SVC V3.1.x Configuration Requirements and Guidelines

Since SVC will be used either in a new SAN, or attached to an existing SAN, the number of different configurations in which it will be used is very large. It is not therefore practical to enumerate or to test all the combinations of all supported SAN devices and fabrics This note describes the configuration rules that apply to the product, and which have been used to determine which configurations to test during product development. Customer configurations must adhere to these rules.

Figure 1 shows a conceptual block diagram of SVC nodes attached to a SAN fabric which comprised SVC nodes, hosts and RAID controllers connected via a switch.

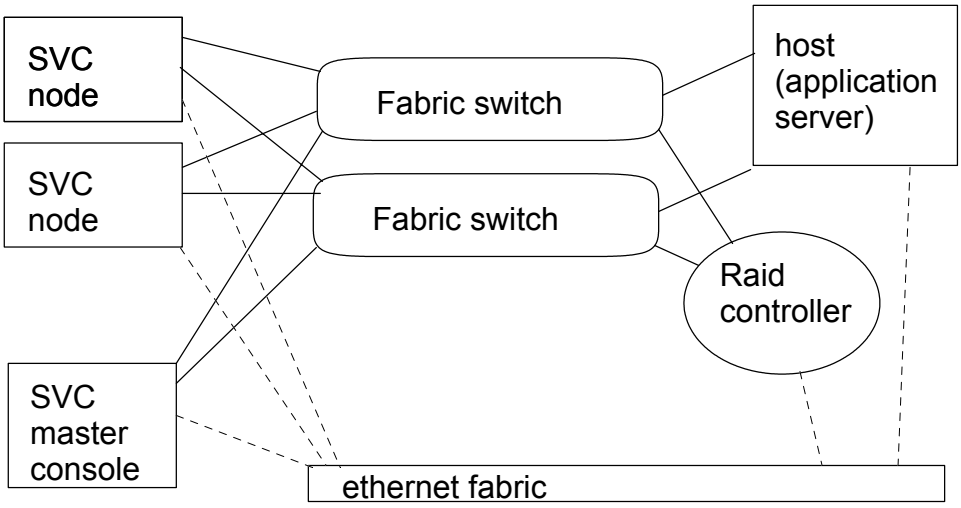


Figure 1

The following general points should be noted:

- Mixed speed fabrics are supported, though it is a requirement that all SVC ports in a single SVC cluster operate at the same speed.
- The SAN shown is a fault tolerant one without a single point of failure. SVC does support SAN configurations which are not redundant but these are not recommended.
- The Fibre Channel switch is zoned to permit the hosts to see the SVC nodes and the SVC nodes to see the RAID Controllers. The SVC nodes within a cluster must be able to see each other. The application hosts must not be allowed to see the RAID controller Luns that are being managed by SVC.
- Each SVC node presents a Vdisk to the SAN via four ports. Since each Vdisk is accessible from the two SVC nodes in an IO Group, this means that a host HBA port may see up to eight paths to each LU presented by SVC. The hosts must run a multipathing device driver to resolve this back to a single device. See the SVC support statement for each of the supported operating systems for details of the specific multipathing software supported.
- As well as a Fibre Channel connection each device has an ethernet connection for

configuration and error reporting. These connections are aggregated together through an ethernet hub (or switch, though performance is not an issue so a hub would be adequate).

Configuration Rules for SVC

The following definitions are used in this section

ISL Hop Count
This is a 'hop' on an Inter Switch Link and is defined as follows: "Considering all pairs of N-ports (endnodes) in a fabric, and measuring distance only in terms of ISL links in the fabric, the ISL Hop Count is the number of ISL links traversed on the shortest route between the pair of nodes that are farthest apart"

Oversubscription
Assuming a symmetrical network, and given a specific workload applied evenly from all initiators and directed evenly to all targets; oversubscription is the ratio of the sum of the traffic on the initiator N-port connection(s) to the traffic on the most heavily loaded ISL or ISLs where there is more than one in parallel between these switches. "Symmetrical" means that all the initiators are connected at the same level, and all the controllers are connected at the same level. SVC makes this calculation more interesting, because it puts its backend traffic onto the same network, and this backend traffic varies by workload, so 100% read hit will give a different oversubscription to 100% write miss." If you have an oversubscription of 1 or less then the network is non-blocking.

Redundant SAN
A SAN configuration in which any one single component may fail, and connectivity between the devices within the SAN is maintained, possibly with degraded performance. This is normally achieved by splitting the SAN into two independent counterpart SANs.

Counterpart SAN
A non-redundant portion of a redundant SAN. A Counterpart SAN provides all the connectivity of the redundant SAN, just without the redundancy. SVC will typically be connected to a Redundant SAN made out of two Counterpart SANs.

Local Fabric
Since SVC supports MetroMirror (remote copy), there may be significant distances between the components in the local cluster and those in the remote cluster. The Local Fabric comprises those SAN components (switches, cables etc.) that connect the components (nodes, hosts, switches) of the local cluster together.

Remote Fabric
Since SVC supports MetroMirror (remote copy), there may be significant distances between the components in the local cluster and those in the remote cluster. The Remote Fabric comprises those SAN components (switches, cables etc.) that connect the components (nodes, hosts, switches) of the remote cluster together.

Local/Remote Fabric interconnect
These are the SAN components that are used to connect the Local and Remote Fabrics together. This link between the local and remote fabrics is subject to less stringent rules than the links within the local cluster -specifically the use of some distance extenders is supported - see the SVC statement of supported hardware.

SVC Fibre Channel port fan in
This is the number of hosts that can see any one SVC port. Some controllers e.g. ESS, recommend limiting the number of hosts that use each port, to prevent excessive queuing at that port. Clearly if the port fails or the path to that port fails, the host may failover to another port and the fan in criteria may be exceeded in this degraded mode.

Illegal configuration
An illegal configuration will refuse to operate and will generate an error code to indicate the cause of

the illegality.

Unsupported configuration
An unsupported configuration might well operate satisfactorily but IBM will not guarantee being able to fix problems that might be experienced by the user. Error logs will not in general be produced.

Valid configuration
A configuration which is neither illegal nor unsupported.

Degraded
A valid configuration which has suffered a failure but which continues to be neither unsupported nor illegal. Typically a repair action which be taken on a Degraded configuration to restore it to a valid configuration.

