IBM System Storage SAN Volume Controller

Configuration Requirements and Guidelines

Version 4.2.0

1 June 2007

Configuration Rules for SVC

The SAN fabric configuration rules for SVC are documented in the publication "SAN Volume Controller Configuration Guide version 4.2.0" in chapter 2, "Configuring the SAN fabric".

The following points are particularly important to observe:

Host queue depth: It is important that hosts have their queue depths set as per Chapter 1: SAN Volume Controller Overview on the SVC Configuration Guide. If they are not correctly set, in certain degraded states, under a heavy load, the SAN can become flooded. This can stop SVC nodes from communicating with each other, possibly leading to lease expiries.

Performance of large fabrics: Whilst SVC supports up to 1024 hosts, care must be taken when planning a large configuration. Having determined the peak IO load, the user should ensure that the bandwidth of the switches is sufficient *even in a degraded mode* (e.g. loss of a redundant fabric). Users should consult their switch vendor for switch performance information.

Oversubscription for ISLs carrying Host to SVC traffic: For any ISLs which are only used for Host -> SVC traffic IBM recommends an ISL Oversubscription Ratio of no more than 7:1. If a higher oversubscription ratio is chosen this is likely to lead to congestion on the ISLs causing severe performance degradation and the possibility of IO errors on the host.

Oversubscription for ISLs carrying SVC to SVC or SVC to Storage traffic: IBM recommends that fabrics be configured to avoid any SVC to SVC or SVC to storage traffic flowing across ISLs. If this recommendation is not followed, then the oversubscription for these ISLs must not exceed 7:1, however IBM strongly recommends that a thorough SAN design analysis to avoid any congestion in the SAN. If congestion should occur on ISLs, this can cause performance degradation within the SVC, and the possibility of IO Errors on the hosts.

Oversubscription for ISLs in mixed-speed fabrics: Oversubscription calculations with respect to ISLs should take into account the speed of the links. Thus for ISLs running at X Gb/s, if the hosts run at Y Gb/s this means that the acceptable port oversubscription is 7*(X/Y). Note that the speed of the SVC ports is not relevant. For example, if the ISLs run at 4 Gb/s and the hosts at 2 Gb/s then we can tolerate an oversubscription of 14 host ports for every ISL port.

ISL Trunking: It should be noted that ISL trunking will reduce the likelihood of any ISL congestion by load balancing across multiple ISLs. IBM therefore recommends the use of this feature.

Meshed SANs: Meshed SANs are supported with SVC 4.1.0 and later.

Host zoning: For clusters with more that 64 hosts attached, the fabric must have an effective zoning configuration such that each HBA port is visible to just one port on each SVC node in the IO group(s) that this host will access. This is most easily achieved by using a separate zone for each HBA port and placing the relevant SVC node ports in these zones. Each HBA port will then appear in just one zone (unless additional zones are used to connect the HBA port to non SVC storage).

Split IO groups

This section discusses the specialised operation of an SAN Volume Controller cluster in SAN fabrics with long distance fibre links where the components within the SVC cluster a distributed over a large area. **This mode of operation is not normally recommended**.

1. An SVC cluster may be connected, via the SAN fabric switches, to application hosts, storage controllers or other SVC clusters, via short wave or long wave optical fibre channel connections with a distance of up to 300m (short wave) or 10 km (long wave) between the cluster and the host, other clusters and the storage controller. Longer distances are supported between SVC clusters when using inter-cluster Metro Mirror or Global Mirror.

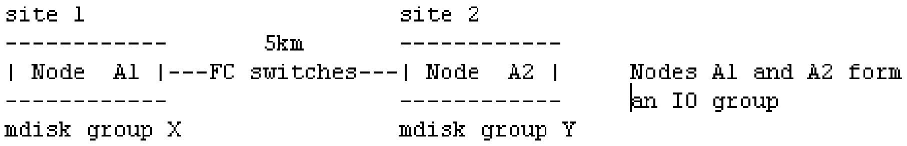
2. A cluster should be regarded as a single entity for disaster recovery purposes. This includes the backend storage that is providing the quorum disks for that cluster. This means that the cluster and the quorum disks should be co-located. Locating the components of a single cluster in different physical locations for the purpose of disaster recovery is not recommended, as this may lead to issues over maintenance, service and quorum disk management, as described below.

3. All nodes in a cluster should be located close to one another, within the same set of racks and within the same room. There may be a large optical distance between the nodes in the same cluster. However, they must be physically co-located for convenience of service and maintenance.

4. All nodes in a cluster must be on the same IP subnet. This is because the nodes in the cluster must be able to assume the same cluster or service IP address.

5. A node must be in the same rack as the UPS from which it is supplied.

Whilst splitting a single cluster into two physical locations might appear attractive for disaster recovery purposes, there are a number of practical difficulties with this approach. These difficulties, which do not apply in the case of the standard, two cluster solution, largely arise over the difficulty of managing a single quorum disk in a cluster that is distributed over two different physical locations. Consider the following configuration:



```
site 1                        site 2
------------       5km        ------------
| Node  A1 |---FC switches---| Node  A2 |      Nodes A1 and A2 form
------------                  ------------      an IO group
mdisk group X                 mdisk group Y
```

We have a node from each IO group at both sites and we set up Metro Mirror or Global Mirror relationships so that the primary vdisks at site 1 come from mdisks in group X (i.e. mdisks at site 1) and the secondary vdisks at site 2 come from mdisks in group Y (i.e. mdisks at site 2). It would appear that this arrangement will provide a means of recovering from a disaster at one or other site i.e. if site 1 fails, we have a live IO group (albeit it in degraded mode) at site 2 to perform the I/O workload. There are however a number of issues with this arrangement:

1. If either site fails, we only have a degraded IO group at the other site with which to continue I/O. Performance therefore during a disaster recovery is significantly impacted, since throughput of the cluster is reduced and the cluster caching is disabled.

