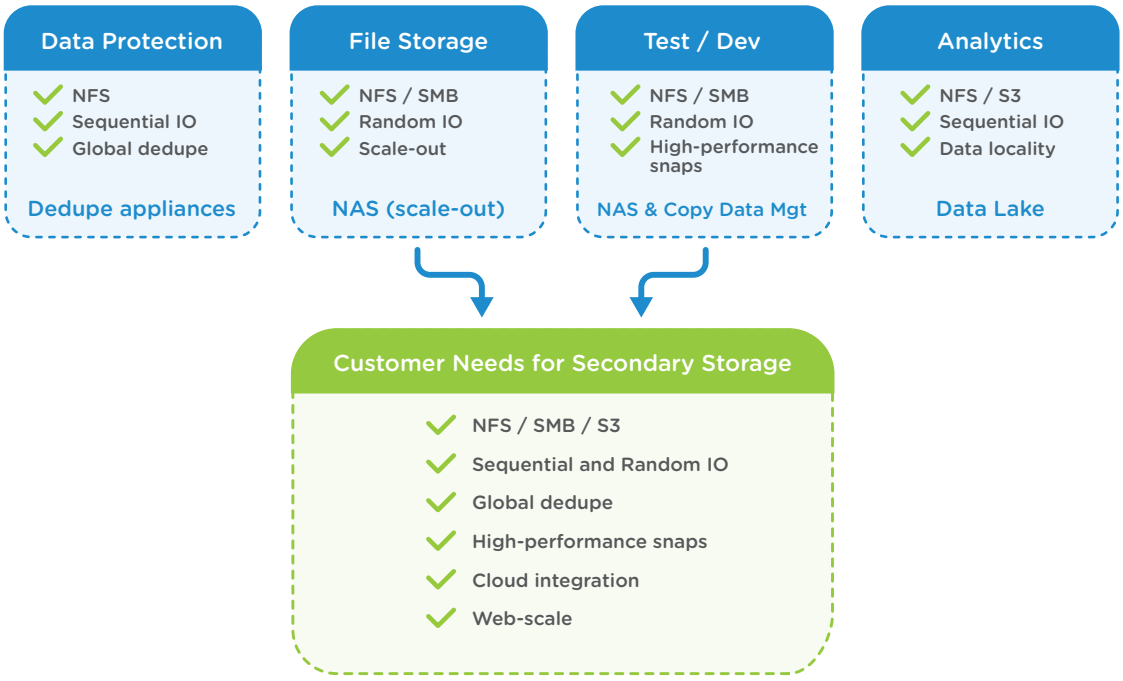
Over the past 10 years, storage has evolved around two very different directions. On the one hand, enterprise storage focused on providing standardized file interfaces (NFS and SMB), on 'scale-up' hardware, and snapshots for resiliency. On the other hand, cloud storage, developed by hyperscale companies like Google and Amazon, focused on delivering scale-out solutions on commodity hardware, strong resiliency to hardware failures, but relying on proprietary protocols and APIs for data access.

What enterprises need today is the best of both worlds. They need to support standardized interfaces like NFS and SMB protocols, to interoperate with existing applications. But they have also reached a level of scale where the 'scale-up' appliances are just not cutting it anymore. Instead, enterprises need to move to software-defined, web-scale solutions on commodity hardware, just like cloud storage. Web-scale provides multiple advantages like 'pay-as-you-grow' consumption, always-on availability, non-disruptive upgrades (instead of forklift upgrades), simpler management, and lower costs.



Busting the Silos Requires a More Flexible File System

Enterprise storage is deployed in isolated silos because each individual use case has spawned a purpose-built storage file system with specific features. For example, purpose-built backup appliances (PBBA) provide in-line variable-length deduplication to maximize space efficiency, but at the expense of random IO performance. Test/dev filers, such as NetApp, provide much better random IO performance and great snapshots, but can't afford the performance overhead of inline deduplication.



