# V1.2.x Configuration Requirements and Guidelines

## Supported SVC Configurations

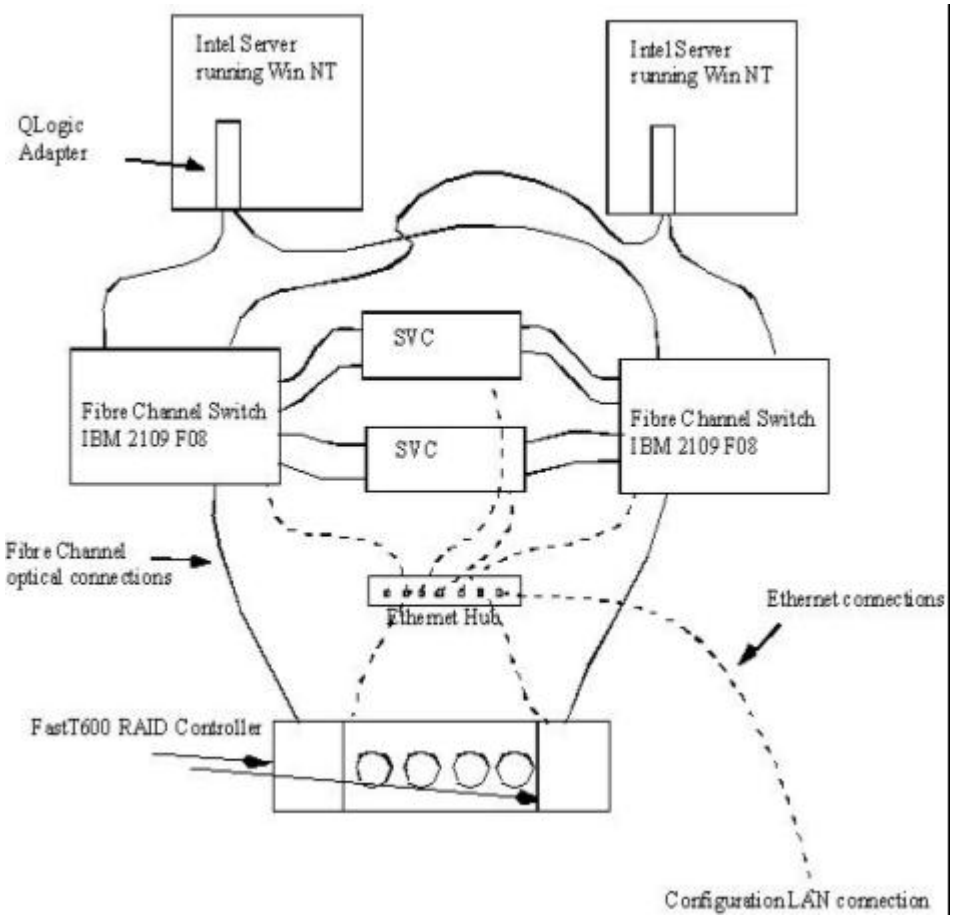
Since SVC will be used either in a new SAN, or attached to an existing SAN, the number of different configurations in which it will be used is very large. It is not therefore practical to enumerate or to test all the combinations of all supported SAN devices and fabrics  This note describes the configuration 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: A simple example of an SVC cluster serving virtualised storage to application hosts*

The following general points should be noted:

The Fibre Channel SAN connections between SVC and the switches are optical fibre running at 2Gb/s, however SVC is also supported in 1Gb/s FC fabrics. All SVC nodes within a single cluster must however run 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 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 the specific exceptions for Blade Centre as described in the SVC hardware support statement. Where a pair of counterpart fabrics (say Fabric A and Fabric B) are used to provide a redundant SAN, it is supported for Fabric A to be comprised of one vendors switches and Fabric B to use a different vendors switches so long as each individual fabric complies with SVC configuration rules.

**One or two fabrics per cluster:** SVC supports configurations containing either one or two independent SAN fabrics.

While SVC could function in configurations with three or four counterpart SANs, the Master Console does not support this, and such configurations are not tested in SVC product development. Therefore they are unsupported for SVC.

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 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.

**Ctlr 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. For example in a DS4000 (FAStT) configuration in which the DS4000 (FAStT) contains two redundant controllers, only 2 controller minihubs are normally used. This means that Controller A in the DS4000 (FAStT) is connected to counterpart SAN A, and controller B in the DS4000 (FAStT) is connected to counterpart SAN B. Operation with direct connections between host and controller is unsupported.

**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 see . Whether this is supported by SVC for a particular storage controller (and the restrictions on this support) is detailed in the appendix of this specification devoted to the specific controller. 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.

**For balanced I/O:** Each SVC node must have the SAME NUMBER of ports that allow it to 'see' the back end storage ports. To ensure the node is not a single point of failure, a minimum of TWO ports per node (one from each SVC HBA on that node) should be able to see the backend storage ports. THREE ports per node is an acceptable configuration, but FOUR ports per node gives maximum bandwidth and is preferred. Single or multiple SAN fabrics may be used to allow the SVC nodes to see the back end storage ports. When choosing ports on the storage device, consideration should be given to providing sufficient bandwidth and eliminating any single point of failure. It is important that every node sees the SAME SET of ports on each back end storage device. If the number of ports on the backend controller visible by the SVC node ports reduces, SVC will regard the device as 'degraded' and will log errors that request 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. It is the responsibility of the user to confirm that their particular configuration falls within these support rules.

**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 supplied cable which joins together the signal and power cables.

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

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:**

Switch zones containing Host HBAs must not contain Host HBAs in dissimilar hosts or dissimilar HBAs in the same host. A configuration that breaks this rule is unsupported.

- (e.g. if you have AIX and NT hosts, they need to be in separate zones).

- Here dissimilar means that the hosts are running different operating systems or are different hardware platforms. Different levels of the same operating system are regarded as “similar”.

- This is a SAN interoperability issue rather than a SVC requirement. It exists because the behaviour of heterogeneous hosts/adapters has been seen to interfere with one another. This configuration rule is consistent with best practice guidelines from SAN Central and switch vendors.

Although IBM tests to the 40 initiators per host zone rule, 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 even though IBM has tested to the higher limit. 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. In future releases of SVC (e.g. 2.1) a rule has been added to restrict zones containing Host HBAs must contain no more than 40 initiators in total including the SVC ports which act as initiators.

SVC supports Hosts (or partitions of a host) which have between 1 and 4 FC ports.

Note also, in subsequent releases of SVC that support attaching more than 64 hosts to a cluster, it is required that switch zoning be used to ensure that each host FC port is zoned to exactly 1 FC port of each SVC node in the cluster. For configurations smaller than this it is recommended that hosts be zones this way but it is not mandatory.

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.

To obtain the best overall 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.

**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 SDD. SDD itself actually supports more paths than this however larger numbers of paths have not 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 above 10km 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 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.

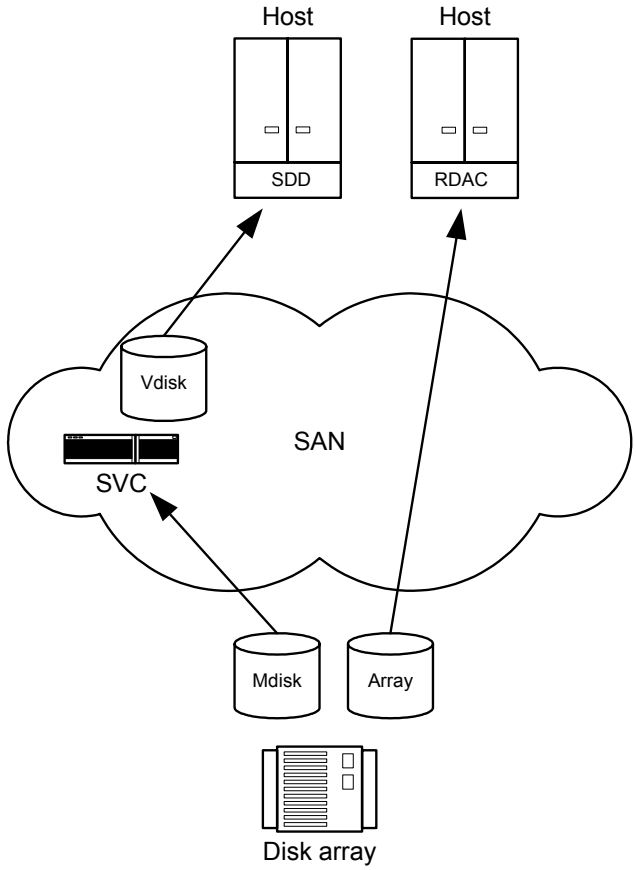
- These rules may permit a cluster to be spread over more than 10km but note that the there is only ever one quorum disk and it is physically resident in one place so geographically dispersed clusters bring attendant risks of loss of quorum resulting in loss of availability.

