IBM Spectrum NAS, IBM Spectrum Scale and IBM Cloud Object Storage

Silverton Consulting, Inc. StorInt™ Briefing

Introduction

Unstructured data has exploded in volume over the past decade. Unstructured data, media files and other data can be created just about anywhere on the planet using almost any smart device available today. As the amount of unstructured data grows exponentially, customers using this data need to be able to take advantage of the right storage solutions to support all of their file and object data requirements.

IBM® recently added a new storage system to their Spectrum product family, IBM Spectrum Network Attached Storage (NAS). IBM Spectrum NAS adds another software-defined file storage system to IBM's current unstructured data storage solutions, IBM Spectrum Scale™ and IBM Cloud Object Storage (COS). Below, we will discuss the three systems and supply some guidance on when and where to use each of them.

IBM Spectrum NAS

IBM Spectrum NAS supports traditional data center environments as a software-defined, scale-out file storage solution. File storage systems commonly support access protocols such as CIFS/SMB (Common Internet File System/Server Message Block) or NFS (Network File System). Spectrum NAS supports SMB1 (CIFS), SMB2 and SMB3 for Windows™ data and NFS3, NFS4 and NFS4.1 for Linux®/Unix® data. All the file storage in your IT shop can be hosted on IBM Spectrum NAS.

Spectrum NAS was designed to be hardware agnostic, meaning it can run on just about any current X86 server in your data center. A group of Spectrum NAS nodes operate together as a cluster file system and can scale from a minimum of four nodes and has been validated to over 1PB configurations. Spectrum NAS can also be run as a virtual machine under standard hypervisors and is supported to run under VMware vSphere®.

In addition, customers can configure IBM Spectrum NAS individual nodes for high IO performance using all-flash storage or configure the nodes for less performance using a combination of memory and solid-state drive (SSD) data cache, or memory alone as the data cache, with disk for backend storage.

Because of the configuration flexibility of Spectrum NAS, customers can readily use new or existing servers with local direct-access storage (DAS) as new Spectrum NAS nodes. In this way, there may be no additional hardware expense to deploy Spectrum NAS storage, making Spectrum NAS an economical storage choice.

Although both node capacity and IO performance can differ throughout the cluster, IBM Spectrum NAS consumes cluster capacity in a balanced fashion. As such, it ensures that an equal percentage of storage is used on each node in the cluster.

Silverton Consulting
Strategy, Storage & Systems

Spectrum NAS supports from 2+1 to 8+2 (data + parity) data protection schemes. However, an 8+2 data protection scheme would require at least 10 nodes.

As a scale-out file system, Spectrum NAS enables customers to add nodes to the cluster at any time to increase capacity or improve IO performance. IO performance should scale linearly with the number of nodes added, assuming sufficient network bandwidth. Accordingly, growing a cluster from, say, 10 to14 nodes with equivalent hardware should increase IO performance by up to 40%.

When new nodes are added to the cluster, Spectrum NAS can be configured to instantly rebalance storage use across the cluster. Instead of waiting until enough new data is written to balance cluster capacity, the system immediately moves old data to any new nodes until the cluster reaches equal usage across all nodes.

Spectrum NAS can support a single file system that spans all nodes in a cluster. Alternatively, it can support multiple file systems for multitenant environments.

Spectrum NAS also offers **synchronous** and **asynchronous replication** to another cluster for disaster recovery. With synchronous replication, the data in one cluster is instantly copied to another cluster within a metropolitan area. With asynchronous replication, two Spectrum NAS clusters can be located much farther apart and still provide disaster recovery services. However, a delay may occur between the time when data is available at the source cluster and the time when it's copied to the replication cluster.

IBM Spectrum NAS use cases

Spectrum NAS is ideal for data centers that support user file directories. These environments call for many files per user and often grow incrementally over time. The economical capacity of Spectrum NAS together with its scalable and configurable IO performance can help here. For example, it supports SMB and NFS file access out of the box, making it easy for traditional data center use.

Data centers that run enterprise applications that can make use of file storage [e.g., database, customer relationship management (CRM), or enterprise resource planning (ERP)] are also a good fit for Spectrum NAS deployments.