To effectively consolidate secondary storage silos, enterprises need a file system which is simultaneously able to handle the requirements of multiple use cases. It must provide standard NFS, SMB and S3 interfaces, strong IO performance for both sequential and random IO, in-line variable length deduplication, and scalable snapshots. And it must provide native integration with the public cloud to support a multicloud data fabric, enabling enterprises to send data to the cloud for archival or more advanced use cases like disaster recovery, test/dev and analytics. All of this must be done on a web-scale architecture to manage the ever-increasing volumes of data effectively.

SpanFS™: A Unique File System Designed to Consolidate Secondary Storage at Web-Scale

Span Everything

To enable enterprises to take back control of their secondary data at scale, Cohesity has built a completely new file system: SpanFS™. SpanFS is designed to effectively consolidate and manage all secondary data, including backups, files, objects, test/dev, and analytics data, on a web-scale platform that spans from the edge to the cloud.

SpanFS is designed to span everything, including physical and logical constructs:

Physical:

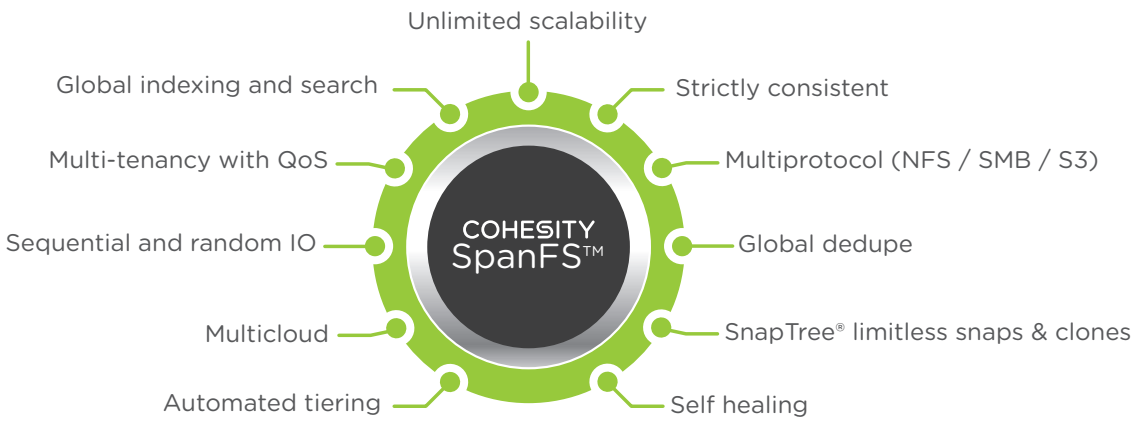
- Nodes: SpanFS provides unlimited scale across many hyperconverged nodes. SpanFS is completely distributed and doesn't have a master node. It scales linearly and dynamically rebalances data as nodes are added or removed. It provides always-on availability, non-disruptive upgrades, and a pay-as-you-grow consumption model.
- Storage Tiers: SpanFS spans across tiers of persistent storage technologies including SSD, HDD, and even remote cloud storage. SpanFS effectively utilizes the most appropriate tier based on IO profiles.
- Geographic Locations and Cloud: SpanFS interconnects remote offices, branch offices, core data centers, and public clouds into one data fabric. SpanFS can be used to build a multicloud data fabric and span data across multiple private and public clouds.

Logical:

- Workloads: SpanFS supports data protection, files, objects, test/dev copies, and analytics data. It supports all the key capabilities required by these use cases including globally distributed NFS, SMB and S3 storage, unlimited snapshots, global deduplication, encryption, replication, global indexing and search, and good performance for both sequential and random IO.
- Namespaces: SpanFS can divide physical storage pools into separate shared namespaces (or View Boxes) that have common data reduction, availability or archive policies.
- Tenants: SpanFS supports multiple tenants (or partitions) with strong QoS capabilities, data isolation between tenants, separate encryption keys, and role-based access control.

Combining the Best of Enterprise and Cloud Stacks

To achieve this objective, SpanFS is designed to combine the best of enterprise and cloud stacks. It's the only file system in the industry that simultaneously provides NFS, SMB and S3 interfaces, global deduplication, and unlimited snaps and clones, on a web-scale platform.

