IBM System Storage SAN Volume Controller

Guidelines for configuring SAN Volume Controller Split I/O Group Clustering

Information Center Errata

Version 6.3.0

Nov 18, 2011

Contents

Introduction

This guide provides errata information that pertains to release 6.3.0 of the IBM *System Storage SAN Volume Controller Information Center*

Who should use this guide

This errata should be used by anyone configuring SAN Volume Controller in a split site (I/O Group) environment..

Last Update

This document was last updated: Nov 18, 2011.

Change History

The following revisions have been made to this document:

Revision Date	Sections Modified
November 18	New publication

Table 1: Change History

SVC split I/O group clustered system configuration

To provide protection against failures that affect an entire location, such as a power failure, you can use a configuration that splits each SAN Volume Controller I/O group across two physical locations. Volume mirroring provides consistent copies of data in both locations. A third, independent site is required for quorum disk placement. Such a configuration is referred as *SAN Volume Controller split I/O group* clustered system.

However, you should consider that split I/O group clusters might exhibit substantially reduced performance, dependent on the distance between the sites.

Bandwidth Requirements

The connection between SVC nodes must guarantee a minimum bandwidth of 4Gbit/s or 2 times the peek host write I/O workload which ever is higher. To avoid performance bottlenecks each SVC node requires 2 ports operating at full speed (e.g. 2 x 8Gbit/s) worth of bandwidth to other nodes within the same cluster.

For example an SVC Split I/O Group (no matter how many nodes) with a total of 4Gbit/s of inter-site bandwidth on the private SANs is a valid configuration as long as the peek host write I/O workload does not exceed 200Mbytes/s.

To operate an 8 node cluster at its maximum performance will require 64Gbit/s of inter-site bandwidth on the private SANs.

Failure domains (sites)

In a split I/O group cluster configuration, the term *site* is used as synonym for an independent failure domain. A *failure domain* is a part of the SVC cluster within a boundary such that any failure (like power failure, fire, and flood) within that boundary is contained within the boundary and the failure does not propagate or affect parts outside of said boundary.

An SVC split I/O group cluster spans three independent failure domains: two failure domains with SVC nodes and storage systems for customer data (*production sites*), and a third failure domain with a storage system for the active quorum disk (*quorum site*).

Failure domains are typically areas or rooms in the data center, buildings on the same campus, or even buildings in different towns. Different kinds of failure domains protect against different types of faults. For example:

- If each site is an area with separate electrical power source within the same data center, the SAN Volume Controller cluster can survive the failure of any single power source.
- If each site is a different building, the SAN Volume Controller cluster can survive the loss of any single building (for example, because of power failure or fire).

If configured properly, the SVC cluster continues to operate after the loss of one failure domain. The key prerequisite is that a failure domain hosts only one node from each I/O group. Placement of whole I/O groups from the same SVC cluster in different sites is not a split I/O group configuration and does not provide higher availability.

In all cases, the SAN Volume Controller cluster does not guarantee that it can operate after the failure of two failure domains.

Split I/O group and Metro or Global Mirror

An SVC split I/O group cluster is designed to continue operation after the loss of one failure domain. It cannot guarantee that it still operates after the failure of two failure domains. IBM recommends to use Metro or Global Mirror to a second SVC cluster for extended disaster recovery. You configure and manage Metro or Global Mirror partnerships that include a split I/O group cluster in the same way as other remote copy relationships. SAN Volume Controller supports SAN routing technology (including FCIP links) for intercluster connections that use Metro Mirror or Global Mirror.

The partner SVC cluster should not be located in a production site of the SVC split I/O group cluster. However, it may be co-located with the storage system that provides the active quorum disk for the split I/O group cluster.

Connectivity between production sites

You have to connect the two production sites by Fibre Channel links. These Fibre Channel links provide paths for SVC node-to-node communication as well as for host access to SVC nodes.