2. The disaster recovery solution is asymmetric. Thus, it is not possible to run applications on both sites and allow either to suffer a failure. One site must be regarded as the primary site and the other is there to provide a recovery site. Consider the situation where the quorum disk is at site 2 (i.e. in mdisk group Y). If site 1 fails, then site 2 retains quorum and can proceed and act as a disaster recovery site. However, if site 2 were to fail, then site 1 cannot act as a disaster recovery site, since site 1 will only see half the nodes in the cluster and will not be able to see the quorum disk. The cluster components at site 1 will no longer form an active cluster (error code 550). It is not possible to communicate with the nodes at site 1 in this state and all I/O will immediately cease. An active cluster can only start

operating at site 1 if the quorum disk reappears or if a node from site 2 becomes visible. And in that case, it is likely, or at least possible, that site 2 might be able to resume operations anyway.

From the discussion above, it can be seen that the split cluster configuration can only provide asymmetric disaster recovery facilities, with substantially reduced performance. This is unlikely to be satisfactory for most production disaster recovery situations.

Splitting a cluster might be thought to be useful if the quorum disks are at a third "site", such that a disaster will only take down one of these three sites. However, even a three site configuration will have significant limitations, since SVC will not be able to change the path it uses to communicate with a quorum disk under all circumstances. Therefore, to be tolerant of a single site failure, it is necessary to ensure that the path to the quorum disk from a node in one site does not go through a switch in the second site before reaching the quorum disk in the third site. For example the following arrangement is acceptable:

```
site 1                       site 2
------------      5km        ------------
| Node  A1 |---FC switches---| Node  A2 |     A1/A2 are one I/O group
------------       |         ------------     and B1/B2 are the other
------------       |         ------------
| Node  B1 |       |         | Node  B2 |     Quorum device is at a
------------       |         ------------     third site
mdisk group X  Quorum Disk   mdisk group Y
                 site 3
```

On the other hand, the following configuration is unlikely to perform satisfactorily.

```
                                              A1/A2 are one i/o group
site 1              5km       site 2          and B1/B2 are the other
 FC switch --------------------- FC switch
     |     ------         ------     |        Nodes at site 1 are
------------     \       /     ------------   connected to one switch
| Node  A1 |      \     /      | Node  A2 |
------------       \   /       ------------   Nodes at site 2 are
------------       |   |       ------------   connected to another
| Node  B1 |       |   |       | Node  B2 |   switch
------------       |   |       ------------
 mdisk group X  Quorum Disk    mdisk group Y  Quorum device is at a
                  site 3                      third site and is
                                              connected to both sites
```

Note: All of the above discussion applies to a split cluster where there is significant optical distance between the nodes within a single cluster. Long distance (up to 10 km) connection of remote hosts or remote controllers are supported in an SVC cluster, as the issues of quorum disk and inter IO group links mentioned above are not relevant. Thus, the following configuration is acceptable:

```
site 1                       site 2
------------      10km       ---------------------
| Node  A1 |---FC switches---| application hosts |
------------                 ---------------------
------------
| Node  A2 |
------------
```

IBM recommends the use of two cluster configurations for all production disaster recovery systems. Customers who wish to use split cluster operation should contact their IBM Regional Advanced Technical Specialist for specialist advice relating to their particular circumstances.

Maximum Configurations

The following table shows the maximum configurations supported by a SVC cluster irrespective of the number of nodes it comprises, unless otherwise stated. Not all maxima can be supported simultaneously.

Objects	Maximum Number	Comments
SVC Cluster		
SAN Volume Controller nodes	8	Arranged as four I/O groups
I/O Groups	4	Each containing two nodes
Fabrics	4	The number of counterpart SANs which are supported
Fibre channel logins per SVC Node Port	512 - Cisco, McDATA and Brocade fabrics 256 - CNT	The number of FC ports which can log into a single SVC Port. These include other SVC ports, storage ports and host ports. This is controlled by zoning on the switches
Managed Disks		
Managed disks (mdisks)	4096	The maximum number of logical units which can be managed by SVC. This number includes disks which have not been configured into managed disk groups
Managed disk groups	128	
Mdisks per mdisk group	128	
Mdisk size	2 TB	
TotalStorage manageable by SVC	2.1 PB	If extent size of 512 Mb is used
Virtual Disks		
Virtual disks (vdisks) per cluster	4096	Includes managed-mode vdisks and image-mode vdisks. Maximum requires an 8 node cluster
Vdisks per I/O group	1024	
Vdisks per mdisk group	N/A	Cluster limit applies
Vdisk size	2 TB	
Vdisks per host object	512	The limit may be different based on host operating system. See Host Attachment Guide for details
SDD	512 SAN Volume Controller vpaths per host	One vpath is created for each vdisk mapped to a host. Although the SAN Volume Controller permits 512 vdisks to be mapped to a host, the SDD limit can be exceeded by either: • Creating two (or more) host objects for one physical host and mapping more than 512 vdisks to the host

		using the multiple host objects • Creating two (or more) clusters and mapping more than 512 vdisks to the host using the multiple clusters Note: Both of these operations are unsupported for SDD
SDDPCM (on AIX)	12,000 vpaths per host	
Vdisks-to-host mappings	20,000	
Hosts / Servers		
Host IDs per cluster	1024 – Cisco, Brocade and McDATA fabrics 155- CNT 256 - Qlogic	A Host ID is a collection of worldwide port names (WWPNs) which represents a host. This is used to associate SCSI LUNs with vdisks See Also – Host IDs per I/O group below For Brocade support, please see ***Note 2*** below this table
Host ports per cluster	2048 - Cisco, McDATA and Brocade fabrics 310 – CNT 512 - Qlogic	
Host IDs per I/O group	256 - Cisco, McDATA and Brocade fabrics N/A – CNT 64 - Qlogic	
Host ports per I/O group	512 - Cisco, McDATA and Brocade fabrics N/A – CNT 128 - Qlogic	

Host ports per host ID	512	
Copy Services		
Metro Mirror or Global Mirror relationships per cluster	1024	
Metro Mirror or Global Mirror consistency groups	256	
Metro Mirror and Global Mirror vdisk per I/O group	40 TB	The total size of all Metro Mirror source and target and all Global Mirror source and target vdisks in an I/O group must not exceed 40 TB
FlashCopy® targets per source	16	
FlashCopy® mappings	3855	Calculated by noticing that 240 source vdisks with 16 FC mappings plus one more with 15 mappings uses up all 4096 vdisks per cluster
FlashCopy® mappings per consistency group	512	
FlashCopy consistency groups	128	
FlashCopy vdisk per I/O group	40 TB	The total size of all FlashCopy targets of source vdisks in an I/O group must not exceed 40 TB
SVC Nodes		
Concurrent SCSI tasks (commands) per node	10,000	
Concurrent commands per FC port	2048	
Storage Controllers		
Storage controller WWNNs	64	Some storage controllers have a separate WWNN per port e.g. Hitachi Thunder
Storage controller WWPNs	256	
LUNs per storage controller WWNN	4096	
WWNNs per storage controller	4	The number of WWNN per storage controller (Usually 1)
WWPNs per WWNN	16	The maximum number of FC ports per worldwide node name