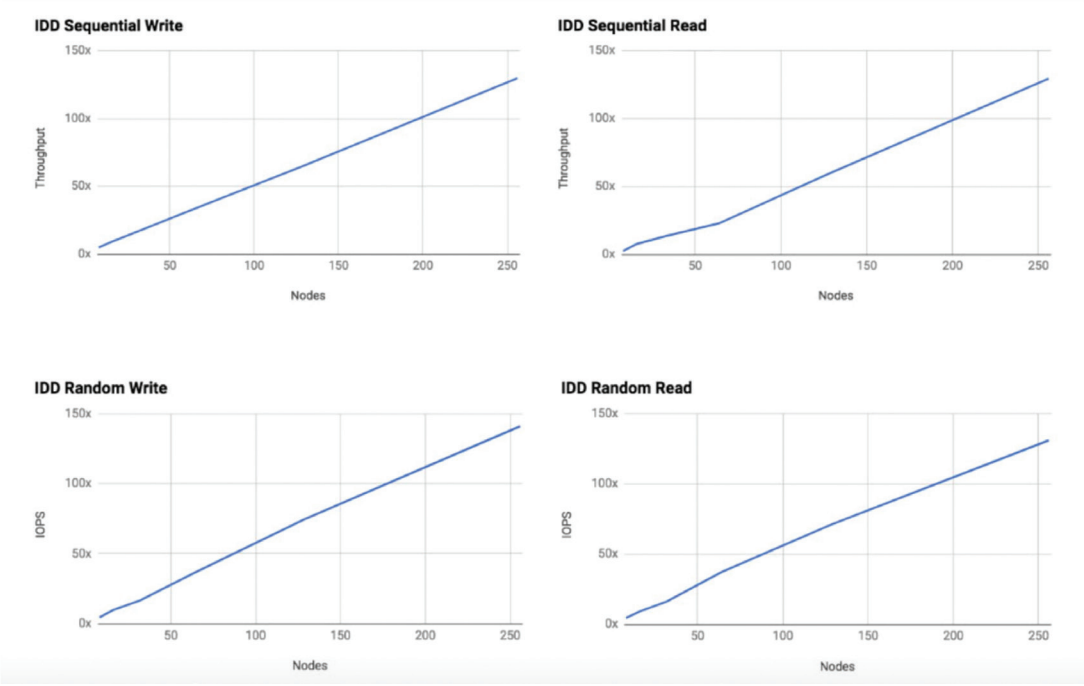


Requirement	SpanFS	SpanFS Implementation
Web-scale	✓	Truly distributed file system with no master node or single point of bottleneck. Always-on availability and data resiliency with erasure coding or replication. Non-disruptive upgrades. Dynamic scalability.
Multiprotocol NFS, SMB and S3	✓	Distributed volumes with multiprotocol NFS, SMB and S3 access
Strict Consistency	✓	Strict Consistency across nodes within a cluster ensures data resiliency and that reads always return most recent data
Global deduplication	✓	Global deduplication across all nodes, variable length, inline or post-process
Unlimited distributed snaps and clones	✓	Unlimited, distributed snaps and clones with no performance impact
Automated tiering	✓	Automatic tiering of data between SSD, HDD, and cloud storage
Software-defined on x86	✓	Software-defiend solution that can be deployed on Cohesity or 3rd party x86 nodes
Global indexing and search	✓	Indexing of all file and object metadata, global search.
Public cloud integration	✓	Integration with all the leading public clouds for archival, tiering and replication
Replication	✓	Unlimited geo replication.
Encryption	✓	Software-defined AES-256, FIPS compliant encryption of data at rest and in-flight
Multitenancy w/ QoS	✓	Built-in multitenancy with strong QoS support, data isolation, separate encryption keys, and role-based access control
High IO performance for sequential and random IO	✓	Auto-detect IO profile and place data on most appropriate media (SSD, HDD or cloud disks)

Unlimited Scalability

SpanFS is designed to maximize scalability. Everything in the file system is completely distributed across all the nodes in a cluster. There is no master node and no single point of bottleneck. The data and the IO are dynamically balanced across all the nodes, and individual nodes can be added or removed to adjust capacity or performance with no downtime. The system provides always-on availability and the data remains available even in the event of multiple node failures. Software updates are completely non-disruptive and done with rotating node updates.

