

# IBM Spectrum Virtualize and IBM Spectrum Scale in an Enhanced Stretched Cluster Implementation

Angelo Bernasconi

Cristiano Beretta

Walter Bernocchi

Giorgio Richelli



Storage





# IBM Spectrum Virtualize and IBM Spectrum Scale in an Enhanced Stretched Cluster Implementation

Business continuity and continuous application availability are among the top requirements for many organizations. Advances in virtualization, storage, and networking have made enhanced business continuity possible. Information technology solutions can now manage both planned and unplanned outages, and provide the flexibility and cost efficiencies that are available from cloud-computing models.

This IBM® Redpaper™ publication describes the following products and topics:

- ▶ IBM Spectrum™ Virtualize, which is the software that is at the core of the IBM SAN Volume Controller
- ▶ IBM Spectrum Scale, which was previously known as IBM General Parallel File System (GPFS™) or IBM Elastic Storage™
- ▶ IBM Spectrum Virtualize (SAN Volume Controller component) and IBM Spectrum Scale, which are together in an Enhanced Stretched Cluster (ESC)
- ▶ An example implementation
- ▶ Test results
- ▶ Preferred practices when using IBM Spectrum Virtualize and IBM Spectrum Scale together in an Enhanced Stretched Cluster

This paper is aimed at technical professionals who are familiar with IBM Spectrum Virtualize (either SAN Volume Controller or the IBM Storwize® family), IBM Spectrum Scale (previously known as IBM General Parallel File System (GPFS) or IBM Elastic Storage), and SAN Volume Controller Enhanced Stretched Cluster (ESC). If you are unfamiliar with these products, see *Implementing the IBM System Storage SAN Volume Controller V7.4*, SG24-7933 and *IBM SAN Volume Controller Enhanced Stretched Cluster with VMware*, SG24-8211.

# IBM Spectrum Virtualize

The IBM Spectrum Virtualize industry-leading storage virtualization enhances existing storage to improve resource utilization and productivity to achieve a simpler, more scalable, and cost-efficient IT infrastructure.

SAN Volume Controller delivers the functions of IBM Spectrum Virtualize, part of the IBM Spectrum Storage™ family, and has been improving infrastructure flexibility and data economics for more than 10 years. Its innovative data virtualization capabilities provide the foundation for the entire IBM Storwize family. SAN Volume Controller provides the latest storage technologies for unlocking the business value of stored data, including virtualization and IBM Real-time Compression™. In addition, the system includes the new SAN Volume Controller Data Engine to help support the massive volumes of data that are created by today's demanding enterprise applications. SAN Volume Controller delivers unprecedented levels of efficiency, ease of use, and dependability for organizations of all sizes.

SAN Volume Controller is a storage virtualization system that enables a single point of control for storage resources. It helps support improved business application availability and greater resource use. The objective is to manage storage resources in your IT infrastructure and to ensure that they are used to the advantage of your business. These processes are done quickly, efficiently, and in real time, while also avoiding increases in administrative costs.

SAN Volume Controller combines hardware and software into an integrated, modular solution that is highly scalable. An I/O Group is formed by combining a redundant pair of storage engines that are based on System x server technology. Highly available I/O Groups are the basic configuration element of a SAN Volume Controller cluster.

SAN Volume Controller configuration flexibility means that your implementation can start small and then grow with your business to manage very large storage environments.

SAN Volume Controller helps increase the amount of storage capacity that is available to host applications. It does so by pooling the capacity from multiple disk systems within the SAN.

In addition, SAN Volume Controller combines various IBM technologies that include thin provisioning, automated tiering, storage virtualization, Real-time Compression, clustering, replication, multiprotocol support, and a next-generation graphical user interface (GUI). Together, these technologies enable SAN Volume Controller to deliver extraordinary levels of storage efficiency.

Because it hides the physical characteristics of storage from host systems, SAN Volume Controller help applications continue to run without disruption while you change your storage infrastructure. This advantage helps your business increase its availability to customers.

For more information about SAN Volume Controller, see <http://www-03.ibm.com/systems/storage/software/virtualization/svc/index.html> and *Implementing the IBM System Storage SAN Volume Controller V7.4*, SG24-7933.

Figure 1 shows the SAN Volume Controller structure and components.

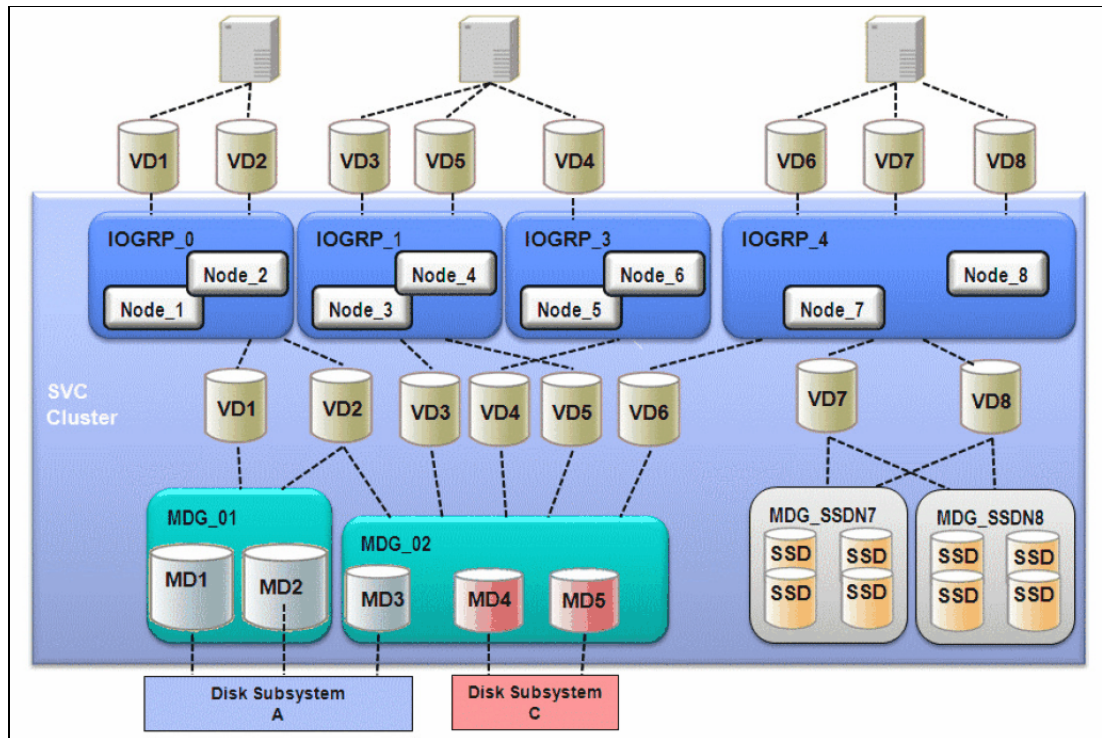


Figure 1 SAN Volume Controller structure

## SAN Volume Controller Enhanced Stretched Cluster solution

When SAN Volume Controller was first introduced, the maximum supported distance between nodes within an I/O Group was 100 m.

SAN Volume Controller V5.1 introduced support for the Stretched Cluster configuration, where nodes within an I/O Group can be separated by a distance of up to 10 km.

With Version 6.3, which was released in October 2011, SAN Volume Controller began supporting Stretched Cluster configurations where nodes can be separated by a distance of up to 300 km in specific configurations.

With Stretched Cluster, the two nodes in an I/O Group are separated by distance between two locations. A copy of the volume is stored at each location. This configuration means that you can lose either the SAN or power at one location and access to the disks remains available at the alternative location. Using this configuration requires clustering software at the application and server layer to fail over to a server at the alternative location and resume access to the disks. The SAN Volume Controller keeps both copies of the storage in synchronization, and the cache is mirrored between both nodes. Therefore, the loss of one location causes no disruption to the alternative location.

As with any clustering solution, avoiding a split-brain situation (where nodes no longer can communicate with each other) requires a tie-break mechanism. SAN Volume Controller is no exception. The SAN Volume Controller uses a tie-break mechanism that is facilitated through the implementation of a quorum disk. The SAN Volume Controller uses three quorum disks from the Managed Disks that are attached to the cluster to be used for this purpose. Usually the management of the quorum disks is transparent to the SAN Volume Controller users. However, in an Enhanced Stretched Cluster (ESC) configuration, the location of the quorum disks must be assigned manually to ensure that the active quorum disk is in a third location. This configuration must be done to ensure the survival of one location if a failure occurs at another location.