**Number of 5125 UPS required:** With the addition of 8 node cluster support in SVC 1.2.1 the requirement that each 5125 UPS only powers two nodes (in distinct I/O groups). Therefore for 6 and 8 node support, four 5125 UPSes are required.

**Vendor Interop mode:** Brocade and McData switches may be configured in " Vendor Interoperability Mode" or in "Native Mode" . Cisco switches are not currently supported in "Vendor Interoperability 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.

## Split controller configurations




I

*Figure 2: RAID controller shared between SVC and host*

In this configuration, a RAID array presents LUs to both SVC (which treats the LU as an Mdisk) and to another host. SVC presents Vdisk(s) created from the Mdisk to another host, as shown in the diagram above.

## Split Host configuration

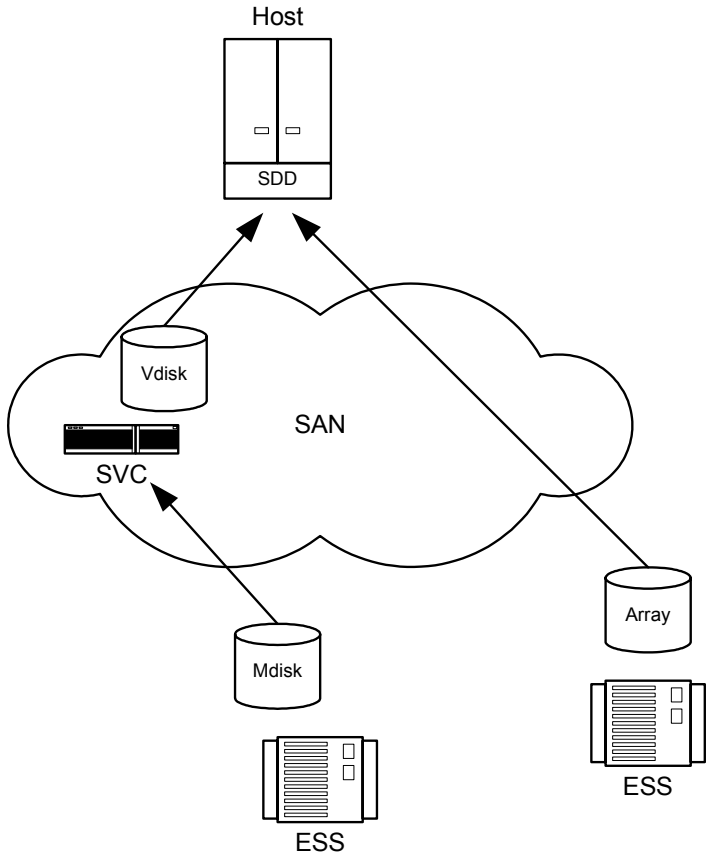



*Figure 3: Split host configuration*

It is also possible to split a host so that it accesses some of its LUNs via SVC and some directly. In this case the multipathing software used by the controller to be accessed directly must be compatible with SVC 's multipathing software. Such a configuration is shown in Figure 3.

## Split controller and split host configuration

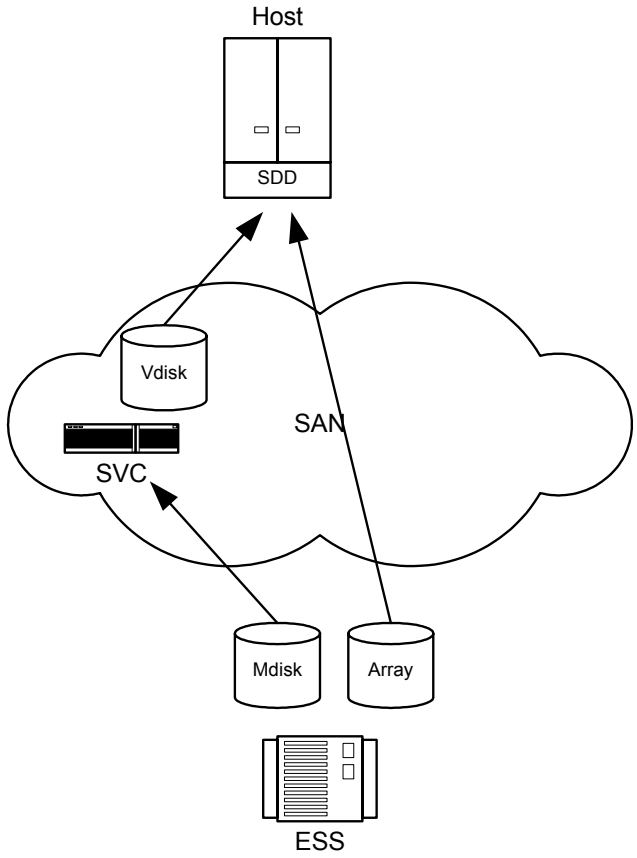



*Figure 4: Split controller and split host configuration*

In the case where the RAID controller uses multipathing software which is compatible with SVC ' s multipathing software it is possible to configure a system such that some LUNS are mapped direct to the host and others are accessed via SVC.

One example would be ESS which uses the same multipathing driver as SVC.

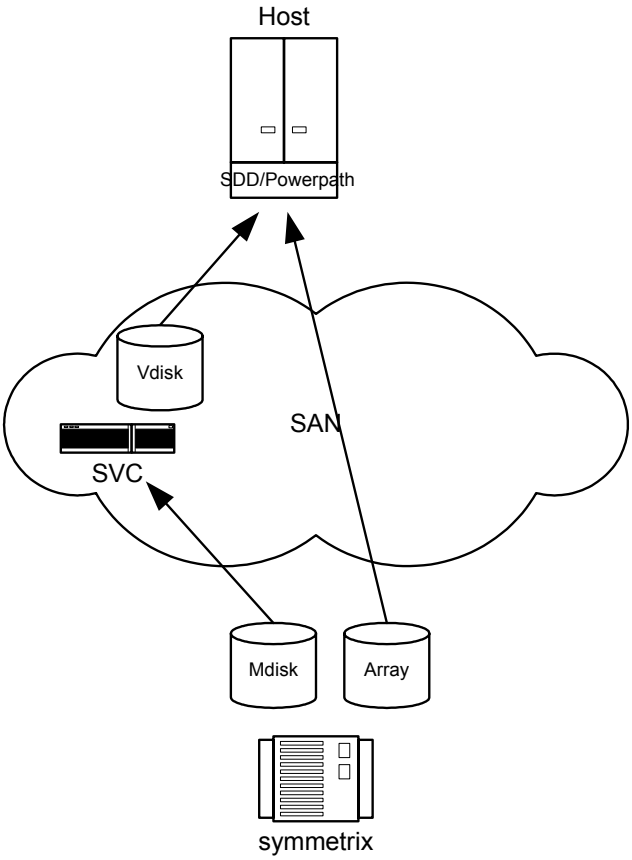



*Figure 5: Unsupported use of PowerPath and SDD on the same host*

Figure 5 however shows a configuration that is not supported because SDD and Powerpath cannot be loaded onto a host together.

## 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.

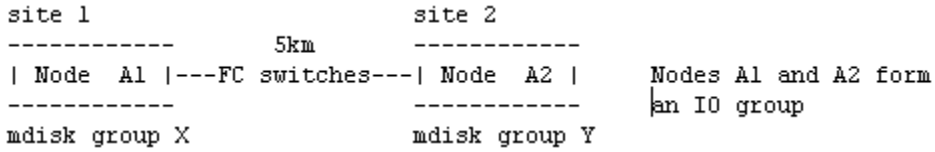
2. A cluster should be regarded as a single entity for disaster recovery purposes. This includes the backend storage that is providing the quorum disks for that cluster. This means that the cluster and the quorum disks should be co-located. Locating the components of a single cluster in different physical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

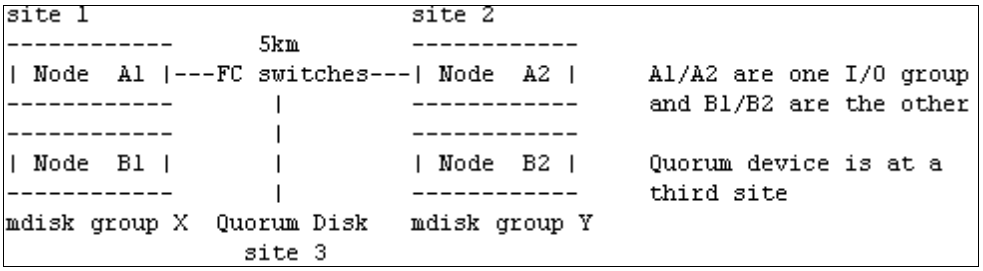
1. If either site fails, we only have a degraded IO group at the other site with which to continue I/O. Performance therefore during a disaster recovery is significantly impacted, since throughput of the cluster is reduced and the cluster caching is disabled.