Channel extender
A device for long distance communication connecting other SAN fabric components. Generally these may involve protocol conversion to ATM or IP or some other long distance communication protocol.

Mesh Configuration
A mesh configuration is an arrangement of switches in which the topology in which four or more switches are connected together in a loop but within that loop are paths which "short circuit the loop" . Thus for example 4 switches connected together in a loop is a small dual core fabric but adding ISLs for one or both diagonals makes it a mesh.

A SAN configuration containing SVC will be a valid SVC configuration if it meets all of the following criteria:

Ports per node : A SVC node always contains 2 HBAs, each of which presents 2 ports. If an HBA fails, this remains a valid configuration, and the node operates in degraded mode. If an HBA is physically removed from a SVC node, then the configuration is unsupported.

Nodes per IO Group : SVC nodes must always deployed in pairs, two nodes per IO group. If a node fails or is removed from the configuration, the remaining node operates in a degraded mode, but is still a valid configuration.

Switch support: The SAN contains only supported switches as described in the SVC hardware support statement. Operation with other switches is unsupported.

Mixed switch vendor fabric : Within an individual SAN fabric, switches must be from the same manufacturer with specific exceptions:

1. **Mixed vendors:** If more than one counterpart SAN is used to make a redundant SAN fabric, then configurations with different vendor's switches in each counterpart SAN are supported.

2. **BladeCentre internal switches:** BladeCentre hosts internal switch may be of a different vendor type to that used in the main part of the counterpart SAN to enable - details are described in the SVC hardware support statement.

3. **CISCO MDS Family Interoperability :** SVC supports the Interoperability modes of the Cisco MDS9000 family of switch/director products specifically when MDS9000 is also connected to Brocade and McDATA a switch/director products and the multivendor fabric zones are connected using MDS Interoperability Mode 1, 2 or 3 subject to the following constraint : The SVC nodes making up an SVC cluster should ALL be attached to the Cisco part of the counterpart fabric or they should ALL be attached to the McDATA or Brocade part of the counterpart fabric. This avoids the situation where a single fabric has SVC nodes in a cluster where some nodes are connected to Cisco switch ports and some to Brocade or McDATA switch ports. The Cisco Interoperability modes can be used to allow SVC to virtualise storage controllers and host ports attached via Cisco Interoperability modes.

One to four fabrics per cluster : SVC supports configurations containing between one and four independent SAN fabrics.

Note: The SVC Master Console contains only two FC ports. It cannot therefore be connected to more than 2 independent SANs. This means that the version of Tivoli SAN manager (TPC for fabric) which is included in the Master Console software will not support such a configuration. Alternative arrangements must therefore be made by customers deploying 3 and 4 SAN environments to ensure that these SAN environments can be adequately managed and debugged.

A SAN which is comprised of a single switch, or a number of switches connected together as a non-redundant fabric is a supported (but not recommended) configuration. In these circumstances there is a possibility that a failure of the SAN fabric may cause loss of access to data - this would be seen as a single point of failure. If the SAN is comprised of two or more independent switches (or networks of switches or a director class switch which appears to behave like a very reliable single switch) so as to make a redundant SAN fabric, then if one SAN fabric fails, the configuration is supported and in a degraded mode.

Various different classes of non redundant configuration are possible:

> ***No redundancy*** *:* A single hardware failure will cause loss of access to data. Note that failure of any switch in a fabric may cause disruption to all other switches in the same fabric and hence may cause temporary loss of communication between any pair of devices on that fabric regardless of whether their communications are routed through the failing switch.
>
> ***Redundant hardware****:* Some switches provide a degree of hardware redundancy such that the failure of some components of the switch does not prevent the switch from continuing operation. With careful design it might be possible to use such switches with SVC such that there is no single point of hardware failure. This kind of configuration still has a single point of failure if a software problem occurs.
>
> ***Virtual SANs (VSANs) :*** . Cisco switches allow switch hardware to be virtualised to create multiple virtual SANs using the same hardware. Virtual SANs provide a higher degree of protection against single points of failure caused by software problems with the exception of software problems with the virtualisation code. Virtual SANs can be used in combination with redundant hardware to provide a configuration that is almost as robust as using a counterpart SAN.

SVC port - switch port connection: On the FC SAN the SVC nodes must always be connected to SAN switches and nothing else. Each node must be connected to each of the counterpart SANs within the redundant SAN fabric. Operation with direct connections between host and node or controller and node is unsupported. Specifically direct connection from one SVC port to another SVC port is illegal.

Controller port - switch port connection: On the FC SAN backend storage must always be connected to SAN switches and nothing else. Multiple connections are allowed from the redundant controllers in the backend storage to improve data bandwidth performance.

It is not mandatory to have a connection from each redundant controller in the backend storage to each counterpart SAN.

Operation with direct connections between host and SVC node or between SVC node and controller is unsupported.

SVC port and associated controller port connection to switches in a greater than 64 host environment: In a large configuration where the number of SVC attached hosts is greater than 64, all SVC ports and all storage ports must be connected to core switches only. In configurations with multiple core switches it is supported to connect the SVC ports and storage ports to multiple core

switches.

Split controller: Certain storage controllers can be configured to safely share resources between an SVC and direct attached hosts. This configuration is described as split controller. Whether this is supported by SVC for a particular storage controller, and the restrictions on this support, is detailed in the SAN Volume Conroller Configuration Guide. In all cases it is critical that the controller and/or SAN is configured so that SVC cannot access LUs that a host can also access. This can be arranged by controller LUN mapping/masking. If this is not guaranteed then data corruption can occur.

Split Controller between SVCs: Where SVC supports a controller being split between SVC and a host as described in rule SVC also supports configurations in which a controller is split between two SVC clusters. In all cases it is critical that the controller and/or SAN is configured so that one SVC cannot access LUs that the other SVC can also access. This can be arranged by controller LUN mapping/masking. If this is not guaranteed then data corruption can occur. This configuration is not recommended because of the risk of data corruption.