Cohesity tested the scalability of SpanFS by running Cohesity DataPlatform Cloud Edition in Microsoft Azure. The cluster was scaled from 8 to 256 nodes, and as the graphs show, IO throughput scaled almost linearly for both sequential and random IO.

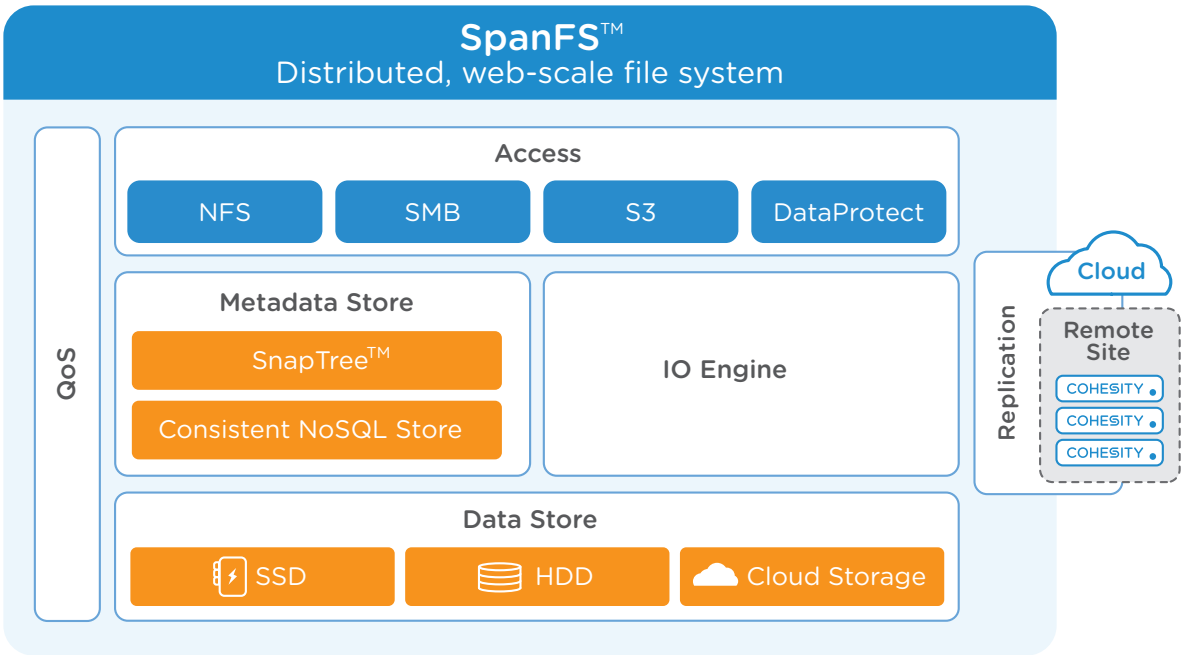


Random and sequential read/write performance of SpanFS scaling from 8 to 256 nodes in Microsoft Azure