Note 1: Fabric and Device Support

A statement of support for a particular fabric configuration here is reflects the fact that SVC has been tested and is supported for attachment to that fabric configuration. Similarly a statement that SVC supports attachment to a particular backend device or host type reflects the fact that SVC has been tested and is supported for that attachment. SVC is only supported however for attachment to particular devices in a given fabric vendor if IBM and that fabric vendor both support that attachment. It is the user's responsibility to verify that this is true for the particular configuration of interest as it is impossible to list individual 'support' or 'no support' statements for every possible intermix of front end and backend devices and fabric types.

Note 2: Support for Brocade fabrics

Support for Brocade fabrics with up to 1024 hosts is available on SVC v4.1.0.3 and later with the following restrictions.

1. Fabrics that use M14, B64 or M48 switches in the core are supported. Any other supported Brocade switches may be used as edge switches in this configuration. The SVC ports and backend storage must all be connected to the core switches.

2. The core M48 and B64s must be running at least the 5.1.0c firmware

3. The core M14s must be running at least the 5.0.5a firmware.

Support of Brocade fabrics with up to 256 hosts is available on SVC v3.1.0.3 and later with the following restrictions.

1. Only core-edge fabrics that use M14 or M48 switches in the core are supported. Any other supported Brocade switches may be used as edge switches in this configuration. The SVC ports and backend storage must all be connected to the core switches.

2. Each SVC port must not see more than 256 N port logins. Error code 1800 is logged if this limit is exceeded on a Brocade fabric.

3. Each I/O group may not be associated with more than 64 host objects.

4. A host object may be associated with one or more I/O groups - if it is associated with more than one I/O group it counts towards the max 64 total in all of the I/O groups it is associated with.

Note 3: Example bitmap usage

Flash Copy: Each I/O group supports up to 40TB of target vdisks. The target vdisks may be in any I/O group. For the purpose of this limit vdisks are rounded up to a multiple of 8GB so 512 vdisks of 24.1GB will use all the bitmap space even though 512*24.1GB is less than 40TB.

Example 1: You can make 1 copy of 40TB of vdisks in an I/O group to another 40TB of vdisks anywhere in the cluster

Example 2: You can make 1 copy of 160TB of vdisks (40TB per I/O group) to another 160TB of vdisks anywhere in the cluster

Example 3: You can make 10 copies of 4TB of vdisks in an I/O group onto another 40TB of vdisks anywhere in the cluster

Metro Mirror and Global Mirror: Each I/O group supports up to 40TB of primary+secondary vdisk (that is 40TB shared between primary and secondary vdisks, not 40TB each). The 40TB can be split in any ratio between primary and secondary vdisk. For the purpose of this limit vdisks are rounded up to a multiple of 8GB. Metro Mirror and Global Mirror share the same bitmap memory - therefore the sum of primary Metro Mirror vdisks + primary Global Mirror vdisks + secondary Metro Mirror vdisks + secondary Global Mirror vdisks per I/O group is limited to 40TB.

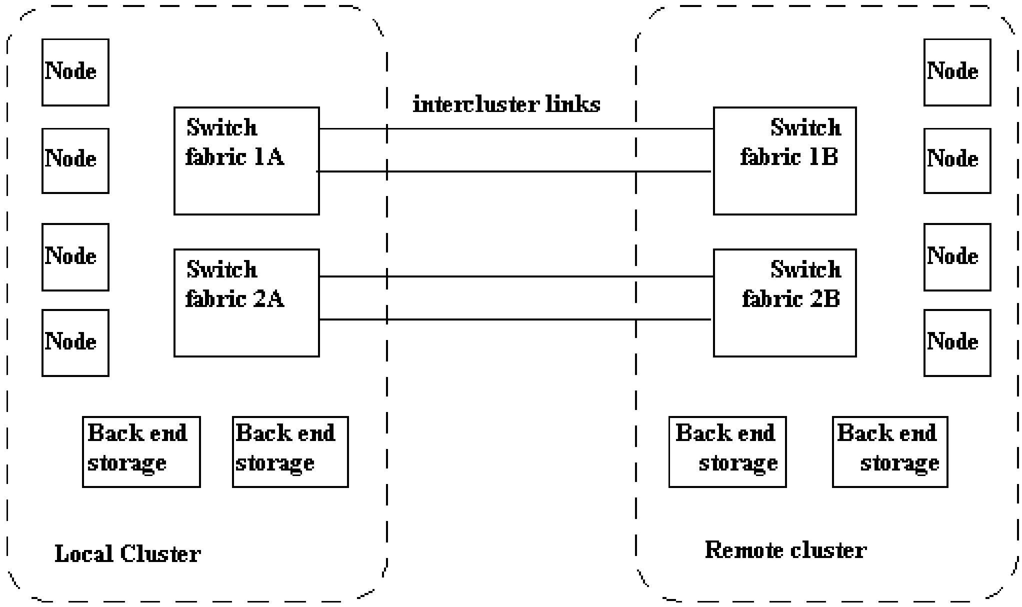
Example 1: You can use Metro Mirror or Global Mirror to copy 40TB of vdisks per I/O group to a secondary cluster for disaster recovery. You can also make a FlashCopy of your data at your primary and/or secondary cluster for backup

Example 2: You can use intra-cluster Metro Mirror or Global Mirror to copy 20TB of vdisks per I/O group to another 20TB of vdisks in the same I/O group.

Metro Mirror and Global Mirror SVC Configurations

SVC supports both intra-cluster and inter-cluster Metro Mirror and Global Mirror. From the intra-cluster point of view, any single cluster is a reasonable candidate for Metro Mirror or Global Mirror operation. Inter-cluster operation on the other hand will need a pair of clusters, separated by a number of moderately high bandwidth links. Such a configuration is shown in the figure below. Note that intra-cluster Metro and Global Mirror is between vdisks in the same IO group only.