SVC split I/O group clusters support two different approaches for node-to-node intra-cluster communication between production sites:

- (a) Attach each SVC node to the Fibre Channel switches in the local as well as in the remote production site directly. Thus, all node-to-node traffic can be done without passing inter-site ISLs. That is referred as *split I/O group configuration without ISLs* between SVC nodes (see section **Node-to-node paths without ISLs** for details).
- (b) Attach each SVC node only to local Fibre Channel switches and configure ISLs between production sites for SVC node-to-node traffic. That is referred as *split I/O group configuration with ISLs* between SVC nodes (see section **Node-to-node paths with ISLs** for details).

SVC version 6.3.0 or later is required for configurations with ISLs between SVC nodes.

For both types of configuration, wavelength-division multiplexing (WDM) devices are supported with certain requirements. FCIP or iSCSI connections are not supported for paths between SVC nodes.

You have to connect the quorum site to both production sites by Fibre Channel or FCIP links (see section **Connectivity to quorum site**). The links to the quorum site provide paths only for the SVC nodes to access the storage system with the quorum disk.

Node-to-node paths without ISLs

The most simple split I/O group configuration is build by attaching each SVC node directly to the Fibre Channel switches in the local and the remote production site. SVC version 5.1.0 or later is required for support. You configure the SVC split I/O group cluster according to the rules below:

- The minimal SAN configuration consists of one Fibre Channel switch per production site as two separate fabrics. For highest reliability, two switches per production site are recommended. Single fabric configurations are not supported for split I/O group clusters.
- As with every SVC cluster, you can use ISLs for host-to-node (with up to 3 hops) or for node-to-storage (at most 1 hop) access. However, configure the SAN zones so that ISLs are not used in paths between SVC nodes.
- Attach two ports of each SVC node to the Fibre Channel switches in the production site where the node resides.
- Attach the remaining two ports of each SVC node to the Fibre Channel switches in the other production site.
- Connect each storage system at the production sites to Fibre Channel switches in the site where the storage system resides.
- Connect the storage system with the active quorum disks to Fibre Channel switches in both production sites.

Restriction: Do not connect a storage system in one site directly to a switch fabric in the other site. Connectivity between SVC ports at one site to storage at the other site that use ISLs to travel between the locations are not affected by this restriction.