SpanFS: Under the Hood

Architecture

SpanFS is a completely new file system designed specifically for secondary storage consolidation. At the topmost layer, SpanFS exposes industry-standard, globally distributed NFS, SMB, and S3 interfaces. It also provides a proprietary interface to store backup data protected with Cohesity DataProtect. The IO Engine manages IO operations for all the data written to or read from the system. It detects random vs. sequential IO profiles, splits the data into chunks, performs deduplication, and directs the data to the most appropriate storage tier (SSD, HDD, cloud storage) based on the IO profile. Random IO is placed on the SSD tier, sequential IO is sent straight to HDD or SSD based on QoS, and colder data may be sent to the cloud if cloud storage is in use. To keep track of and manage the data sitting across nodes, Cohesity also had to build a completely new metadata store. The metadata store incorporates a consistent, distributed NoSQL store for fast IO operations at scale, and SnapTree™ provides a distributed metadata structure based on B+ tree concepts. SnapTree is unique in its ability to support unlimited, frequent snapshots with no performance degradation. SpanFS has QoS controls built at all layers of the stack to support workload and tenant-based QoS, and can replicate, archive and tier data to another Cohesity cluster or to the cloud. The file system is distributed on hyperconverged nodes, built with commodity x86 servers, available from Cohesity or 3rd party hardware partners such as Cisco and Hewlett-Packard Enterprise. SpanFS can also be deployed in the public cloud, on cloud VMs, and is available in public cloud service catalogs.



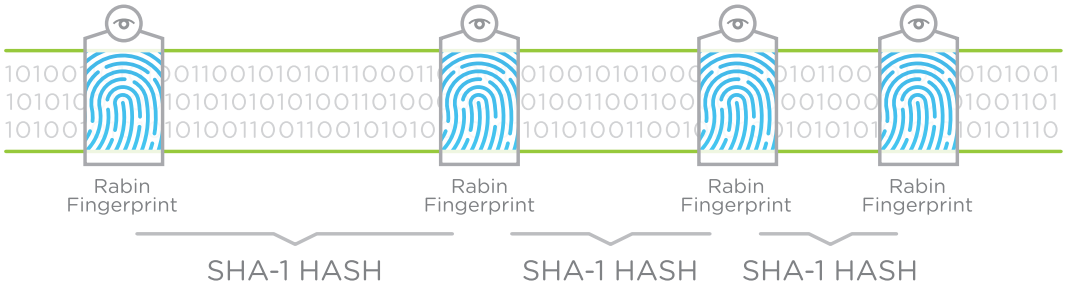
The Components

The access layer: NFS, SMB and S3 protocols: SpanFS exposes industry-standard NFS, SMB and S3 protocols. Any number of volumes or object buckets can be configured simultaneously on a single Cohesity cluster. The volumes are completely distributed with no single choke point. The data is spread out across all the nodes in the cluster. Volumes are accessed through a virtual IP mount point, and user access a ıd IO are distributed across the nodes using the virtual IP address. Each of these volumes benefits from all the unique SpanFS capabilities such as global deduplication, encryption, replication, unlimited snapshots, and file / object level indexing and search.

The IO Engine: The IO Engine manages read and write IOs as well as data operations like deduplication. The IO Engine automatically detects whether the workload is sequential or random in nature, and directs IO to the appropriate data path and media tier based on the profile. Sequential IOs may go straight to HDDs or may use the SSDs based on QoS policies. Random IOs are directed to a distributed data journal that resides on SSDs. This mechanism enables SpanFS to effectively manage both sequential and random IOs with high throughput and low latency. The IO Engine splits the data into chunks and spreads the chunks across nodes to maximize performance

and capacity utilization. Each chunk is protected against node failures either by replicating the chunks across nodes or by using erasure coding across nodes. The IO Engine is also responsible for performing data operations that are required prior to storing the chunks of data, such as variable-length deduplication and indexing.

Deduplication is performed using a unique, variable-length data deduplication technology that spans an entire cluster, resulting in significant savings across a customer's entire storage footprint. SpanFS creates variable-length chunks of data, which optimizes the level of deduplication no matter the type of file. In addition to providing global data deduplication, Cohesity allows customers to decide if their data should be deduplicated in-line (when the data is written to the system), post-process (after the data is written to the system), or not at all.



The Data Store: The data store is responsible for storing data on HDD, SSD, and cloud storage. The data is spread out across the nodes in the cluster to maximize throughput and performance, and is protected either with multi-node replication or with erasure coding. Sequential IOs may go straight to HDDs or to SSDs based on QoS policies. Random IOs are directed to a distributed data journal that resides on SSDs. As the data becomes colder, the data store can tier the data down from SSD to HDD. And hot data can be up-tiered to SSD. Customers may set up cloud storage, in which case the data store can automatically move colder chunks of data to reside in the cloud, and bring the chunks back to HDD or SSD once they become hot again.