IBM Spectrum Scale

IBM Spectrum Scale was designed for fast file IO performance, extreme file throughput and large file repositories and is typically used in high-performance computing (HPC) simulation, large-scale analytics and machine learning environments. Spectrum Scale is a software-defined, parallel file system that can scale in performance or capacity by adding nodes to the cluster. Highly configurable, Spectrum Scale nodes can be metadata alone, storage alone or combined

Silverton Consulting
Strategy, Storage & Systems

metadata/storage. Spectrum Scale can support a very large number of nodes in its cluster or deployed as a shared filesystem across storage rich compute nodes.

IBM Spectrum Scale has multiple deployment options. For highest performance it requires the use of host software that needs to be installed in all clients using the system, but, once installed, parallel, distributed access is available across the cluster. For these clients using the GPFS network protocols, the data appears as local (POSIX) data. Similarly, native Hadoop® Data File System (HDFS) is supported with the use of the client connector. Further, protocol nodes can be added to the cluster to support SMB and NFS file protocols, Amazon Web Services® Simple Storage Service (AWS® S3), OpenStack® Cinder block and Swift object storage access to storage without the need for IBM Spectrum Scale client software to be installed on the clients. Spectrum Scale was designed to be able to support thousands of clients, billions of files and an exabyte (EB, 10^{18} bytes) of storage in a single file system and with multiple file systems in a cluster, up to yottabyte (YB, 10^{24} bytes) scale.

Similar to Spectrum NAS, IBM Spectrum Scale supports both synchronous and asynchronous replication, but Spectrum Scale also offers multisite access to its storage via **Active File Management (AFM)**. With AFM, Spectrum Scale caches central site file data at remote sites and maintains data consistency across both central and remote sites. In this fashion, the central and remote sites can access the same file data stored on the cluster.

IBM Spectrum Scale supports multiple storage pools and provides automated policy management to migrate data from one pool to another. Each storage pool generally has similar performing storage and can be composed of block, NAS or tape library storage. Data can be automatically migrated over time to slower, less expensive storage while still being accessible to all clients across the cluster. Spectrum Scale also supports archiving to cloud storage or tape.

IBM Spectrum Scale use cases

Spectrum Scale emerged as file storage for large scientific applications running in HPC environments and remains an ideal storage solution for these environments. HPC data centers typically have high throughput requirements and run technical applications such as weather simulations, fluid dynamic models and seismic field analyses. Most of these applications require a lot of data and have to quickly move all that data into and out of racks of servers during processing. Such extreme throughput is one of the hallmarks of IBM Spectrum Scale.

Big data analytics and machine learning applications also have an insatiable need for large data repositories as well as fast load/unload of all their data. IBM Spectrum Scale is ideal for these environments because of its native support of HDFS, high throughput and large storage capacity support.

Silverton Consulting
Strategy, Storage & Systems

Furthermore, with its high throughput performance, IBM Spectrum Scale can be used as a backup target for very large data repositories. When used in this fashion, customers can benefit from Spectrum Scale's high-availability configurations with high file IOPS performance during recovery/restore operations.

IBM Cloud Object Storage

IBM COS is available both on IBM Cloud and on-premises environments. In contrast to IBM Spectrum NAS and Spectrum Scale, COS is an object storage system.

Objects are just another form of unstructured data that exist in a flat hierarchy. Objects are identified by object IDs and are accessed using a RESTful, web-like interface. COS supports its own RESTful API as well as AWS S3 and OpenStack Swift RESTful APIs.

IBM COS is also software defined and consists of a cluster of nodes. COS nodes come in three types: Manager, Accessor and Slicestor. In the IBM Cloud, COS nodes exist behind the scenes, and customers remain unaware of the number or types of COS nodes. On-premises deployments can benefit from customizing configurations to meet different performance or availability demands.

IBM COS was designed to support billions of objects spread across multiple data centers and zettabytes (ZB, 10^{21} bytes) of capacity. COS uses **Information Dispersal Algorithms (IDAs)** to distribute and protect data across sites and regions.