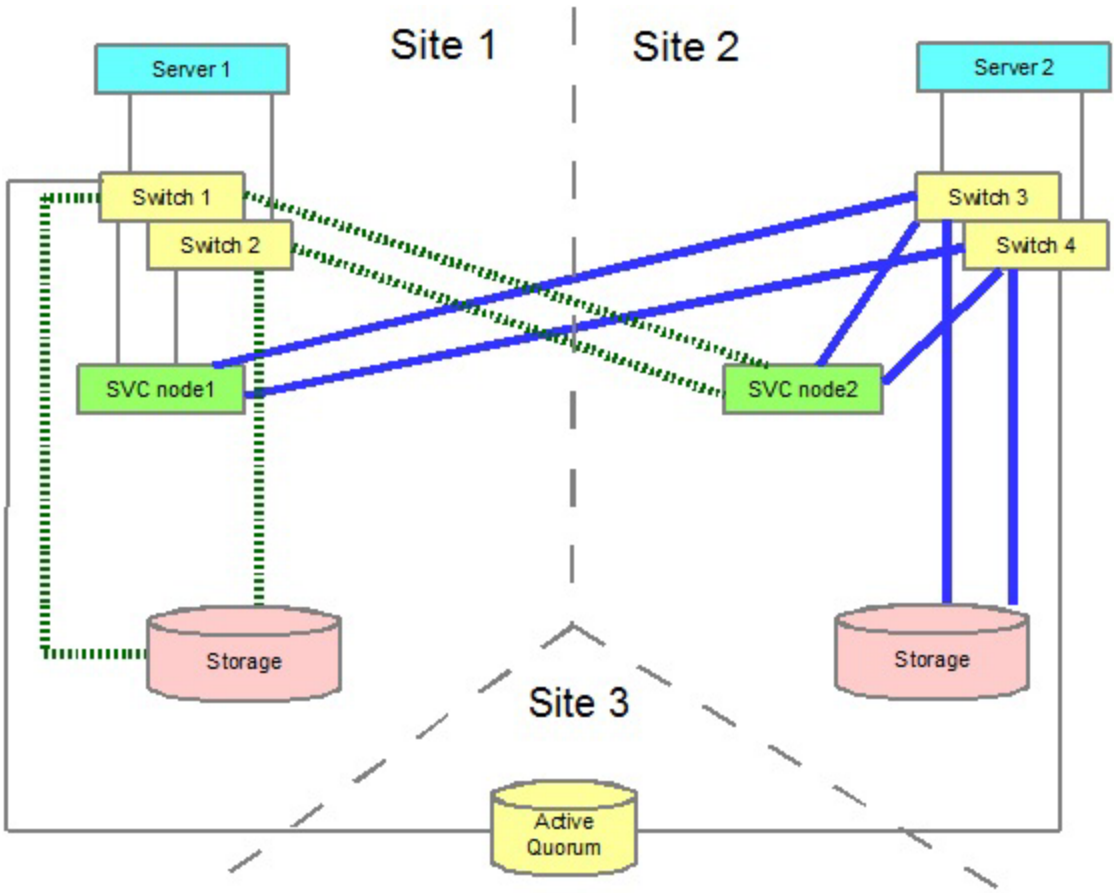

Figure 1: Zoning example

The exact requirements for the links between an SVC node and the SAN switches in the production site depend on the SAN Volume Controller version.

Up to SAN Volume Controller version 6.1.0

The requirements below apply to split I/O group clustered systems with SVC versions 5.1.0 to 6.1.0.

- The cable length must not exceed 10 km.
- The maximum cable length for short-wave SFPs depends on Fibre Channel link speed and cable type as shown as an example in Table 1.

	Maximum cable length (meters)			
	OM 1 (62.5 micron)	OM 2 (50 micron)	OM 3 (50 micron)	OM 4 (50 micron)
2 GBps	150	300	500	
4 GBps	70	150	380	400
8 GBps	21	50	150	190

Table 1: Maximum distances for short-wave Fibre Channel links

- For longer distances as shown in Table 1 , single-mode long-wave Fibre Channel connections must be used. Long-wave SFPs can be purchased as an optional SAN Volume Controller component.
- With SVC nodes of model CF8 or CG8, configure the Fibre Channel switch ports with nodes attached to 41 F_Port buffer-to-buffer credits for distances beyond 2 km.
- With SVC nodes that are not capable to support 8 Gbit/s, the link speed has to be set to 2 Gbit/s for cable lengths beyond 4 km.
- Passive wavelength-division multiplexing (WDM) devices (that do not require electrical power) that rely on SFPs with different wavelengths (colored SFPs) can be used with the following requirements:

  - The WDM vendor supports the colored SFPs for usage in the WDM device.
  - The Fibre Channel switch vendor supports the colored SFPs.
  - IBM supports the WDM device for SVC Metro Mirror or Global Mirror.

  To purchase colored SFPs for passive WDM, please contact you WDM vendor.

- Do not use powered devices to provide distance extension (active WDM) for the SAN Volume Controller to switch connections.

SAN Volume Controller version 6.2.0 and newer

The same requirements as for version 6.1.0 apply (see section **Up to SAN Volume Controller version 6.1.0**). Additionally, with SVC nodes of model CF8 or CG8, the maximum cable length can be extended up to 40 km according to Table 2 below.

Minimum cable length	Maximum cable length	Maximum link Speed
>= 0 km	= 10 km	8 Gb/s
> 10 km	= 20 km	4 Gb/s
> 20 km	= 40 km	2 Gb/s

Table 2: Maximum cable length and according link speed

SAN Volume Controller version 6.3.0 and later

The same requirements as for versions 6.1.0 (see section **Up to SAN Volume Controller version 6.1.0**) and 6.2.0 (see section **SAN Volume Controller version 6.2.0 and newer**) apply. Additionally, active wavelength-division multiplexing (WDM) devices (that require electrical power for their operation) can be used for paths between SVC nodes with the requirements below:

- The Fibre Channel switch vendor supports the WDM device.

- IBM supports the WDM device for SVC Metro Mirror or Global Mirror.