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.

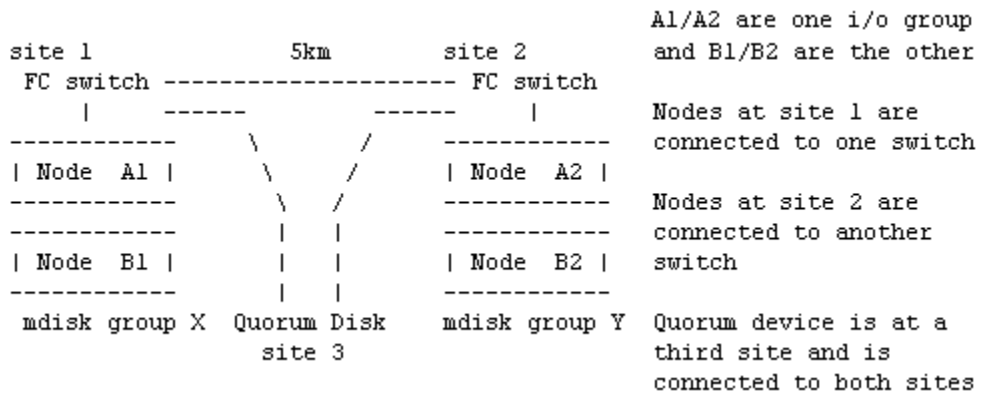
**From the discussion above, it can be seen that the split cluster configuration can only provide asymmetric disaster recovery facilities, with substantially reduced performance. This is unlikely to be satisfactory for most production disaster recovery situations.**

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

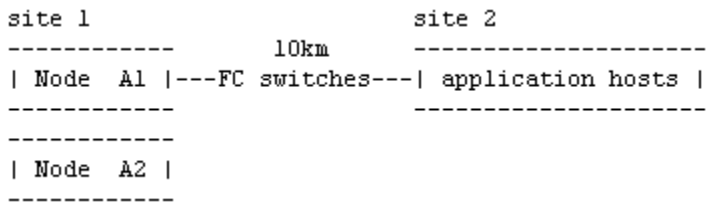
second site before reaching the quorum disk in the third site. For example the following arrangement is acceptable:




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.**

# Configuration Recommendations

This section discusses some further issues on configuring the SAN that SVC is attached to. These are recommendations rather than rules. If a customer does not follow these recommendations they will still be operating with a valid configuration.

When configuring a Cluster with ISLs between the nodes in the Cluster all ISLs on the same fabric are treated as a single point of failure. With reference to configuration (a) If either Link 1 or Link 2 fails the cluster will continue normal operation. If either Link 3 or Link 4 fails the cluster will continue normal operation. If either ISL 1 or ISL 2 fails then communication between Node A and B will fail for a period of time leading to a lease expiry on one or other node even thought there is still a connection between the nodes.

To ensure that no fibre channel link failure causes nodes to fail when there are ISLs between nodes it is necessary to use a configuration with redundant fabric. See configuration b - if any one of the ISL links fails then the cluster will continue normal operation.

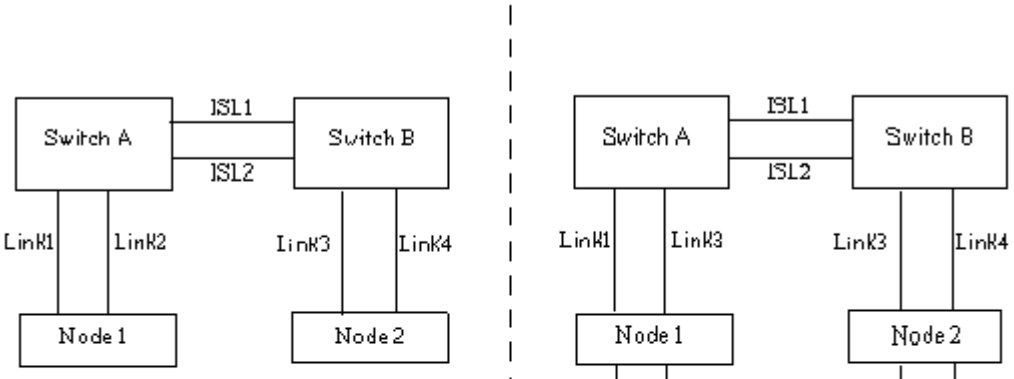
When attaching SVC to a SAN fabric containing core directors and edge switches it is preferable to connect the SVC ports to the core directors and to connect the host ports to the edge switches.

In such a fabric the next priority for connection to the core directors is the storage controllers, leaving the host ports connected to the edge switches.

If an SVC node can see another SVC node through multiple paths some of which use ISLs and some of which do not then zoning should be used where possible to ensure that the SVC to SVC communication does not travel over an ISL. Likewise if SVC can see a storage controller through multiple paths, some of which travel over ISLs and some of which do not then consideration should be given to using zoning to restrict communication to those paths which do not travel over ISLs.

**Homogeneous MDisk Groups** : In configuring Managed Disks into Managed Disk Groups a Managed Disk Group may span more than one RAID Controller. Whilst this is a supported configuration, it should be noted that should any of the RAID Controllers in a MDG go offline, then Virtual Disks whose storage is on that MDG will be taken off-line also. Clearly the probability of losing access to Virtual Disks increases the more RAID Controllers there are in a given MDG, so it is recommended to limit the number.

**Trunking:** Where multiple ISLs are used in parallel between switches we strongly recommend the use of trunking. This is because it is very easy for an individual ISL to become more loaded than the others leading to congestion and dropped frames.




**Configuration (a)** **Configuration (b)**

*Figure 6: Recommended configuration when using ISLs between nodes in a Cluster*

## SAN Fabric Topologies

The configuration rules cited above give a wide range of flexibility in constructing an valid SVC SAN. This section discusses some of the arrangements that we expect customers to use. Some of the questions that need to be resolved when constructing a suitable SAN are:

- Will the fabric provide enough point to point bandwidth capability?
- Will the fabric provide enough ports in the SAN to connect up all the backend storage, the SVC nodes and the front end test hosts?
- How can legacy products such as older 16 port switches be incorporated usefully into a fabric?

The following types of network might be used to address these issues:

- Simple, single switch fabrics, arranged in dual redundant fabric pairs. .
- Meshed networks, that combine a number of small switches into a highly redundant configuration. See Figure 7 for a simple implementation of this arrangement. Mesh configurations are not supported for SVC.