The links between fabrics at either site have certain requirements that must be validated.

For more information about Enhanced Stretched Cluster prerequisites, see *IBM SAN Volume Controller Enhanced Stretched Cluster with VMware*, SG24-8211 or the following website:

[http://www-01.ibm.com/support/knowledgecenter/STPVGU\\_7.2.0/com.ibm.storage.svc.sole.720.doc/svc\\_stretchedclusteroverview.html?lang=en](http://www-01.ibm.com/support/knowledgecenter/STPVGU_7.2.0/com.ibm.storage.svc.sole.720.doc/svc_stretchedclusteroverview.html?lang=en)

The prerequisites at the time of writing were introduced with SAN Volume Controller V7.2 and are still valid for SAN Volume Controller V7.4, which was used in this paper.

SAN Volume Controller is a flexible solution. You can use the storage controller of your choice at any of the three locations, and with SAN Volume Controller they can be from different vendors. Also, this is all possible by using the base SAN Volume Controller virtualization license with no additional charge.

SAN Volume Controller uses two major I/O functions that were introduced beginning with Version 4.3: thin provisioning (Space-Efficient Virtual Disks (SEV)), and Virtual Disk Mirroring (VDM). VDM is a mechanism by which a single volume has two physical copies of the data on two independent Managed Disk Groups (storage pools and storage controllers). This feature provides these capabilities:

- ▶ The ability to change the extent size of a volume.
- ▶ A way to migrate between storage controllers, or split off a copy of a volume for development or test purposes.
- ▶ A method to increase redundancy and reliability of lower-cost storage controllers.
- ▶ A temporary mechanism to add a second copy to a set of volumes to enable disruptive maintenance to be run to a storage controller without any loss of access to servers and applications.

Another capability that is provided by VDM is the ability to *split* the cluster while still maintaining access to clustered servers and applications.

For example, imagine that you have two servers that act as a cluster for an application. These two servers are in different rooms and power domains, and are attached to different fabrics. You also have two storage controllers, one in each room. You want to mirror data between the controllers, and at the same time provide access to users when you lose power, or access to disks within one of the rooms. This process can now be done through the implementation of the SAN Volume Controller Enhanced Stretched Cluster configuration.

The solution in this paper focuses on the SAN Volume Controller and VMware environment. However, the SAN Volume Controller Enhanced Stretched Cluster configuration can be applied to any other operating system and environment. These systems include a native Microsoft cluster, IBM AIX® Power HA, IBM PowerHA® System Mirror for iSeries, and a Linux cluster.

All the Enhanced Stretched Cluster benefits and protection criteria apply, and use data protection and business continuity requirements regardless of the operating system your application is using.

For more information about the interoperability of the SAN Volume Controller Enhanced Stretched Cluster configuration, see the following website:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004946>

## Understanding SAN Volume Controller quorum disk

The quorum disk fulfills two functions for cluster reliability:

- ▶ Acts as a tie-breaker in split-brain scenarios.
- ▶ Saves critical configuration metadata.

The SAN Volume Controller quorum algorithm distinguishes between the active quorum disk and quorum disk candidates. There are three quorum disk candidates. At any time, only one of these candidates acts as the active quorum disk. The other two are reserved to become active if the current active quorum disk fails. All three quorum disks are used to store configuration metadata, but only the active quorum disk acts as tie-breaker for split-brain scenarios.

**Requirement:** A quorum disk must be placed in each of the three failure domains. Set the quorum disk in the third failure domain as the active quorum disk.

The following items should be considered when you use a SAN Volume Controller quorum disk:

- ▶ If the DR feature is disabled, the quorum selection algorithm operates as it would with SAN Volume Controller V7.1 and prior versions.
- ▶ When the DR feature is enabled and automatic quorum disk selection is also enabled, three quorum disks are created, one in each site, that is, in sites 1, 2, and 3.
- ▶ If a site has no suitable MDisks, then less than three quorum disks are automatically created. For example, if SAN Volume Controller can create only two quorum disks, then only two are used.
- ▶ If a user is controlling the quorum by using the **chquorum** command, then the choice of quorum disk that the user selects must also follow the one disk per site rule.
- ▶ If a user uses **chquorum** to assign manually quorum disks and configures the topology as stretched, then SAN Volume Controller ignores any quorum disk that is not assigned to a site. SAN Volume Controller chooses only quorum disks that are configured to site 3 as the active quorum disk and chooses only quorum disks that are configured to site 1 or 2 as stand-by quorum disks.
- ▶ If a user does not have a quorum disk that is configured at each site, then the user may restrict when, or if, T3 recovery procedure is possible, and how resilient the cluster is to site failures. Without access to a quorum disk, SAN Volume Controller cannot continue I/O when one copy of a mirrored volume goes offline.

**Note:** For stretched clusters that are implemented with the DR feature enabled, the recommendation is to configure manually quorum devices to track which MDisk is chosen, and to select the MDisks you want to be your quorum disks.

## SAN Volume Controller cluster state and voting

The cluster state information on the active quorum disk is used to decide which SAN Volume Controller nodes survive if exactly half the nodes in the cluster fail at the same time. Each node has one vote, and the quorum disk has half the votes for determining cluster quorum.

The SAN Volume Controller cluster manager implements a dynamic quorum, which means that following a loss of nodes, if the cluster can continue operating, it dynamically alters the voting set that defines which nodes must be present to allow more node failures to be tolerated. In this way, the voting set is continually updated to match the set of nodes that is present. This process enables servicing of the cluster.

The cluster manager determines the dynamic quorum from the current voting set and a quorum disk (if it is available). If nodes are added to a cluster, they are added to the voting set. When nodes are removed, they are also removed from the voting set. Over time, the voting set, and hence the nodes in the cluster, can completely change. The process of updating the voting set for dynamic quorum is automatic and is done concurrently.

The cluster can migrate onto a completely separate set of nodes from the set on which it started. Within a SAN Volume Controller cluster, the quorum is defined in one of the following ways:

- ▶ Starting with SAN Volume Controller V7.2 Enhanced Stretched Cluster, the system continues to maintain the voting set with a dynamic quorum as for previous versions. But, to provide greater resiliency if there are planned or unplanned failures of nodes, the voting rules are changed.
- ▶ In particular, all the voting set nodes of a site plus the quorum disk are enough to achieve quorum, even if that voting set of nodes is less than half the nodes in the system.

A human vote, by using the **overridequorum** command, also is enough to establish quorum in this case.

To prevent unwanted behavior of the cluster, the voting rules, if there is no quorum disk, require that there are more nodes present than the largest site's voting set.

For example, if a two-I/O Group, four-node system has one node down for service, then one site has two nodes and the other site has one node.

If the inter-site link fails, then either of these sites can establish quorum by using the quorum disk. Alternatively, a user can use the **overridequorum** command to force a DR feature start, even on the site with one node.

As a further example, if there is an eight-node cluster with one node down for service and there is a failure that causes connectivity loss to the quorum disk and some nodes, then five nodes are needed to continue cluster operation.

Figure 2 on page 9 summarizes the behavior of the SAN Volume Controller cluster as a result of failures that affect the site or failure domains.



Failure Domain 1 1	Node	Failure Domain 2 Node 2	Failure Domain 3 Quorum disk	Cluster Status
Operational		Operational	Operational	Operational, optimal
Failed		Operational	Operational	Operational, Write cache disabled
Operational		Failed	Operational	Operational, Write cache disabled
Operational		Operational	Failed	Operational, Active Quorum disk moved
Operational, Link to Failure Domain 2 has failed, Split Brain	Split Brain	Operational, Link to Failure Domain 2 has failed, Split Brain	Operational	The node that accesses the active quorum disk first remains active and the partner node goes offline. If this is the beginning of a rolling disaster and the node who win the Quorum race goes offline too, then the surviving site can be restored with <code>overridequorum</code> command.
Operational		Failed	Failed	Stopped, then the surviving site can be restored with <code>overridequorum</code> command.
Failed		Operational	Failed	Stopped, then the surviving site can be restored with <code>overridequorum</code> command.