Node-to-node paths with ISLs

With SAN Volume Controller versions 6.3.0 and later, you can use ISLs in paths between SVC nodes of the same I/O group. The cable distance between the two production sites must not exceed 300 km. However, because of potential performance impacts, IBM does not recommend to configure SVC split I/O group clusters beyond 100 km.

Using ISLs for node-to-node communication requires configuring two separate SANs:

- One SAN is dedicated for SVC node-to-node communication. This SAN is referred as *private SAN* below.
- One SAN is dedicated for host attachment, storage system attachment, and SVC Metro Mirror or Global Mirror. This SAN is referred as *public SAN* below.

Each SAN consists of at least one fabric that spans both production sites. At least one fabric of the public SAN includes also the quorum site. You can configure private and public SANs using different approaches:

- Dedicated Fibre Channel switches for each SAN, or
- switch partitioning features, or
- virtual or logical fabrics.

To implement private and public SANs with dedicated switches, any combination of supported switches can be used. For the list of supported switches and for supported switch partitioning and virtual fabric options please see the SVC interoperability website:

> Support for SAN Volume Controller (2145) website at
> http://www.ibm.com/storage/support/2145

Configure the SANs according to the rules below:

- Two ports of each SVC node are attached to fabrics of the public SANs.
- Two ports of each SVC node are attached to fabrics of the private SANs.
- A single trunk between switches is required for the private SAN.
- Hosts and storage systems are attached to fabrics of the public SANs. Links used for SVC Metro Mirror or Global Mirror are parts of public SANs.
- At least one fabric of the public SAN includes also the quorum site.
- ISLs of the fabrics belonging to the private SAN must not be shared and must not be over-subscribed.
- Passive wavelength-division multiplexing (WDM) devices (that do not require electrical power) that rely on SFPs with different wavelengths (colored SFPs) can be used with the following requirements:

- The WDM vendor supports the colored SFPs for usage in the WDM device.
- The Fibre Channel switch vendor supports the colored SFPs and the WDM device for ISLs.
- IBM supports the WDM device for SVC Metro Mirror or Global Mirror.

To purchase colored SFPs for passive WDM, please contact you WDM vendor.

- Active wavelength-division multiplexing (WDM) devices (that require electrical power for their operation) can be used with the requirements below:

  - The Fibre Channel switch vendor supports the WDM device for ISLs.
  - IBM supports the WDM device for SVC Metro Mirror or Global Mirror.

Quorum disk

A quorum disk is a managed disk (MDisk) that contains a reserved area that is used exclusively for system management. For general information about SVC cluster quorum disks see SVC Information Center.

In a split I/O group configuration, the active quorum disk must be located at a third site (the quorum site) as an independent failure domain. If communication is lost between the production sites, the site with access to the active quorum disk continues to process transactions. If communication is lost only to the active quorum disk (without impact on node-to-node communication), an alternative quorum disk at another site can become the active quorum disk.

Although a system of SAN Volume Controller nodes can be configured to use up to three quorum disks, only one quorum disk can be elected to resolve a situation where the system is partitioned into two sets of nodes of equal size. The purpose of the other quorum disks is to provide redundancy if a quorum disk fails before the system is partitioned.

Quorum disk placement

Generally, when the nodes in a system have been split among sites, configure the SAN Volume Controller system this way:

- Site 1: Half of SAN Volume Controller system nodes + one quorum disk candidate.
- Site 2: Half of SAN Volume Controller system nodes + one quorum disk candidate.
- Site 3: Active quorum disk.
- Disable the dynamic quorum configuration by using the `chquorum` command with the `-override yes` option. Important: The `chquorum` command has to be launched separately for the active quorum disk as well as for each quorum disk candidate.

This configuration ensures that a quorum disk is always available, even after a single site failure.

Several scenarios cause the SVC clustered system to change the active quorum disk. These changes may result in reduced availability of the system. The following scenarios describe examples that result in changes to the active quorum disk:

- Scenario 1:

  1. Either site 3 is powered off or connectivity from any other site to site 3 is broken.
  2. The system selects a quorum disk candidate at site 1 or 2 to become the active quorum disk.
  3. Either site 3 is powered on or connectivity to the site is restored.
  4. Assuming that the system was correctly configured initially, SAN Volume Controller automatically recovers the configuration when the power or connectivity is restored.

