COHESITY

Cohesity Data Protection White Paper

Table of Contents

Executive summary

Cohesity provides the only hyper-converged platform that eliminates the complexity of traditional data protection solutions by unifying your end-to-end data protection infrastructure – including target storage, backup, replication, disaster recovery, and cloud tiering. Cohesity DataPlatform provides scale-out, globally deduped, highly available storage to consolidate all your secondary data, including backups, files, and test / dev copies. Cohesity also provides Cohesity DataProtect, a complete backup and recovery solution fully converged with Cohesity DataPlatform. It simplifies backup infrastructure and eliminates the need to run separate backup software, proxies, media servers, and replication. This paper specifically focuses on the business and technical benefits of Cohesity DataPlatform for the data protection use case. It is intended for IT professionals interested in learning more about Cohesity's technology differentiation and advantages it offers for data protection - (i) Eliminate complexity with a unified platform for end-to-end data protection (ii) Ensure fast Recovery Points and near-instantaneous Recovery Times (iii) Increase storage efficiency with scale-out, globally deduped storage (iv) Integrate seamlessly with all the leading cloud providers to leverage economics of public cloud services.

Challenges in data protection fueling the next wave of innovation

A robust data protection strategy must meet the following criteria: (i) Provide local data protection from accidentally deleted files, application crashes, data corruption, and viruses (ii) Enable fast recovery of individual files and applications (iii) Retain data for long periods of time to satisfy compliance and regulatory requirements (iv) Provide off-site data protection and reliable Disaster Recovery.

Over the last two decades, we have experienced tremendous innovations in data protection technology. Each innovation made huge advances in backup & recovery SLAs with much greater simplicity when compared to previous generations of products (Figure 1).
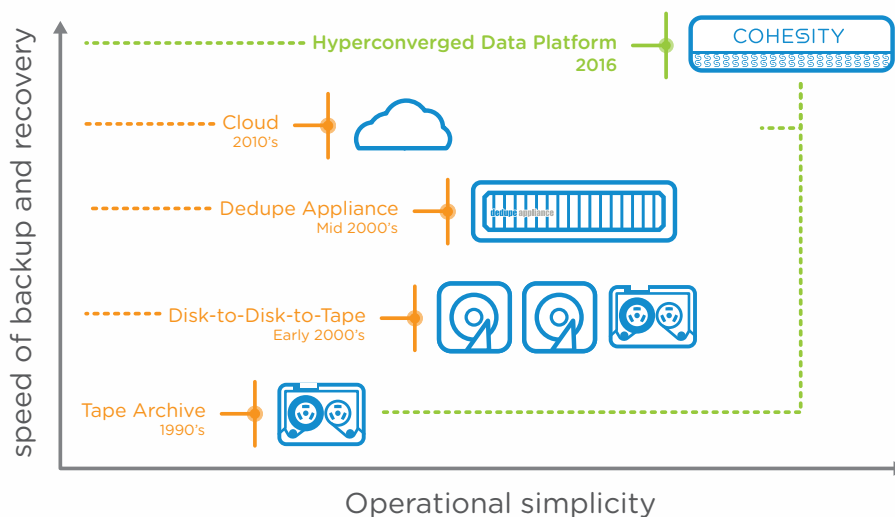

Figure 1: Evolution of data protection technologies. Cohesity is driving innovation through a Hyperconverged Data Platform that improves the Data protection metrics of speed and economics.

Tribune Media, a global media company, had separate data management, protection, and storage solutions that resulted in integration challenges and management complexity. Tribune achieved the following benefits by deploying Cohesity DataProtect

Key Benefits

- Simplify data protection with integrated backup and recovery running directly on the Cohesity DataPlatform
- Instantly search-and-recover data with the Cohesity indexing engine
- Scale data protection across the enterprise with Cohesity's web-scale DataPlatform architecture

" Cohesity gives us an easy way to manage, store, and extract value from our data — all from a single platform that scales as we need it. Cohesity is dramatically simplifying the way we do business. "

– David Giambruno,
CIO and Senior VP of IT

COHESITY

Originally data protection focused on backing up data from internal hard drives to tape drives and libraries. With advances in disk densities and inevitable reduction in costs, backup broadened to include hard disk drives (HDD) as the backup target. HDDs also improved the granularity of allowing individual items to be recovered thereby improving restore times and processes. The next wave of disk-based backup introduced a tiered backup architecture that combined disk and tape leading to Disk-to-Disk-to-Tape (D2D2T). The continued growth of disk as a backup target offered new opportunities to exploit the advantages of disk over tape leading to deduplication and compression technologies driving efficiencies in physical storage. Vendors began packaging purpose built backup appliances to further simplify deployment and maintenance. In the last few years, external cloud based storage as a backup target has entered the scene driving cost optimization efforts through a shift from upfront capital investment to an operational pay-as-you-grow model. Cohesity provides the next revolutionary wave of innovation that combines the advantages of disk-based deduplication, scale-out architecture, unified data protection, and native cloud integration. The net result is that Cohesity can deliver very fast SLAs while simplifying the end-to-end data protection environment, at a substantially lower cost per protected TB.