Full connectivity to controller ports: All SVC nodes in a SVC cluster must be able to see the same set of back end storage ports on each back end controller. Operation in a mode where two nodes see a different set of ports on the same controller is degraded and the system will log errors requesting a repair action. This could occur if inappropriate zoning was applied to the fabric. It could also occur if inappropriate LUN masking is used. This rule has important implications for back end storage such as DS4000 (FAStT) which impose exclusivity rules on which HBA WWNs a storage partition can be mapped to

Uniform SVC port Speed: The connections between the switches and the SVC nodes run at either 1Gb/s or 2Gb/s, and are made with optical fibre. However all of the FC ports on SVC nodes in a single cluster will run at one speed. Operation with different speeds running on the node to switch connections in a single cluster is illegal (and is impossible to configure).

Mixed fabric speed support : Mixed speeds are permitted within the fabric. The user may use lower speeds to extend distance or to make use of 1 Gb/s legacy components.

Local ISL hops: The local or remote fabric should not contain more than 3 ISL hops within each fabric. Operation with more ISL hops is unsupported. When a local and a remote fabric are connected together for MetroMirror (remote copy) purposes, then the ISL hop count between a local node and a remote node may not exceed 7. This means that some ISL hops may be used in cascaded switch link between local and remote clusters, provided that the local or remote cluster internal ISL hop count is less than 3.

Local/Remote ISL Hop : Where all three allowed ISL hops have been used within the local/remote fabrics (rule), then the Local/Remote Fabric Interconnect must be a single ISL hop between a switch in the local fabric and a switch in the remote fabric. If less than three ISL hops are used in the local/remote fabric then more described in rule.

ISL Oversubscription: Where ISLs are used, each ISL link oversubscription may not exceed 6. Operation with higher values is unsupported.

Node to UPS cable: The nodes must be connected to the UPS using the IBM supplied cable which joins together the signal and power cables.

Node to UPS: The UPS must be in the same rack as the nodes.

Switch support: The switch configuration in a SVC SAN must be legal with respect to the switch manufacturer's configuration rules. This may impose restrictions on the switch configuration, e.g. it may be a switch manufacturer 's requirement that no other manufacturer' s switches are present in the SAN. Operation outside the switch manufacturer 's rules is not supported.

Controller Zones: Switch zones containing controller ports must not contain more than 40 ports. A configuration that breaks this rule is unsupported.

SVC Zones: The switch fabric must be zoned so that the SVC nodes can see the backend storage and the front end host HBAs. Usually the front end host HBAs and the backend storage will not be in the same zone. The exception to this would be where split Host and split controller configuration is in use as described in this document.

- It is permissible to zone the switches in such a way that particular SVC ports is used solely for inter node communication, or for communication to host or for communication to back end storage. This is possible since each SVC contains 4 ports. In any case, each SVC node must still remain connected to the full SAN fabric.

- In MetroMirror (remote copy) configurations, additional zones are required that contain both the local nodes and the remote nodes but normally nothing else. It is valid for the local hosts to see the remote nodes or for the remote hosts to see the local nodes.

The SVC zones must ensure that every port of every SVC node can see at least one port belonging to every other node in the cluster.

The SVC zones must ensure that the nodes in the local SVC cluster do not see SVC nodes in any cluster other than the remote cluster. The situation where more than two clusters can see each other over the fibre channel must be avoided. It is permissible to have one or two hot spare nodes which are not members of any cluster and which are zoned to see the clusters.

Host Zones: The configuration rules for Host zones different depending on the number of hosts that will access the SVC cluster. For small configurations a simple set of rules apply. For larger configurations these rules become more complex and in some ways more restrictive. Customers who therefore plan to migrate from an initial small configuration to a larger configuration at a later time should be careful to plan ahead and ensure that their initial setup configuration takes future expansion into account.

- ***Homogeneous HBA port host zones:*** Switch zones containing Host HBAs must not contain Host HBAs in dissimilar hosts or dissimilar HBAs in the same host., (Here, dissimilar means that the hosts are running different operating systems use different hardware platforms, or use different HBA vendors e.g. if you have AIX and NT hosts, they need to be in separate zones; similarly, if you have Qlogic and Emulex adapters in the same host, then they also need to be in separate zones).

- ***Multiple HBA port zoning for small cluster.***: For configurations with less than 64 hosts, Switch zones containing Host HBAs must contain no more than 40 initiators in total including the SVC ports which act as initiators. Thus a valid zone would be 32 host ports plus 8 SVC ports This rule exists because there is a concern that situations may occur in a SAN where the order N^2 scaling of number of RSCN with number of initiators per zone can cause operational problems in that SAN. A configuration that breaks this rule is unsupported. Note, using multiple initiator host zones for small configurations means that a rezoning activity will be required if the configuration subsequently grows beyond 64 hosts.

  Switch vendors sometimes recommend configurations which have fewer initiators per zone than this. If the switch vendor recommends fewer ports per zone for a particular SAN then the stricter rules imposed by the FC vendor take precedence over the SVC rules.

- ***Single HBA port zoning for clusters attaching more than 64 hosts.***: Each HBA port must be placed into a separate zone. Also included in each zone must be exactly one port from each SVC node in the IO group(s) that this host will access. Configurations which do not follow this are not supported. For configurations smaller than this, it is recommended, but not mandatory, that hosts be zoned to facilitate migration to a larger configuration later. In large configurations, with greater than 64 hosts, the number of host HBAs in a single zone must be restricted to minimise the interaction between devices when fabric changes occur. A configuration that breaks these rules is unsupported. This is a SAN interoperability issue rather than a SVC requirement and is consistent with the best practice guidelines from IBM

Solution Central and switch vendors.

- ***HBAs per host:*** SVC does not specify the number of host Fibre Channel ports or HBAs that a host or a partition of a host can have. The host multipathing device driver will have a specified number of ports that it supports and SVC will support this number, provided the other configuration rules specified here are met.

- ***Balanced load across HBA ports:*** To obtain the best performance from a host with multiple FC ports the zoning should ensure that each FC port of a host is zoned with a different group of SVC ports.

- ***Balanced Host load across SVC ports:*** To obtain the best performance of the subsystem and to prevent overloading, the workload to each SVC port should be equal. This will typically involve zoning approximately the same number of host FC ports to each SVC FC port.

- ***Number of SVC ports within a host zone:*** If it is intended that the configuration is to support more than 256 hosts, then hosts should only be zoned and associated with IO groups that they are using. This is because each SVC port supports up to 512 logins (see maximum configurations table), and in large configurations these logins should only be consumed if they are actively being used for IO purposes.