  In this scenario, the system returns automatically to same level of high availability as before without administrator intervention.

- Scenario 2:

  1. The storage system that is hosting the active quorum disk at site 3 is removed from the configuration.
  2. If possible, the system automatically configures a new quorum disk candidate at site 1 or 2.
  3. The system selects a quorum disk candidate at site 1 or 2 to become the active quorum disk.
  4. A new storage system is added to site 3.
  5. The SAN Volume Controller administrator must reassign all three quorum disks to ensure that the active quorum disk is now located at site 3 again.

  In this scenario, the system does not return automatically to the same level of high availability. Without administrator intervention, the system would not survive a failure of the site hosting the active quorum disk after step 3.

The storage system that provides the quorum disk in a split I/O group configuration at the third site must support “extended quorum” disks. Storage systems that provide extended quorum support are listed at the following Web site:

> Support for SAN Volume Controller (2145) website at
> http://www.ibm.com/storage/support/2145

Connectivity to quorum site

You have to connect the quorum site with both production sites via Fibre Channel or FCIP routers. Maximal one ISL hop is supported for connectivity between the nodes and the storage system with the quorum disk.

Figure 2 illustrates an example configuration. The storage system that hosts the third-site quorum disk is attached directly to a switch at both production sites using long-wave Fibre Channel connections. If either production site fails, you must ensure that the remaining site has retained direct access to the storage system that hosts the quorum disks.

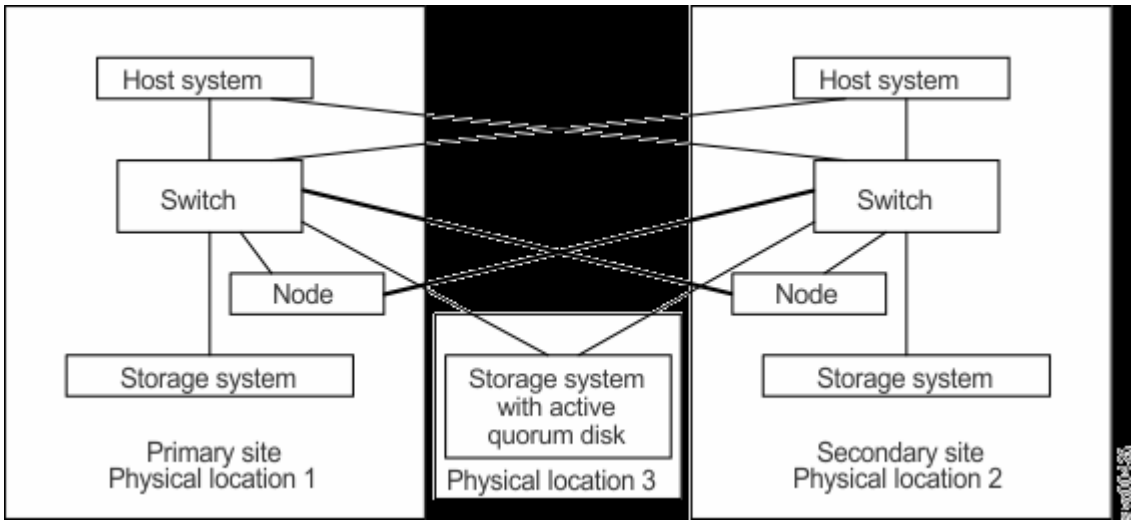

Figure 2: Example configuration

An alternative configuration can use two additional Fibre Channel switches at the third site with connections from one switch to the primary site and the other switch to the secondary site. Using only a single switch at the third site can lead to the creation of a single fabric rather than two independent and redundant fabrics and can introduce single points of failure that can defeat the objectives of creating a split site configuration.

Like for every managed disk, all SVC nodes need access to the quorum disk via the same storage system ports. If a storage systems with active/passive controllers (like IBM DS3000/4000/5000 or IBM FAStT) is attached to a fabric, then the storage system must be connected with both internal controllers to this fabric.