Before we delve into the details of the Cohesity data protection solution, it is important to understand the metrics across which these incumbent technology advances fall short in meeting the rigors of data protection. IT executives increasingly recognize that the key driver of exponential data growth is redundant copies of primary corporate data created by the various point-solution tools used to protect, share, and analyze information. IDC quantifies this as the copy data ratio which is measured as the total data in the environment over the total amount of production data. The increase in this ratio for enterprises stems from data residing in backup storage systems, in disaster recovery environments, in test and development clusters, and in archival for long term retention; all of which quickly multiplies across all applications. Gaining visibility and control over these data copies is becoming paramount to gaining operational efficiency and agility. Figure 2 is an illustration of the data protection architecture that is common across enterprises today. The challenge is that the different tiers of storage and backup solutions work in siloed hardware and software form factors with their corresponding data management and protection software. The solutions are mostly passive in nature, becoming expensive insurance policies that sit idle until a restore request is received. Also, in many cases customers only test these data backup copies infrequently and long gaps between testing increase the chance of issues being found when data needs to be actually recovered – at which point it may be too late.

GS1, global logistics provider, had data growth rates of 30% with stringent retention requirements. With a limited time window for completing backups, they were using expensive primary storage to accomplish this task. By leveraging Cohesity, they were able to

- Achieve faster backup and restore that has shrunk RTO and RPO windows by over 90%
- Single pane of glass to manage consolidated secondary storage

> " Faced with the complex task of managing massive amounts of data that must be backed up and secured, we are excited about Cohesity's vision for converging data onto a unified, scale-out solution. Cohesity provides simpler data protection and powerful insight, which is what we want from our next-generation solution for storage and recovery. We look forward to scaling this environment as we need, one node at a time, so we can spend less time managing our data infrastructure. "
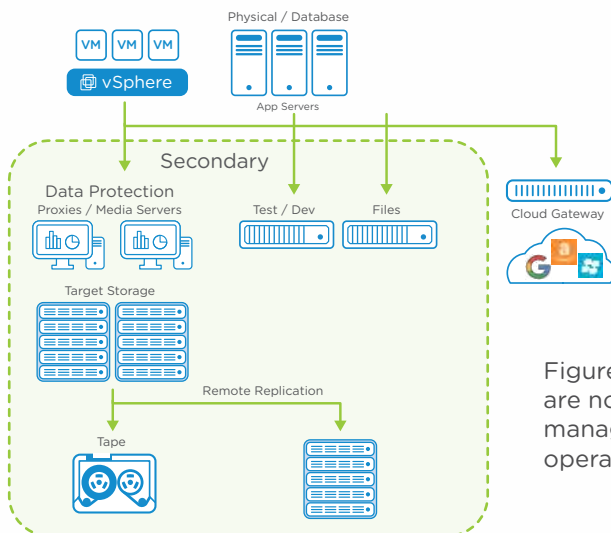>
> – Sase Janki, SVP of Technology

Figure 2: Challenges with current data protection solutions include but are not limited to multiple siloed hardware and software, heterogeneous management tools, islands of storage subsystems, and complexity of operational maintenance.

When we look further at the incumbent data protection solutions through the lens of speed, complexity and costs, the assessment is as follows:

Speed of Backup & Recovery measured against delivery on SLAs: The 24/7 nature of enterprise is leading customers to reduce Recovery Time Objectives (RTO) and Recovery Point Objectives (RPO) to a minimum – ideally in the order of minutes, not hours or days. In spite of the high investment across various products, enterprises are still incurring downtime leading to several million dollars in penalties and lost productivity. This business impact is exacerbated by loss of customer and employee confidence if the data protection plan for critical services does not achieve its objectives. In trying to reduce RTO and RPO; customers and vendors create expensive, complex and one-off solutions that require ongoing maintenance to deliver on the "always-on-enterprise" precept.

Complexity of data protection environments: The solutions used to protect the data were designed more than 10 years ago, and haven't kept up with today's environments.   Data protection typically consists of a complex patchwork of point products for target storage, backup software, media servers, proxies, replication, and disaster recovery. The net result is that organizations are spending a lot of resources on data protection, simply as an insurance policy – maintaining copies of data that will be unproductive until a failure or disaster happens.

Total cost of ownership: Organizations have invested in multiple solutions across backup, replication, DR, and archival to create and manage multiple copies of their data. This leads to expensive data migrations, forklift upgrades, and sub-optimal capacity planning strategies to accommodate future growth. In addition to the complexities associated with managing the independent hardware and multiple software elements, data protection solutions often have different procurement methods and licensing terms, making it extremely difficult for organizations to budget and procure. Even as customers are juggling with all these elements, they are also looking at cloud services as an extensible part of an overall storage solution because of the promise of economic and scalability advantages.