- ***Migration considerations:*** To reduce or avoid the rezoning required when the configuration size passes the 64 host threshold it is strongly recommended that

    a. As new hosts are added to an existing configuration, these hosts are added in single host zones.
    b. As new host ports are added to an existing configuration they are zoned so that they can see just one port of each node in each SVC IO group. Configurations that attach or intend to attach more than 256 hosts, should limit the number of IO groups that each host accesses.
    c. Any new configuration that is likely to grow to larger than 64 hosts is zoned using single HBA port zones from day 1.

Guidelines for zoning iSCSI hosts and SVC: In a conventional Fibre Channel SAN, there will normally be a number of SAN paths between a particular SVC IO group and the server HBA ports that use the SVC vdisks supplied by that IO group. A multipathing device driver is run on the server to resolve these multiple paths into a single logical device that the server can perform I/O to. The multipathing device driver also provides failover and path recovery functions that deal with scenarios where the SAN fabric paths change or fail.

The present iSCSI solution however only supports a single path between the iSCSI host NIC and the SVC vdisk and there is no multipathing driver in the iSCSI host. This means that there is no recovery from errors and is not possible to concurrently upgrade the SVC firmware while maintaining connectivity from an iSCSI host system. As such it is inappropriate for the SVC to present the vdisk at multiple ports in the Fibre Channel SAN, and to prevent this, the user must select a single SVC port in each SVC IO group that is to be associated with each iSCSI host. Zoning is then applied in the MDS switch so that each iSCSI host can see only one SVC port in each SVC IO group. If multiple iSCSI hosts are in use, the hosts should be evenly spread across the ports in each SVC IO group. The SVC svctask mkvdiskhostmap command should then used to ensure that each SVC vdisk is mapped to a single NIC in the server.

Supported Controllers: SVC is configured to manage LUs exported only by RAID controllers as defined in the SVC hardware support statement. Operation with other RAID controllers is illegal. Whilst it is possible to use SVC to manage JBOD LUs presented by supported RAID controllers, it should be noted that SVC itself provides no RAID function, so such LUs would be exposed to data loss in the event of a disk failure.

Supported Hosts: SVC is configured to export virtual disks to host FC ports on HBAs as defined in the SVC hardware support statement. Operation with other HBAs is unsupported.

Maximum host paths per LU . For any given vdisk, the number of paths through the SAN from the SVC nodes to a host must not exceed 8. Configurations in which this number is exceeded are unsupported.

- SVC has 4 ports/node with 2 nodes in an IO group, Thus without any zoning the number of paths to a vdisk would be 8* (number of host ports.)

- This rule exists to limit the number of paths that need to be resolved by the host Multipathing Device Driver. While Multipathing Device Drivers often support more paths than this no more than 8 paths have been fully tested with SVC.

No Mesh: SVC is not supported on SANs which are created from a mesh of switches.

Link Length: Within the local SAN, FC link distances above10km are not supported. It is permissible for multiple links within the local cluster to be 10km. This applies to ISLs and switch to N port links.

The optical connections supported between host and switch, controller and switch, and switch ISL should be determined by the fabric rules imposed by the vendors of the components used to connect the cluster.
SVC supports Short wave optical fibre and long wave optical fibre connections between the SVC nodes and the switch as described in the SVC hardware support statement.

The supported length of the local to remote Metro Mirror fabric link is defined in the SVC hardware support statement. This will depend on the nature of the intercluster link technology and any extender technology that is supported.

Geographical spread of cluster: Node to node distances are limited by the rules above and in particular it should be noted that all nodes in the cluster need to be connected to the same IP subnet to ensure cluster failover operation.

- These rules may permit a cluster to be spread over more than 10km but note that there is only ever one quorum disk and it is physically resident in one place so geographically dispersed clusters bring attendant risks of loss of quorum resulting in loss of availability.

Number of 2145UPSs required: With the addition of 8 node cluster support in SVC 1.2.1 the requirement was introduced that each 2145UPS only powers two nodes (in distinct I/O groups). Therefore for 6 and 8 node support, four 2145UPSs are required.

Vendor Interop mode: Brocade, McDATA and Cisco switches may be configured in " Vendor Interoperability Mode" or in "Native Mode" . CNT do not have a "native mode".

SAN Timeouts: SVC has only been tested with timeout values of R_A_TOV = 10 seconds and E_D_TOV =2. These are the default timeouts for fabrics. Operation with values other than these is not supported.

DSL and IOD in Brocade fabrics: The recommended settings for DSL and IOD in Brocade fabrics are the same for SVC as for other IBM SAN products: DSL should be off and IOD should be on.

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.

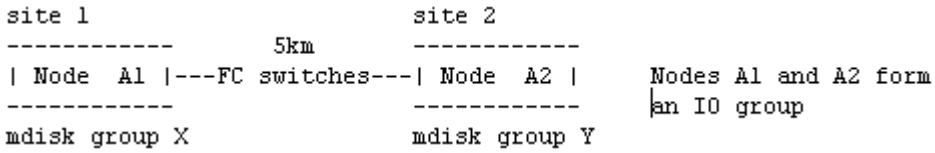
2. A cluster should be regarded as a single entity for disaster recovery purposes. This includes the backend storage that is providing the quorum disks for that cluster. This means that the cluster and the quorum disks should be co-located. Locating the components of a single cluster in different physical 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:



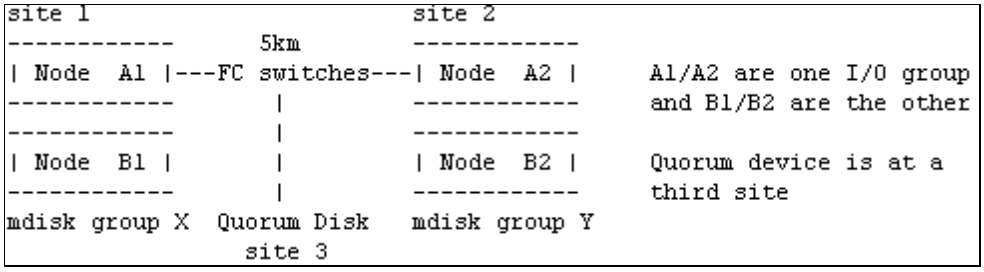
We have a node from each IO group at both sites and we set up Metro Mirror relationships so that the primary vdisks at site 1 come from mdisks in group X (i.e. mdisks at site 1) and the secondary vdisks at site 2 come from mdisks in group Y (i.e. mdisks at site 2). It would appear that this arrangement 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.