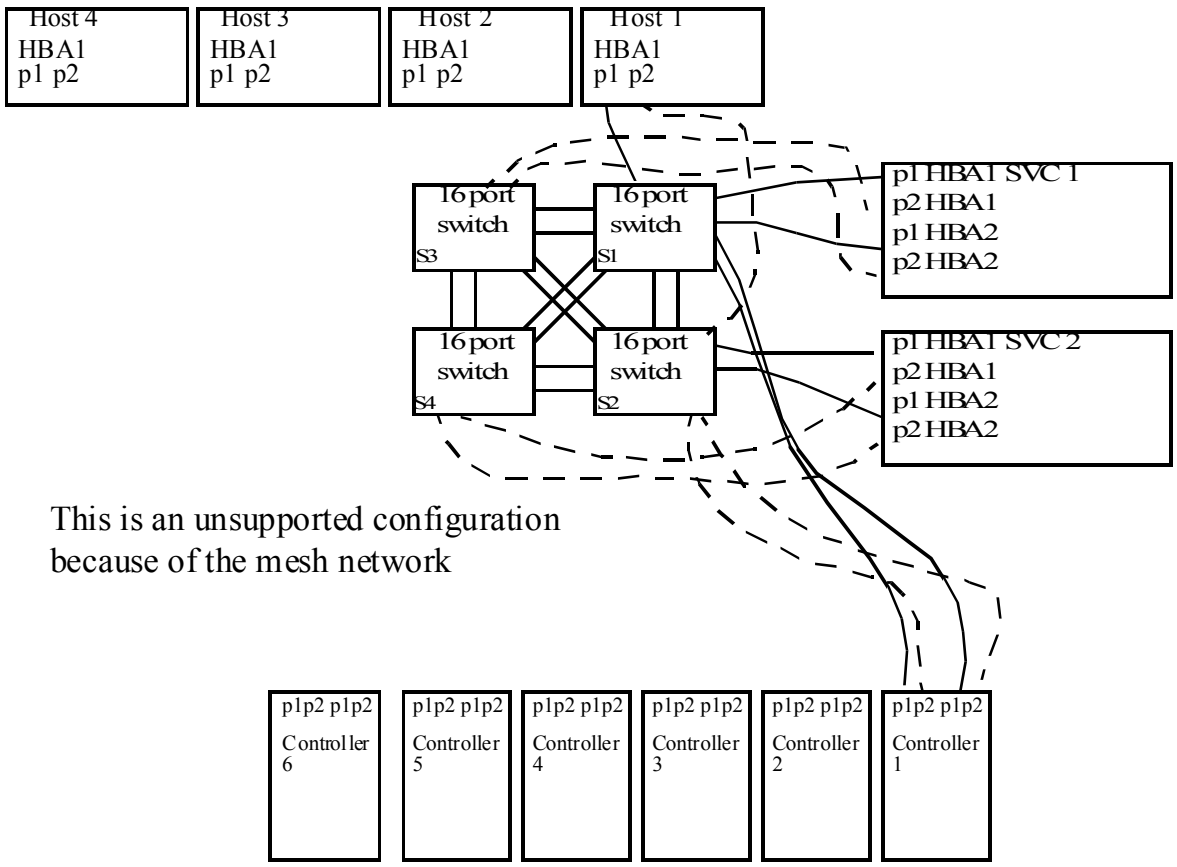



*Figure 7: Mesh configuration - this is an unsupported SVC configuration*

- Tiered and cascaded networks provide a better utilisation of switch ports within a switch network than a 4 way mesh. In this arrangement the switches cascade down the fabric. See Figure 8 for an example of the tiered configuration.

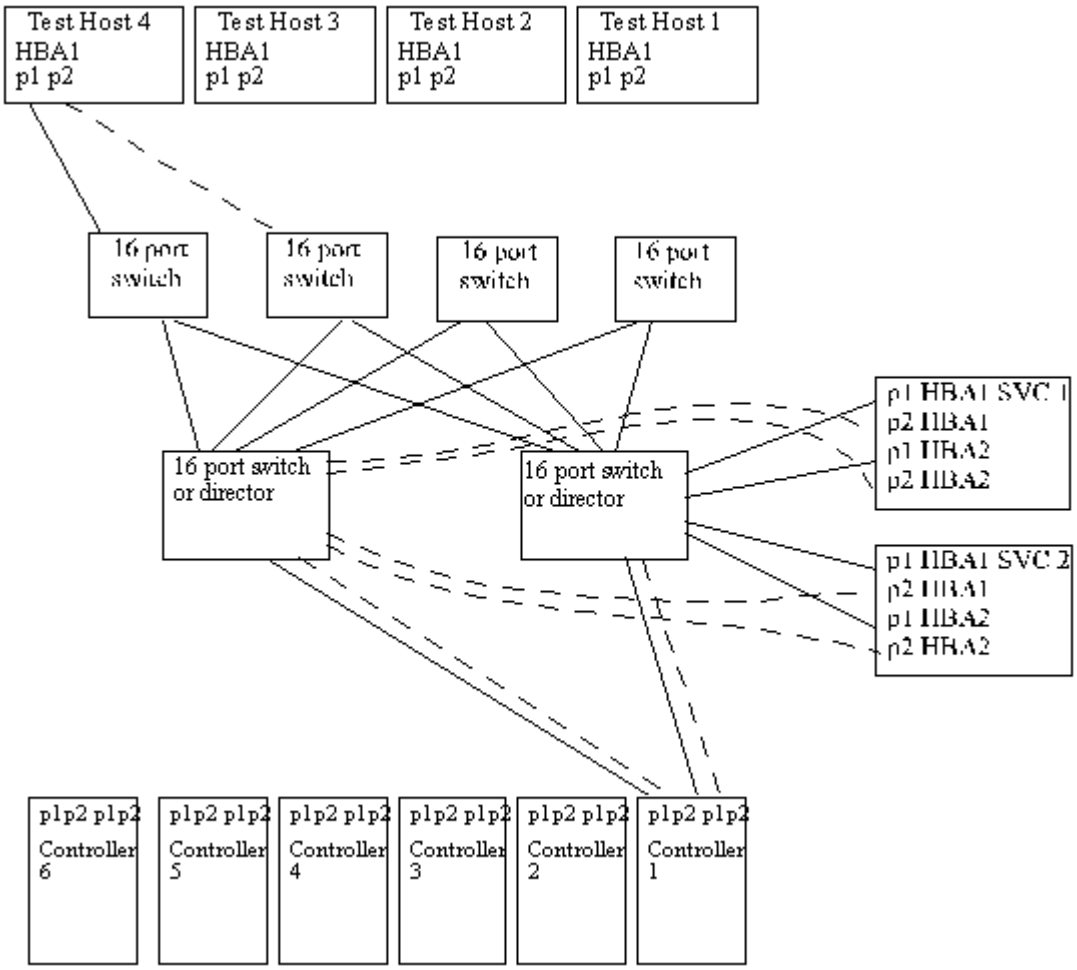



*Figure 8: Fault tolerant, tiered switch fabric, 2 node configuration*

*For clarity, only wiring for one test host and one back end is controller, is shown*

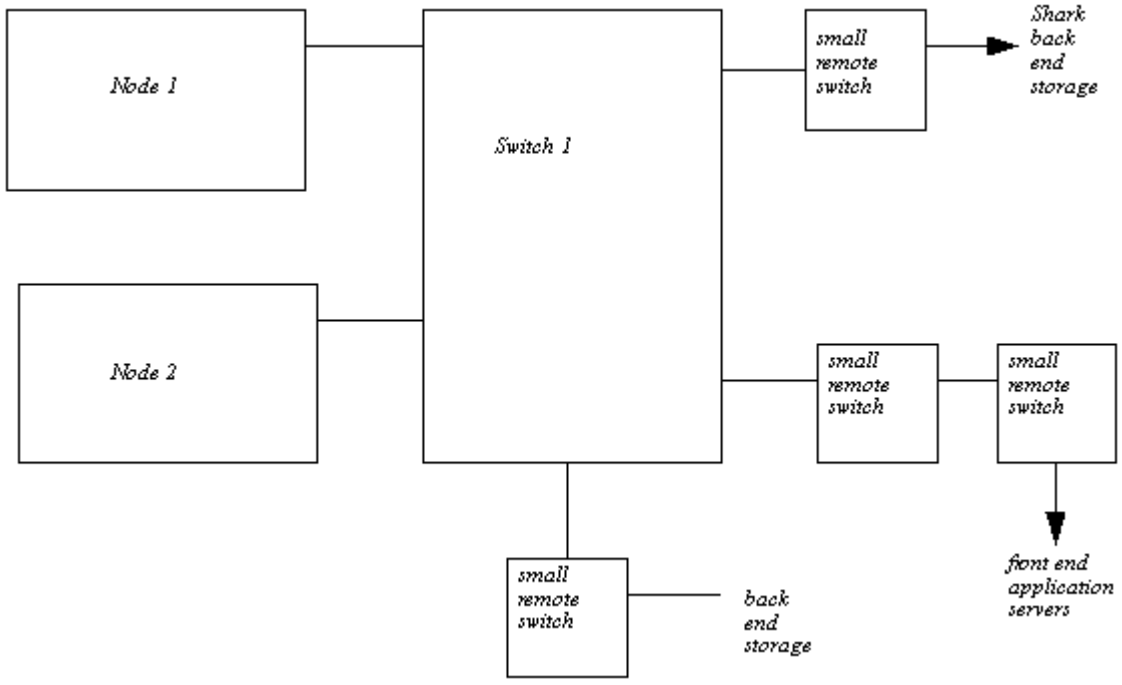



*Figure 9: Cascaded non redundant switches*

*This type of configuration is used when devices are distributed across a number of machines rooms in different places in the same large buildi fashion is required simply because of distance issues (and the need to have a few ISLs between switches in different rooms, rather than many this is a non-redundant configuration - to protect against switch failure some switches and paths would need to be duplicated.*

- Extending this idea further, a core-edge network gives a symmetrical network in which simple switches are arranged around the periphery of the network while a more powerful switch coordinates activity between switches inside the fabric. Thus simple switches appear on the edge of the network and more powerful (sometimes director class switches) appear in the core. See Figure 10 below for an examples of these configurations.