Metro Mirror or Global Mirror configuration using dual redundant fabrics



This contains two redundant fabrics. Part of each fabric exists at the local and remote cluster. There is no direct connection between the two fabrics.

Technologies for extending the distance between two SVC clusters can be broadly divided into two categories:

Fibre Channel Extenders

Fibre channel extenders simply extend a fibre channel link by transmitting fibre channel packets across long links without changing the contents of those packets. Examples include:

- FCIP extenders implemented in Cisco MDS 9500 series switches
- CNT Ultranet Edge Storage Router
- Any Multiprotocol Router **only when used in FCIP tunnelling mode**
    - For example the Brocade Multiprotocol Router when used in FCIP tunnelling mode
- DWDM/CWDM and Longwave GBIC extenders
- Ciena CN2000 distance extenders

Fibre channel extenders of all types are supported for use with SVC when planned, installed and operated as described below under “Configuration Requirements for Long Distance Links”.

SAN Routers

SAN Routers extend the scope of a SAN by providing "virtual nPorts" on two or more SANs. The router arranges that traffic at one virtual nPort is propagated to the other virtual nPort but the two fibre channel fabrics are independent of one another. Thus nPorts on each of the fabrics cannot directly log into each other.

Due to the more complex interactions involved, IBM explicitly tests products of this class for interoperability with SAN Volume Controller. The current list of supported SAN routers can be found in the supported hardware list on the SAN Volume Controller support web site at http://www.ibm.com/storage/support/2145.

These are supported for use with SVC when planned, installed and operated as described below under “Configuration Requirements for Long Distance Links”.

Configuration Requirements for Long Distance Links

IBM has tested a number of fibre channel extender and SAN router technologies with SVC. These must be planned, installed and tested such that the following requirements are met:

- **For SVC 4.1.0.x**, the round-trip latency between sites must not exceed 68ms (34ms one-way) for fibre channel extenders, or 20ms (10ms one-way) for SAN routers.

- **For SVC 4.1.1.x** and later, the round-trip latency between sites must not exceed 80ms (40ms one-way). For Global Mirror this would allow a distance of between primary and secondary sites of up to 8000 km using a planning assumption of 100km per 1ms of round trip link latency.

  Note: The latency of long distance links depends upon the technology used to implement them. A point to point dark fibre-based link will typically provide a round trip latency of 1ms per 100km or better, other technologies will provide longer round trip latencies and this will affect the maximum supported distance.

- The configuration must be tested with expected peak workloads.

- When Metro Mirror or Global Mirror is being used, a certain amount of bandwidth will be required for SVC inter-cluster heartbeat traffic. The amount of traffic depends on how many nodes are in each of the two clusters. The table below shows the amount of traffic, in megabits per second, generated by different sizes of cluster.
  These numbers represent the total traffic between the two clusters, when no I/O is taking place to mirrored vdisks. Half of the data is sent by one cluster, and half by the other cluster. The traffic will be divided evenly over all available inter-cluster links; therefore, if

you have two redundant links, half of this traffic will be sent over each link, during fault-free operation.

SVC inter-cluster heartbeat traffic (Megabits per second):

		Cluster 2			
		2 nodes	**4 nodes**	**6 nodes**	**8 nodes**
Cluster 1	**2 nodes**	2.6	4.0	5.4	6.7
	4 nodes	4.0	5.5	7.1	8.6
	6 nodes	5.4	7.1	8.8	10.5
	8 nodes	6.7	8.6	10.5	12.4

- The bandwidth between sites must be at least sized to meet the peak workload requirements while maintaining the maximum latency specified above. The peak workload requirement must be evaluated by considering the average write workload over a period of 1 minute or less plus the required synchronisation copy bandwidth. With no synchronisation copies active and no write IO disks in Metro Mirror or Global Mirror relationships the SVC protocols will operate with the bandwidth indicated in the table above, but the true bandwidth required for the link can only be determined by considering the peak write bandwidth to Virtual Disks participating in Metro Mirror or Global Mirror relationships and adding to it the peak synchronisation copy bandwidth.

- If the link between the sites is configured with redundancy so that it can tolerate single failures then the link must be sized so that the bandwidth and latency statements continue to hold true even during such single failure conditions.

- The channel must **not** be used for links between nodes in a single cluster. Using long distance links within a single cluster is not supported and can lead to IO errors and loss of access.

- The configuration is tested to simulate failure of the primary site (to test the recovery capabilities and procedures) including eventual failback to the primary site from the secondary.

- The configuration is tested to confirm that any failover mechanisms in the inter-cluster links interoperate satisfactorily with SVC.

- All other SVC configuration requirements are met. Particular attention is drawn to the rules surrounding interoperability between fibre channel switches from different vendors. The presence of a fibre channel extender does not change the restrictions on the interoperability of different vendors' products in SVC configurations. The fibre channel extender should be treated as a normal link when reasoning about interoperability rules.

The bandwidth and latency measurements must be made by or on behalf of the client and are not part of the standard installation of SVC by IBM. IBM recommends that these measurements are made during installation and that records are kept. Testing should be repeated following any significant changes to the equipment providing the inter-cluster link.

Limitations on host to cluster distances and use of the intercluster link for host traffic

Two common questions arise when designing an SVC disaster recovery configuration relating to host to SVC distances. These are:

How far away can a host be from the SVC cluster that supplies its vdisk?

Can a host server access vdisks or hosts that are in the 'remote cluster' by means of the intercluster link?

The answer to the first question is that SVC imposes no special limit on the fibre channel optical distance between SVC nodes and host servers. A server may therefore be attached to an edge switch in a core-edge configuration while the SVC cluster would be at the core. SVC supports up to 3 ISL hops in the fabric and this means that the server and the SVC cluster may be separated by up to 5 actual fibre channel links, four of which can be 10km long, if long-wave SFPs are used. (The SVC nodes themselves have short wave SFPs and must therefore be within 300m of the switch they are attached to). The following configuration is therefore supported:

```
      10km ---- 10km  ---- 10km  ---- 10km ---- 10km
host1------|sw|-------|sw|-------|sw|------|sw|------host2
           ----       ----       ----      ----
            |300m
           svc_1
```

Here the optical distance between cluster svc_1 and host2 is just over 40km.