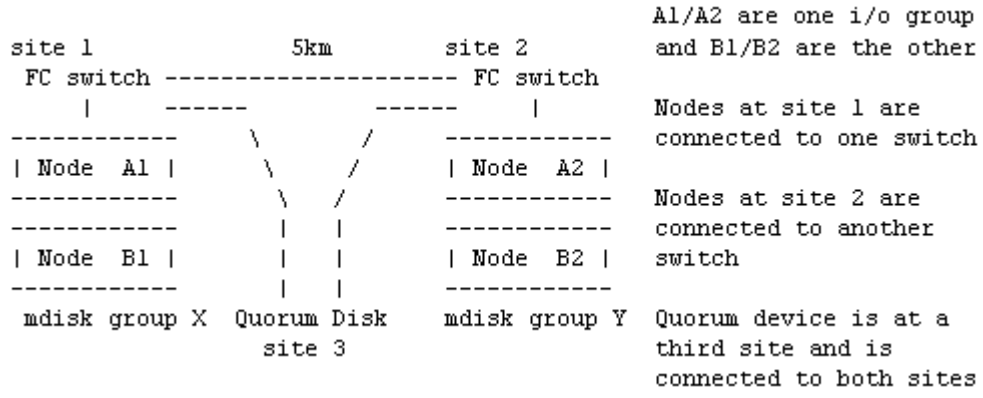
2. The disaster recovery solution is asymmetric. Thus, it is not possible to run applications on both sites and allow either to suffer a failure. One site must be regarded as the primary site and the other is there to provide a recovery site. Consider the situation where the quorum disk is at site 2 (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 re-appears 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.

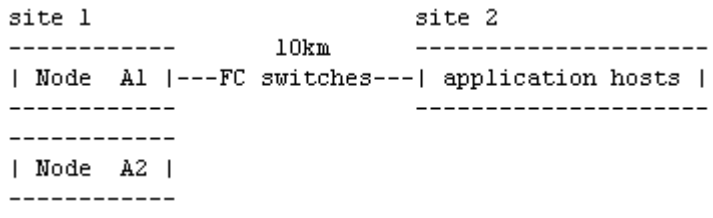
Splitting a cluster might be thought to be useful if the quorum disks are at a third "site", such that a disaster will only take down one of these three sites. However, even a three site configuration will have significant limitations, since SVC will not be able to change the path it uses to communicate 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:



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



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:



IBM recommends the use of two cluster configurations for all production disaster recovery systems. Customer 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 and McDATA fabrics 256 - for Brocade and 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 and McDATA fabrics 155 - CNT 256 - Brocade ***Note 1***	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 256 - Brocade support, please see ***Note 1*** appended to this table
Host ports per cluster	2048 - Cisco and McDATA fabrics 512 - Brocade 310 - CNT	A host port is a fibre channel HBA port in a host See Also – Host ports per I/O group below
Host IDs per I/O group	256 - Cisco and McDATA fabrics 64 - Brocade N/A - CNT	
Host ports per I/O group	512 - Cisco and McDATA fabrics 128 - Brocade N/A - CNT	
Host ports per host ID	512	
Copy Services		
Metro Mirror relationships per cluster	1024	
Metro Mirror consistency groups	256	
Metro Mirror vdisk per I/O group	16 TB	The total size of all Metro Mirror source and target vdisks in an I/O group must not exceed 16 TB
FlashCopy® mappings	2048	
FlashCopy® mappings	512	

per consistency group		
FlashCopy consistency groups	128	
FlashCopy vdisk per I/O group	16 TB	The total size of all FlashCopy source vdisks in an I/O group must not exceed 16 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

For GA support of Brocade fabrics with up to 256 hosts, we now impose the following restrictions at SVC V3.1.0.3:

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 ore 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.

Calculating Oversubscription for SVC configurations

A SAN that is using SVC will have three main traffic flows.

Host - Node: The most significant workload will be between the SVC nodes and the hosts. For calculating oversubscription we will assume that all hosts are generating as much I/O as possible and that this I/O is distributed evenly between the SVC ports (of course in practice this isn't true which is why we can support oversubscription values greater than 1).

Node - Controller: The second most significant workload will be between SVC nodes and storage controllers. SVC will attempt to distribute the workload to storage controllers evenly across all available paths between the SVC nodes and the storage controller.

Node - Node: The other flow of traffic across a SAN with SVC will be traffic between nodes. These traffic flows will be between each pair of nodes in an I/O group. For the purpose of calculating oversubscription we will ignore this traffic flow because:

In a “typical” 70% read, 30% write I/O workload this traffic only accounts for 13% of the traffic on the SAN and hence is relatively insignificant.

The more traffic there is between SVC nodes the less traffic there can be between SVC nodes and hosts or storage controllers.

SVC nodes use load balancing techniques to determine which path to use across the fabric to communicate with another SVC node. If ISLs become congested and there are alternative paths then SVC is likely to use the alternative paths in preference to the ISLs.

So for each switch in the fabric we have:

Oversubscription = (amount of traffic from local hosts to remote switches + amount of traffic from remote hosts to local nodes + amount of traffic from local nodes to storage controllers on remote nodes + amount of traffic from remote nodes to storage controllers on the local switch) / Number of ISLs

Use the following formula to make this calculation

I = Number of ISL between this switch and the rest of the fabric

HL = Number of host server ports attached to this switch which are zoned to SVC ports on other switches

HR = Number of host server ports attached to other switches which are zoned to SVC ports on this switch

NL = Number of SVC ports attached to this switch

NR = Number of SVC ports attached to other switches

SL = Number of storage controller ports attached to this switch which are zoned to SVC ports on other switches

SR = Number of storage controller ports attached to other switches which are zoned to SVC ports on this switch

The ISL oversubscription for each switch is calculated as:

((HL*NR/(NR+NL))+(HR*NL/(NR+NL))+(NL*SR/(SR+SL))+(NR*SL/(SR+SL))) / I

This calculation assumes the use of trunking. It is strongly recommended that trunking be used where multiple ISLs are used in parallel.