IBM COS IDA takes erasure coding to a multisite level. Data is sliced and then dispersed across multiple devices in multiple clusters at one or more sites. By using multisite dispersal, no one site has a complete copy of all of the object's data, and access to multiple sites is needed to read or write the object. However, IBM COS data can be dispersed within a single data center as well.

Using IDA together with IBM COS internal data integrity checking and automated self-repair capabilities, customers can configure COS data for 15 nines or more (99.9999999999999%) of data durability and up to eight nines of availability (99.999999%). We believe this level of data durability and availability is higher than any other storage system on the market today.

IBM COS use cases

Even though it can be configured to support object storage over as few as two sites and under 100TB, COS is ideal for cloud-like environments. With its multisite data slicing protection schemes, it's a natural for cloud-native applications that span regions. These applications can run in the cloud or anywhere with Internet access and use IBM COS data wherever it resides.

Silverton Consulting
Strategy, Storage & Systems

Because of the high data durability, high availability and cloud economics of IBM COS, it's also a good fit for long-term data protection of large archives. IBM COS archive data can be made resilient to any single or multiple site failure with no interruption to data access.

Finally, with its ability to store billions of objects, COS is a natural for environments with large data lake applications that need to process increasingly large amounts of data and have only modest performance requirements.

When to use IBM Spectrum NAS, Spectrum Scale and COS

Differences between the IBM unstructured data solutions above are summarized in the table below.

	Spectrum NAS	**Spectrum Scale**	**COS**
Access protocols	SMB and NFS	Natively via Scale client Via protocol server: SMB, NFS, Object and HDFS	Object
Multisite support	Replication	Replication and AFM	Information dispersal
Node types	Combined metadata and storage nodes	Metadata nodes, storage nodes, protocol nodes, client nodes and combined nodes	Manager, Accessor and Slicestor nodes
Optimal environment	User file services	Simulation and Analytic applications	Cloud-native apps and archive data
Other environments	Enterprise apps and vSphere	Machine learning, big data and high throughput backup	Large resilient content repositories, or file sync and share
Capacity range	TB to <1PB	TB to <1EB or YB for multiple file systems	TB to ZB
Typical file/object counts	Millions	Billions	Billions+
Deployment options	Onsite	Onsite or central-remote sites, cloud	Onsite or cloud
Performance	Configurable IOPS	Configurable high IOPS and throughput	Medium IOPS and high throughput
Tiering	Tier within the cluster	Tier within the cluster, to cloud, and to tape	Federation with cloud service

As the table shows, the three storage solutions have many differences, probably the most significant of which are the optimal and other environments supported.

Silverton Consulting
Strategy, Storage & Systems

Customers that need only traditional data center user file services or applications are probably best served with IBM Spectrum NAS, customers that are running simulation and analytics (HPC), machine learning or big data analytics applications are probably best served with IBM Spectrum Scale and customers that need to use cloud storage, large content repositories or large file sync and share solutions are probably best served with IBM COS.

Other characteristics to consider include capacity ranges, file/object counts and deployment options. Customers that need a TB to PB of file data probably should consider IBM Spectrum NAS. Customers that have over a million files or objects probably should consider Spectrum Scale or COS.

In choosing between Spectrum Scale and COS for object storage, Spectrum Scale should be used only when there's a need for extreme throughput or high IOPS performance, whereas COS should be used if latency and IOP requirements are more modest.

Moreover, if very large (>PB) file storage repositories, extreme throughput or multisite access to the same file data is needed, Spectrum Scale is the best file storage solution. For sites with smaller file repositories, Spectrum NAS is best.

Summary

IBM has a long history of supporting the storage and management of unstructured data with their storage systems. The latest addition, IBM Spectrum NAS, complements IBM's current file and object storage lineup and is a great addition to IBM's Spectrum Scale and IBM Cloud Object Storage solutions. However, each is best suited to particular application environments. Understanding when and where to use each storage solution is key.

Regardless, any customer using unstructured data – basically every IT, HPC and cloud customer in the world – should be able to easily take advantage of one of these three storage solutions to support all of their file and object data requirements.

Silverton Consulting
Strategy, Storage & Systems