The second question relates to a two cluster configuration where the hosts in a local cluster may wish to read/write the vdisks belonging to a remote cluster, or they may wish to exchange heartbeats (if they are in a host cluster) with hosts located at the remote site. An extreme example would be:

```
      10km ---- 10km  ----  10km ---- 10km ----
host1------|sw|-------|sw|-------|sw|------|sw|---intercluster link ---
           ----       ----       ----      ----     40 ms latency     |
                                             |                        |
                                           svc_1                      |
                                                                      |
              ---------------------------------------------------------
              |
              |
              |
              |       ---- 10km  ----  10km ---- 10km ---- 10km
              --------|sw|-------|sw|-------|sw|------|sw|------ host2
                      ----       ----       ----      ----
                        |
                      svc_2
```

Here host1 may access vdisks supplied by cluster svc_2 as well as exchange heartbeats with host2. This configuration is supported; however, since the intercluster link is now being used for multiple purposes, sufficient bandwidth must be provisioned to satisfy all three possible sources of load listed below:

- SVC remote copy requirements for Global or Metro Mirror data transfers and SVC cluster heartbeat traffic
- Local host to remote vdisk I/O traffic or remote host to local vdisk I/O traffic (e.g. when verifying that a particular mirrored copy is useable)
- Local host to remote host heartbeat traffic (e.g. HACMP/XD heartbeat)

For example, if the local host to remote vdisk traffic is allowed to consume too much inter-cluster link bandwidth, this is likely to impact the latency seen by hosts that access SVC vdisks that are participating in remote copy operations. In the worst case scenario, this could lead to the Global

Mirror link tolerance threshold being exceeded and the Global Mirror relationships stopping due to bandwidth congestion caused by too much local host to remote vdisk I/O.

Global mirror guidelines

When using SVC Global Mirror, all components in the SAN must be capable of sustaining the workload generated by application hosts, as well as the Global Mirror background copy workload. If this is not true, then Global Mirror may automatically stop your relationships to protect your application hosts from increased response times. Therefore, it is important to configure each component correctly. The requirements and guidelines below will allow you to do so.

In addition, you should use a SAN performance monitoring tool, such as IBM TotalStorage Productivity Center, which will allow you to continuously monitor the SAN components for error conditions and performance problems. This will assist you to detect potential issues before they impact your disaster recovery solution.

Long-distance link requirements

The long-distance link between the two clusters must be provisioned to allow for:

- **the peak application write workload to the Global Mirror source vdisks, plus**
- **the customer-defined level of background copy.**

The long-distance link must have a round-trip latency of 80ms or less.

The peak application write workload should be ideally be determined by analysing SVC performance statistics.

Statistics should be gathered over a typical application IO workload cycle, which may be days, weeks, or months depending on the environment in which SVC is used. These statistics should be used to find the peak write workload which the link must be able to support.
Characteristics of the link may change with use: for example, the latency may increase as the link is used to carry an increased bandwidth. The user should be aware of the link’s behaviour in such situations, and ensure that the link remains within the specified limits. If the characteristics are not known, testing should be performed to gain confidence of the link’s suitability.

Users of Global Mirror should consider how to optimise the performance of the long-distance link. This will depend upon the technology used to implement the link. For example, when transmitting FC traffic over an IP link, it may be desirable to enable jumbo frames to improve efficiency.

SVC configuration requirements

IBM TotalStorage Productivity Center (TPC) or an equivalent SAN performance analysis tool must be configured to monitor your SAN. SVC produces many performance statistics which are useful when external problems cause Global Mirror to report an error. TPC provides an easy way to monitor and analyse these statistics; if you do not use TPC, you should use another tool which provides equivalent functionality.

The Global Mirror partnership’s background copy rate must be set to a value appropriate to the link and secondary back-end storage.

It is supported to use Global Mirror and Metro Mirror between the same two clusters. However, customers intending to do this should read the section “Using both Metro Mirror and Global Mirror between two clusters” below.

It is not supported for cache-disabled vdisks to participate in a Global Mirror relationship.

The gmlinktolerance parameter of the remote copy partnership must be set to an appropriate value. (See “Using the gmlinktolerance configuration option” below for more details). The default value is 300 seconds (5 minutes), which will be appropriate for most customers. During SAN maintenance, the user should either:

- Reduce application IO workload for the duration of the maintenance (such that the degraded SAN component(s) are capable of the new workload), or
- Disable the gmlinktolerance feature, or increase the gmlinktolerance value (meaning that application hosts may see extended response times from Global Mirror vdisks), or
- Stop the Global Mirror relationships

If the gmlinktolerance value is increased for maintenance lasting *x* minutes, it should be only be re-set to the normal value *x* minutes after the end of the maintenance activity. If gmlinktolerance is disabled for the duration of the maintenance, it should be re-enabled once the maintenance is complete.

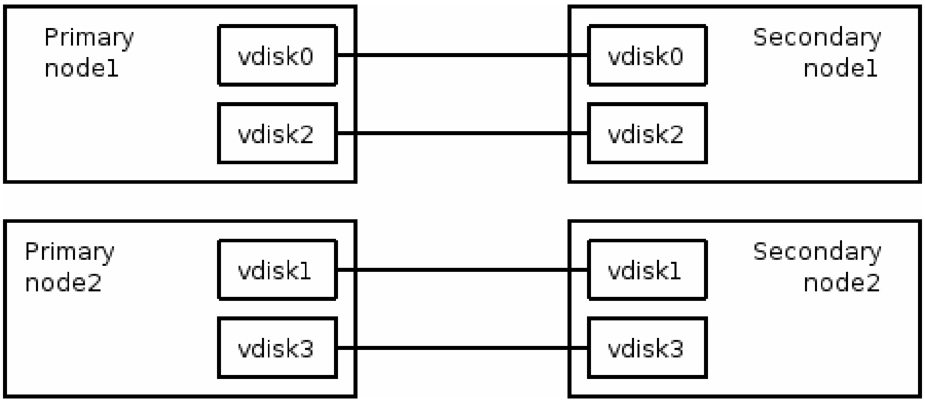
Global Mirror vdisks should have their preferred nodes evenly distributed between the nodes of the clusters
Each vdisk within an IO group has a preferred node property that can be used to balance the IO load between nodes in that group. This property is also used by Global Mirror to route IO between clusters.