Example: Large SAN for use with SVC

Figure 15 shows the key points of an overall design for a large SAN using SVC. For clarity only two SVC nodes, one controller and one host have been shown but in reality such a SAN would perhaps contain an

8 node SVC cluster several large storage controllers and many tens if not hundreds of hosts.

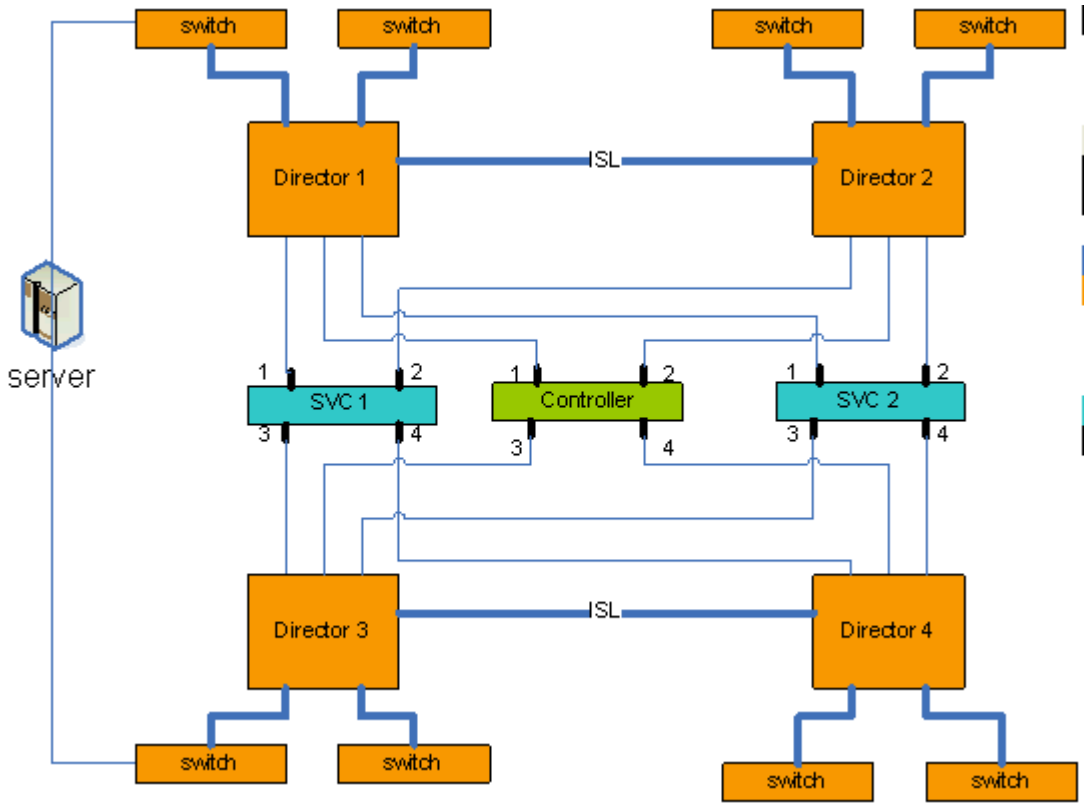


Figure 15: A large SAN configuration

For optimum performance four directory class switches have been used and these are configured into two separate SAN fabrics. Note that the SVC ports and controller ports are connected to the directors for maximum bandwidth whereas the host ports are connected to the edge switches.

In this configuration traffic between SVC nodes does not need to pass over an ISL, nor does traffic between SVC ports and controller ports. Note however that SVC is not aware of the presence of ISLs and could choose for traffic to flow across them unless zoning is used to prevent this. Thus a zone should be created which contains port 1 of each SVC node, forcing all port 1 inter SVC node traffic to remain within director 1. Similarly three further zones should be created for ports 2 ports 3 and ports 4 of the SVC

nodes. By also including the controller port 1 in the SVC port 1 zone it is possible to contain the SVC to controller traffic to within director 1. Similarly for the other controller ports. These zones are the practical embodiment of configuration recommendations and in section .

Note that when configuring zoning the user should be careful not to include SVC ports that should not communicate with each other in any of the host or controller zones since this would allow them to communicate with each other using the host or controller zone.

The design shown it a particularly good one for a large SVC based SAN because:

- There is no contention between Host traffic and controller traffic. The SVC to controller traffic has a non blocking high bandwidth path.
- SVC to SVC traffic does not contend with host or controller traffic.
- The ISL between director 1 and director 2 and between 3 and 4 should not carry very much traffic at all. No SVC to SVC or SVC to controller traffic travels over this link.

MetroMirror (remote copy) SVC Configurations

SVC supports both intra-cluster and inter-cluster Metro Mirror. From the intra-cluster point of view, any single cluster is a reasonable candidate for Metro 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 Figure 16 below. . Note that Intra-cluster Metro Mirror is between Vdisks in the same IO group only.