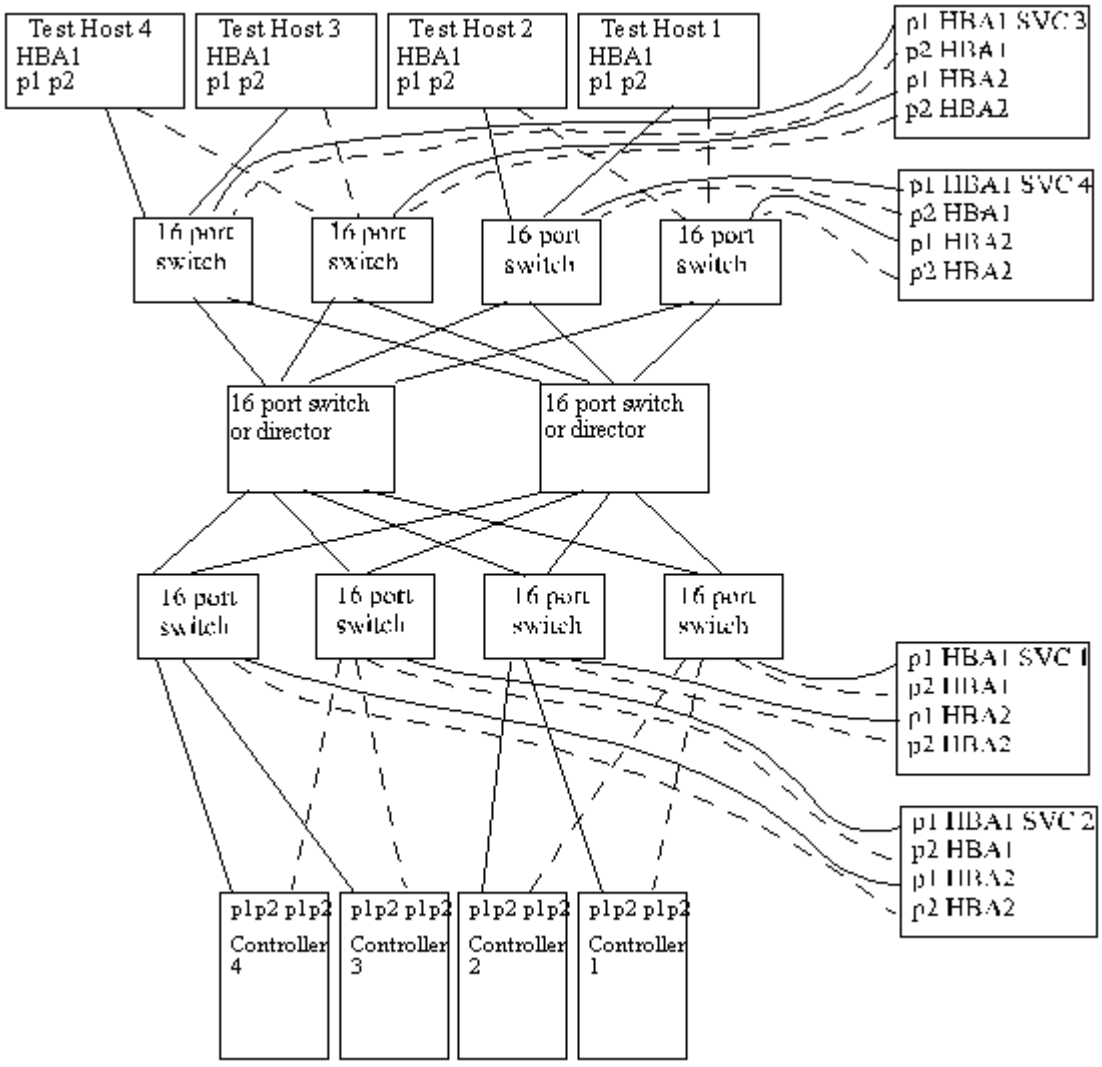



*Figure 10: 4 node Core and Edge Network*

*Note the central switches in the core are not connected together (horizontally), nor are the simple switches around the edge. Because each de*
*a single switch failure without loss of connectivity.*

# 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 Configurations

## SVC in a 16 port switch SAN

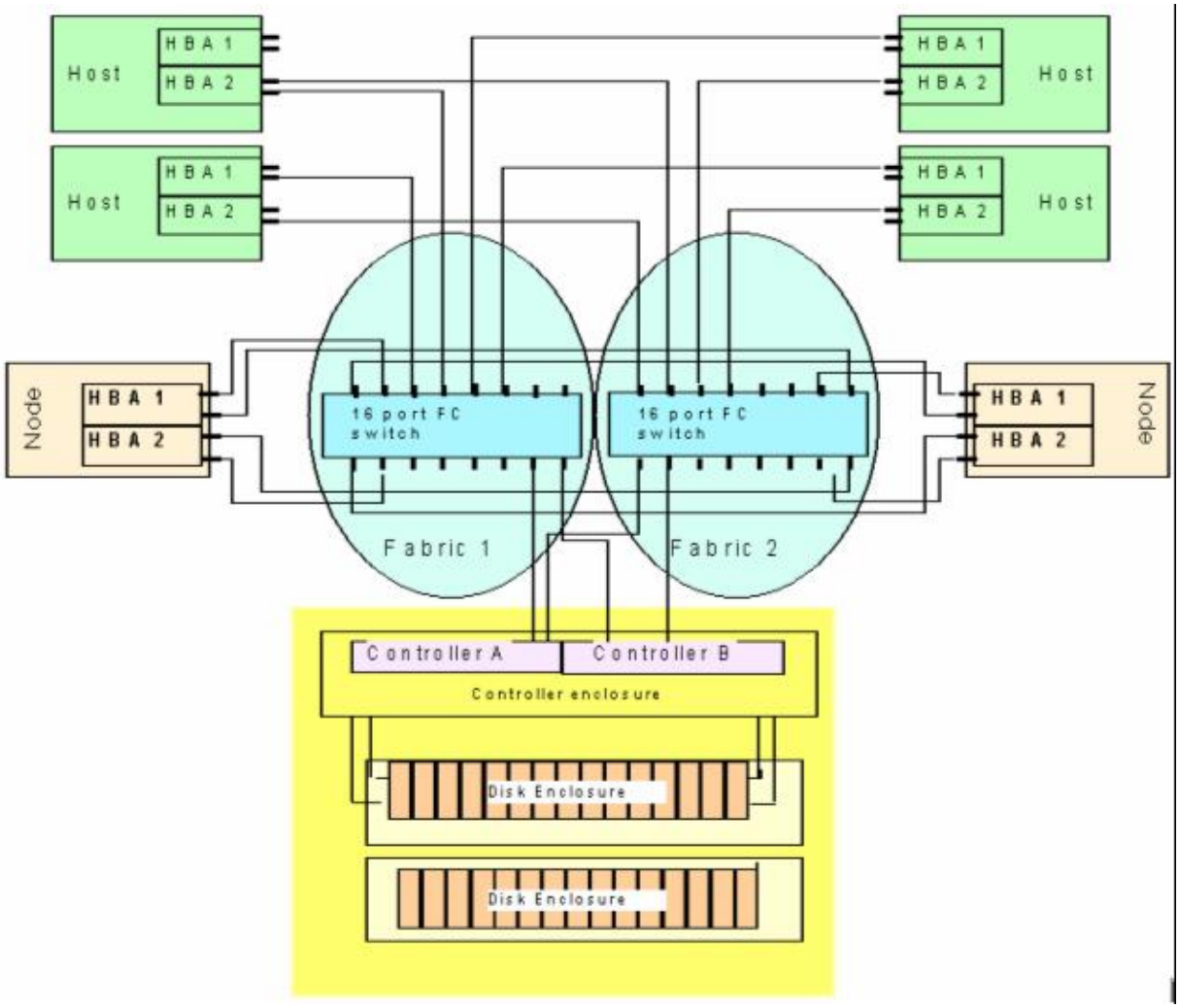


*Figure 11: A simple 2 node SVC configuration with dual SAN fabrics, employing 2 16 port switches*

A Configuration of SVC in a SAN with 16 port switches would be typical for two SVC nodes and up to four RAID controller pairs. The SVC nodes and RAID controllers would utilise eight of the ports on each switch, leaving eight for connection to Hosts. Clearly this ratio may be adjusted to allow for more RAID controllers and fewer hosts.