In a fully available system, the node that receives a write for a vdisk will normally be that vdisk's preferred node. If that vdisk is in a Global Mirror relationship, the node is responsible for sending that write to the preferred node of the secondary vdisk. The primary preferred node is also responsible for sending any writes relating to background copy - again these are sent to the preferred node of the secondary vdisk.

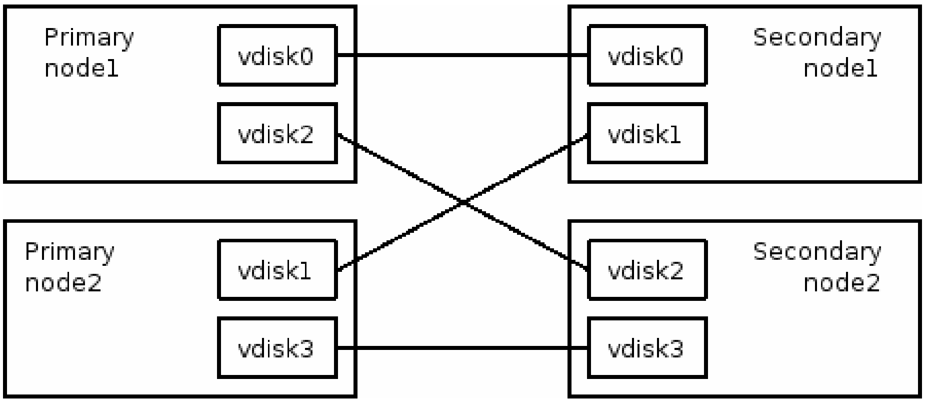
The preferred node property, unless provided on the command-line, alternates between the nodes of an IO group as vdisks are created within it. Note that it cannot be changed after creation.

Each node of the secondary cluster has a fixed pool of Global Mirror system resource for each node of the primary cluster. In other words, each secondary node has a separate queue for I/O from each of the primary nodes. This queue is of a fixed size, and is the same for every node. If preferred nodes for the vdisks of the secondary cluster are set such that every combination of primary node and secondary node is used, Global Mirror performance will be maximised.

To help illustrate the point, consider an example where a user creates 4 vdisks on both their primary and secondary clusters without specifying the preferred node. They then create four Global Mirror relationships between corresponding vdisks - vdisk0->vdisk0, vdisk1->vdisk1 etc. The diagram below shows a configuration where the resources of secondary node 1 that are allocated for communication with primary node 2 are not used. The same is true of the resources of secondary node 2 used for communication with primary node 1.



In contrast, the following diagram shows a system where all Global Mirror resources are being used because each secondary node is communicating with every primary node. This has been achieved by careful selection of the preferred node of the vdisks on each cluster.



For a configuration that has more than one IO group on either cluster, there are many more primary-secondary node combinations that must be considered when balancing the load.

Back-end storage controller requirements

The capabilities of the storage controllers at the secondary cluster must be provisioned to allow for:

- The peak application workload to the Global Mirror vdisks, plus
- The customer-defined level of background copy, plus
- Any other IO being performed at the secondary site.

The performance of applications at the primary cluster can be limited by performance of the back-end storage controllers at the secondary cluster. To maximise the amount of IO that applications may make to Global Mirror vdisks:

- Global Mirror vdisks at the secondary cluster should be in dedicated mdisk groups (which contain no non-Global Mirror vdisks)
- Storage controllers should be configured to support the Global Mirror workload that is required of them. This might be achieved by:
    - Dedicating storage controllers to only Global Mirror vdisks
    - Configuring the controller to guarantee sufficient quality of service for the disks being used by Global Mirror
    - Ensuring that physical disks are not shared between Global Mirror vdisks and other IO (e.g., by not splitting an individual RAID array)

- Mdisks within a Global Mirror mdisk group should be similar in their characteristics (e.g., RAID level, physical disk count, disk speed). This is true of all mdisk groups, but is particularly important to maintain performance when using Global Mirror.

Using TotalStorage Productivity Center to monitor Global Mirror performance

It is important to use a SAN performance monitoring tool to ensure that all SAN components are performing correctly. While this is useful in any SAN environment, it is particularly important when using an asynchronous mirroring solution such as SVC Global Mirror. IBM TotalStorage Productivity Center (TPC) is capable of constantly monitoring key performance measures, and alerting the user if thresholds are exceeded. Performance statistics should be gathered at the highest possible frequency; for TPC, this is currently 5 minutes. Note that if your vdisk or mdisk configuration is changed, you should restart your TPC performance report, to ensure that performance is correctly monitored for the new configuration. If using TPC, the following should be monitored:

- *Port to Remote Node Send Response Time* should be less than 80ms (the maximum latency supported by SVC Global Mirror). A figure in excess of this suggests that the long-distance link has excessive latency, which should be rectified. One possibility to investigate is that the link is operating at maximum bandwidth.
- Sum of *Port to Local Node Send Response Time* and *Port to Local Node Send Queue Time* should be less than 1ms for the primary cluster. A figure in excess of this may indicate that an IO group is reaching its IO throughput limit, which may limit performance. For the same reason, *CPU Utilization Percentage* should be below 50%.
- Sum of *Backend Write Response Time* and Write Queue Time for Global Mirror mdisks at the secondary cluster should be less than 100ms. A longer response time may indicate that the storage controller is overloaded. If the response time for a given storage controller is outside of its specified operating range, this should be investigated for the same reason.
- Sum of *Backend Write Response Time* and *Write Queue Time* for Global Mirror mdisks at the *primary* cluster should also be less than 100ms. If response time is greater than this, then application hosts may see extended response times if SVC's cache becomes full.
- *Write Data Rate* for Global Mirror mdisk groups at the secondary cluster indicates the amount of data being written by Global Mirror. If this figure approaches either the inter-cluster link bandwidth, or the storage controller throughput limit, then you should be aware that further increases may cause overloading of the system, and monitor appropriately.