Figure 2 SAN Volume Controller Stretched Cluster behavior

## Quorum disk requirements

The storage controller that provides the quorum disk in an Enhanced Stretched Cluster configuration in the third site must be supported as an *extended quorum disk*. Storage controllers that provide extended quorum support are listed at the following website:

<http://www.ibm.com/storage/support/2145>

**Requirement:** Quorum disk storage controllers must be Fibre Channel or FCIP-attached. They must be able to provide less than 80 ms response times with an ensured bandwidth of greater than 2 MBps.

**Important:** Here are the quorum disk candidate requirements for the SAN Volume Controller Enhanced Stretched Cluster configuration:

- ▶ The SAN Volume Controller Enhanced Stretched Cluster configuration requires three quorum disk candidates.
- ▶ The active quorum disk must be assigned to the failure domain or site 3.
- ▶ Dynamic quorum selection must be disabled by using the `chquorum` command.
- ▶ Quorum disk candidates and the active quorum disk assignment must be done manually by using the `chquorum` command.

## IBM Spectrum Scale

IBM Spectrum Scale™ is a scalable, high-performance data and file management solution (based on IBM General Parallel File System™ (GPFS), also formerly known as IBM Elastic Storage). IBM Spectrum Scale provides world-class storage management with extreme scalability, flash accelerated performance, and automatic policy-based storage tiering from flash to disk to tape. IBM Spectrum Scale reduces storage costs while improving security and management efficiency in cloud, big data, and analytics environments.

Today's never-ending data growth is challenging traditional storage and data management solutions. New applications are generating massive amounts of unstructured data, such as video, audio, and text files, and data must be managed across traditional and cloud platforms. Being able to balance traditional workloads with new workloads and data types puts pressure on IT administrators to deliver application performance and reduce data access bottlenecks that delay schedules and waste expensive resources.

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. 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 V4.1 introduces revolutionary new features that clearly demonstrate the IBM commitment to providing groundbreaking storage solutions:

- ▶ 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

IBM Spectrum Scale removes data-related bottlenecks by providing parallel access to data, eliminating single filer choke points or hot spots. IBM Spectrum Scale also simplifies data management at scale by providing a single namespace that can be scaled simply and quickly by adding more scale-out resources, which eliminates “filer sprawl” and its associated problems.

IBM Spectrum Scale empowers geographically distributed organizations by expanding that single global namespace (literally globally) by placing critical data close to everyone and everything that needs it, no matter where they are in the world. Speeding data access to stakeholders around the world accelerates schedules and improves productivity.

IBM Spectrum Scale management and data lifecycle automation bridges the ever-widening data growth / budget chasm, bringing storage costs into line and making backup, restore, and disaster recovery integral components of the solution. As part of the IBM Spectrum Storage family, it is integrated with IBM Spectrum Protect™ and IBM Spectrum Archive™. IBM Spectrum Scale can uniquely manage the full data lifecycle, delivering geometrically lower-cost savings through policy-driven automation and tiered storage management.

IBM Spectrum Scale is part of IBM's market-leading software defined storage.

Spectrum Scale is storage software that runs on virtually any hardware platform and supports almost any block storage device. IBM Spectrum Scale is now available on Linux, AIX, and Windows based systems, and IBM recently announced that IBM Spectrum Scale also is available for the Linux on z Systems platform.

IBM Elastic Storage Server is an optimized storage solution bundled as hardware and software with exceptional performance and ease of management. Elastic Storage Server (ESS) provides unsurpassed end-to-end data availability, reliability, and integrity with unique technologies, including IBM Spectrum Scale RAID.

As a cloud service, IBM Spectrum Scale, which is delivered as a service, brings high performance, scalable storage, and integrated data governance for managing large amounts of data and files in the IBM SoftLayer cloud.

For more information about IBM Spectrum Scale, see the following website:

[http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=STGE\\_DC\\_ZQ\\_USEN&htmlfid=DCW03057USEN&attachment=DCW03057USEN.PDF#loaded](http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=STGE_DC_ZQ_USEN&htmlfid=DCW03057USEN&attachment=DCW03057USEN.PDF#loaded)

## Understanding IBM Spectrum Scale Quorum

Every application has different reliability requirements, from scientific scratch data to mission-critical fraud detection systems. IBM Spectrum Scale supports various reliability levels, depending on the needs of the application. When you design an IBM Spectrum Scale cluster, consider what type of events you need your system to “survive” and how automatic you want to the recovery to be. Any discussion of reliability in an IBM Spectrum Scale cluster starts with quorum.

The worst type of failure in a cluster is called *split-brain*. Split-brain happens when you have multiple nodes in a cluster that continue operations independently, with no way to communicate with each other. This situation cannot happen in a cluster file system because without coordination your file system can become corrupted. Coordination between the nodes is essential to maintaining data integrity. To keep the file system consistent, a lone node cannot be permitted to continue to write data to the file system without coordinating with the other nodes in the cluster. When a network failure occurs, some nodes must stop writing. Who continues and who stops is determined in IBM Spectrum Scale by using a mechanism that is called *quorum*.

Maintaining quorum in an IBM Spectrum Scale cluster means that most of the nodes that are designated as quorum nodes can successfully communicate. In a three-quorum node configuration, two nodes must be communicating for cluster operations to continue. When one node is isolated by a network failure, it stops all file system operations until communications are restored so that no data is corrupted by a lack of coordination.

In an IBM Spectrum Scale environment, there are two different quorums: the *cluster quorum* and the *filesystem quorum*. The former then can be configured by using two different techniques:

- ▶ Odd number of quorum nodes
- ▶ Tiebreaker disks

The goal of this paper is not to describe in detail how the IBM Spectrum Scale quorum can be configured, but to help you understand how IBM Spectrum Scale and SAN Volume Controller in Enhanced Stretched Cluster configuration can work together.

For more information about IBM Spectrum Scale quorum and GPFS reliability, see the following website:

<http://www-03.ibm.com/systems/resources/configure-gpfs-for-reliability.pdf>

## SAN Volume Controller Enhanced Stretched Cluster and IBM Spectrum Scale

SAN Volume Controller in its Enhanced Stretched Cluster Configuration (ESC) together with IBM Spectrum Scale can supply a rock-solid storage infrastructure, joining the SAN Volume Controller virtualization and business continuity functions and IBM Spectrum Scale unlimited scalability.

IBM Spectrum Scale has its own resilience and it can replicate (mirror) a single file, a set of files, or the entire file system, and you can change the replication status of a file at any time by using a policy or command. You can replicate metadata (file inode information), file data, or both.

In addition to replication, IBM Spectrum Scale provides an erasure-code based native RAID software implementation within IBM Spectrum Scale. Using conventional dual-ported disks in a JBOD configuration, IBM Spectrum Scale Native RAID implements sophisticated data placement and error correction algorithms to deliver high levels of storage reliability, availability, and performance.

Many customers have already implemented a storage infrastructure based on the IBM SAN Volume Controller in ESC configuration. When implementing an IBM Spectrum Scale into their environment, they want use the solid and reliable storage infrastructure that is already present in the data center. Thus, this paper shows how SAN Volume Controller and Spectrum Scale can work together and use each of their unique resilience and performance capabilities.

For more information about IBM Spectrum Scale and its reliability, see *Configuring GPFS for Reliability High availability for your enterprise applications*, found at:

<http://www-03.ibm.com/systems/resources/configure-gpfs-for-reliability.pdf>

The following IBM developerWorks® website also has information about this topic:

<http://tinyurl.com/mtcfx8w>

Figure 3 shows our example implementation high-level design.