SnapTree™: In legacy storage, snapshots form a linked chain, with each link containing the changes from the prior snapshot. Every time a new snapshot is done, an additional link is added to the chain. As the chain grows, the performance overhead required to access the data increases proportionally because the file system must traverse the chain to access the data. Hence snapshots introduce performance overhead and are limited in scope.

SnapTree™ introduces a completely new approach to managing metadata at scale, and enables SpanFS to provide unlimited snapshots with no performance overhead. SnapTree is based on a B+ tree metadata structure, but adds multiple innovations including:

- Distributes the tree across nodes
- Provides concurrent access from multiple nodes
- Supports the creation of instantaneous clones and snaps
- Garbage collects unreferenced nodes in the background using Map-Reduce
- Ensures consistent access performance regardless of the number of snapshots and clones
- Stores only one value per leaf node, as opposed to multiple values in traditional B+ trees. This avoids unnecessary snapshotting of multiple values
- Supports a variable fan-out factor that increases further down the tree. This avoids making any given sub-part of the tree too hot while at the same time keeps tree balancing costs low

With SnapTree™, Views (volumes) and files are represented by a tree of pointers to the underlying data. The root node represents the View or individual files. The root node points to some intermediary nodes, which in turn point to the leaf nodes which contain the location of the chunks of data. Customers can take snapshots of entire Views (volumes) or individual files within the Views. As snapshots are taken, the number of hops from the root to the leaves does not increase. Customers can take snapshots as frequently as desired – without ever experiencing performance degradations.

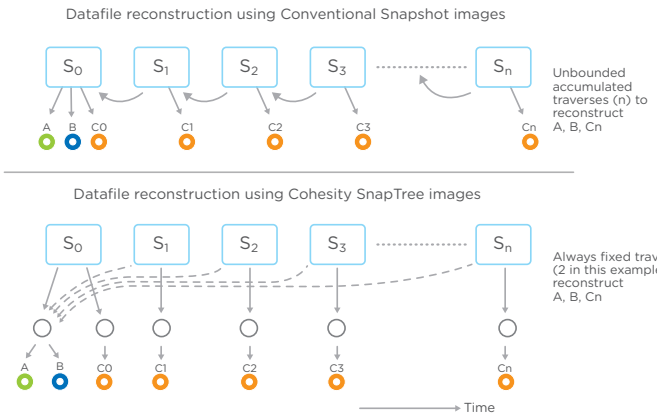


Fig 1. Cohesity SnapTree technology helps create snapshots without incurring the recovery penalty of traversing the entire snapshot chain seen in traditional snapshot architecture.

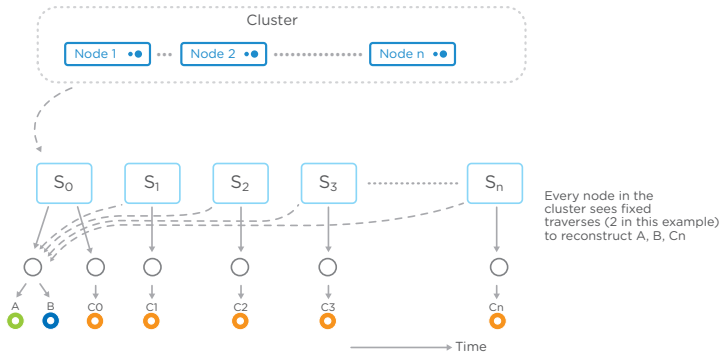


Fig 2. Distributed-Redirect-on-Write' (DROW) snapshot ensures that every node sees the same nested structure of the SnapTree

Because SnapTree is implemented on a distributed file system (Fig 2), every node sees the same nested structure of the chain with a fixed depth independent of where the actual data is stored in the cluster. Keeping the snapshots fully hydrated improves the recovery times of any snapshot from t_0 to t_n because it does not incur the time penalty of traversing the entire chain of changes. Each of these snapshot clones is fully hydrated so that businesses can achieve fast RTO and near-continuous RPO objectives. SnapTree is available as part of Cohesity DataPlatform.