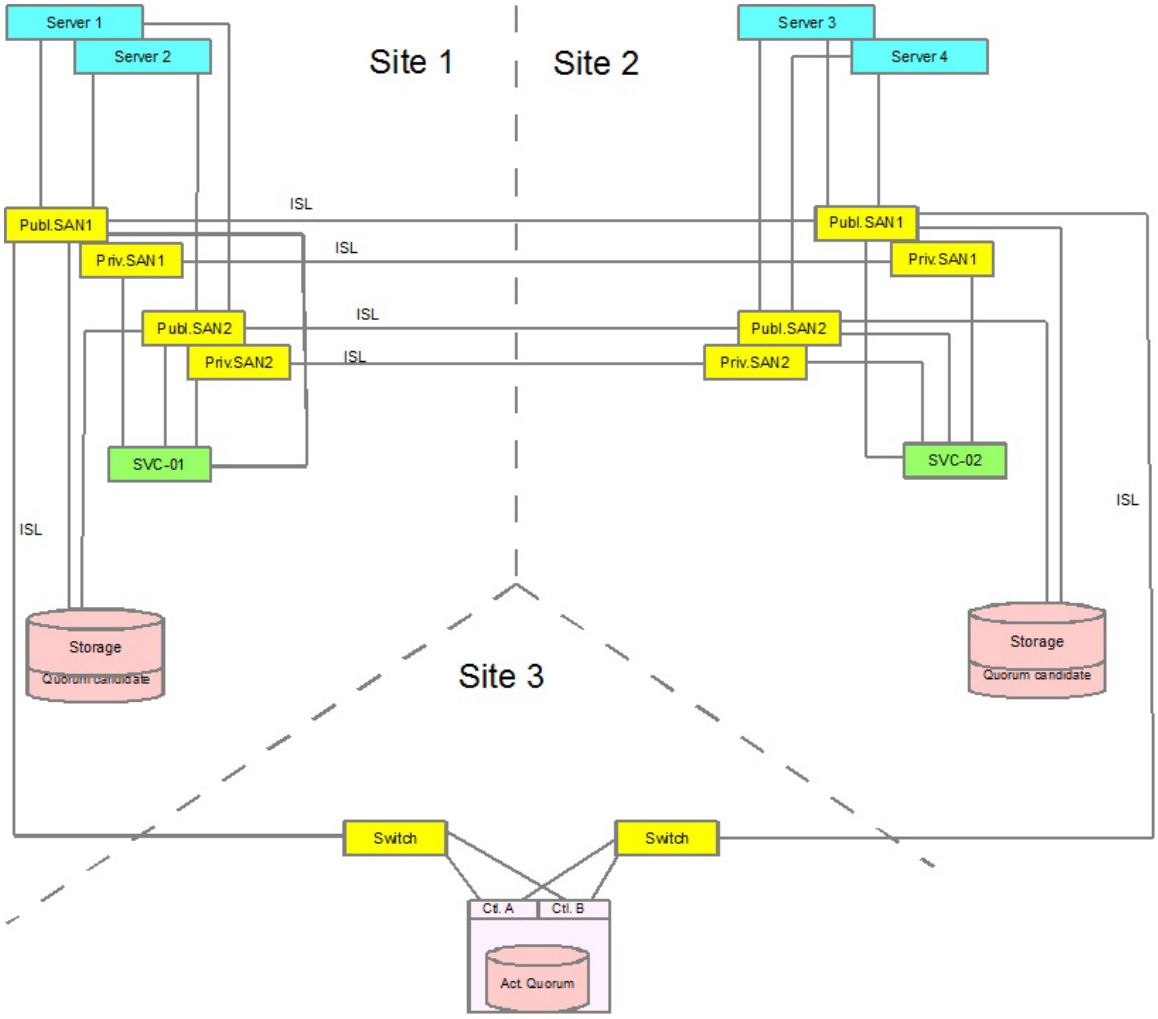

Figure3: Split I/O group with ISLs, DS3/4/5K connected to both fabrics

You can use distance extension via FCIP or WDM for quorum site connectivity with the requirements as described below. In any case, the connections have to be reliable. It is strictly required that the links from both production sites to the quorum site are independent and do not share any long-distance equipment.

