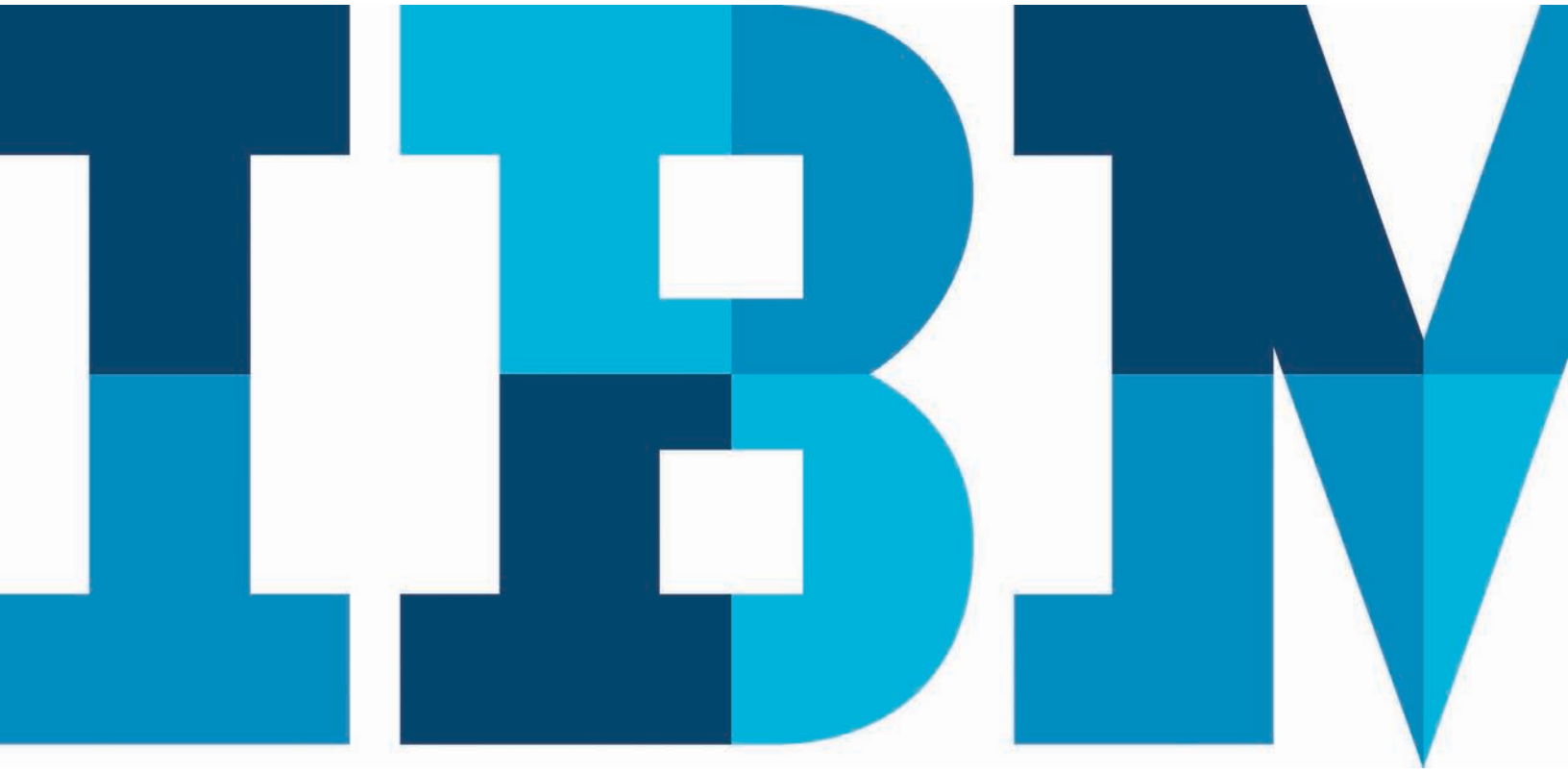
IBM Platform Computing
Thought Leadership White Paper

February 2015

An introduction to IBM Spectrum Scale

A fast, simple, scalable and complete storage solution for today's data-intensive enterprise



IBM

Contents

Introduction

Big data. Cloud storage. It doesn't matter what you call it—today's exponential growth of data, transactions and digitally aware devices is demanding larger and larger amounts of unstructured data storage.

IBM has taken on this challenge with a new software defined storage solution, IBM® Spectrum Scale™. IBM Spectrum Scale was formerly IBM General Parallel File System (IBM GPFS™), also formerly known as code name IBM Elastic Storage™. A high-performance enterprise platform for optimizing data and file management, Spectrum Scale is used extensively across industries worldwide. Spectrum Scale simplifies data management with integrated tools designed to help organizations manage petabytes of data and billions of files—as well as control the cost of managing these ever-growing data volumes.

Long considered a pioneer in big data storage, IBM leads the industry in advanced storage technologies that enable companies to store large quantities of file data.[1] The latest version of Spectrum Scale continues this tradition and marks a significant milestone in the evolution of big data management. Part of the IBM Spectrum Storage™ family, Spectrum Scale 4.1 introduces revolutionary new features that clearly demonstrate the IBM commitment to providing groundbreaking storage solutions, including:

- File encryption and secure erase
- Transparent flash cache
- Network performance monitoring
- Active File Management (AFM) parallel data transfers
- Network File System (NFS) version 4 support and data migration
- Backup and restore improvements
- File Placement Optimizer (FPO) enhancements[2]

This paper outlines the features available today in Spectrum Scale that organizations can use to manage file data. This functionality includes core Spectrum Scale concepts such as striped data storage, cluster configuration options such as direct storage access and network-based block I/O, storage automation technologies such as information lifecycle management (ILM) tools, and more.

Spectrum Scale overview

Since its introduction in 1998, Spectrum Scale (formerly GPFS) has been field-proven on some of the world's most powerful supercomputers,[3] providing highly reliable and efficient use of infrastructure bandwidth. The Mira supercomputer at Argonne National Laboratory—the fifth-fastest supercomputer in the world[4]—features a 768,000-core IBM Blue Gene® cluster with Spectrum Scale that supports scientific research, including complex modeling in the fields of material science, climatology, seismology and computational chemistry. Spectrum Scale is full-featured software defined storage with management tools for advanced storage virtualization, integrated high availability, automated tiered storage and the performance to effectively manage very large quantities of file data.