Introducing Cohesity DataPlatform for data protection

Cohesity was founded on the core vision of eliminating the fragmentation in data protection and putting an end to the above mentioned shortcomings that plague incumbent solutions. Cohesity DataPlatform is the only hyper-converged platform designed to manage all your secondary data. Inspired by web scale architectures, Cohesity provides a scale-out, globally deduped, highly available platform to consolidate all your secondary data, including backups, files and test/dev copies. Cohesity DataProtect provides a complete backup and recovery solution, fully converged with Cohesity DataPlatform. It simplifies backup infrastructure and eliminates the need to run separate backup software, proxies, media servers, and replication.
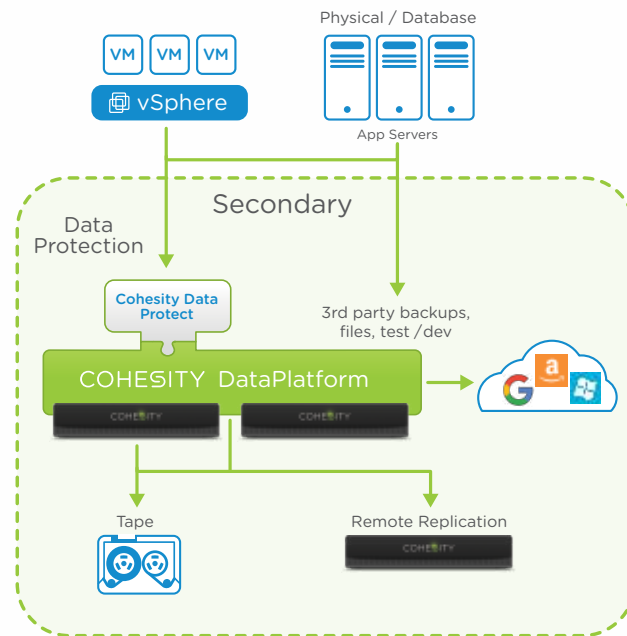

Figure 3: Cohesity data protection for virtual and physical environments with integrated, policy-driven backup and recovery. The solution also supports native cloud integration for tiering, archival, and replication to public cloud services such as Google Cloud Storage Nearline, Microsoft Azure and Amazon S3/Glacier.

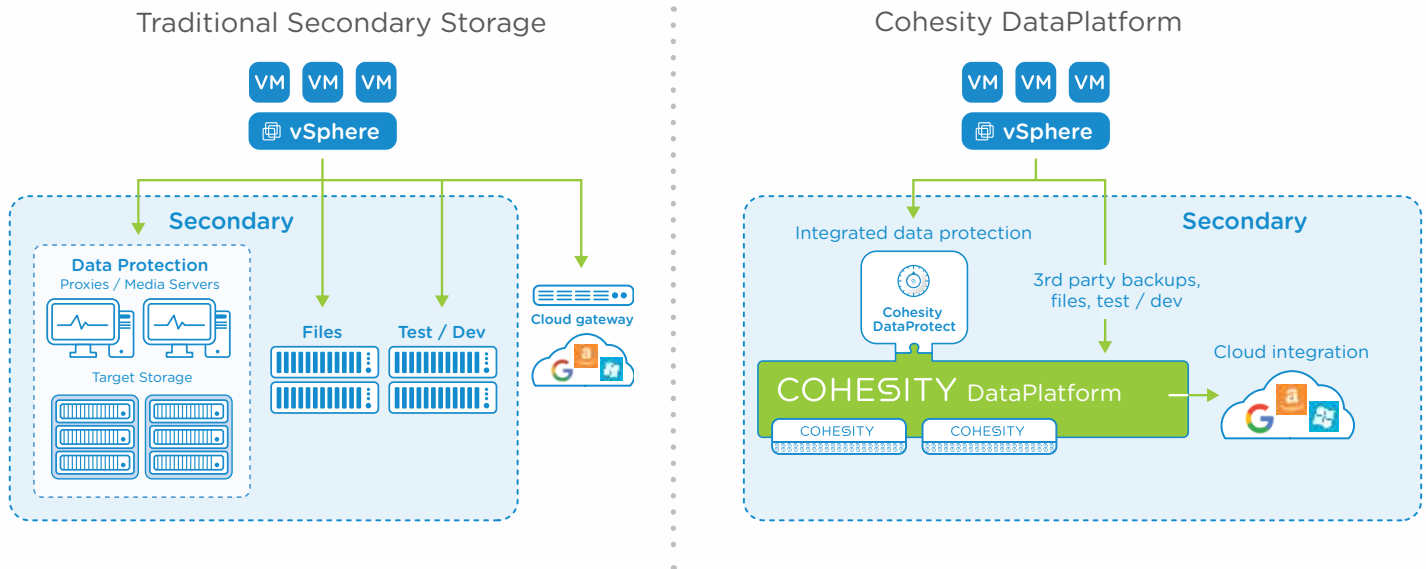The Cohesity Data Platform and Cohesity Data Protect provides the following benefits for data protection:

Figure 4: Cohesity DataPlatform and DataProtect for virtual and physical environments with integrated, policy-driven backup and recovery. The solution also supports native cloud integration for tiering, archival, and replication to public cloud services such as Google Cloud Storage Nearline, Microsoft Azure and Amazon S3/Glacier.

Eliminate complexity with a unified platform for end-to-end data protection: Cohesity provides an end-to-end backup and recovery solution that is fully converged on the Cohesity platform. The Cohesity solution integrates backup, restore, replication, DR and failover/failback orchestration, and eliminates the need for complicated point products. Data protection is managed through a set of policies that specify application SLA requirements including RPO, retention policies, off-site replication and cloud or tape archival. Managing data protection becomes as simple as assigning policies to Virtual Machines and applications based on business requirements. Data can also be replicated across Cohesity clusters in multiple sites to offer protection from disaster scenarios. This is all achieved with a single click user interface, drastically reducing the time and overhead when compared with traditional data protection solutions.