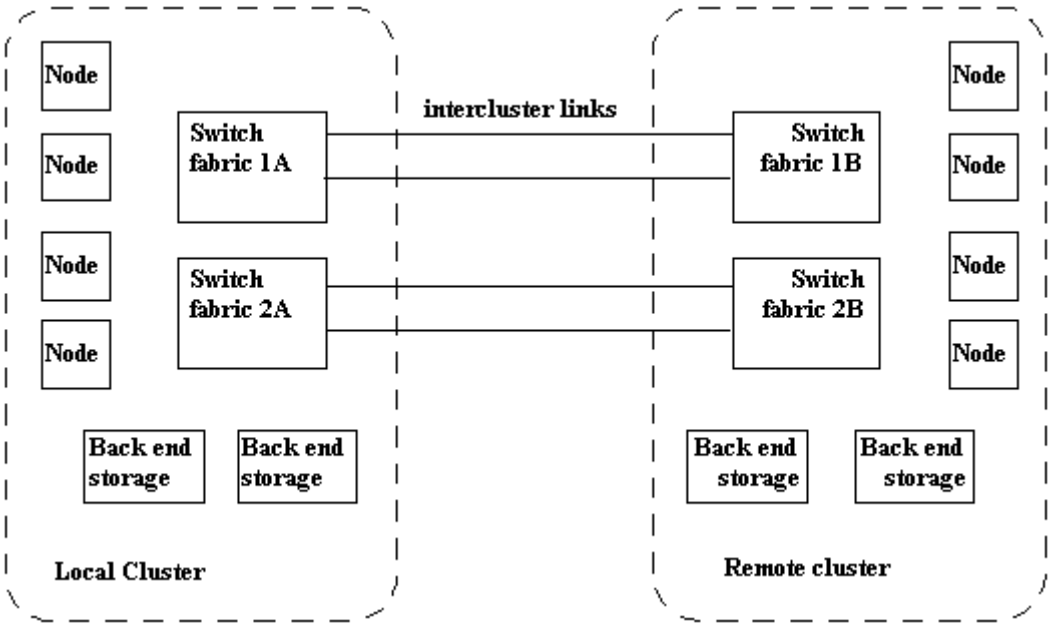


Figure 16: Metro Mirror configuration using dual redundant fabrics

This contains 2 redundant fabrics. Part of each fabric exists at the local and remote cluster. There is no direction connection between the tw fabrics.

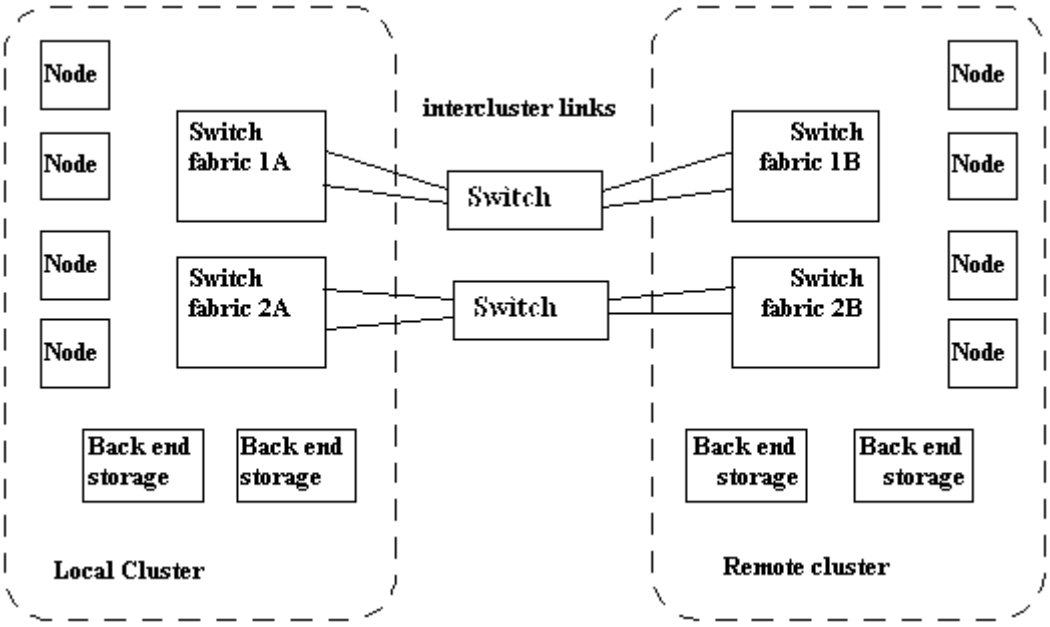


Figure 17: Metro Mirror configuration with intercluster switching - is this a supported way of extending the distance between two clusters (provided that don 't exceed the ISL hop limit of 7 for local to Metro Mirror links).

Even with multiple switches in the intercluster links, the max distance allowed between local and remote clusters is 10km

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 tunneling mode.
DWDM/CWDM and Longwave GBIC extenders.

IBM has tested a number of such fibre channel extender technologies with SVC up to a maximum one way latency of 34ms and will support any fibre channel extender technology provided that it is planned, installed and tested meeting the following requirements:

- The one way latency between sites must not exceed 34ms. Note that 1ms equates to approximately 100 to 150km but this will depend on the type of equipment used and the configuration.
- The bandwidth between sites must be sized to meet the peak workload requirements while maintaining the maximum latency of 34ms.
- 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 channel extenders between links within a single cluster is not supported and can lead to IO errors and

loss of access. Channel extenders can be used only for inter-cluster links. The configuration is tested with expected peak workloads **.**

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

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.

Currently (Oct 2005) IBM supports the following:

McDATA 1620 and 2640

These are supported up to a 1-way latency of 10ms. Note that 1ms equates to approximately 100 to 150km but this will depend on the type of equipment used and the configuration.

Cisco MDS 9000 series Inter vSan Routing.

SVC supports the use of inter vSAN routing in configurations using the Cisco MDS 9000 family of fabric switches. Successful test results have been obtained with packet latencies of up to 10ms. Distance will depend on the type of network and number of hops but could typically be 100-150kms per ms.

Configuring a balanced Storage Subsystem

SVC Virtualisation allows a very large number of Virtual Disks to be configured to use a small number of backend MDisks. Since the Virtual Disks can be mapped to a large number of individual Hosts, each of which can supply a significant number of parallel IOs, SVC can be used to critically overload a backend storage subsystem. The impact of this will range from poor performance for slight overload conditions to loss of availability for significant overload conditions when host systems begin to time IO out before it can be serviced.

In order to avoid these situations refer to the San Volume Controller Configuration Guide.

Support for between 256 and 1024 Host Objects per cluster

SVC 3.1 introduces support for 1024 hosts per SVC cluster. In order to configure 1024 host objects

however each host must be associated with only one IO Group so that each IO group is associated with 256 hosts. SVC continues to support configurations in which each host is associated with all four IO groups but in such a configuration the maximum number of configurable hosts remains 256. It is possible to mix the two approaches because each IO group is limited to 256 hosts irrespective of whether those hosts are associated with 0,1,2 or 3 other IO groups. Policed limits have been added or adjusted for the number following quantities:

- Host Objects per Cluster
- Host Objects per IO group
- Host WWPNs per cluster
- Host WWPNs per IO group
- Host WWPNs per Host

The association of a host with an IO group is explicit in the configuration model. New commands and options have been added in Section to define the association between a host and an IO group.