Spectrum Scale allows a group of computers concurrent access to a common set of file data over a storage area network (SAN) infrastructure, a TCP/IP network or a combination of connection types. The computers can run a mix of IBM AIX®, Linux or Microsoft Windows operating systems. Spectrum Scale provides storage management, information lifecycle management tools, centralized administration and shared access to file systems from remote Spectrum Storage clusters providing a global namespace.

A Spectrum Scale cluster can be a single node in a tiered storage solution, two nodes providing a high-availability platform supporting a database application, or thousands of nodes used for applications such as weather-pattern modeling. The largest existing configurations—like Mira—can exceed hundreds of thousands of cores.

Spectrum Scale was designed from the beginning to support high-performance parallel workloads and has since proven effective for a variety of applications. Today, it is installed in clusters supporting high-performance computing applications from climate modeling to tornado simulation, with databases such as IBM DB2®, in big data MapReduce analytics, gene sequencing, digital media and scalable file serving. These applications are used across many industries, including financial, retail, digital media, biotechnology, science and government.

Spectrum Scale continues to push technology limits in demanding large environments. You may not have multiple petabytes of data today, but chances are you will eventually. And when you do, you can be assured that Spectrum Scale has already been tested in these situations. This proven leadership is what makes Spectrum Scale a solid solution for any size application.

The file system

A Spectrum Scale file system is built from a collection of storage devices that contain the file system data and metadata. A file system can be built from a single disk or contain thousands of disks storing petabytes of data. Each file system can be accessible from all nodes within the cluster. There is no practical limit on the size of a file system. The architectural limit for a single file system is more than a yottabyte. Some Spectrum Scale customers use single file systems up to 18 PB in size, while others utilize file systems containing billions of files.

Application interfaces

Applications access files through standard POSIX file system interfaces. Since all nodes see all file data, any node in the cluster can concurrently read or update a common set of files—enabling applications to scale out easily. Spectrum Scale maintains the coherency and consistency of the file system using sophisticated byte-range locking, token (distributed lock) management and journaling. This approach means applications using standard POSIX locking semantics do not need to be modified to run successfully on Spectrum Scale.

A MapReduce connector is included with Spectrum Scale, enabling applications to access file data from IBM Platform™ Symphony or Hadoop MapReduce applications. The MapReduce connector can be used independently from the cluster architecture, giving you the flexibility to optimize for your environment.

OpenStack includes Cinder drivers for Spectrum Scale, enhancing support for virtual machine (VM) storage. The OpenStack Object Store project, known as Swift, offers cloud storage software that enables customers to store and retrieve large volumes of object data with a simple application programming interface (API). To learn more about how Spectrum Scale integrates with Swift, documented in an IBM Redbook®, visit: www.redbooks.ibm.com/abstracts/redp5113.html?Open

In addition to standard interfaces, Spectrum Scale offers a unique set of extended interfaces that can be used to provide advanced application functionality. Using these extended interfaces, an application can determine the storage pool placement of a file, create a file clone and manage quotas.

Performance and scalability

Spectrum Scale provides unparalleled I/O performance for unstructured data by:

- Striping data across multiple disks attached to multiple nodes
- Employing high-performance metadata (inode) scans
- Supporting a wide range of file system block sizes to match I/O requirements
- Utilizing advanced algorithms to improve read-ahead and write-behind I/O operations
- Using block-level locking, based on a sophisticated scalable token management system, to provide data consistency while allowing multiple application nodes concurrent access to files

When creating a Spectrum Scale file system, raw storage devices are assigned to the file system as Network Shared Disks (NSD). Once an NSD is defined, all of the nodes in the Spectrum Scale cluster can access the disk, using a local disk connection or the NSD network protocol for shipping data over a TCP/IP or InfiniBand connection.

Spectrum Scale token (distributed lock) management coordinates access to NSDs, helping ensure the consistency of file system data and metadata when different nodes access the same file. Token management responsibility is dynamically allocated among designated manager nodes in the cluster. Spectrum Scale can assign one or more nodes to act as token managers for a single file system, allowing greater scalability for large numbers of files with high transaction workloads. In the event of a node failure, token management responsibility is transparently moved to another node.

All data stored in a Spectrum Scale file system is striped across all storage devices within a storage pool—whether the pool contains 2 or 2,000 storage devices. When storage devices are added to a storage pool, existing file data can be redistributed across the new storage to improve performance. Data redistribution can be scheduled or can be done organically when there is a high data change rate. When redistributing data, a single node can be assigned to perform the task (to control the impact on a production workload). Alternately, all nodes in the cluster can participate in data movement (in order to complete the operation as quickly as possible).

Intelligent software is required to take advantage of high-performance storage. Spectrum Scale includes many I/O optimizations, including automatically recognizing typical access patterns, such as sequential, reverse sequential and random I/O.

Along with distributed token management, Spectrum Scale provides scalable metadata management by allowing all nodes of the cluster accessing the file system to perform file metadata operations. This feature distinguishes Spectrum Scale from other cluster file systems, which typically have a centralized metadata server handling fixed regions of the file namespace. A centralized metadata server can often become a performance bottleneck for metadata-intensive operations, limiting scalability and possibly introducing a single point of failure. Spectrum Scale solves this problem by enabling all nodes to manage metadata.

Administration

Spectrum Scale provides an administration model that is easy to use and consistent with standard file system administration practices, while providing extensions for the clustering aspects of Spectrum Scale. These functions support cluster management and other standard file system administration functions such as user quotas, snapshots and extended access control lists (ACLs).

Spectrum Scale administration tools simplify cluster-wide tasks. A single command can perform a file system function across the entire cluster and most can be issued from any node in the cluster. Optionally, you can designate a group of administration nodes that can be used to perform all cluster administration tasks, or authorize only a single login session to perform administrative commands cluster-wide. This approach allows for higher security by reducing the scope of node-to-node administrative access.