Using SVC Global Mirror delay simulation

SVC Global Mirror should be tested using the same type (bandwidth, latency) of inter-cluster link that will be used in production. If you wish to test using two clusters at the same site, without a long-distance link, there are two possibilities:

- Use dedicated Fibre-Channel delay simulation hardware to artificially delay Fibre-Channel communications between the two clusters
- Use the *gminterdelaysimulation* or *gmintradelaysimulation* cluster settings at the primary SVC cluster. These specify a number of milliseconds of round-trip delay to add to each Global Mirror communication between the two clusters, for inter-cluster and intra-cluster Global Mirror relationships respectively. See the Command-Line Interface Guide (and Errata, if available) for further information on these settings.

Using Flash Copy to create a consistent image before restarting a Global Mirror relationship

When a consistent relationship is stopped, for example, by a persistent IO error on the inter-cluster link, the relationship enters the *consistent_stopped* state. IO at the primary site continues but the updates are not mirrored to the secondary site. Restarting the relationship will begin the process of synchronizing new data to the secondary disk. While this is in progress, the relationship will be in the *inconsistent_copying* state. This means that the Global Mirror secondary vdisk will not be in a usable state until the copy has completed and the relationship has returned to a consistent state.
Therefore, it is highly advisable to create a Flash Copy of the secondary vdisk before restarting the relationship. Once started, the Flash Copy will provide a consistent copy of the data, even while the Global Mirror relationship is copying. If the Global Mirror relationship does not reach the synchronized state (if, for example, the inter-cluster link experiences further persistent IO errors), then the Flash Copy target can be used at the secondary site for disaster recovery purposes.

The SVCTools package available on the IBM Alphaworks contains an example script which demonstrates how this Flash Copy process might be managed. See the “copymanager” script and documentation within this package for details. The package is available from:
http://www.alphaworks.ibm.com/tech/svctools/download

Using both Metro Mirror and Global Mirror between two clusters

Global Mirror and Metro Mirror relationships may exist simultaneously between two clusters. However, consideration should be given to the performance implications of this configuration, as write data from all mirroring relationships will be transported over the same inter-cluster link(s).

Metro Mirror and Global Mirror respond differently to a heavily-loaded, poorly performing link. Metro Mirror will usually maintain the relationships in a copying or synchronized state, meaning that primary host applications will start to see poor performance (as a result of the synchronous mirroring being used). Global Mirror, on the other hand, offers a higher level of write performance to primary host applications. With a well-performing link, writes are completed asynchronously. If link performance becomes unacceptable, the link tolerance feature automatically stops Global Mirror relationships, to ensure that performance for application hosts remains within reasonable limits.

Therefore, with active Metro Mirror and Global Mirror relationships between the same two clusters, Global Mirror writes may suffer degraded performance, if Metro Mirror relationships consume most of the inter-cluster link's capability. If this degradation reaches a level where hosts writing to Global Mirror would experience extended response times, the Global Mirror relationships may be stopped when the link tolerance threshold is exceeded. If this occurs, please refer to the "diagnosing and fixing 1920 errors triggered by the gmlinktolerance function" section of this document.

Migrating a Metro Mirror relationship to Global Mirror

Can I/O to the mirrored vdisks be stopped during the migration?
If so, it is recommended that the Global Mirror relationship is created as “synchronized”. This means that SVC will assume that the data on both vdisks is the same, and so no synchronization is required. This will mean that creating the relationship will be much faster (taking just a few seconds).

To do this, firstly cease all host IO to the primary vdisk. Then verify that the Metro Mirror relationship is consistent, and delete the relationship. Next, create the Global Mirror relationship between the same two vdisks and use the "-sync" flag to the "svctask mkrcrelationship" command (or the equivalent option in the SVC Console GUI). This tells SVC to assume that the data on the secondary vdisk is already identical to the primary vdisk, and so the relationship will be in "consistent_synchronized" state when it is created. Once the relationship is created, the relationship may be started and host IO may be resumed.
If following this approach, note that if the Metro Mirror relationship is not consistent when stopped, or if any host IO takes place between stopping the Metro Mirror relationship and creating the Global Mirror relationship, the data on the secondary vdisk will be invalid, and those changes will never be mirrored to the secondary disk. Therefore care should be taken to avoid this.

If I/O cannot be stopped, the data on the secondary vdisk will become out-of-date. Therefore, when the Global Mirror relationship is started, the secondary vdisk will be inconsistent, and will remain so until all of the recent changes have been copied to the remote site.
If you do not require a consistent copy at the remote site during these copies, then simply delete the Metro Mirror relationship, and create and start a Global Mirror relationship between the same two vdisks. Remember that the data on the secondary vdisk will not be usable until the synchronization is complete, which could take a long time, depending on your link capabilities and the amount of data being copied. The inter-cluster partnership's background copy bandwidth should be set such that the inter-cluster link is not overloaded.

If you *do* require a consistent copy at the secondary site, you may preserve the Metro Mirror vdisk after deleting the Metro Mirror relationship, by creating the Global Mirror relationship with a different secondary vdisk. If at a later point you need a consistent copy, the preserved vdisk can be used. Alternatively, you may start a Flash Copy of the Metro Mirror target vdisk. Once the Flash Copy is in place, you may create the Global Mirror relationship with the same secondary vdisk; the Flash Copy target will be your consistent copy.

Using the gmlinktolerance function

SVC 4.1.1 introduces a new cluster configuration option, called "gmlinktolerance". This is set using the *svctask chcluster –gmlinktolerance <value>* CLI command, or the SVC Console GUI. It represents the number of seconds for which the primary SVC cluster will tolerate slow response times from the secondary cluster – for example, because the long-distance link is congested, or the secondary back-end storage is overloaded.

If the poor response times continue beyond this period, then an error will be logged ("1920: A Global Mirror or Metro Mirror (Remote Copy) relationship has stopped due to poor performance."), and one or more Global Mirror relationships will automatically be stopped. This is to protect the application hosts at the primary: during normal operation, application hosts will see minimal impact to response times, due to Global Mirror's asynchronous replication. However, if Global Mirror experiences degraded response times from the secondary cluster for an extended period of time, then IO will begin to queue at the primary cluster, resulting in extended response time to application hosts. Therefore, the gmlinktolerance feature will eventually stop Global Mirror relationships, and the application host response time will return to normal.

Once a 1920 error has occurred, your Global Mirror auxiliary vdisks will no longer be consistent_synchronized, until you rectify the cause of the error and restart your Global Mirror relationships. For this reason, you should monitor the cluster (perhaps using SNMP trap monitoring) so that the administrator is informed when this happens.