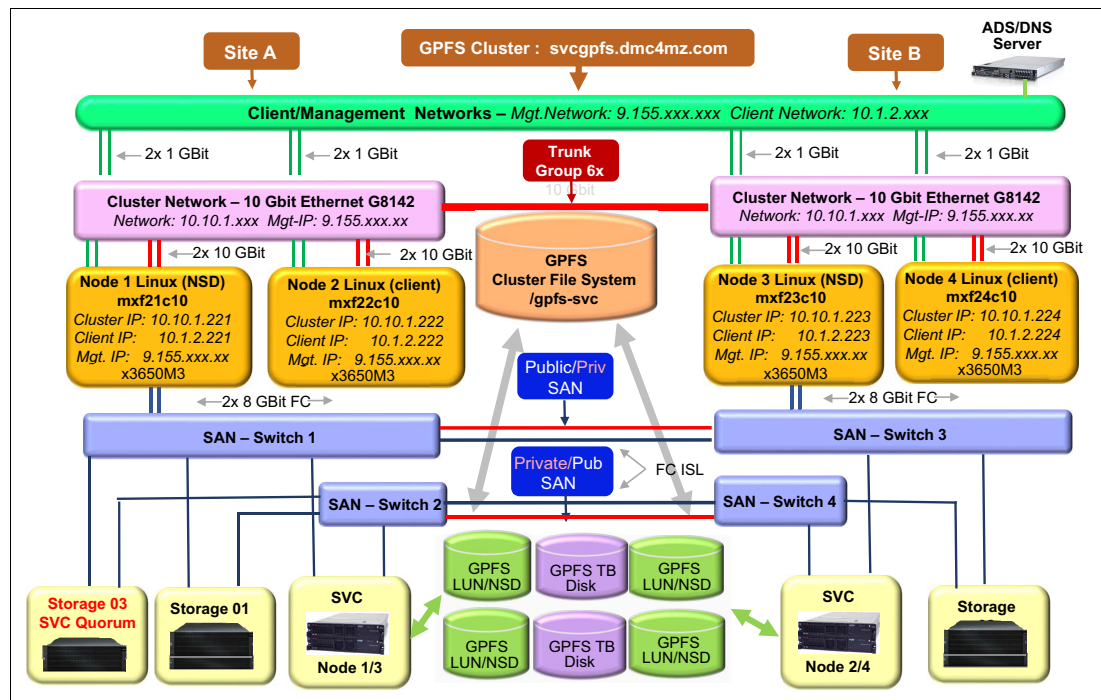


Figure 3 High-level design

Our example implementation includes the following items:

- ▶ Two 2145-DH8 SAN Volume Controller nodes at each site (nodes 1 and 3 and nodes 2 and 4)
- ▶ Two Storwize V7000 systems as SAN Volume Controller back-end storage (Storage01 and Storage02)
- ▶ One DS3400 as SAN Volume Controller cluster Active Quorum Disk
- ▶ Four FC switches that are partitioned to configure a public and private SAN for SAN Volume Controller ESC
- ▶ Two Linux nodes as the IBM Spectrum Scale client (mxf22c10 and mxf24c10)
  - x3650-M3
  - 48 GB RAM
  - Two 10 Gbps network adapters
  - SLES 11 SP3
  - IBM Spectrum Scale V4.1
- ▶ Two Linux nodes as the IBM Spectrum Scale NSD Server (mxf21c10 and mxf23c10)
  - x3650-M3
  - Two 146 GB SAS disks
  - 64 GB RAM
  - Six cores
  - One dual-port FC 8 Gb adapter
  - One dual-port Qlogic/Brocade CNA (Cluster-network - 10 Gb)
  - One dual-port Qlogic/Brocade CNA (Client-network - 10 Gb)
  - SLES 11 SP3
  - IBM Spectrum Scale V4.1

The goal of this configuration is to create a solution where IBM Spectrum Scale and the SAN Volume Controller ESC cluster behavior can be predicted if there is a split-brain condition of the SAN Volume Controller or IBM Spectrum Scale at the same time or at a different time or different sequence.

In this configuration, any single failure that is related to the IBM Spectrum Scale nodes or SAN Volume Controller infrastructure is handled by the high-availability (HA) solution itself.

In our example, the choice to locate the SAN Volume Controller Active Quorum Disk in the same site where 50% of the IBM Spectrum Scale nodes are located and 50% of the SAN Volume Controller resources are located was done so that we can predict what site remains available if there is a split-brain and a consequent SAN Volume Controller and IBM Spectrum Scale quorum race.

In this configuration, the site that is named Campus\_1 remains online because it is the only one that can reach the SAN Volume Controller Active Quorum Disk.

For the IBM Spectrum Scale in this example, we choose to create a tie-breaker disk that is supplied by the SAN Volume Controller.



```

id name                status mdisk_count vdisk_count capacity extent_size free_capacity virtual_capacity
used_capacity real_capacity overallocation warning easy_tier easy_tier_status compression_active
compression_virtual_capacity compression_compressed_capacity compression_uncompressed_capacity
parent_mdisk_grp_id parent_mdisk_grp_name child_mdisk_grp_count child_mdisk_grp_capacity type encrypt
0 Q_Campus_1_Active_34 online 1 0 19.00GB 1024 19.00GB 0.00MB
0.00MB 0.00MB 0 0 auto balanced no 0.00MB
0.00MB 0.00MB 0.00MB 0 Q_Campus_1_Active_34
0 0.00MB parent no
1 Q_Campus_2_stby_3 online 1 0 2.00TB 1024 2.00TB 0.00MB
0.00MB 0.00MB 0 0 auto balanced no 0.00MB
0.00MB 0.00MB 1 Q_Campus_2_stby_3
0 0.00MB parent no
2 Q_Campus_1_stby_35 online 1 0 1.00GB 1024 1.00GB 0.00MB
0.00MB 0.00MB 0 0 auto balanced no 0.00MB
0.00MB 0.00MB 2 Q_Campus_1_stby_35
0 0.00MB parent no
3 GPFS_Campus_1 online 2 6 4.72TB 1024 3.92TB 820.00GB
820.00GB 820.00GB 16 0 auto balanced no 0.00MB
0.00MB 0.00MB 3 GPFS_Campus_1
0 0.00MB parent no
4 GPFS_Campus_2 online 2 6 3.00TB 1024 2.20TB 820.00GB
820.00GB 820.00GB 26 0 auto balanced no 0.00MB
0.00MB 0.00MB 4 GPFS_Campus_2
0 0.00MB parent no

```

IBM\_2145:SVC\_ITALY:superuser>lsmdisk

```

idname statusmode mdisk_grp_idmdisk_grp_name capacityctrl_LUN# controller_nameUID
tier encrypt
0 GPFS_1_1 online managed 3 GPFS_Campus_1 1.7TB 0000000000000001 V7000_35_N2
600507680286001fd8000000000000200000000000000000000000000000000 enterprise no
1 GPFS_1_2 online managed 3 GPFS_Campus_1 3.0TB 0000000000000002 V7000_35_N2
600507680286001fd8000000000000300000000000000000000000000000000 enterprise no
2 Quorum_1 online managed 2 Q_Campus_1_stby_35 2.0GB 0000000000000000 V7000_35_N2
600507680286001fd8000000000000400000000000000000000000000000000 enterprise no
3 Quorum_2 online managed 1 Q_Campus_2_stby_3 2.0TB 0000000000000002 V7000_3_N1
6005076802820013d00000000000011100000000000000000000000000000000 enterprise no
4 GPFS_2_1 online managed 4 GPFS_Campus_2 3.0TB 0000000000000001 V7000_3_N1
6005076802820013d0000000000001100000000000000000000000000000000 enterprise no
5 GPFS_2_2 online managed 4 GPFS_Campus_2 2.0GB 0000000000000000 V7000_3_N1
6005076802820013d00000000000011200000000000000000000000000000000 enterprise no
6 Quorum_0 online managed 0 Q_Campus_1_Active_34 20.0GB 0000000000000000 DS3400
600a0b80003913b500002c6c54e54cb900000000000000000000000000000000 enterprise no

```

IBM\_2145:SVC\_ITALY:superuser>lsvdisk

```

id name IO_group_id IO_group_name status mdisk_grp_id mdisk_grp_name capacity type FC_id FC_name
RC_id RC_name vdisk_UID fc_map_count copy_count fast_write_state se_copy_count
RC_change compressed_copy_count parent_mdisk_grp_id parent_mdisk_grp_name
0 GPFS_0 0 io_grp0 online many many 200.00GB many
600507680C8100C4800000000000000 0 2 empty 0 no 0
many many
1 GPFS_1 0 io_grp0 online many many 200.00GB many
600507680C8100C48000000000000001 0 2 empty 0 no 0
many many
2 GPFS_2 1 io_grp1 online many many 200.00GB many
600507680C8100C48000000000000002 0 2 empty 0 no 0
many many
3 GPFS_3 1 io_grp1 online many many 200.00GB many
600507680C8100C48000000000000003 0 2 empty 0 no 0
many many
4 GPFS_tb_01 0 io_grp0 online many many 10.00GB many
600507680C8100C48000000000000004 0 2 empty 0 no 0
many many