Rolling upgrades allow you to upgrade individual nodes in the cluster while the file system remains online. Rolling upgrades are supported between two major version levels of Spectrum Scale (and service levels within those releases). For example, you can mix GPFS Version 3.5 nodes with Spectrum Scale 4.1 nodes while migrating between releases.

Quotas enable the administrator to manage file system usage by users and groups across the cluster. Spectrum Scale provides commands to generate quota reports by user and group and on a sub-tree of a file system called a fileset. Quotas can be set on the number of files (inodes) and the total size of the files. For greater granularity in a single file system, you can define user and group per fileset quotas. In addition to traditional quota management, the policy engine can be used to query the file system metadata and generate customized space usage reports.

A Simple Network Management Protocol (SNMP) interface allows monitoring by network management applications. The SNMP agent provides information on the state of the Spectrum Scale cluster and generates traps when events occur in the cluster. For example, an event is generated when a file system is mounted or if a node fails. The SNMP agent runs on Linux and AIX. You can monitor a heterogeneous cluster as long as the agent runs on a Linux or AIX node.

Using callbacks, you can customize the response to a cluster event. A callback is an administrator-defined script that is executed by Spectrum Scale when an event occurs—for example, when a file system is unmounted or a file system is low on free space. Callbacks can be used to create custom responses to Spectrum Scale events and integrate these notifications into various cluster monitoring tools.

Spectrum Scale provides support for the Data Management API (DMAPI)—an IBM implementation of the X/Open data storage management API. This DMAPI interface allows vendors of storage management applications such as IBM Spectrum Protect™ (formerly Tivoli® Storage Manager [TSM]), IBM Spectrum Archive (formerly IBM Linear Tape File System™ [LTFS]) and IBM High Performance Storage System (HPSS) to provide Hierarchical Storage Management (HSM) support for Spectrum Scale.

Spectrum Scale supports POSIX and NFS v4 ACLs. NFS v4 ACLs can be used to serve files using NFS v4, but can also be used in other deployments—for example, to provide ACL support to nodes running Windows. To provide concurrent access from multiple operating system types, Spectrum Scale allows you to run mixed POSIX and NFS v4 permissions in a single file system and map user and group IDs between Windows and Linux/UNIX environments.

Spectrum Scale is often used as the base for a scalable NFS file service infrastructure. File systems may be exported to clients outside the cluster through NFS. The Spectrum Scale clustered NFS (cNFS) feature adds data availability to NFS clients by providing NFS service continuation if an NFS server fails. This functionality allows a Spectrum Scale cluster to deliver scalable file service by enabling simultaneous access to a common set of data from multiple nodes. The clustered NFS tools include monitoring of file services and IP address failover. Spectrum Scale cNFS supports NFS v3 and NFS v4.

Data availability

Spectrum Scale delivers a number of features that—along with a high-availability infrastructure—help ensure a reliable enterprise-class storage solution. Robust clustering features and support for synchronous and asynchronous data replication make Spectrum Scale fault-tolerant and can help provide continued data access even if cluster nodes or storage systems fail.

Spectrum Scale software includes the infrastructure to handle data consistency and availability, and does not rely on external applications for cluster operations such as node failover. In a Spectrum Scale cluster, all nodes see all data, and all cluster operations can be conducted through any node in the cluster. All nodes are capable of performing all tasks—and are not limited by who owns the data or is connected to the disks. The tasks a node can perform are determined by license type and cluster configuration.

As part of the product's built-in availability tools, Spectrum Scale continuously monitors the health of the file system components. When failures are detected, Spectrum Scale attempts automatic recovery. Version 4.1 automatically detects and frees deadlocks (when possible) and new network performance monitoring tools help to more closely monitor the environment. Extensive journaling and recovery capabilities help maintain metadata consistency when a node holding locks or performing administrative services fails.

Spectrum Scale supports snapshots, enabling you to protect the file system's contents against user error. Snapshots can be used as an online backup capability that allows files to be recovered easily from common problems such as accidental file deletion.

Snapshots preserve a point-in-time version of the file system or a sub-tree of a file system called a fileset. Spectrum Scale implements a space-efficient snapshot mechanism that generates a map of the file system or fileset at the time the snapshot is taken. New data blocks are consumed only when the file system data has been deleted or modified after the snapshot was created. This is accomplished by using a redirect-on-write technique (sometimes called copy-on-write). Snapshot data is placed in existing storage pools, simplifying administration and optimizing the use of existing storage.

Spectrum Scale 4.1 improves snapshot management by optimizing snapshot operations run concurrently on multiple snapshots within a file system, simplifying administration when there are a large number of filesets with snapshots.

Other snapshot enhancements include a new Fileset Snapshot Restore tool that restores the active fileset data and attributes to the point in time when the snapshot was taken. (For more information on backup and restore improvements in Spectrum Scale 4.1, see "What's new in Spectrum Scale Version 4.1.")

Data replication

For increased data availability and protection, Spectrum Scale offers synchronous data replication of file system metadata and data. Spectrum Scale utilizes a flexible replication model that lets you replicate a file, a set of files or an entire file system. The replication status of a file can be changed at any time using a command or a policy. Synchronous replication allows for continuous operation even if a path to a storage device, the storage device itself or even an entire site fails.

Synchronous replication is location aware, so you can optimize data access when replicas are separated across a wide area network (WAN). Spectrum Scale understands which copy of the data is "local," so read-heavy applications can achieve local data read performance even when data is replicated over a WAN. Synchronous replication works well for many workloads by replicating data across storage devices within a data center or a campus or across geographical distances using high-quality WAN connections. With the ability to use two- or three-way replication, you can choose the right level of protection for your environment.

When WAN connections are not high-performance or are unreliable, an asynchronous approach to data replication is required. For this type of environment, you can use a Spectrum Scale feature called Active File Management (AFM). AFM is a distributed disk-caching technology developed by IBM Research that allows the expansion of the Spectrum Scale global namespace across long geographical distances. It can be used to provide high availability between sites or to provide local "copies" of data distributed to one or more Spectrum Scale clusters.