Ensure fast Recovery Points and near-instantaneous Recovery Times: Cohesity supports zero-cost, highly scalable snapshots and clones. Businesses can take a large number of clones at any time interval with uncapped retention policies, without ever affecting performance or consuming additional space. Leveraging the Cohesity snapshots and fast, parallel data ingest, businesses can protect applications as often as desired and reduce RPOs to as low as 15 minutes. Virtual Machines can be recovered instantly by creating a clone of the backup VM and running that clone directly on the Cohesity platform. If needed, the clone can be moved back to primary storage using Storage vMotion. Data protection is further enhanced through an indexing engine that rapidly indexes all the VMs being backed up, the files within the VMs, and all associated metadata. This has the benefit of easily mining backup data with a simple text-based search and restore to quickly find VMs and files. This restore can also place the file(s) in the original source location further reducing the burden associated with managing restore processes.

Leverage favorable economics of public cloud services: The Cohesity solution supports long term data retention to external tape and public cloud services such as Google Cloud Storage Nearline, Microsoft Azure and Amazon S3/Glacier. Customers can leverage the public cloud as an extension of the on-prem Cohesity infrastructure in one of three ways (i) CloudArchive–archiving the older local snapshots in the Cohesity cluster to cloud for long-term retention. As restores from archivals are typically few and far apart, the cost associated to recover data from the cloud can be kept under check. (ii) CloudTier – using cloud as an extension to Cohesity's built-in storage tiers provides the ability to algorithmically decide when to down-tier or up-tier the data between Cohesity cluster and the Cloud. (iii) CloudReplicate – Replicating the latest data to a Cohesity cluster in the cloud to provide Disaster Recovery and off-site data protection.

Increase storage efficiency with scale-out, globally deduped storage: Cohesity provides a scale-out, globally deduped storage platform. With the ability to add or remove individual nodes at any time, the Cohesity cluster automatically scales

up or down by rebalancing the data and its associated metadata to ensure redundancy. The shared-nothing scale-out model eliminates the cost and complexity associated with individual dedupe appliances. Cohesity enables a pay-as-you grow, linear scaling model and increases space efficiency by deduplicating across an entire Cohesity cluster – not just one appliance. The platform increases availability by maintaining data availability even in the event of a complete node failure. Finally, the Cohesity approach eliminates the need to perform data migrations for forklift upgrades.

Provide tight integration with VMware vSphere: Cohesity's native data protection software enables businesses to easily protect data for virtual and physical environments, dramatically reducing cost and complexity for businesses. Cohesity DataProtect is tightly integrated with VMware vCenter for businesses to instantly see a full list of the virtual environment and choose which virtual machines to protect. These virtual machines can then be protected with easy-to-use protection profiles that can be customized based on SLA, retention period, or application group. The process leverages available VMware APIs for Data Protection (VADP), eliminating the need to install in-guest agents across the virtualized infrastructure. As new virtual machines are added, they are auto discovered and included in the protection policy that meets the desired SLAs. Easy and intuitive policy administration through a single pane of glass provides a plug-and-play experience for managing daily operations.

Support physical applications and custom backup scripts: In addition, Cohesity provides out-of-the-box adapters for physical applications including SQL, Windows, Linux and other widely used enterprise applications. Customers also have the option to supply custom backup scripts that leverage native APIs (such as Oracle RMAN, rsync for physical server, dumpbin for MySQL) to move the backup file onto the Cohesity platform. Such broad support for virtual and physical environments has the benefit of having Cohesity schedule, report, replicate, and archive the backup datasets across heterogeneous applications.

Support for 3rd party backup applications: Cohesity provides customers with choice of backup software. Cohesity DataProtect provides the fully integrated backup and recovery solution on Cohesity DataPlatform. But Cohesity also presents itself as an NFS target and integrates seamlessly with all the existing backup products. This gives customers a choice of deploying the Cohesity solution in either of two modes (Figure 4): (i) Using Cohesity DataProtect or (ii) Leveraging its highly efficient, web-scale storage capabilities as a backup target in conjunction with third-party data protection software (such as Veeam, Commvault, NetBackup). In either mode, you can buy what you need today and scale up as your data grows with the benefit of leveraging the latest hardware technology in the market.