```

```

5 GPFS_tb_02 1          io_grp1      online many          many          10.00GB many
600507680C8100C48000000000000005 0          2          empty          0          no          0
many          many
IBM_2145:SVC_ITALY:superuser>ls site
id site_name
1 Campus_1
2 Campus_2
3 Quorum
IBM_2145:SVC_ITALY:superuser>ls quorum
quorum_index status id name          controller_id controller_name active object_type override
0          online 2 Quorum_1 0          V7000_35_N2   no   mdisk      yes
1          online 6 Quorum_0 4          DS3400_SVC_Q  yes  mdisk      yes
2          online 3 Quorum_2 2          V7000_3_N1   no   mdisk      yes
IBM_2145:SVC_ITALY:superuser>ls host
id name          port_count iogrp_count status
0 GPFS_mxf21mz 2          4          online
1 GPFS_mxf23mz 2          4          online
IBM_2145:SVC_ITALY:superuser>ls hosvdiskmap 0
rbash: ls hosvdiskmap: command not found
IBM_2145:SVC_ITALY:superuser>ls hostvdiskmap 0
id name          SCSI_id vdisk_id vdisk_name vdisk_UID          IO_group_id IO_group_name
0 GPFS_mxf21mz 0          0          GPFS_0      600507680C8100C48000000000000000 0          io_grp0
0 GPFS_mxf21mz 1          1          GPFS_1      600507680C8100C48000000000000001 0          io_grp0
0 GPFS_mxf21mz 2          4          GPFS_tb_01 600507680C8100C48000000000000004 0          io_grp0
0 GPFS_mxf21mz 0          2          GPFS_2      600507680C8100C48000000000000002 1          io_grp1
0 GPFS_mxf21mz 1          3          GPFS_3      600507680C8100C48000000000000003 1          io_grp1
0 GPFS_mxf21mz 3          5          GPFS_tb_02 600507680C8100C48000000000000005 1          io_grp1
IBM_2145:SVC_ITALY:superuser>ls hostvdiskmap 1
id name          SCSI_id vdisk_id vdisk_name vdisk_UID          IO_group_id IO_group_name
1 GPFS_mxf23mz 0          0          GPFS_0      600507680C8100C48000000000000000 0          io_grp0
1 GPFS_mxf23mz 1          1          GPFS_1      600507680C8100C48000000000000001 0          io_grp0
1 GPFS_mxf23mz 2          4          GPFS_tb_01 600507680C8100C48000000000000004 0          io_grp0
1 GPFS_mxf23mz 0          2          GPFS_2      600507680C8100C48000000000000002 1          io_grp1
1 GPFS_mxf23mz 1          3          GPFS_3      600507680C8100C48000000000000003 1          io_grp1
1 GPFS_mxf23mz 3          5          GPFS_tb_02 600507680C8100C48000000000000005 1          io_grp1

```

## IBM Spectrum Scale configuration

The Spectrum Scale cluster is composed of four nodes:

- ▶ site1: mxf23c10.dmc4mz.com and mxf24c10.dmc4mz.com
- ▶ site2: mxf21c10.dmc4mz.com and mxf22c10.dmc4mz.com

To simulate a real environment in our example, we configured two IBM Spectrum Scale NSD servers (mxf23c10 and mxf21c10) and assigned the IBM Spectrum Scale LUNs to them only. We used the other two servers (mxf22c10 and mxf24c10) as Spectrum Scale clients.

You can run commands on any node in the IBM Spectrum Scale cluster, so in our example, the following commands were run from one IBM Spectrum Scale NSD server (mxf21):

- ▶ **mmfsccluster**
- ▶ **mmgetstate**
- ▶ **mmgetstate -a**
- ▶ **mmgetstate -L**
- ▶ **mmgetstate -v**
- ▶ **mmfsmount all**



- ▶ **mmlsmount all -L**
- ▶ **mmlsnode**
- ▶ **mmlsnode -N mx23c10,mx21c10,mx22c10,mx24c10**
- ▶ **mmlsdisk gpfs-svc**
- ▶ **mmlsdisk gpfs-svc -L**
- ▶ **mmlsnsd -a**
- ▶ **mmlsnsd -aLv**
- ▶ **mmlsnsd -avX**
- ▶ **mmlsfs all**

Example 2 shows our example IBM Spectrum Scale configuration.

*Example 2 NSD server IBM Spectrum Scale configuration output*

```
mx21:~ # mmlscluster
GPFS cluster information
=====
```

```
GPFS cluster name:      svcgpfs.dmc4mz.com
GPFS cluster id:       3039273354341853589
GPFS UID domain:      svcgpfs.dmc4mz.com
Remote shell command: /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:      CCR
```

Node	Daemon node name	IP address	Admin node name	Designation
1	mx21c10.dmc4mz.com	10.10.1.221	mx21c10.dmc4mz.com	quorum-manager
2	mx22c10.dmc4mz.com	10.10.1.222	mx22c10.dmc4mz.com	
3	mx23c10.dmc4mz.com	10.10.1.223	mx23c10.dmc4mz.com	quorum-manager
4	mx24c10.dmc4mz.com	10.10.1.224	mx24c10.dmc4mz.com	

```
mx21:~ # mmgetstate
```

Node number	Node name	GPFS state
1	mx21c10	active

```
mx21:~ # mmgetstate -a
```

Node number	Node name	GPFS state
1	mx21c10	active
2	mx22c10	active
3	mx23c10	active
4	mx24c10	active

```
mx21:~ # mmgetstate -L
```

Node number	Node name	Quorum	Nodes up	Total nodes	GPFS state	Remarks
1	mx21c10	1	2	4	active	quorum node

```
mx21:~ # mmgetstate -v
```

Node number	Node name	GPFS state
-------------	-----------	------------

```

1      mx21c10      active
mx21:~ #
mx21:~ # mm1smount all
File system gpfs-svc is mounted on 4 nodes.
mx21:~ #
mx21:~ # mm1smount all -L

```

File system gpfs-svc is mounted on 4 nodes:

```

10.10.1.223      mx23c10
10.10.1.221      mx21c10
10.10.1.222      mx22c10
10.10.1.224      mx24c10

```

```

mx21:~ # mm1snode
GPFS nodeset      Node list

```

```

-----
      svcgpfs      mx22c10 mx23c10 mx24c10 mx21c10
mx21:~ # mm1snode -N mx23c10,mx21c10,mx22c10,mx24c10
mx21c10.dmc4mz.com
mx22c10.dmc4mz.com
mx23c10.dmc4mz.com
mx24c10.dmc4mz.com

```

```

mx21:~ #
mx21:~ #mx21:~ #
mx21:~ # mm1sdisk gpfs-svc
disk      driver  sector  failure holds  holds  storage
name      type   size    group metadata data  status  availability pool
-----
tb02_svc_951 nsd      512      -1 no      no  ready  up      system
vd01_svc_893 nsd      512      -1 yes   yes  ready  up      system
vd02_svc_952 nsd      512      -1 yes   yes  ready  up      system
vd03_svc_951 nsd      512      -1 yes   yes  ready  up      system
vd04_svc_890 nsd      512      -1 yes   yes  ready  up      system

```

```

mx21:~ # mm1sdisk gpfs-svc -L
disk      driver  sector  failure holds  holds
storage
name type size group metadata data status  availability disk id pool
remarks

```

```

-----
tb02_svc_951 nsd  512  -1 no  no  ready  up  1 syste
m      desc
vd01_svc_893 nsd  512  -1 yes  yes  ready  up  2 syste
m      desc
vd02_svc_952 nsd  512  -1 yes  yes  ready  up  3 syste
m      desc
vd03_svc_951 nsd  512  -1 yes  yes  ready  up  4 syste
m      desc
vd04_svc_890 nsd  512  -1 yes  yes  ready  up  5 syste
m      desc

```

```

Number of quorum disks: 5
Read quorum value:      3
Write quorum value:     3