If an optional Service Node is included in the configuration then the Service Node will attach to one FC port on each of the switches.

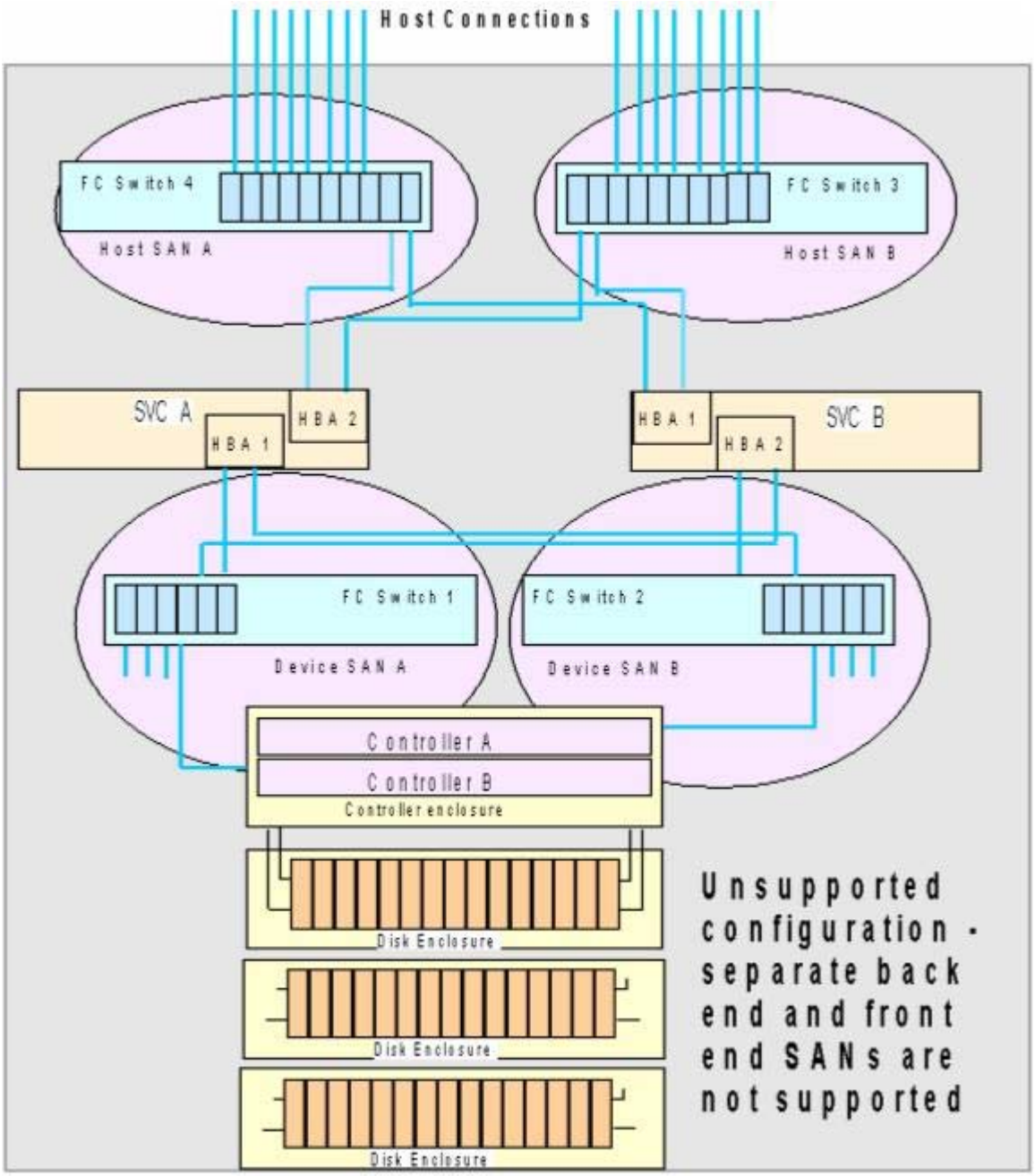



*Figure 12: SVC configuration using separate back end and front end SAN - this is unsupported.*

## SVC in a 32 port switch SAN

To connect sufficient RAID controllers and hosts to make use of more than two SVC nodes, SVC needs to be in a SAN with 32 port switches. Even so, in a SAN with a pair of 32 port switches, with four pairs of SVC nodes, 16 ports on each switch would be connected to SVC nodes, leaving 16 to share between hosts and RAID controllers. shows an example of this type of configuration.

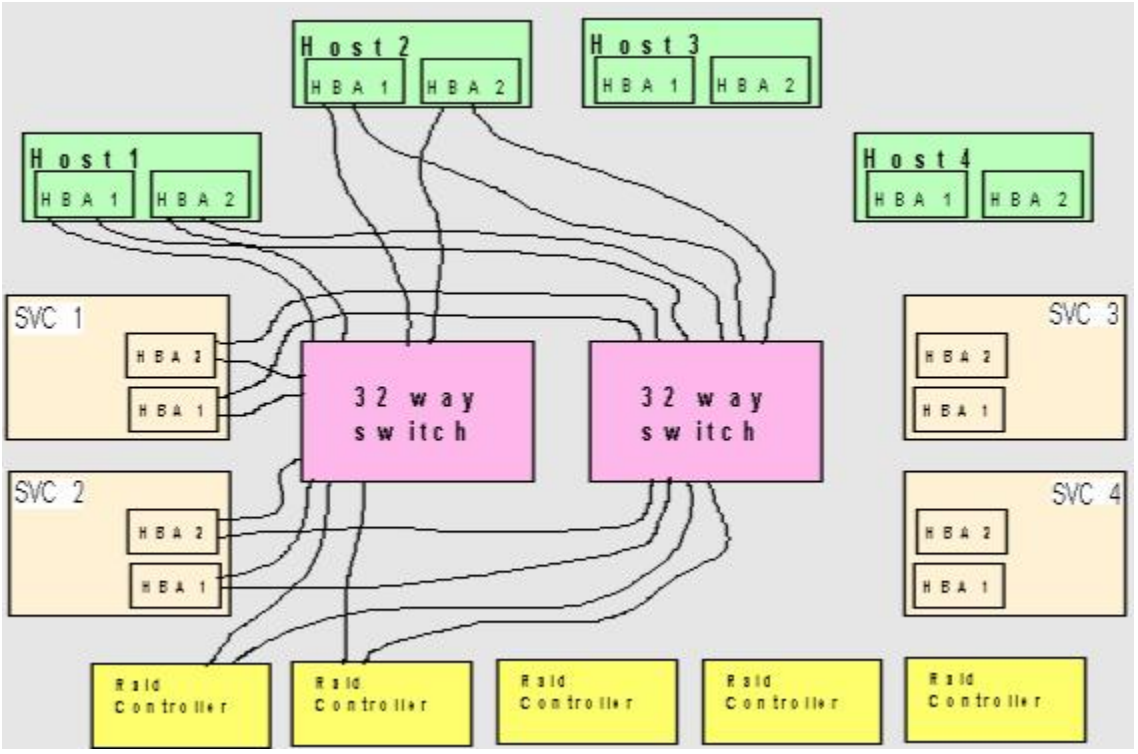



*Figure 13: :Dual switch, redundant SAN fabric, 4 node SVC configuration, using 32 port switches (partially wired)*

## SVC in a mixed vendor SAN

Since SVC may be attached to two different, separate SANs, it would be possible to have different vendor switches in a SVC configuration provided that each fabric contains only switches from a single vendor. In this configuration the two counterpart SANs have different vendor switches. This arrangement is not supported in SVC 1.2 but is supported in SVC 2.1 and later.

## SVC in a SAN with director class switches

When a large number of RAID controllers and test hosts are to be connected to a SVC cluster, it is possible that a director class switches will be employed within the SAN. Director class switches can offer internal redundancy so it may be possible to replace a SAN that employs multiple switches to provide two redundant SAN fabrics with a single director class switch. While this may give adequate SAN network redundancy this does not protect against physical damage (e.g. flood, fire) which would destroy the entire function in the director class switch. It may be more attractive therefore to use tiered networks of smaller switches or a core-edge topology, with multiple switches in the core. The redundancy in the network can then be physically distributed across a wide area protecting against physical damage.