Cohesity Technical Advantages

Cohesity DataPlatform is a scale-out, hyperconverged platform for secondary data. It is deployed on Hyperconverged Nodes – each one providing a combination of compute, SSDs and HDDs. Cohesity DataPlatform provides built-in dedupe, compression, SnapTree snapshots, data indexing, replication, encryption, and cloud integration. Cohesity DataProtect is fully integrated on the platform and provides an end-to-end backup and recovery suite.
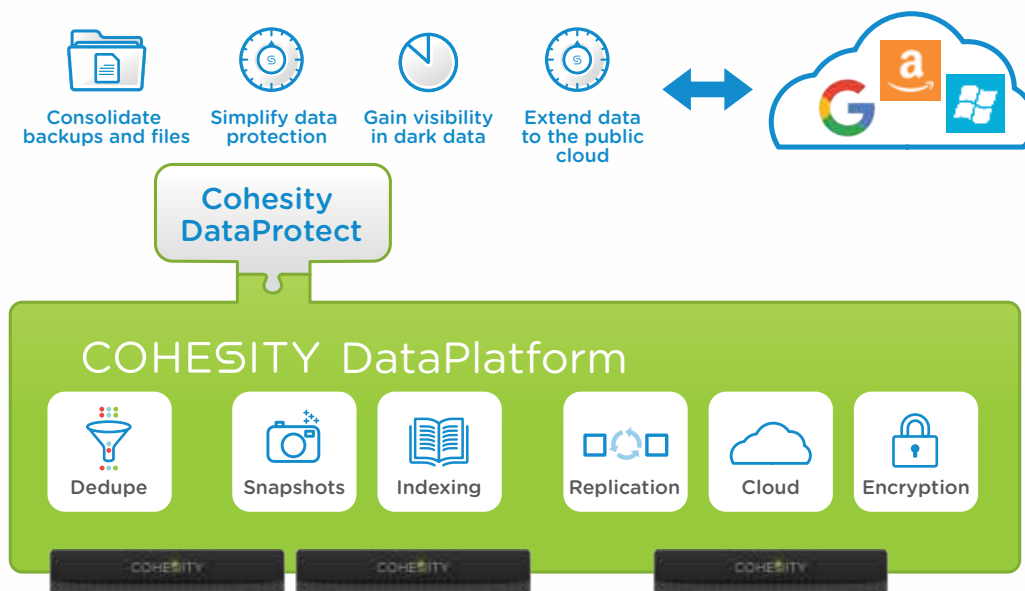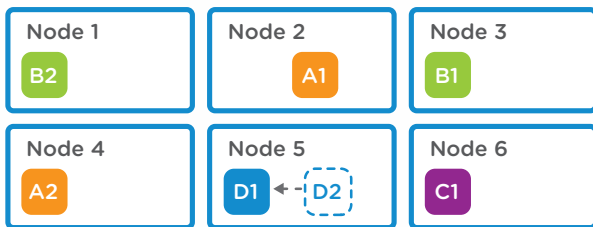

Figure 5: Cohesity Architecture

Cohesity DataPlatform: Cohesity Data Platform is the scale-out, hyperconverged operating environment that provides the storage target for data protection. Cohesity DataPlatform is deployed on Cohesity Hyperconverged Nodes - hyperconverged x86 servers that provide compute and storage. Customers can start with as low as a three-node cluster and scale out one-node-at-a-time, to accommodate future capacity and performance requirements.

Cohesity DataPlatform includes a true distributed file system with a shared-nothing architecture, inspired by web-scale file systems (Google File System). The data is split into chunks and spread out across the nodes, with 2 copies of each chunk for redundancy and availability. The ingest engine ensures that data is optimally placed onto the SSD or spinning disk tier that best suits the profile of the incoming data stream. In order to ensure that performance of the primary datastore is impacted minimally, the ingest engine offers adaptive data throttling that modulates the backup ingest performance over the production workloads at the vCenter or datastore level.

Global deduplication and compression: Data deduplication is a storage efficiency feature that frees up storage capacity by eliminating redundant data blocks. Different vendors implement deduplication at a file-level and/or a block-level of different sizes, which only works well across a single storage pool or within a single object (e.g. application or VM). Cohesity leverages a unique, variable-length data deduplication technology that spans an entire cluster, resulting in significant savings across a customer's entire storage footprint (Figure 7). With variable-length deduplication, the size is not fixed. Instead, the algorithm divides the data into chunks of varying sizes based on the data characteristics. The chunk size is varied in real time in response to the incoming data which results in greater data reduction than fixed-size deduplication. The efficiency benefit of variable-length deduplication compounds over time as additional backups are retained. Cohesity also allows customers to decide if their data should be deduplicated in-line (when the data is written to the system) or post-process (after the data is written to the system) to optimize the backup protection jobs against backup time windows. Cohesity also provides compression of the deduped blocks to further maximize space efficiency.

Node level deduplication only maintains 1 copy of block D; D2 is a pointer to D1. No dedupe is achieved for blocks A & B.

Cluster level deduplication only maintains 1 copy of blocks A, B & D. A2, B2 & D2 are just pointers to A1, B1 & D1. This results in greater efficiencies in terms of utililzation.

Traditional dedupe implementation
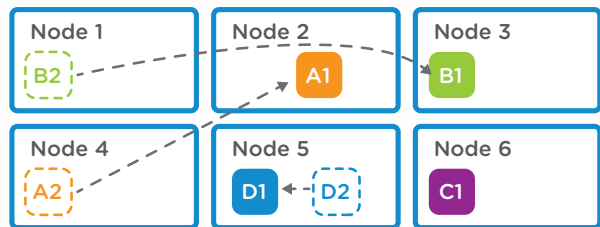
Cohesity dedupe implementation

Figure 6: Cohesity's global deduplication across all nodes in a cluster results in less storage consumed compared to just node level deduplication used in several other in market data protection solutions

Cohesity SnapTree™ for managing data copies: In legacy storage solutions, snapshots (of a file system at a particular given point in time) form a chain, tracking the changes made to a set of data and form the basis for organizing and storing copies of data. Every time a change is captured, a new link is added to the chain. As these chains grow with each and every snapshot, the time it takes to retrieve data on a given request grows because the system must re-link the chain to access that data. Cohesity's patented SnapTree™ technology creates a tree of pointers that limits the number of hops it takes to retrieve blocks of data, regardless of the number of snapshots that have been taken. SnapTree uses a B+ tree data structure such that access to any point in the tree takes a fixed number of hops no matter how many snapshots there are, without having to rebuild any chain linkage. Because SnapTree is implemented on a distributed file system, 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 t0 to tn because it does not incur the time penalty of traversing the entire chain of changes (Figure 6). This capability is available with the Integrated Cohesity data protection.