```

```

mx21:~ #
mx21:~ # mm1snsd -a

```

```

File system  Disk name  NSD servers

```

```
-----
gpfs-svc      tb02_svc_951 mxf23c10.dmc4mz.com,mxf21c10.dmc4mz.com
gpfs-svc      vd01_svc_893 mxf21c10.dmc4mz.com,mxf23c10.dmc4mz.com
gpfs-svc      vd02_svc_952 mxf21c10.dmc4mz.com,mxf23c10.dmc4mz.com
gpfs-svc      vd03_svc_951 mxf23c10.dmc4mz.com,mxf21c10.dmc4mz.com
gpfs-svc      vd04_svc_890 mxf23c10.dmc4mz.com,mxf21c10.dmc4mz.com
(gpfs disk)   tb01_svc_890 mxf23c10.dmc4mz.com,mxf21c10.dmc4mz.com
```

```
mxvf21:~ # mmlsnsd -aLv
```

File system	Disk name	NSD volume ID	NSD servers
gpfs-svc	tb02_svc_951	0A0102DF54F4BD8D	mxf23c10.dmc4mz.com,mxf21c10.dmc4mz.com
gpfs-svc	vd01_svc_893	0A0102DD54F4BD8E	mxf21c10.dmc4mz.com,mxf23c10.dmc4mz.com
gpfs-svc	vd02_svc_952	0A0102DD54F4BD91	mxf21c10.dmc4mz.com,mxf23c10.dmc4mz.com
gpfs-svc	vd03_svc_951	0A0102DF54F4BD95	mxf23c10.dmc4mz.com,mxf21c10.dmc4mz.com
gpfs-svc	vd04_svc_890	0A0102DF54F4BD97	mxf23c10.dmc4mz.com,mxf21c10.dmc4mz.com
(free disk)	tb01_svc_890	0A0102DF54F4C1C6	mxf23c10.dmc4mz.com,mxf21c10.dmc4mz.com

```
mxvf21:~ # mmlsnsd -avX
```

Disk name	NSD volume ID	Device	Devtype	Node name	Remarks
tb01_svc_890	0A0102DF54F4C1C6	/dev/dm-3	dmm	mxf21c10.dmc4mz.com	server node
tb01_svc_890	0A0102DF54F4C1C6	/dev/dm-0	dmm	mxf23c10.dmc4mz.com	server node
tb02_svc_951	0A0102DF54F4BD8D	/dev/dm-0	dmm	mxf21c10.dmc4mz.com	server node
tb02_svc_951	0A0102DF54F4BD8D	/dev/dm-3	dmm	mxf23c10.dmc4mz.com	server node
vd01_svc_893	0A0102DD54F4BD8E	/dev/dm-4	dmm	mxf21c10.dmc4mz.com	server node
vd01_svc_893	0A0102DD54F4BD8E	/dev/dm-1	dmm	mxf23c10.dmc4mz.com	server node
vd02_svc_952	0A0102DD54F4BD91	/dev/dm-2	dmm	mxf21c10.dmc4mz.com	server node
vd02_svc_952	0A0102DD54F4BD91	/dev/dm-5	dmm	mxf23c10.dmc4mz.com	server node
vd03_svc_951	0A0102DF54F4BD95	/dev/dm-1	dmm	mxf21c10.dmc4mz.com	server node
vd03_svc_951	0A0102DF54F4BD95	/dev/dm-4	dmm	mxf23c10.dmc4mz.com	server node
vd04_svc_890	0A0102DF54F4BD97	/dev/dm-5	dmm	mxf21c10.dmc4mz.com	server node
vd04_svc_890	0A0102DF54F4BD97	/dev/dm-2	dmm	mxf23c10.dmc4mz.com	server node

```
mxvf21:~ #mxf21:~ # mmlsfs all
```

```
File system attributes for /dev/gpfs-svc:
```

```
=====
flag          value          description
-----
-f            65536         Minimum fragment size in bytes
-i            4096          Inode size in bytes
-I            32768         Indirect block size in bytes
-m            2             Default number of metadata replicas
-M            2             Maximum number of metadata replicas
-r            2             Default number of data replicas
-R            2             Maximum number of data replicas
-j            cluster       Block allocation type
-D            nfs4          File locking semantics in effect
-k            all           ACL semantics in effect
-n            32           Estimated number of nodes that will mount file syst
em
-B            2097152      Block size
```

-Q	none	Quotas accounting enabled
	none	Quotas enforced
	none	Default quotas enabled
--perfilesset-quota	no	Per-fileset quota enforcement
--filesetdf	no	Fileset df enabled?
-V	14.10 (4.1.0.4)	File system version
--create-time	Mon Mar 2 20:52:31 2015	File system creation time
-z	no	Is DMAPI enabled?
-L	4194304	Logfile size
-E	yes	Exact mtime mount option
-S	no	Suppress atime mount option
-K	whenpossible	Strict replica allocation option
--fastea	yes	Fast external attributes enabled?
--encryption	no	Encryption enabled?
--inode-limit	635392	Maximum number of inodes
--log-replicas	0	Number of log replicas
--is4KAligned	yes	is4KAligned?
--rapid-repair	yes	rapidRepair enabled?
--write-cache-threshold	0	HAWC Threshold (max 65536)
-P	system	Disk storage pools in file system
-d	tb02_svc_951;vd01_svc_893;vd02_svc_952;vd03_svc_951;vd04_svc_890	Disks in file system
-A	yes	Automatic mount option
-o	none	Additional mount options
-T	/gpfs/gpfs-svc	Default mount point
--mount-priority	0	Mount priority

In addition, in our example, we run the following Linux commands, as shown in Example 3:

- ▶ **lsscsi -g --transport**
- ▶ **multipath -l**

*Example 3 Linux command example*

```

mx21:~ # lsscsi -g --transport
[0:0:17:0] disk /dev/sda /dev/sg0
[0:0:18:0] disk /dev/sdb /dev/sg1
[0:2:0:0] disk /dev/sdc /dev/sg2
[1:0:0:0] cd/dvd ata: /dev/sr0 /dev/sg3
[9:0:0:0] disk fc:0x500507680c120952,0x481900 /dev/sdd /dev/sg4
[9:0:0:1] disk fc:0x500507680c120952,0x481900 /dev/sde /dev/sg5
[9:0:0:3] disk fc:0x500507680c120952,0x481900 /dev/sdf /dev/sg6
[9:0:1:0] disk fc:0x500507680c121893,0x482100 /dev/sdg /dev/sg7
[9:0:1:1] disk fc:0x500507680c121893,0x482100 /dev/sdh /dev/sg8
[9:0:1:2] disk fc:0x500507680c121893,0x482100 /dev/sdi /dev/sg9
[9:0:2:0] disk fc:0x500507680c121890,0x3e2500 /dev/sdab /dev/sg10
[9:0:2:1] disk fc:0x500507680c121890,0x3e2500 /dev/sdac /dev/sg11
[9:0:2:2] disk fc:0x500507680c121890,0x3e2500 /dev/sdad /dev/sg12
[9:0:3:0] disk fc:0x500507680c120951,0x3e1900 /dev/sdah /dev/sg22
[9:0:3:1] disk fc:0x500507680c120951,0x3e1900 /dev/sdai /dev/sg23
[9:0:3:3] disk fc:0x500507680c120951,0x3e1900 /dev/sdaj /dev/sg24
[10:0:0:0] disk fc:0x500507680c120952,0x481900 /dev/sdp /dev/sg16
[10:0:0:1] disk fc:0x500507680c120952,0x481900 /dev/sdq /dev/sg17
[10:0:0:3] disk fc:0x500507680c120952,0x481900 /dev/sdr /dev/sg18
[10:0:1:0] disk fc:0x500507680c121893,0x482100 /dev/sds /dev/sg19
[10:0:1:1] disk fc:0x500507680c121893,0x482100 /dev/sdt /dev/sg20

```

```

[10:0:1:2] disk fc:0x500507680c121893,0x482100 /dev/sdu /dev/sg21
[10:0:2:0] disk fc:0x500507680c121890,0x3e2500 /dev/sdae /dev/sg13
[10:0:2:1] disk fc:0x500507680c121890,0x3e2500 /dev/sdaf /dev/sg14
[10:0:2:2] disk fc:0x500507680c121890,0x3e2500 /dev/sdag /dev/sg15
[10:0:3:0] disk fc:0x500507680c120951,0x3e1900 /dev/sdak /dev/sg25
[10:0:3:1] disk fc:0x500507680c120951,0x3e1900 /dev/sdal /dev/sg26
[10:0:3:3] disk fc:0x500507680c120951,0x3e1900 /dev/sdam /dev/sg27