- Each IO Group is associated with between 0 and 256 host objects.
- Each Host object is associated with between 0 and 4 IO groups.

Hosts cannot be associated with recovery IO group.

A host mapping cannot exist between a vdisk and a host which is not associated with the IO group of the vdisk. This is policed on all commands which create or change the relationships between hosts, IO groups, and vdisks. Vdisks can be put in recovery IO group irrespective of host mappings.

All existing SVC 2.1 action command lines will continue to behave exactly as they did in SVC 2.1

When creating a Host object or upgrading an existing cluster from SVC 2.1 the default is for a host to be associated with all four IO groups.

Zoning Requirements

SVC 3.1 will support no more than 512 Fibre Channel logins per Fibre Channel port. This count includes logins from host ports, other SVC nodes, logins to storage controllers and the fibre channel nameserver. Clearly a configuration with 1024 hosts containing 2048 host ports must not be zoned so that each host HBA sees a port from each SVC node. The configuration rules required that in these large configurations the SAN will be zoned so that each host HBA port can see one SVC port in each node in the IO group associated with that host.

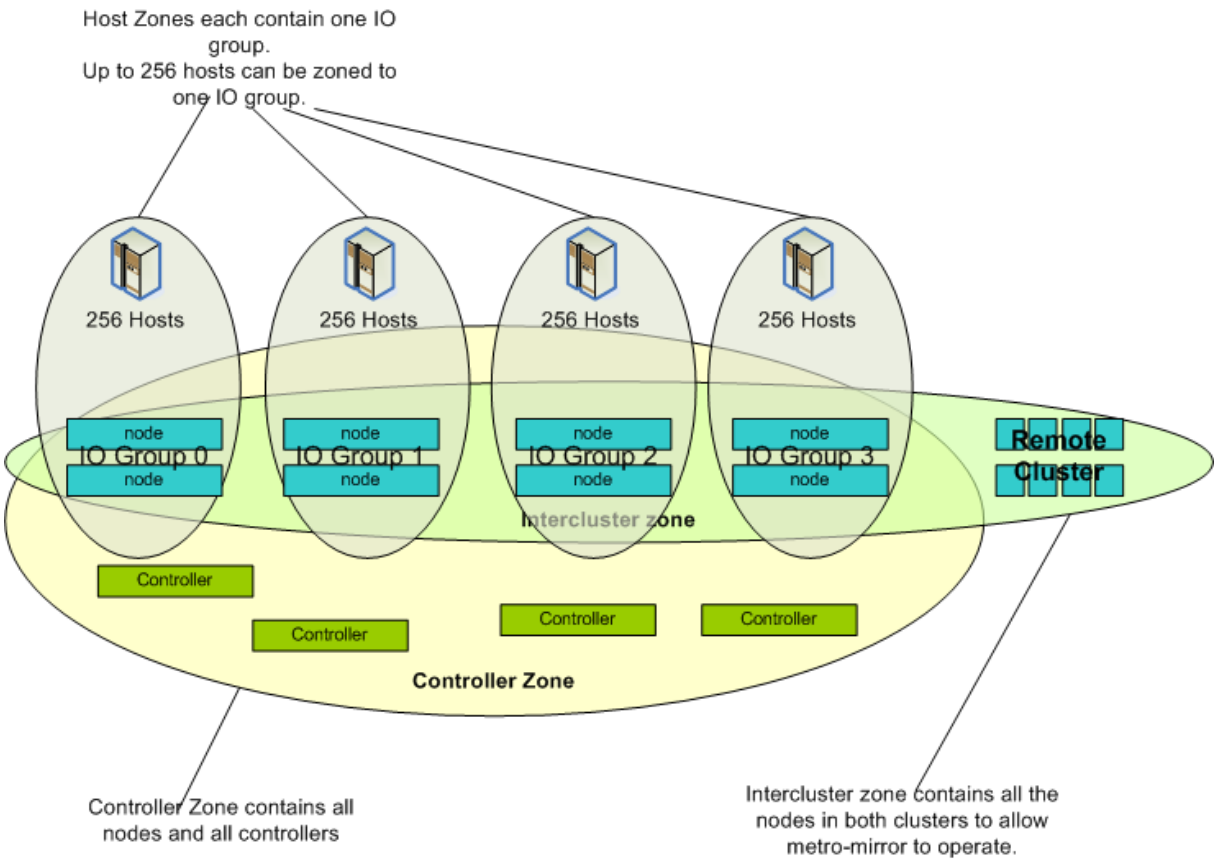
It is **legal** for a host which is not associated with an IO group to be zoned so that it can login to ports in nodes in that IO Group. This only becomes illegal if the maximum number of logins per port is exceeded.

SVC cannot explicitly police the number of logins per FC port because SVC cannot control the SAN zoning. SVC will however log error ID 073006 if any port has more than 512 logins.

It is expected that the user would use the new svcinfo lsfabric tool to list the logins seen by each SVC port to list the logins for each port so that the zoning error can be remedied.

If between 512 and 1024 logins exist to an SVC FC port SVC should continue to function normally. This configuration is not supported and not tested but the internal login table in SVC does have space for up to 1024 logins per node. Above 1024 logins SVC will not be able to record the logins in its internal tables and so some logins will be dropped. The logins dropped could be important logins such as those between nodes making the cluster unstable.

The figure below shows an example zoning arrangement for a 1024 host configuration. The diagram represents only one counterpart SAN in reality both counterpart SANs would be zoned equivalently. The diagram shows that the hosts are arranged into four groups of 256 hosts each and each group is zoned to one IO group. The hosts are zoned separately so that they do not see each other as per the configuration guidelines above. Note that all storage controllers must be zoned to all node, as must the remote cluster. Note that the nodes have a requirement to be zoned together but this requirement is satisfied by the controller zone in the diagram below.


Host Zones each contain one IO group.
Up to 256 hosts can be zoned to one IO group.
256 Hosts
256 Hosts
256 Hosts
256 Hosts
node
node
IO Group 0
node
node
IO Group 1
node
node
IO Group 2
node
node
IO Group 3
node
Remote Cluster
Intercluster zone
Controller
Controller
Controller
Controller
Controller Zone
Controller Zone contains all nodes and all controllers
Intercluster zone contains all the nodes in both clusters to allow metro-mirror to operate.

.