Consistent NoSQL Store: The metadata store uses a distributed NoSQL store that stores the metadata on the SSD tier. It's optimized for fast IO operations, provides data resiliency across nodes, and is continually balanced across all the nodes. However, the key-value store by itself provides only 'eventual consistency'. To achieve strict consistency, the NoSQL store is complemented with Paxos algorithms. With Paxos, the NoSQL store provides strictly consistent access to the value associated with each key.

QoS: Quality of Service is designed into every component of the system. As data is processed by the IO Engine, Metadata Store, or Data Store, each operation is prioritized based on QoS. High priority requests are moved ahead in subsystem queues, and are given priority placement on the SSD tier.

Replication and Cloud: SpanFS can replicate data to another Cohesity cluster for disaster recovery, and archive data to 3rd party storage like tape libraries, NFS volumes and S3 storage. SpanFS has also been designed to interoperate seamlessly with all the leading public clouds (AWS, Microsoft Azure, Google Cloud). SpanFS makes it simple to use the cloud in three different ways. CloudArchive enables long-term archival to the cloud, providing a more manageable alternative to tape. CloudTier supports data bursting to the cloud. Cold chunks of data are automatically stored in the cloud, and can be tiered back to the Cohesity cluster once they become hot. Finally, CloudReplicate provides replication to a Cohesity Cloud Edition cluster running in the cloud. The Cohesity cluster in the cloud manages the data to provide instant access for disaster recovery, test/dev, and analytics use cases.

Beyond Copy Data Management: Instant Provisioning of Zero-Cost Clones

SpanFS enables enterprises to provision data instantly to support secondary storage use cases. SpanFS doesn't just do Copy Data Management – it eliminates the need to make data copies. Users can instantly provision clones of backup data, files, objects, or entire Views and present those clones to support a variety of use cases. For example, the clones can be used for instant recovery, as test/dev copies, or to support analytics. All these use cases can be supported directly on SpanFS serving as the active storage system, or the data can first be moved to another storage system such as primary storage or test/dev storage.

The snapshots and clones are very efficient. They don't consume space until data is modified, in which case they only need to store the deltas from the original copy. They can be created instantly without having to move data between storage devices. This is in stark contrast to the inefficiency of traditional secondary storage, where full copies of data are created between storage silos wasting lots of storage capacity, time, and IO bandwidth.

Summary

Cohesity set out on a bold mission to redefine secondary storage, enabling enterprises to take back control of their data by consolidating all secondary storage on a single web-scale platform that spans from the edge to the cloud. To achieve that objective, Cohesity designed SpanFS, a web-scale, distributed file system that provides unlimited scale across any number of industry-standard x86 nodes. SpanFS manages data across private datacenters and public clouds, spans media tiers, and covers all secondary storage use cases including data protection, file and object storage, cloud integration, test/dev, and analytics.

SpanFS is designed to combine the best of enterprise and cloud stacks. It's the only file system in the industry that simultaneously provides distributed NFS, SMB and S3 interfaces, global deduplication, and unlimited snapshots and clones, on a web-scale platform with non-disruptive upgrades and dynamic "pay-as-you-grow" capacity expansion.

SpanFS includes a completely new metadata store engineered from scratch for secondary storage consolidation. It's based on a consistent NoSQL store and SnapTree, which provides unlimited, zero-cost, distributed snapshots and clones with no performance overhead.

SpanFS is a unique file system in the industry – the only file system that spans all secondary storage use cases at web-scale, bringing together the best of enterprise and cloud stacks.

2132018

Cohesity, Inc.
Address 300 Park Ave., Suite 800, San Jose, CA 95110
Email contact@cohesity.com www.cohesity.com
@cohesity