## Example Large SAN for use with SVC

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.

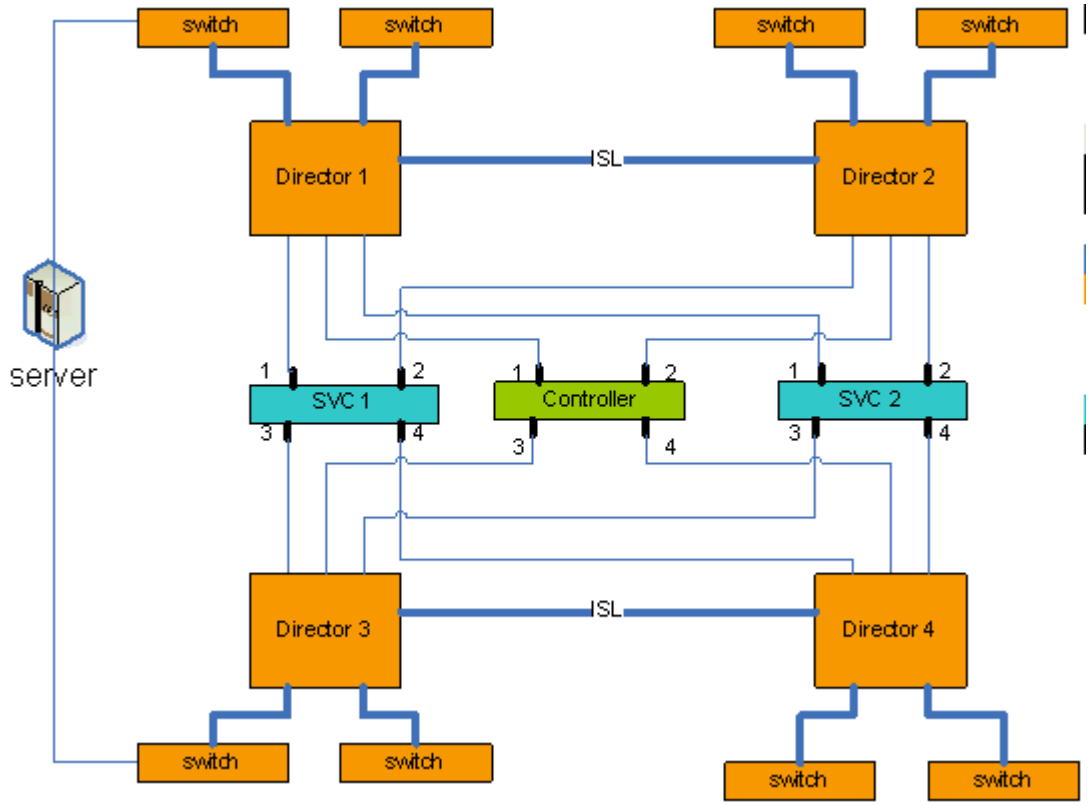



*Figure 15: A large SAN configuration*

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 director1 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.

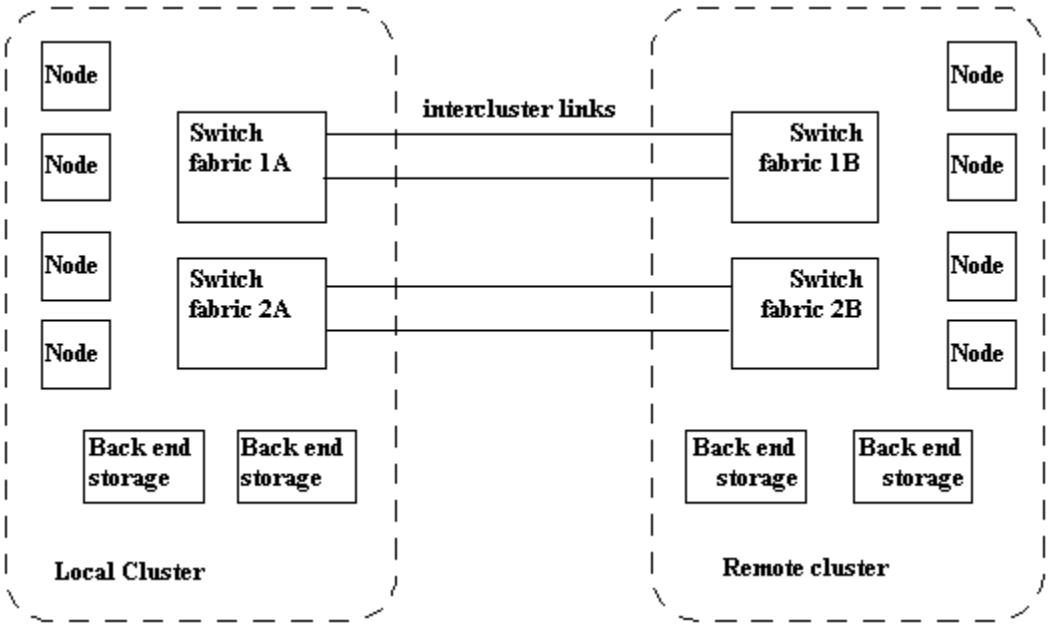


*Figure 16: Metro Mirror configuration using dual redundant fabrics*

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

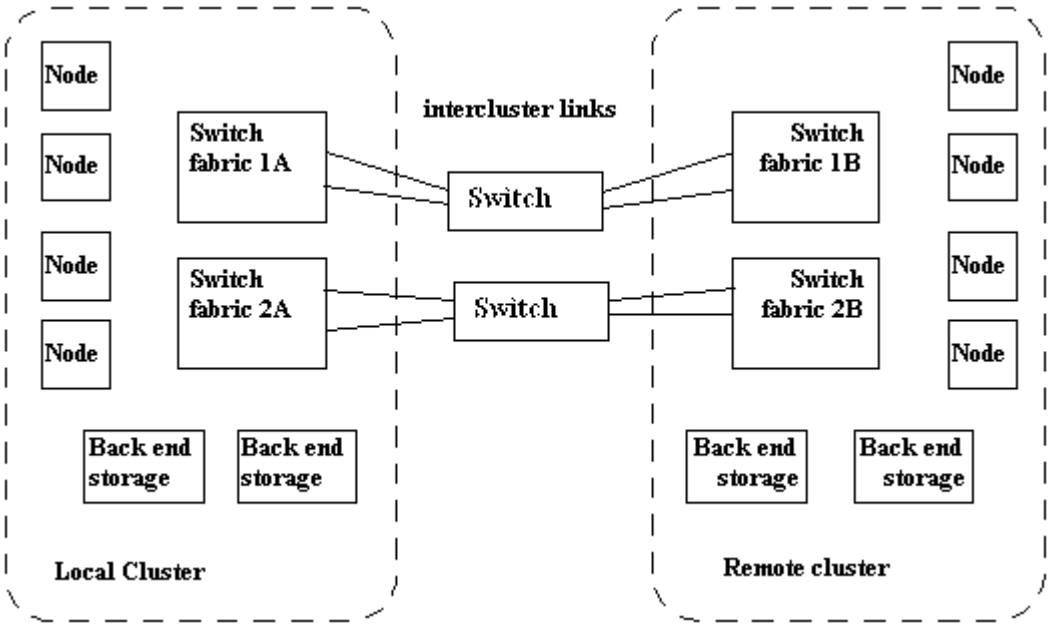



*Figure 17: Metro Mirror configuration with intercluster switching - is this a supported way of extending the distance between two clusters (provided that w*

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

# Back End Storage Configuration Recommendations

# 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 it is important to follow the configuration recommendations presented here.

The following calculations allow the user to calculate the IO rate capability of each vdisk.

Typically there are two major aspects in the IO capability of a vdisk; the IO capability of the MDisks and the number of flash copies taking place. The IO capacity of the MDisks is determined by taking into account the capability of the underlying Hard Disk Drives (HDDs), the effect of RAID and the capability of the storage controller to service IOs.

**MDisks**

The following assumes that:

- All MDisk represent entire arrays. Arrays have not been split.
- All the MDisks in the Managed Disk Group are provided by controllers of the same type.
- Virtual Disks are striped evenly across all of the MDisks in the Managed Disk Group.
- MDisks provided by one storage controller are contained entirely within one Managed Disk Group.
- All the MDisks provided by a controller are of the same RAID level.