FCIP routers can be used for quorum disk connections with SVC version 6.3.0 or later with the following requirements:

- The FCIP router device is supported for SVC remote mirroring (Metro Mirror or Global Mirror).
- The maximal round trip delay does not exceed 80 ms, that means 40 ms each direction.
- A minimal bandwidth of 2 MByte/s is guaranteed for node-to-quorum traffic.

Recommendation: It is a good practice to configure FCIP links so that they do not carry ISLs (to avoid fabric topology changes in case of IP errors).

Connections using **iSCSI** are not supported.

Passive wavelength-division multiplexing (WDM) devices (that do not require electrical power) can be used for quorum disk connections with SVC version 6.2.0 or later. Passive WDM relies on SFPs with different wavelengths (referred as *colored SFPs*) for fiber sharing. The requirements below apply:

- The WDM vendor supports the colored SFPs for usage in the WDM device.
- The Fibre Channel switch vendor supports the colored SFPs for ISL.
- IBM supports the WDM device for SVC Metro Mirror or Global Mirror.
- The SFPs must comply with the SFP/SFP+ power and heat specifications

To purchase colored SFPs for passive WDM, please contact you WDM vendor.

Active wavelength-division multiplexing (WDM) devices (that require electrical power for their operation) can be used for quorum disk attachment with SVC version 6.3.0 and later with the requirements below:

- The Fibre Channel switch vendor supports the WDM device for ISLs.
- IBM supports the WDM device for SVC Metro Mirror or Global Mirror.

It is not required to UPS-protect FCIP routers or active WDM devices that are used only for the node-to-quorum communication.

Quorum disk located in Metro Mirror or Global Mirror partner site

You can co-locate the storage system that provides the active quorum disk with another SVC cluster that acts as remote copy partner.

The quorum disk connection may use the same links and devices as the Metro Mirror or Global Mirror connection. The storage system providing the quorum disk must support "extended quorum" disks. Storage systems that provide extended quorum support are listed at the following Web site:

> Support for SAN Volume Controller (2145) website at
> http://www.ibm.com/storage/support/2145

Furthermore, if this storage system is attached to the remote copy partner cluster too, then it must be supported for split attachment. Please see the SVC Information Center website, chapter "Configuring and servicing external storage systems," if your storage system is supported for split attachment:

> SVC Information Center website at
> http://publib.boulder.ibm.com/infocenter/svc/ic/index.jsp

General SAN configuration rules

The rules below apply to all SVC split I/O group configurations with or without ISLs in node-to-node paths:

- Avoid using inter-switch links (ISLs) in paths between SAN Volume Controller nodes and external storage systems. If this is unavoidable, do not oversubscribe the ISLs because of substantial Fibre-Channel traffic across the ISLs. Because ISL problems are difficult to diagnose, switch-port error statistics must be collected and regularly monitored to detect failures.
- A SAN Volume Controller node must be located in the same rack as the 2145 UPS or 2145 UPS-1U that supplies its power.
- SAN Volume Controller nodes in the same cluster must be connected to the same Ethernet subnet. Ethernet port 1 on every SAN Volume Controller node must be connected to the same subnet or subnets, Ethernet port 2 (if used) of every node must be connected to the same subnet (this may be a different subnet from port 1). The same principle applies to other Ethernet ports.
- Using a single switch at the third site can:
  - lead to the creation of a single fabric rather than two independent and redundant fabrics,
  - can introduce single points of failure that can defeat the objectives of creating a split site configuration.
- Prior to SVC 6.1 some service actions require physical access to all SAN Volume Controller nodes in a cluster. If nodes in a split cluster (prior SVC 6.1) and separated by more than 100 meters, service actions might require multiple service personnel.
- For every storage system, create one zone that contains SAN Volume Controller ports from every node and all storage system ports, unless otherwise stated by the zoning guidelines for that storage system. However, do not connect a storage system in one site directly to a switch fabric in the other site. Instead, connect each storage system only to switched fabrics in the local site. (In split I/O group configurations with ISLs in node-to-node paths, these fabrics belong to the public SAN).