The gmlinktolerance function can be disabled. This is achieved by setting the gmlinktolerance value to '0'. Most customers should have gmlinktolerance enabled in day-to-day operation to protect application hosts from extended response times. You should satisfy yourself that extended response times will not be a problem for your applications and should monitor performance using a tool such as TotalStorage Productivity Center to ensure that performance is sufficient.

It might be appropriate to disable gmlinktolerance in these circumstances:

- During SAN maintenance windows where degraded performance is expected from SAN components, and application hosts can withstand extended response times from Global Mirror vdisks.
- During other periods when application hosts can tolerate extended response times, and it is expected that the gmlinktolerance feature would otherwise stop the Global Mirror relationships. For example, if you are testing using an IO generator which is configured to stress the back-end storage, gmlinktolerance may detect the resulting high latency and stop the Global Mirror relationships. Disabling gmlinktolerance will prevent this, at the risk of exposing the test host to extended response times.

Diagnosing and fixing 1920 errors triggered by the gmlinktolerance function

A 1920 error ("A Global Mirror or Metro Mirror (Remote Copy) relationship has stopped due to poor performance."**) indicates that one or more of the SAN components is unable to provide the performance required by the application hosts**. This may be temporary (e.g. as a result of maintenance activity) or permanent (due to hardware failure, or unexpected host IO workload). A 1920 error should not occur during normal operation, assuming that your configuration complies with the recommendations above.

In order to diagnose the cause of such an error, it is very important that TPC, or your chosen SAN performance analysis tool, is correctly configured and monitoring statistics when the problem occurs. TPC should be set to the minimum available statistics collection interval, which is currently 5 minutes. It is suggested that you follow this procedure to diagnose why the 1920 error occurred; the possible causes are listed in approximate order of likelihood.

If several 1920 errors have occurred, the cause of the *earliest* error should be diagnosed first.

Was maintenance occurring at the time of the error? This might include replacing a storage controller's physical disk, upgrading a storage controller's firmware, performing a code upgrade on one of the SVC clusters. If so, wait until the maintenance has completed. You should then wait for the length of the maintenance window, then restart the Global Mirror relationships. This wait is necessary to prevent a second 1920 error occurring, as the system has not yet returned to a stable state with good performance.

Is the long-distance link overloaded? If your link is not capable of sustaining the short-term peak Global Mirror workload, this will cause a 1920 error. To identify whether this is the case, perform each of these checks:

- *Look at total Global Mirror auxiliary vdisk write throughput before the Global Mirror relationships were stopped*. If this is approximately equal to your link bandwidth, it is very likely that your link is overloaded. This may be due to application host IO, or a combination of host IO and background (synchronization) copy activity.
- *Look at total Global Mirror source vdisk write throughput before the Global Mirror relationships were stopped*. This represents only the IO being performed by the application hosts. If this in itself is approaching the link bandwidth, you may need to either upgrade the link's bandwidth, or reduce the IO that the application is attempting to perform, or choose to mirror fewer vdisks using Global Mirror. If, on the other hand, the auxiliary disks show much more write IO than the source vdisks, this suggests a high level of background copy . Try decreasing the Global Mirror partnership's background copy rate parameter, to bring the total application IO bandwidth and background copy rate within the link's capabilities.

- *Look at total Global Mirror source vdisk write throughput AFTER the Global Mirror relationships were stopped.* If write throughput increases greatly (by 30% or more) when the relationships were stopped, this indicates that the application host was attempting to perform more IO than the link could sustain. While the Global Mirror relationships are active, the overloaded link causes higher response times to the application host, which decreases the throughput it can achieve. Once the relationships have stopped, the application host sees lower response times, and so the true IO workload is seen. In this case, the link bandwidth must be increased, or the application host IO rate must be decreased, or fewer vdisks must be mirrored using Global Mirror.

Are the storage controllers at the secondary cluster overloaded?
If one or more of the mdisks on a storage controller is providing poor service to the SVC cluster, this may cause a 1920 error if this prevents application IO from proceeding at the rate required by the application host. If the *back-end storage controller requirements* above have been followed, it is most likely that such an error has been caused by a decrease in controller performance due to maintenance actions, or a hardware failure of the controller.

Use TPC to obtain the *backend write response time* for each mdisk at the secondary cluster. If the response time for any individual mdisk exhibits a sudden increase of 50 ms or more, or if the response time is above 100ms, then this indicates a problem. You should:

- Check the storage controller for error conditions such as media errors, a failed physical disk, or associated activity such as RAID array rebuilding. If there is an error, you should fix the problem and then restart the Global Mirror relationships.
- If there is no error, you should consider whether the secondary controller is capable of processing the required level of application host IO. It may be possible to improve the performance of the controller by
    - adding more physical disks to a RAID array
    - changing the RAID level of the array
    - changing the controller's cache settings (and checking that the cache batteries are healthy, if applicable)
    - changing other controller-specific configuration parameters.

Are the storage controllers at the primary cluster overloaded?
The performance of the primary backend storage should be analysed using the same steps as for the secondary backend storage. The main effect of bad performance will be to limit the amount of IO that can be performed by application hosts. Therefore, back-end storage at the primary site should be monitored regardless of Global Mirror However, if bad performance continues for a prolonged period, it is possible that a 1920 error will be seen and the Global Mirror relationships will be stopped.

Is one of your SVC clusters overloaded?
Use TPC to obtain the *port to local node send response time* and *port to local node send queue time*. If the total of these two statistics for either cluster is above 1 millisecond, this suggests that the SVC may be experiencing a very high IO load, which may have resulting in a 1920 error. Also check the SVC node CPU utilization; if this figure is in excess of 50%, this may also be contributing to the problem. In either case, please contact your IBM support representative for further assistance.
If CPU utilization is much higher for one node than for the other in the same IO group, this may be caused by having different node hardware types within the same IO group (4F2 mixed with 8F2 or 8F4). If this is the case, then again, please contact your IBM support representative.

Do you have Flash Copies in 'prepared' state at the Secondary?
If the Global Mirror auxilliary vdisks are the sources of a Flash Copy mapping, and that mapping is in 'prepared' state for an extended time, performance to those vdisks may be impacted because the cache is disabled. Starting the flash copy mapping will re-enable the cache, improving their performance for Global Mirror IO.