The raw IO capability of the Managed Disk Group is the sum of the capabilities of its MDisks, which can be found in , where C is the characteristic IO capability of the backend Hard Disk Drives (HDD) and is typically 150 operations per second. For example for 5 RAID-5 MDisks with 8 component disks on a typical backend (C=150), the IO capability will be 5 * (150 * 7) = 5250.

This raw number may be constrained by the IO processing capability of the backend storage controller itself.

**Flash Copies**

The affect of flash copies depends on the type of IO taking place. The effect of a flash copy is effectively adding a number of loaded vdisks to the group. Thus a weighting factor needs to be introduced. The total weighting factor, for F flash copies is given in the table XX. For example in a group with 2 flash copies and random writes to those reads/writes to those vdisks, the weighting factor is 14 * 2 = 28.

| Type of IO (to vdisk) | Impact on IO | Weighting for flash copy |
|---|---|---|
| None/very little | Insignificant | 0 |
| Reads Only | Insignificant | 0 |
| Sequential Reads and Writes | Up to 2x IO | 2*F |
| Random reads and writes | Up to 15x IO | 14*F |
| Random writes | Up to 50x IO | 49*F |

**Overall loading**

To calculate the average IO rate per vdisk, use the following equation:

IO rate = (IO Capability) / (No vdisks + Weighting Factor)

For example, continuing the case above, we have IO capacity of 5250, a weighting of 28 and, say, 20 vdisks. So our IO rate per vdisk is 5250 / (20 + 28) = 110. Note this is average IO rate, so for example, in this example, half the vdisks could be running at 200 iops and the other half at just 20 iops. This would not overload the system, since the average load is 110.

If the average IO rate to the vdisks exceeds this figure, then the system will be overloaded. As an approximate guide, a heavy I/O rate is ~200, medium ~80 and low ~10.

If the backend is being overloaded, consider migrating some of the vdisks into other, less loaded groups.

**Managed Disk Groups**

SVC supports an arbitrary mixture of different storage controller types and RAID levels within a Managed Disk Group. There are various valid usage scenarios which might lead to a customer wanting to mix different RAID levels or storage provided by different controller types in a Managed Disk Group. Whilst such configurations are supported they are however not recommended.

It is recommended that an Managed Disk Groups should contain “similar” MDisks, that is to say they should have similar IO characteristics. It is important that a single MDisk in a groups should not be loaded significantly more than the others. It is therefore *recommended* that the following factors should not be mixed in an Managed Disk Group:

| | |
|---|---|
| Backend type: | Different backends which have widely varying performance should not be mixed. |
| Backend type: | Different backends which have widely varying availability and should not be mixed. |
| RAID type: | Different RAID types have vastly different performance. (e.g. do not mix RAID 0 with RAID 5) |
| RAID size: | The performance of most RAID types is directly related to the number of disks/spindles in the array. (e.g. RAID 5 with 5 spindles should not be mixed with RAID 5 with 4 spindles) |

| | |
|---|---|
| Different Capacities: | MDisks with significantly different capacities should not be mixed as when the group nears being full, on the larger disk and thus on average that disk will endure more IO. |

## Effect of mixing types of MDisks within a Managed Disk Group.

If MDisks of differing types are mixed in a Managed Disk Group then there will be an implication both in terms of Availability and in terms of performance.

**Availability.**

If any MDisk in a Managed Disk Group goes offline or cannot be accessed then the entire Managed Disk Group goes off line and all Virtual Disks associated with that Managed Disk Group go offline. Thus adding a single low availability MDisk into an Managed Disk Group containing high availability MDisks will reduce the availability of all of the storage associated with the Managed Disk Group.

**Performance**

The calculations in assume that the Managed Disk Group contains homogeneous MDisk types and that all Virtual Disks are striped across all MDisks. If these recommendations are not followed then it becomes more difficult to determine whether the MDisks will be overloaded.

If all of the Virtual Disks in the Managed Disk Group are striped evenly across all of the MDisks then the calculations in should be performed by making the assumption that all MDisks in the Managed Disk Group perform as badly as the worst MDisk in the Managed Disk Group.

If the Virtual Disks in the Managed Disk Group are not striped evenly across all of the MDisks in the Managed Disk Group then it becomes difficult to calculate the loading on the MDisks and this should be monitored carefully using the facilities of MDM performance manager.

# Configuration Recommendations for Large SANs.

**Purpose of these recommendations.**

The purpose of these recommendations is to avoid situations where an SVC node reaches its maximum number of queued commands. When this happens, SVC will return Task Set Full Status or will return check conditions or set unit attention conditions to indicate the commands have been discarded and must be retried by the host. Unfortunately many host operating systems do not have helpful recovery algorithms if this situation persists for more than 15 seconds and the result will often be that one or more hosts present errors to applications resulting in application failures. Following these recommendations will avoid this.

Note that this issue is not in any way specific to SVC. All controllers and operating systems have the same issues if the maximum queue depth is reached.

## Calculating a queue depth limit

In calculating the queue depth the following factors should be considered:

Although the maximum number of queued commands is per node and there are two nodes in an IO group, the system must continue to function when one of the nodes in an IO Group is not available. Thus we must consider an IO Group to have the same number of queued commands as a node. **However**

since when a node fails, the number of paths to each disk is halved. In practice this effect can be neglected and we can count nodes rather than IO groups in the calculation below.

If a Vdisk is mapped so that it can be seen by more than one host then each host that it is mapped to may send a number of commands to it.

Multipathing drivers on most hosts round robin IOs amongst the available IO paths. For hosts which do not currently do this it may be the case that this behaviour may change in the future and we need to avoid breaking customer 's configurations when this happens.

If a device driver times out a command, it will typically re-issue that command almost immediately. SVC will have both the original command and the retry in it 's command queue in addition to any ERP commands that are issued.

In order for the maximum queue depth not to be reached it must be the case for an IO group that:

For all Vdisks associated with the IO group and for all Hosts that are mapped to be able to see each Vdisk and for all paths on each host, the sum of the queue depths must be less than 10,000. Because ERPs can consume some number of queued command slots a this number is reduced to 7000 to allow a safety margin.

**Homogeneous queue depth calculation.**

In systems where:

- The available queued commands are to be shared out among all paths rather than giving some hosts additional resources.
- The Vdisks are shared out evenly amongst the IO groups in the cluster.

then the queue depth for each vdisk should be set on the hosts using the following calculation:

q = round up ((n * 7000) / (v * p*c))

q = per device path q-depth setting

n = number of **nodes** in the cluster

v = number of **vdisks** configured in the cluster

p= Number of **paths per Vdisk per host.** A path is a route from a host FC port to an SVC FC port which is recognised by the host as giving access to the vdisk.

c= the **number of hosts** with can concurrently access each vdisk. Very few applications support concurrent access from multiple hosts to a single vdisk.

Examples where multiple hosts have concurrent access to a disk include where SAN Filesystem is in use. It is likely therefore, except in these specific examples, c will typically be 1.

**As an example:**

An 8 node SVC cluster (n=8)  with 4096 vdisks (v=4096) and 1hosts able to access each vdisk (c=1) and where each host has 4 paths to each vdisk (p=4)

Calculation gives a roundup value of ((8*7000) / (4096*4*1)) = 4.

So the queue depth in the operating systems should be set to 4 concurrent commands per path.

**Non Homogeneous queue depth calculation.**

In some cases it will be appropriate to give favoured hosts additional resources to allow them to queue additional commands, or the number of vdisks supported by each IO group may be different. In these cases calculate the queue depth as follows:

Consider each IO group in turn:

For each Vdisk, consider each host to which that Vdisk has a mapping. This gives a set of (Host,Vdisk) pairs. So long as the sum of “queue depth (host,vdisk)” for all pairings is less than 7000 the system should not experience problems due to queue full situations.

**IMPORTANT - When can the recommended queue depth be increased/not set?**

The above calculation assumes that there is a significant probability that all of the hosts will initiate the number of concurrent commands that they are limited to. That is to say that each host is busy. If there are a large number of fairly idle hosts in the configuration which are not going to be initiating very many concurrent hosts then it may be possible to reason that the queue depth does not need to be limited even if the calculation above says that it should.

## How to limit the queue depth

Once the appropriate queue depth limit has been determined as described in . It must be applied. Each operating system has an OS/HBA specific way to limit the queue depth on a per device path basis. An alternative to setting a per path limit is to set a limit on the HBA. Thus if the per path limit is 5 and the host has access to 40 vdisks through two adapters (4 paths) it may be appropriate to place a queue depth limit of (40*(4*5))/2=400 on each adapter. This allows sharing of the queue depth allocation between vdisks.