Datafile reconstruction using Conventional Snapshot images

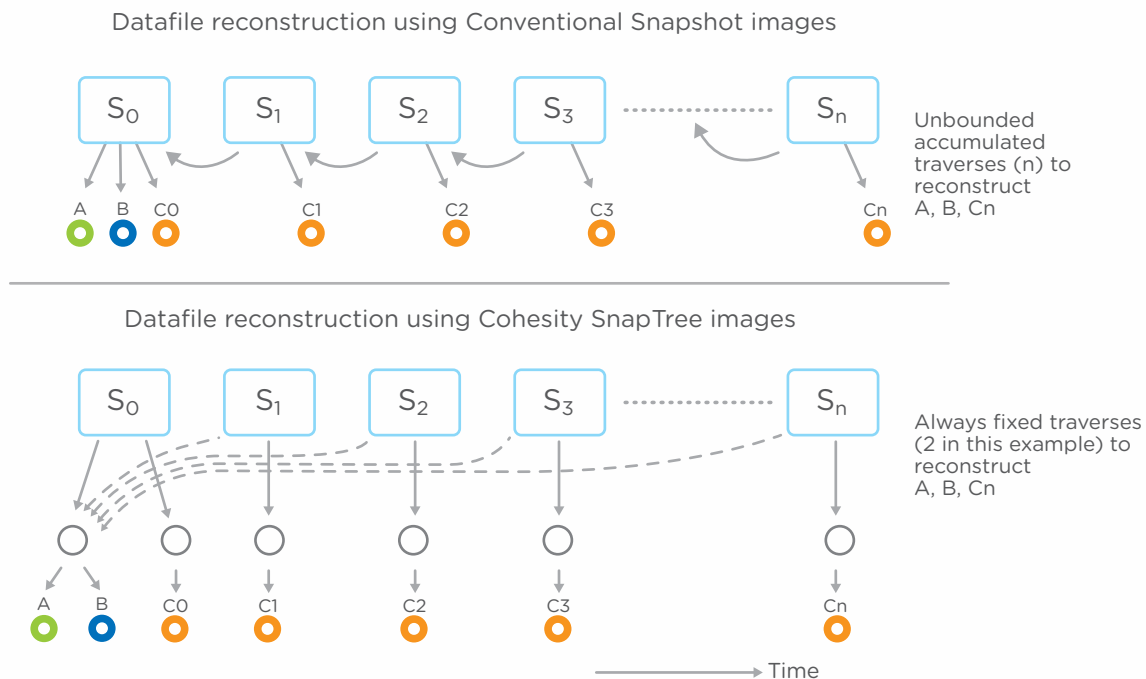Datafile reconstruction using Cohesity SnapTree images

Figure 7: Cohesity SnapTree technology helps create large number of snapshots without incurring the penalty of traversing the entire snapshot chain seen in traditional snapshot architecture

Cohesity indexing engine for rapid search and recovery: Cohesity DataPlatform includes an indexing engine that enables rapid search-and-recover capabilities. This allows users to quickly find and restore files stored within higher-level data objects such as VMs. This indexing engine automatically and rapidly indexes an entire vCenter cluster and all its associated metadata. As virtual machines become protected, Cohesity's indexing engine cracks open the underlying files and indexes the metadata; enabling extremely fast, wild-card search results that are then used for instantaneous granular restores. As organizations scale-out, this indexing engine spans across all nodes in the cluster, leveraging the aggregate power of all CPUs and available memory to rapidly recover files or virtual machines. This greatly improves the RTO and RPO objectives compared to traditional data protection architectures. This capability is available with the integrated Cohesity DataProtect.

Cohesity replication: The backups that are generated in one Cohesity cluster can be replicated to one or more target Cohesity clusters on multiple sites. The ability to replicate backups from primary Cohesity cluster to multiple secondary Cohesity clusters across various geographical sites, helps facilitate disaster recovery needs. The technology supports (i) One-to-one model - A single production cluster can back up to a disaster recovery site (ii) One-to-many model - A single production cluster can backup to multiple disaster recovery sites (iii) Many-to-one model - Multiple Cohesity clusters can back up to a remote cluster (iv) Many-to-many model - Multiple Cohesity clusters can backup to multiple remote clusters across several sites. This technology works in conjunction with global data deduplication to greatly reduce storage requirements across several sites and as such the network bandwidth required for replication of data for DR purposes.

Cohesity CloudTier, CloudArchive, and CloudReplicate: Cohesity DataPlatform integrates natively with all the leading public cloud providers. Cohesity can tier data to the cloud to extend storage capacity of the cluster for colder data. Cohesity can also use the cloud as a replication destination for disaster recovery. And finally, Cohesity can archive data in the cloud for long-term retention.