Spectrum Scale 4.1 includes a number of features that optimize AFM operation. These include improved prefetch performance to prepopulate the AFM cache, support for NSD protocol as a protocol for AFM file transfers and parallel data transfers. (For more information on AFM enhancements in Spectrum Scale 4.1, see "What's new in Spectrum Scale Version 4.1.")

Further boosting cluster reliability, Spectrum Scale employs advanced clustering features designed to help maintain network connections. If a network connection to a node fails, Spectrum Scale automatically attempts to reestablish the connection before marking the node unavailable. In this way, Spectrum Scale can help provide better uptime in environments communicating across a WAN or experiencing network issues. Spectrum Scale 4.1 introduces new network performance monitoring capabilities that can help detect and troubleshoot networking issues that may affect system operation.

Spectrum Scale RAID

Larger disk drives and file systems create considerable challenges for traditional storage controllers. Current RAID-5 and RAID-6–based arrays cannot address the demands of exabyte-scale storage performance, reliability and management. To meet these needs, Spectrum Scale RAID brings parity-based data protection into software, alleviating the need to rely on hardware RAID controllers. The storage devices can be individual disk drives or any other block device, eliminating the need for a storage controller.

Spectrum Scale RAID employs a de-clustered approach to RAID. This de-clustered architecture reduces the impact of drive failures by spreading data over all of the available storage devices, improving application I/O and storage recovery performance. Spectrum Scale RAID delivers high reliability through an 8+2 or 8+3 Reed-Solomon–based RAID code that divides each block of a file into eight parts and associated parity. This algorithm scales easily, starting with as few as 11 storage devices and growing to over 500 per storage pod. Spreading the data over many devices provides more predictable storage performance and fast recovery times measured in minutes rather than hours in the case of a device failure. In addition to performance improvements, Spectrum Scale RAID provides advanced checksum protection to ensure data integrity. Checksum information is stored on disk and verified all the way to the Spectrum Scale client.

This RAID software is delivered as part of the IBM Elastic Storage Server (ESS) system. For more information, visit: **ibm.com**/systems/storage/spectrum/ess

Information lifecycle management (ILM) toolset

Spectrum Scale can help you achieve data lifecycle management efficiencies through policy-driven automation and tiered storage management. The use of storage pools, filesets and user-defined policies allow you to better match the cost of your storage to the value of your data.

Storage pools are used to manage groups of storage devices within a file system. Using storage pools, you can create tiers of storage by grouping storage devices based on performance, locality or reliability characteristics. For example, one pool could contain high-performance solid-state devices and another, more economical 7,200 rpm disk storage. Pools created using direct access storage media are called internal storage pools. When data is placed in or moved between internal storage pools, all data management is carried out by Spectrum Scale without affecting the namespace or user access to the file.

In addition to internal storage pools, Spectrum Scale supports external storage pools. External storage pools are used to interact with an external storage management application, including Spectrum Protect, Spectrum Archive and HPSS. When moving data to an external pool, Spectrum Scale handles all the metadata processing, and then hands the data to the external application for storage on alternate media (such as tape). When using Spectrum Protect, Spectrum Archive or HPSS, data can be retrieved from the external storage pool on demand (as a result of an application opening a file), or retrieved in a batch operation using a command or policy.

A fileset is a sub-tree of the file system namespace and provides a way to partition the namespace into smaller, more manageable units. Filesets provide an administrative boundary that can be used to set quotas, take snapshots, define AFM relationships and function in user-defined policies to control initial data placement or data migration. Data within a single fileset can reside in one or more storage pools. Where the file data resides and how it is managed once it is created is based on a set of rules in a user-defined policy.

There are two types of user-defined policies in Spectrum Scale: file placement and file management. File placement policies determine the storage pool into which file data is initially placed. File placement rules are defined using attributes of a file known when a file is created, such as file name, fileset or the user who is creating the file. For example, a placement policy may be defined that states:

'*Place all files with names that end in .mov onto the near-line SAS-based storage pool* and *place all files created by the CEO onto the solid-state drive-based storage pool*'

or

'*Place all files in the fileset 'development' onto the SAS-based storage pool*'

Once files exist in a file system, file management policies can be used for file migration, deletion, changing file replication status or generating reports.

You can use a migration policy to transparently move data from one storage pool to another without changing the file's location in the directory structure. Similarly, you can use a policy to change the replication status of a file or set of files, allowing fine-grained control over the space used for data availability.

You can use migration and replication policies together. For example, a policy may say:

'*Migrate all of the files located in the subdirectory/database/payroll which end in *.dat and are greater than 1 MB in size to storage pool #2 and un-replicate these files*'

File deletion policies allow you to prune the file system, deleting files as defined by policy rules. Reporting on the contents of a file system can be accomplished through list policies. List policies let you quickly scan the file system metadata and produce information listing selected attributes of candidate files.

File management policies can be based on more file attributes than placement policies because once a file exists there is more known about it. For example, file placement policies can utilize attributes such as last access time or size of the file. This approach may result in policies such as:

'*Delete all files with a name ending in .temp that have not been accessed in the last 30 days*'

or

'*Migrate all files owned by Sally that are larger than 4 GB to the high-density storage pool*'

Rule processing can be further automated by including attributes related to a storage pool instead of a file using the threshold option. Using thresholds, you can create a rule that moves files out of the high-performance pool if it is more than 80 percent full, for example. The threshold option comes with the ability to set high, low and pre-migrate thresholds. Pre-migrated files exist on disk and tape at the same time. This method is typically used to enable disk access to the data and allow disk space to be freed up quickly when a maximum space threshold is reached.

Spectrum Scale begins migrating data when the high threshold is reached and continues until the low threshold is reached. If a pre-migrate threshold is set, Spectrum Scale copies data to tape until the pre-migrate threshold is reached. This method permits continued access to data in the internal pool until it is quickly deleted to free up space the next time the high threshold is reached. Thresholds let you fully utilize your highest-performance storage and automate the task of making room for new high-priority content.

Policy rule syntax is based on the SQL 92 syntax standard and supports multiple complex statements in a single rule to enable powerful policies. Multiple levels of rules can be applied to a file system, and rules are evaluated in order for each file when the policy engine executes, allowing a high level of flexibility.