```

```

mxf21:~ #

```

```

mxf21:~ # mxf21:~ # multipath -l

```

```

mpathl (3600507680c8100c48000000000000003) dm-1 IBM ,2145

```

```

size=200G features='0' hwhandler='0' wp=rw

```

```

|-+- policy='service-time 0' prio=0 status=active

```

```

| |- 9:0:0:1 sde 8:64 active undef running

```

```

| `~ 10:0:0:1 sdq 65:0 active undef running

```

```

`-+- policy='service-time 0' prio=0 status=enabled

```

```

| |- 9:0:3:1 sdai 66:32 active undef running

```

```

| `~ 10:0:3:1 sda1 66:80 active undef running

```

```

mpathk (3600507680c8100c48000000000000002) dm-0 IBM ,2145

```

```

size=200G features='0' hwhandler='0' wp=rw

```

```

|-+- policy='service-time 0' prio=0 status=active

```

```

| |- 9:0:3:0 sdah 66:16 active undef running

```

```

| `~ 10:0:3:0 sdak 66:64 active undef running

```

```

`-+- policy='service-time 0' prio=0 status=enabled

```

```

| |- 9:0:0:0 sdd 8:48 active undef running

```

```

| `~ 10:0:0:0 sdp 8:240 active undef running

```

```

mpathj (3600507680c8100c48000000000000001) dm-4 IBM ,2145

```

```

size=200G features='0' hwhandler='0' wp=rw

```

```

|-+- policy='service-time 0' prio=0 status=active

```

```

| |- 9:0:1:1 sdh 8:112 active undef running

```

```

| `~ 10:0:1:1 sdt 65:48 active undef running

```

```

`-+- policy='service-time 0' prio=0 status=enabled

```

```

| |- 9:0:2:1 sdac 65:192 active undef running

```

```

| `~ 10:0:2:1 sdaf 65:240 active undef running

```

```

mpathi (3600507680c8100c48000000000000004) dm-5 IBM ,2145

```

```

size=10G features='0' hwhandler='0' wp=rw

```

```

|-+- policy='service-time 0' prio=0 status=active

```

```

| |- 9:0:2:2 sdad 65:208 active undef running

```

```

| `~ 10:0:2:2 sdag 66:0 active undef running

```

```

`-+- policy='service-time 0' prio=0 status=enabled

```

```

| |- 9:0:1:2 sdi 8:128 active undef running

```

```

| `~ 10:0:1:2 sdu 65:64 active undef running

```

```

mpathh (3600507680c8100c48000000000000000) dm-3 IBM ,2145

```

```

size=200G features='0' hwhandler='0' wp=rw

```

```

|-+- policy='service-time 0' prio=0 status=active

```

```

| |- 9:0:2:0 sdab 65:176 active undef running

```

```

| `~ 10:0:2:0 sdae 65:224 active undef running

```

```

`-+- policy='service-time 0' prio=0 status=enabled

```

```

| |- 9:0:1:0 sdg 8:96 active undef running

```

```

| `~ 10:0:1:0 sds 65:32 active undef running

```

```

mpathg (3600507680c8100c48000000000000005) dm-2 IBM ,2145

```

```

size=10G features='0' hwhandler='0' wp=rw

```

```

|-+- policy='service-time 0' prio=0 status=active

```

```

| |- 9:0:3:3 sdaj 66:48 active undef running

```

```

| `~ 10:0:3:3 sdam 66:96 active undef running

```

```

`-+- policy='service-time 0' prio=0 status=enabled

```

```
|- 9:0:0:3 sdf 8:80 active undef running
^- 10:0:0:3 sdr 65:16 active undef running
```

**Note:** The `multipath -l` command output can be different for each NSD server because each NSD server assigns `mpathxx` to each LUN in a different sequence. Use the LUN UID to verify that a LUN is the same on more than one node.

The IBM Spectrum Scale and the SAN Volume Controller ESC architecture simplifies the IBM Spectrum Scale configuration because IBM Spectrum Scale HA is handled by using the SAN Volume Controller Enhanced Stretched Cluster configuration.

## Test scenarios and preferred practices

In our implementation, we ran some Linux `dd` read and write commands from the IBM Spectrum Scale client to the IBM Spectrum Scale NSD server and observed the infrastructure impact of certain failures, and the IBM Spectrum Scale and SAN Volume Controller ESC behavior.

Table 1 summarizes the failure scenarios that were tested in our implementation.

Table 1 Failure scenarios that were tested

Test description	Event cause	IBM Spectrum Scale and SAN Volume Controller behavior
A SAN Volume Controller node was lost at a site without a SAN Volume Controller Active Quorum Disk.	We set a SAN Volume Controller node to Service Mode.	There was no impact.
A SAN Volume Controller node was lost at a site with a SAN Volume Controller Active Quorum Disk.	We set a SAN Volume Controller node to Service Mode.	There was no impact.
Two SAN Volume Controller nodes were lost at a site with a SAN Volume Controller Active Quorum Disk	We set two SAN Volume Controller nodes at the same site to Service Mode.	There was no IBM Spectrum Scale impact. The SAN Volume Controller volume went into a Degraded State. The Storage Pool and SAN Volume Controller Quorum at this site went offline.
Two SAN Volume Controller nodes were lost at a site without a SAN Volume Controller Active Quorum Disk.	We set two SAN Volume Controller nodes at the same site to Service Mode.	There was no IBM Spectrum Scale impact. The SAN Volume Controller volume went into a Degraded State. The Storage Pool and SAN Volume Controller Quorum at this site went offline.
There was a split-brain with all the SAN Volume Controller Quorum Disks available.	We disabled the public SAN ISL and then disabled the private SAN ISL.	The SAN Volume Controller winning site was the one with the SAN Volume Controller Active Quorum Disk. There was no impact on IBM Spectrum Scale. There was a temporary IBM Spectrum Scale workload I/O wait.

<b>Test description</b>	<b>Event cause</b>	<b>IBM Spectrum Scale and SAN Volume Controller behavior</b>
There was a split-brain with all the SAN Volume Controller Quorum Disks available.	We disabled the private SAN ISL and then disabled the public SAN ISL.	The SAN Volume Controller winning site was the one with the SAN Volume Controller Active Quorum Disk. There is no impact on IBM Spectrum Scale. There was a temporary IBM Spectrum Scale workload I/O wait.
There was a split-brain with all the SAN Volume Controller Quorum Disks available.	We disabled the public SAN ISL and private SAN ISL at the same time.	The SAN Volume Controller winning site was the one with the SAN Volume Controller Active Quorum Disk. There was no impact on IBM Spectrum Scale. There was a temporary IBM Spectrum Scale workload I/O wait.
There was a split-brain with a Rolling Disaster on the SAN and IBM Spectrum Scale network at the same time.	We disabled the public SAN ISL, private SAN ISL, and IBM Spectrum Scale ETH Network PortChannel at the same time.	The SAN Volume Controller winning site was the one with SAN Volume Controller Active Quorum Disk. The IBM Spectrum Scale file system was no longer accessible on the lost site. The IBM Spectrum Scale file system was still mounted and accessible by the clients through the NSD server on the winning site. The IBM Spectrum Scale workload was suspended and waiting for a new File System Manager election for about 150 seconds.
There was a split-brain with a Rolling Disaster on the SAN first, and then on the IBM Spectrum Scale network.	We disabled the public SAN ISL and private SAN ISL first, and then disabled the IBM Spectrum Scale ETH Network PortChannel.	The SAN Volume Controller winning site was the one with the SAN Volume Controller Active Quorum Disk. The IBM Spectrum Scale file system was no longer accessible on the lost site. The IBM Spectrum Scale file system was still mounted and accessible by the clients through the NSD server on the winning site. The IBM Spectrum Scale workload was suspended and waiting for a new File System Manager election for about 150 seconds.
There was a split-brain with a Rolling Disaster on the IBM Spectrum Scale network first, and then on the SAN.	We disabled the IBM Spectrum Scale ETH network first, and then the private SAN ISL and public SAN ISL.	The SAN Volume Controller winning site is the one with the SAN Volume Controller Active Quorum Disk. The IBM Spectrum Scale file system was no longer accessible on the lost site. The IBM Spectrum Scale file system was still mounted and accessible by the clients through the NSD server on the winning site. The IBM Spectrum Scale workload was suspended and waiting for a new File System Manager election for about 150 seconds.