Cohesity encryption engine: Cohesity DataPlatform also provides encryption of data at rest and in transit over the network with AES 256-bit encryption to secure data. Encryption in flight is applicable to data that is replicated to a remote Cohesity cluster or when data is tiered/archived to the cloud from the Cohesity platform. This ensures that data stored on the Cohesity cluster is protected well from malicious attacks.
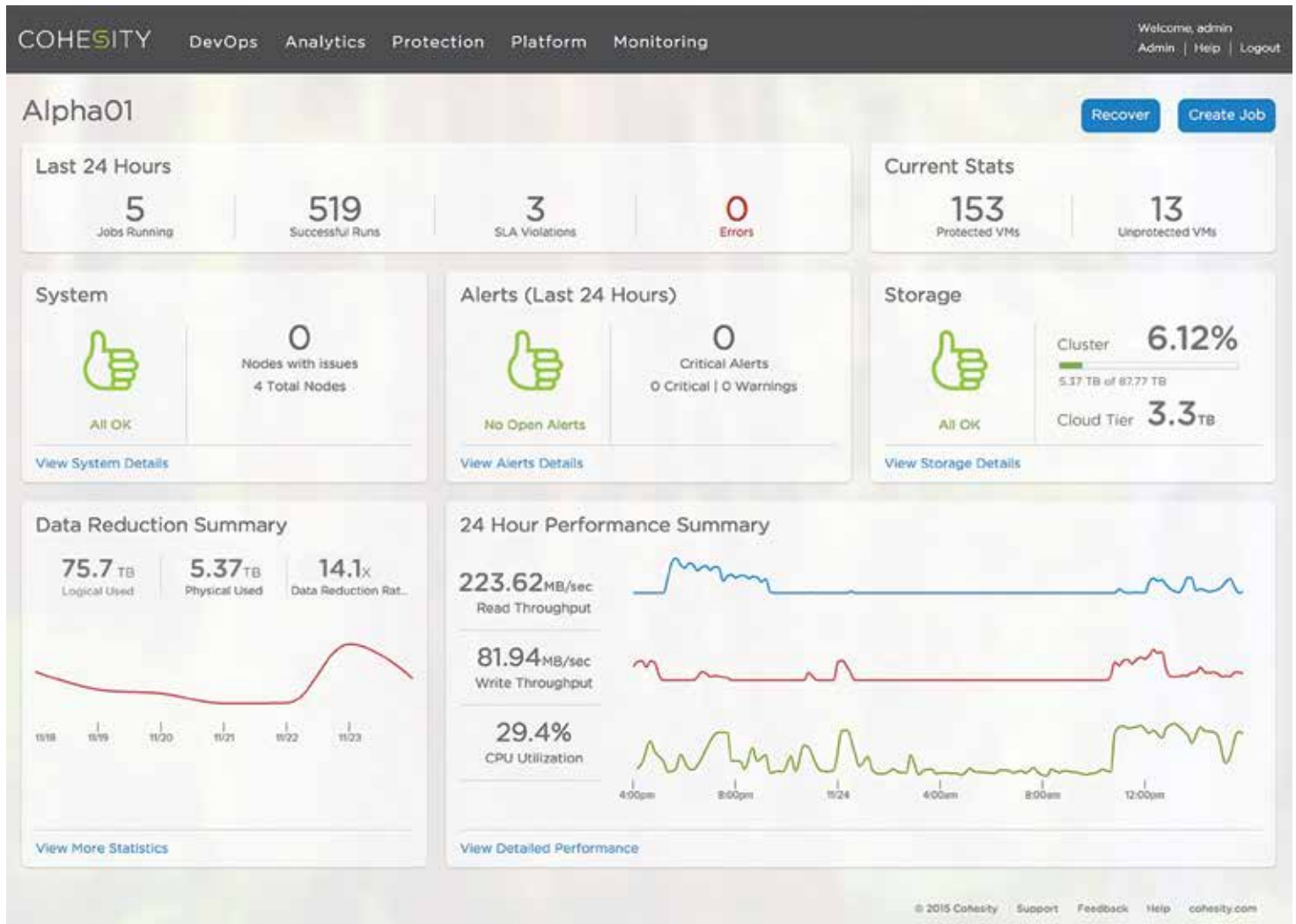

Figure 8: View of Cohesity User Interface and Data Protection

Role-based access control: Administration of the cluster can be achieved through roles assigned within the cluster or can be integrated with Windows Active Directory domains. Default roles exist that provide access for different Cohesity cluster users including Admin, Operator, Viewer, Self Service and Data Security. Each role provides access to cluster workflows such as read-only views and data restore or self service administrators that are only allowed to access specific objects like VMs or physical hosts.

Custom roles can also be created and assigned to users and groups such as cluster Access, Clone, Data Protection, Recovery , Storage and Analytics management.

Cohesity Performance Advantages

Fast backups with highly granular RPOs and instantaneous RTOs

With the unique scale-out architecture, hyperconverged nodes, and SnapTree technology, Cohesity provides significant performance advantages compared to traditional backup and recovery solutions.

Fast backups with parallel ingest: Traditional target storage consists of stand-alone dedupe appliances. These appliances have only one active controller assigned to data ingest. With the Cohesity scale-out architecture, backup jobs and backup tasks can be parallelized across the multiple scale-out nodes of the Cohesity cluster, significantly increasing the ingest throughput.