Spectrum Scale provides unique functionality through standard interfaces such as extended attributes. Extended attributes are a standard POSIX facility. Spectrum Scale includes enhanced support for POSIX extended attributes. In Spectrum Scale, extended attributes are accessible by the high-performance policy engine, allowing you to write rules that use custom file attributes.

Policy-based storage management is not practical without a method to efficiently query the file metadata. Spectrum Scale includes a high-performance metadata scan interface that lets you query the metadata for billions of files in a matter of minutes,[5] making the Spectrum Scale ILM toolset a very scalable way to automate file management. This high-performance metadata scan engine employs a scale-out approach. The identification of candidate files and data movement operations can be performed concurrently by one or more nodes in the cluster. Spectrum Scale can spread rule evaluation and data movement responsibilities over multiple nodes in the cluster to provide a scalable, high-performance rule processing engine.

Big data support

For organizations managing big data workloads, Spectrum Scale File Placement Optimizer (FPO) delivers a set of features that extend Spectrum Scale to work seamlessly in the Hadoop ecosystem. Further enhanced in Spectrum Scale 4.1, Spectrum Scale FPO offers an enterprise-class alternative to the Hadoop Distributed File System (HDFS) for building big data platforms. With Spectrum Scale FPO, you get all the functionality of a traditional file system with additional features designed to support MapReduce and other shared-nothing workloads.

Spectrum Scale FPO is an implementation of a shared-nothing storage architecture that enables each node to operate independently, reducing the impact of failure events across multiple nodes. FPO extends the core Spectrum Scale architecture, providing greater control and the flexibility to leverage data location, reduce hardware costs and improve I/O performance.

Originally developed to support MapReduce workloads, FPO features provide tools to support data locality, shared-nothing storage management and interfacing with Hadoop MapReduce. FPO provides a platform for MapReduce while maintaining full POSIX compliance so you do not need to change the way you edit and manage file data to run MapReduce workloads.

Spectrum Scale 4.1 takes the integration of traditional data storage and shared-nothing a step further, simplifying storage management by allowing you to create an FPO-enabled storage pool. Now, you can add a storage pool for MapReduce and use the Spectrum Scale policy tools to automatically migrate data between traditional shared disk storage and shared-nothing pools.

The I/O performance benefits of MapReduce come from taking advantage of data locality, keeping I/O access within the server instead of sending data across a network. FPO data locality features include the ability to create "chunks" of data you can use to place larger continuous regions of file data within a node for better locality.

In addition to locality, shared-nothing clusters require enhanced failure detection and recovery features to operate efficiently. In Spectrum Scale 4.1, failure recovery is enhanced to provide better data recovery performance. FPO includes a Hadoop MapReduce connector so it can easily communicate with Spectrum Scale.

Storing your data in Spectrum Scale FPO frees you from the architectural restrictions of HDFS. Spectrum Scale FPO provides Hadoop compatibility extensions to replace HDFS in a Hadoop ecosystem, with no changes required to your Hadoop applications.

Cluster configurations

Spectrum Scale is software defined storage, so it supports a variety of hardware configurations—independent of the file system features you use—freeing you to select the hardware that best matches your application requirements. Hardware configuration options can be characterized into three basic categories:

- Shared disk
- Network-based block I/O
- Mixed clusters

Shared disk

A shared disk (SD) architecture is the most basic environment. In this configuration, the storage is directly attached to all machines in the cluster, as shown in Figure 1. With a direct connection to the storage, each shared block device is available concurrently to all the nodes in the Spectrum Scale cluster. Direct access means that the storage is accessible using a SCSI or other block-level protocol using a SAN, InfiniBand, iSCSI, virtual I/O interface or other block-level I/O connection technology.

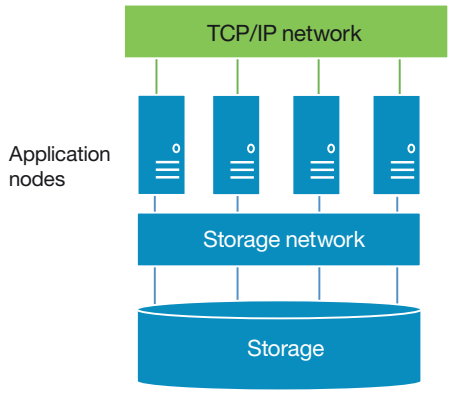


Figure 1. SAN-attached storage.

Figure 1 illustrates a Spectrum Scale cluster where all nodes are connected to a common Fibre Channel SAN, although the storage attachment technology could be InfiniBand, Serial Attached SCSI (SAS), Fibre Channel over Ethernet (FCoE) or any other.

The nodes are connected to the storage using the SAN and to each other using a local area network (LAN). Data used by applications running on the Spectrum Scale nodes flows over the SAN, and Spectrum Scale control information flows among the Spectrum Scale instances over a LAN.

This configuration is optimal when all nodes in the cluster need the highest-performance access to the data. For example, this is a good configuration for providing network file service to client systems using clustered NFS, high-speed data access for digital media applications or a grid infrastructure for data analytics.

Network-based block I/O

As data storage requirements increase and new storage and connection technologies are introduced, a common SAN may not be a sufficient or appropriate choice of storage connection technology. In environments where every node in the cluster is not attached to a SAN, Spectrum Scale makes use of an integrated network block-device capability called the Network Shared Disk (NSD) protocol. Whether using the NSD protocol or a direct attachment to a SAN, the mounted file system looks the same to the application, and Spectrum Scale transparently handles I/O requests.

Spectrum Scale clusters can use the NSD protocol to provide high-speed data access to applications running on TCP/IP-attached nodes. Data is served to these client nodes from one or more NSD servers. In this configuration, disks are attached only to the NSD servers. Each NSD server is attached to all or a portion of the disk collection. With Spectrum Scale, you can define up to eight NSD servers per disk; it is recommended that at least two NSD servers be defined for each disk to avoid a single point of failure.