Test description	Event cause	IBM Spectrum Scale and SAN Volume Controller behavior
<p>There was a split-brain with a Rolling Disaster on the IBM Spectrum Scale network first and then the SAN. Later, even the SAN Volume Controller winning site fails. We restarted the SAN Volume Controller cluster by running <b>overridequorum</b>.</p>	<p>We disabled the public SAN ISL, private SAN ISL, and IBM Spectrum Scale ETH Network PortChannel at the same time. Later, we set the surviving SAN Volume Controller winning site SAN Volume Controller Nodes to Service Mode to simulate a total disaster.</p>	<p>The SAN Volume Controller winning site was the one with the SAN Volume Controller Active Quorum Disk. The IBM Spectrum Scale file system was no longer accessible on the lost site. The IBM Spectrum Scale file system was still mounted and accessible by the clients through the NSD server on the SAN Volume Controller winning site. The IBM Spectrum Scale workload was suspended and waiting for a new File System Manager election for about 150 seconds. After a total disaster, IBM Spectrum Scale stops and is unmounted. After we ran <b>overridequorum</b>, IBM Spectrum Scale was still must to be remounted so that the data is still accessible.</p>

**Note:** The workload test was run by running Linux **dd** commands. Thus, the workload behavior is related to this specific test. Other specific workloads were not tested, hence a specific application configuration was not required to support IBM Spectrum Scale and SAN Volume Controller ESC behavior.

## SAN Volume Controller Enhanced Stretched Cluster preferred practices

Here are the preferred practices that we applied to configure our SAN Volume Controller Enhanced Stretched Cluster example environment in addition to the common SAN Volume Controller ESC preferred practices:

- ▶ We used four SAN Volume Controller nodes to have the best resilience.
- ▶ We used three different storage subsystems.
- ▶ We implemented SAN Volume Controller Volume Mirroring to achieve business continuity from the SAN Volume Controller back-end storage point of view.
- ▶ We used the two storage subsystems in site 1 and 2 to allocate the two SAN Volume Controller standby Quorum Disks.
- ▶ We installed the third storage subsystem acting as an SAN Volume Controller Active Quorum Disk in site Campus 1 where the SAN Volume Controller standby quorum also is. With this set, we were able to predict what site remained online if there was a split-brain. In our implementation, the winning site is always the site Campus 1.
- ▶ We created all IBM Spectrum Scale LUNs in Volume Mirroring, spreading the preferred node for each LUN on all the SAN Volume Controller nodes to get the preferred resilience and performance.
- ▶ We created IBM Spectrum Scale tiebreaker LUNs on SAN Volume Controller in Volume Mirroring by using the SAN Volume Controller nodes in the predicted winner site as the preferred node.



## IBM Spectrum Scale preferred practices

Here are the preferred practices that we applied to our IBM Spectrum Scale example environment in addition to the common IBM Spectrum Scale preferred practices and configuration parameters that can be found at the following website:

<http://tinyurl.com/pdn79w9>

- ▶ We assigned the same number of IBM Spectrum Scale Quorum Managers at each site.
- ▶ IBM Spectrum Scale NSD Network LAN was configured on two different Ethernet switches.
- ▶ You can make IBM Spectrum Scale more tolerant of slow networks and high workload by increasing the IBM Spectrum Scale `minMissedPingTimeout` parameter (for the `mmchconfig` command). However, setting this parameter prevents fast failover if there is a real node failure because a node is not expelled until at least the time that is set in `minMissedPingTimeout` has passed. So, it all depends on what you want:
  - Fast failover, in which case you might get false node expels because of a slow network response.
  - Slow failover, to keep things going while possible even in the face of a system overload or network glitches. This situation then causes longer failover times if there is a real failure.

For our implementation, we set `minMissedPingTimeout` to 120 seconds.

## Authors

This paper was produced by this team of specialists:

**Angelo Bernasconi** is an Executive Certified, Storage, SAN, and Storage Virtualization IT Specialist. He is a member of IBM Italy TEC. Angelo was a Storage FTSS at STG Italy. He has 29 years of experience in the delivery of maintenance, professional services, and solutions for IBM Enterprise customers in z/OS, and for the last 14 years he has focused on open systems. He holds a degree in Electronics and his areas of expertise include storage hardware, SANs, storage virtualization design, solutions, and implementation, DR solutions, and data deduplication. Angelo writes extensively about SAN and storage products in IBM Redbooks publications and white papers.

**Cristiano Beretta** is an IT Specialist working at the Storage Systems Group in Italy. He joined IBM in 2001. His role is a Field Technical Sales Specialist, and his main responsibility is providing technical support for IBM storage solutions to IBM professionals, IBM Business Partners, and IBM clients. He holds a degree in Management, Economics, and Industrial Engineering from the Politecnico di Milano. Before joining IBM, he worked at SGI. His areas of expertise include IBM Systems Disk and Tape products, SAN Volume Controller, network-attached storage (NAS), SANs, IBM AIX®, SGI IRIX, Linux, and Microsoft Windows.

**Walter Bernocchi** holds a degree in Electronics and Computer Science. He joined IBM Italy in 1989 and held various technical positions in the System Test and System Analysis area. Since 2001, he has worked as a Technical Computing Architect and is focused on supporting software-defined infrastructures from deployment to technical validation, and the support of high performance computing cloud solutions, including remote 2D/3D visualization, parallel file systems, virtualization, scheduler, and computational optimization. He has been the visiting lecturer at the University of La Spezia (I) for High Performance Computing architectures.

**Giorgio Richelli** is a Senior IT Specialist. He graduated as an Electronic Engineer in 1985 and since then he has worked for a number of UNIX systems vendors, including IBM, mostly as a High Performance Computing Specialist. He rejoined IBM in 2000 and since 2004, he has worked as a System Architect. Since 2013, he has been a part of Platform Technical Sales.

Thanks to the following people for their contributions to this project:

Jon Tate  
**International Technical Support Organization, San Jose Center**

Scott Fadden  
**IBM US**

## Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience by using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

[ibm.com/redbooks/residencies.html](http://ibm.com/redbooks/residencies.html)

## Stay connected to IBM Redbooks

- ▶ Find us on Facebook:  
<http://www.facebook.com/IBMRedbooks>
- ▶ Follow us on Twitter:  
<http://twitter.com/ibmredbooks>
- ▶ Look for us on LinkedIn:  
<http://www.linkedin.com/groups?home=&gid=2130806>
- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:  
<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>
- ▶ Stay current on recent Redbooks publications with RSS Feeds:  
<http://www.redbooks.ibm.com/rss.html>

# Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

*IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.*

**The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:** INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

## COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

© Copyright International Business Machines Corporation 2015. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

This document REDP-5224-00 was created or updated on July 8, 2015.



Send us your comments in one of the following ways:


- ▶ Use the online **Contact us** review Redbooks form found at:  
[ibm.com/redbooks](http://ibm.com/redbooks)
- ▶ Send your comments in an email to:  
[redbooks@us.ibm.com](mailto:redbooks@us.ibm.com)
- ▶ Mail your comments to:  
IBM Corporation, International Technical Support Organization  
Dept. HYTD Mail Station P099  
2455 South Road  
Poughkeepsie, NY 12601-5400 U.S.A.



## Trademarks

IBM, the IBM logo, and [ibm.com](http://ibm.com) are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

AIX®	IBM Spectrum™	Redbooks®
GPFS™	IBM Spectrum Storage™	Redpaper™
IBM®	PowerHA®	Redbooks (logo)  ®
IBM Elastic Storage™	Real-time Compression™	Storwize®

The following terms are trademarks of other companies:

SoftLayer, and SoftLayer device are trademarks or registered trademarks of SoftLayer, Inc., an IBM Company.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.





REDP-5224-00

ISBN 0730454265

Printed in U.S.A.

Get connected