Traditional Secondary Storage Architecture

Traditional dedupe appliance
Single controller choke point

Cohesity
Parallelized backup jobs / tasks

Target Storage

Target Storage

APP OS   APP OS   APP OS   APP OS

APP OS   APP OS   APP OS   APP OS

vmware®

vmware®

Backup Jobs

Backup Jobs

COHESITY DataPlatform

COHESITY   COHESITY   COHESITY
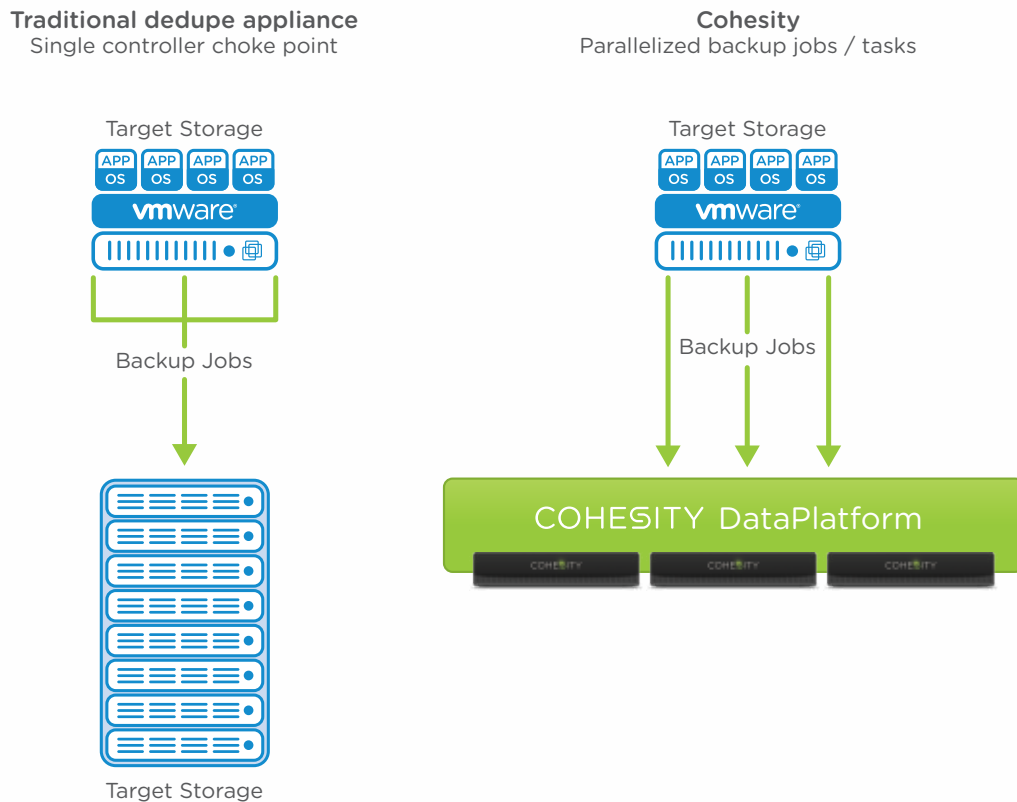
Target Storage

Figure 9: Parallel ingest on Cohesity DataPlatform

Workload-optimized ingest: The Ingest Engine ensures that data is optimally placed onto the SSD or spinning disk tier that best suits the profile of the incoming data stream. This is called the Tier-Optimized Write Scheme (TOWS) wherein spinning disks that prefer sequential I/O write data out-of-place; while SSDs are used for random I/O such that the writes are in-place and committed straight away. This engine provides the basis to interoperate with VMware VADP, physical servers and application connectors; to provide end-to-end data protection for customer environments. In order to ensure that performance of the primary datastore is impacted minimally, the Ingest Engine offers Adaptive data throttling that modulates the backup ingest performance over the production workloads at the vCenter or Datastore level.

Incrementals forever with 15-minute RPOs: Cohesity provides the SnapTree technology for zero-cost, unlimited snapshots and clones. Cohesity maps each incremental backup to an incremental snapshot. This enables highly granular, unlimited snapshots with RPOs as short as 15 minutes, with no negative performance impact and high storage efficiency.

All incrementals available instantly: Each snapshot in SnapTree is instantly available as a fully hydrated copy. With SnapTree, each incremental backup becomes instantly available as a full image, eliminating the need to perform periodic full backups.

Instant restore: Cohesity enables instant VM restore to any point-in-time copy. Upon restore, Cohesity creates an instantaneous clone of the snapshot. The Virtual Machine can be launched directly from the clone, with storage running directly from the Cohesity cluster. This eliminates the need to move data from secondary to primary prior to initiating a restore. If needed, after restore the data can be moved to primary using Storage vMotion.

Cohesity is powering customers into a new era of data protection

Cohesity DataPlatform is helping customers usher in new technology to achieve their data protection business objectives. Customers such as Tribune Media and GS1 had disparate data management, protection, and storage solutions that resulted in integration challenges and management complexity. With Cohesity, they have been able to shrink their RTO and RPO windows by over 90% and have greatly reduced the time spent on managing the infrastructure. In conclusion, organizations can use the integrated data protection capabilities of Cohesity DataPlatform for physical and virtual environments to consolidate disparate hardware and software elements. It is also possible to leverage the highly efficient, web-scale storage capabilities of Cohesity in conjunction with third-party data protection software. The move to Cohesity DataPlatform can provide a robust data protection strategy that delivers improvement in backup/recovery speed and better economics than incumbent solutions.

For more information about Cohesity, please visit www.cohesity.com