The NSD protocol operates over any TCP/IP-capable network fabric. On Linux, Spectrum Scale can use the VERBS RDMA protocol on compatible fabrics (such as InfiniBand) to transfer data to NSD clients. The network fabric does not need to be dedicated to Spectrum Scale, but should provide sufficient bandwidth to meet performance expectations for both Spectrum Scale and applications sharing the bandwidth.

Spectrum Scale can concurrently use multiple networks for NSD protocol communication. You can designate separate IP or InfiniBand networks to communicate concurrently with a common set of file data. This approach provides interaction between new and old systems and allows greater flexibility when you need to increase throughput or add nodes to a Spectrum Scale cluster.

Using multiple subnets means that not all of the NSD clients need to be on a single physical network. For example, you can place groups of clients onto separate subnets that access a common set of disks through different NSD servers, so not all NSD servers need to serve all clients. This approach can reduce networking hardware costs and simplify the topology, ultimately reducing support costs, providing greater scalability and improving overall performance.

An example of the NSD server model is shown in Figure 2.

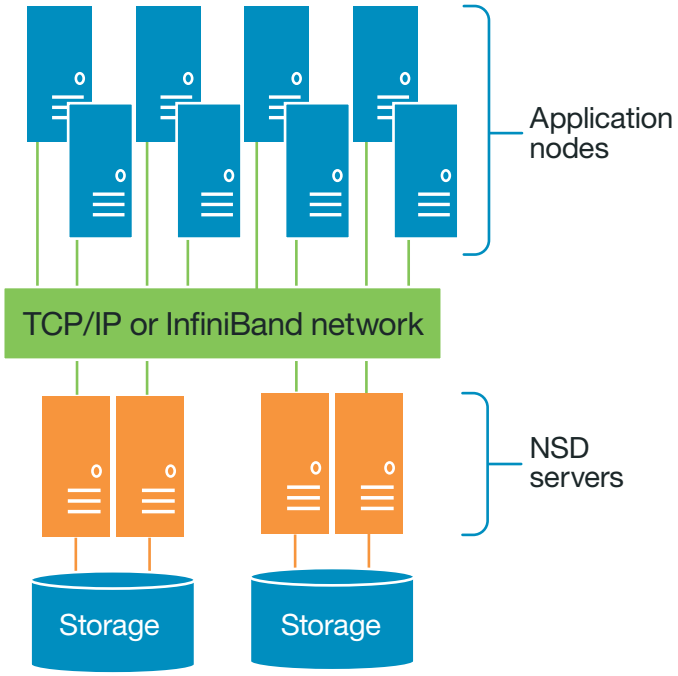


Figure 2. Network block I/O.

In this configuration, some of the nodes are defined as NSD server nodes. The NSD server is responsible for the abstraction of disk data blocks across a TCP/IP or InfiniBand VERBS (Linux only)-based network. The fact that the disks are remote is transparent to the application.

Figure 2 shows an example of a configuration where a set of compute nodes are connected to a set of NSD servers using a network such as Ethernet. In this example, data to the NSD servers flows over the SAN and both data and control information to the clients flow across the LAN. Since the NSD servers serve data blocks from one or more devices, data access is similar to a SAN-attached environment in that data flows from all servers simultaneously to each client. This parallel data access provides the best possible throughput to all clients and has the ability to scale up the throughput to a common data set or even a single file.

The choice of how many nodes to configure as NSD servers is based on performance requirements, network architecture and the capabilities of the storage subsystems. High-bandwidth LAN connections should be used for clusters requiring significant data transfer rates and can include 1 Gbit, 10 Gbit or 40 Gbit Ethernet. For additional performance or reliability, you can use link aggregation (EtherChannel or bonding), networking technologies such as source-based routing or higher-performance networks such as InfiniBand.

The choice between SAN attachment and network block I/O is a performance and economic one. In general, using a SAN provides the highest performance for smaller clusters, but the cost and management complexity of using a SAN for large clusters is often prohibitive. In these cases, network block I/O provides an option.

Network block I/O is well suited to grid computing, where there is sufficient network bandwidth between the NSD servers and the clients. For example, an NSD protocol-based grid is effective for web applications, supply chain management or modeling weather patterns.

Mixed clusters

The last two sections discussed shared disk and network attached Spectrum Scale cluster topologies. You can mix these storage attachment methods within a Spectrum Scale cluster to better match the I/O requirements to the connection technology (see Figure 3).

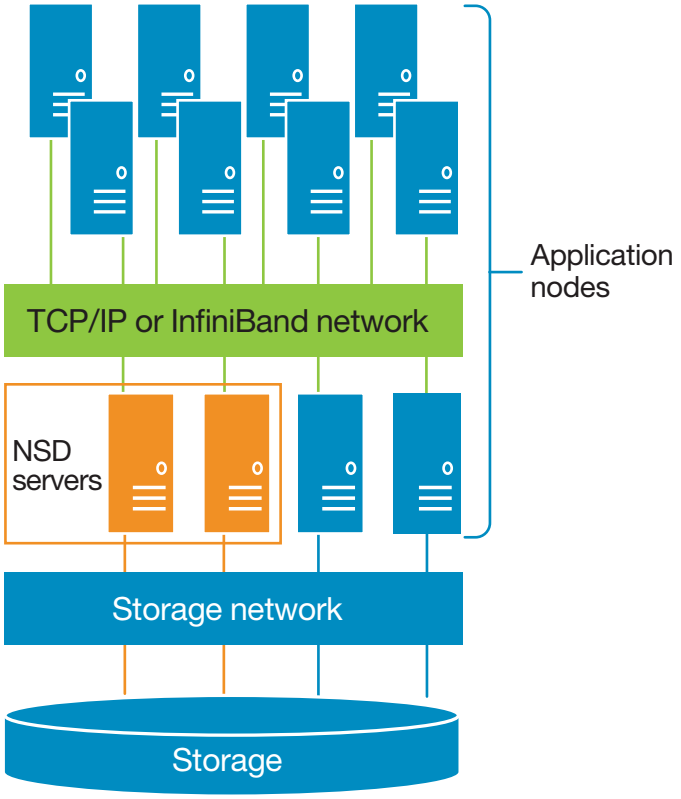


Figure 3. Mixed-cluster architecture.

A Spectrum Scale node always tries to use the most efficient path to the storage. If a node detects a block device path to the data (such as Fibre Channel SAN), it is used. If there is no block device path, then the network is used. This capability can be leveraged to provide additional availability. If a node is SAN-attached to the storage and there is a host bus adapter (HBA) failure, for example, Spectrum Scale can fail over to the network path to the storage. A mixed-cluster topology can provide direct storage access to non-NSD server nodes for high-performance operations, including backups or data ingest.

Sharing data between clusters

There are two methods available to share data across Spectrum Scale clusters: Multicluster and AFM.

Multicluster lets you use the NSD protocol to share data across clusters. With this feature, you can allow clusters to mount file systems that belong to other Spectrum Scale clusters. A multicluster environment allows the administrator to permit access to specific file systems from another Spectrum Scale cluster. This feature is intended to allow clusters to share data at higher performance levels than file-sharing technologies such as NFS or Common Internet File System (CIFS). It is not intended to replace such file-sharing technologies, which are optimized for desktop access or for access across unreliable network links.

Spectrum Scale clusters are typically connected using multicluster to ease administration by separating cluster roles. It is common to have a storage cluster that contains all of the file systems and one or more application clusters sharing one or

more file systems from the storage cluster. Applications clusters often do not include storage, and use only the capacity in the storage cluster.

Multicluster is useful for sharing across clusters within a physical location or across locations. When the remote clusters need access to the data, they mount the file system by contacting the owning cluster and passing required security checks. Data access takes place using the NSD protocol. Once authentication is complete, data consistency and performance accessing the data from a remote cluster is the same as if the node were part of the storage cluster. Multicluster environments are well suited to sharing data across clusters belonging to different organizations for collaborative computing, grouping sets of clients for administrative purposes or implementing a global namespace across locations.

A multicluster configuration allows you to connect Spectrum Scale clusters within a data center, across campus or across reliable WAN links. For sharing data between Spectrum Scale clusters across less-reliable WAN links, or in cases where you want a copy of the data in multiple locations, you can use AFM.

AFM lets you create associations among Spectrum Scale clusters, and automate the location and flow of file data between them. Relationships between Spectrum Scale clusters using AFM are defined at the fileset level. A fileset in a file system can be created as a "cache" that provides a view to a file system in another Spectrum Scale cluster called the home (or target). The term cache may imply transient or temporary data, but in AFM a cached file is the same as any other file in a Spectrum Scale file system. The difference is that the file data is kept in sync with a copy in another file system. File data is moved into a cache fileset on demand. When a file is read, the file data is copied from the home into the cache fileset. Data consistency and file movement into and out of the cache is managed automatically by Spectrum Scale.

Cache filesets can be read-only or writeable. Cached data is locally written and locally read when the data is available. On read, if the data is not in the cache, Spectrum Scale automatically copies the data from the home file system. When data is written into the cache, the write operation completes locally, and then one or more nodes called a *gateway* asynchronously push the changes back to the home. You can define multiple cache filesets for each home data source. The number of cache relationships for each home is limited only by the bandwidth available at the home location. If you want to keep only active files in a read-only cache, you can place a quota on the read-only cache fileset. Placing a quota on a read-only cache fileset causes the data to be cleaned (evicted) out of the cache automatically based on the space available. If a quota is not set, a copy of the file data remains in the cache until it is manually evicted from the cache or deleted at home.

AFM is designed to enable efficient data transfers over WAN connections or high-performance transfers over local network connections. When file data is read from home into cache, the transfer can take place in parallel within a gateway or across multiple gateway nodes.

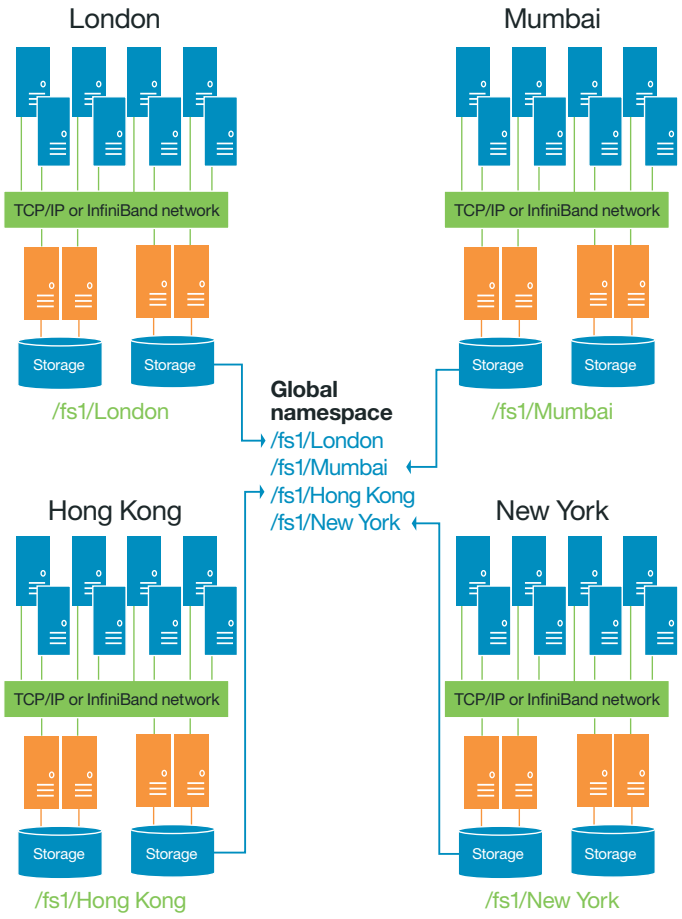


Figure 4. Global namespace using AFM.

Using AFM, you can create a truly worldwide global namespace within a data center, across a campus or between data centers located around the world.

Figure 4 is an example of a global namespace built using AFM. In this example, each location owns one-quarter of the namespace. The other sites use cache filesets that point to the home location to access the data owned by the other clusters. In this example, the data in /fs1/London originates and is mostly used in the London location. Other locations that need occasional access to the London data have cache filesets that point to London as the home. Using the same methodology across all sites provides the same namespace within each Spectrum Scale cluster, giving applications the same path to a file across the entire organization.

This type of global namespace can be achieved using either multicluster or AFM, depending on the quality of the WAN link and the application requirements.

What's new in Spectrum Scale Version 4.1

File encryption and secure erase

Spectrum Scale 4.1 introduces file-level encryption for the protection of data at rest, and secure data deletion. Spectrum Scale file encryption can help protect data from security breaches or unauthorized access, and from being stolen or improperly discarded. Encryption is controlled by policy, and data is not decrypted until it reaches the application node. Each application node can have the same encryption key or a different key, allowing for secure scale-out processing or multi-tenancy on shared storage hardware.

Without encryption, when a file is deleted the space is marked as free, but the data is not overwritten until a new file occupies the same physical storage location. Secure erase provides a fast, simple way to not only delete a file but also make the free space unreadable and therefore securely erased. Spectrum Scale 4.1 complies with publications NIST SP 800-131A, "Recommended Security Controls for Federal Information Systems and Organizations," and FIPS 140-2, "Security Requirements for Cryptographic Modules."

Transparent flash cache

Many applications can benefit from a large amount of local file cache. Local data cache can give applications low-latency access to file data, and reduce the load on shared network and back-end storage. Using system memory for file cache is expensive and has limited capacity compared to persistent data storage devices. Solid-state disks or flash provide an economical way to expand the capacity of local data cache, but using them for file cache requires intelligent software.

Spectrum Scale flash cache addresses the need for expanded local file cache by taking advantage of solid-state disks or flash placed directly in client nodes. The solid-state disks are seamlessly integrated as an extension of the Spectrum Scale file cache. Flash cache transparently acts as an extension of the Spectrum Scale file cache, called the *pagepool*. In this way, Spectrum Scale flash cache can cost-effectively accelerate applications and deliver more than a terabyte of local data cache capacity.

Network performance monitoring

Because software defined storage relies on the underlying server, network and storage hardware, it is critical to keep this infrastructure running at peak performance. To help monitor network performance, Spectrum Scale 4.1 now provides Remote Procedure Call (RPC) latency statistics to help detect and troubleshoot networking issues that may affect Spectrum Scale operation.

AFM enhancements

Spectrum Scale 4.1 contains a number of features that optimize AFM operation and make it easier to move data. While data transfers into an AFM cache are triggered on demand, it is possible for AFM to prefetch (prepopulate) data—so the necessary file data is waiting in the cache before it is needed. Spectrum Scale 4.1 speeds data prefetching, making it easier to prepopulate a greater number of files in a shorter period of time. In this way, prefetch improves the performance of large file transfers.

To move more data more quickly between the cache and home, Spectrum Scale 4.1 also adds parallel data movement within a single fileset. Now, you can move data in parallel between cache and home for a single large file or many smaller files.

To further improve data transfer speed across high-performance networks, AFM now supports the use of the NSD protocol, in addition to NFS. This addition makes it easier to use AFM to leverage flash devices to accelerate applications and move data between Spectrum Scale clusters.

NFS data migration

With Spectrum Scale 4.1, you can now use AFM to migrate data from one cluster to another when upgrading hardware or buying a new system. Because AFM utilizes the NFS protocol, any NFS data source may serve as home—it does not have to be Spectrum Scale. Spectrum Scale data migration can minimize downtime for applications, move file data along with permissions associated with files, and consolidate data from multiple legacy systems into a single, more powerful system.

Backup and restore improvements

In Spectrum Scale 4.1, fileset snapshots can now be restored into the active file system. The tool restores the active fileset data and attributes to the point in time when the snapshot was taken. The *mmbackup* utility has been enhanced to automatically adjust its own work distribution and parallel access to Spectrum Protect based on resource availability and user-provided input parameters.

FPO enhancements

File Placement Optimizer (FPO) enhancements include improved data recovery with the addition of advanced data locality–aware file system recovery. To better support MapReduce and other big data workloads, Spectrum Scale 4.1 also boosts the performance of concurrent directory changes.

Summary

Spectrum Scale (formerly GPFS) was designed from the beginning for high performance, and since its release in 1998 has achieved unmatched scalability, performance and field-proven reliability. The maturity, ease of use and flexibility of Spectrum Scale is demonstrated by the fact that it is used by over 3,000 enterprises in industries ranging from finance to life sciences to automotive design. Spectrum Scale continues to be highly effective in the most demanding data-intensive applications.

Spectrum Scale is designed to help organizations tackle emerging file storage challenges, and Spectrum Scale 4.1 provides a whole new set of tools to help customers address them. Accelerated I/O performance from Spectrum Scale flash cache and enhanced AFM functionality continue to keep pace with the latest workloads, and advanced file encryption and secure erase functionality make it easier for companies to meet strict data compliance standards. IBM is committed to ongoing expansion and enhancement of Spectrum Scale performance and features—and to its continued leadership in scalable, innovative data management solutions.

For more information

To learn more about IBM Spectrum Scale, please contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/systems/storage/spectrum/scale

For more information about the IBM Spectrum Storage family of software defined storage solutions, visit: **ibm.com**/systems/storage/spectrum

Additionally, IBM Global Financing can help you acquire the IT solutions that your business needs in the most cost-effective and strategic way possible. We'll partner with credit-qualified clients to customize an IT financing solution to suit your business goals, enable effective cash management, and improve your total cost of ownership. IBM Global Financing is your smartest choice to fund critical IT investments and propel your business forward. For more information, visit: **ibm.com**/financing

IBM Systems
Route 100
Somers, NY 10589

Produced in the United States of America
February 2015

Actual available storage capacity may be reported for both uncompressed and compressed data and will vary and may be less than stated.

[1] **ibm.com**/systems/technicalcomputing/sc13.html

[2] For detailed information on these features, see pages 10 and 18.

[3] See the top 100 list. Source: Top 500 Super Computer Sites: www.top500.org

[4] www.top500.org/list/2012/06/100/

[5] http://domino.watson.ibm.com/library/CyberDig.nsf/papers/4A50C2D66A1F90F7852578E3005A2034

Please Recycle

DCW03057-USEN-01