BROCADE

Brocade SAN Design Guide

Table of Contents

Preface

No matter how big a single switch might be, eventually a requirement for one more port than that switch supports will arise. This “n+1” problem is solved with networking. All Storage Area Network (SAN) designs should involve planning for networks of switches, even if they do not have network requirements initially. The Brocade SAN Design Guide discusses factors that influence fabric topology and overall SAN architecture, and facilitates storage network design by explaining:

- When and how to connect fabric switches together
- When to leave them unconnected
- The pros and cons of different SAN architectures and fabric topologies

The target audience for this document consists of technical professionals responsible for the design, implementation, and administration of SANs. It focuses on delivering essential information that is necessary to design a Brocade SAN. While many of the design parameters discussed here are wide-ranging, these topics are examined here only to the extent that each affects SAN design.

To obtain additional background on SANs and SAN design, see the book *Building SANs with Brocade Fabric Switches* (published by Syngress press ISBN: 1-928994-30-x). This book offers more detailed information on a wider range of SAN topics.
It is important to have a working knowledge of Brocade SANs to effectively use this guide. The Brocade Educational Services series of courses is an excellent resource for those new to SANs and for more experienced designers who want to keep their skills sharp. You can find out more about Brocade Educational Services on the Brocade web site: *www.brocade.com/education_services*.

Section 1: Overview of SAN Design

The first part of this section is devoted to descriptions of the Brocade product line.

The second part provides a high-level discussion of requirements that can affect SAN design.

The final part is devoted to a discussion of SAN terminology. The terms and definitions provided form the foundation of SAN design terminology used throughout this document. Many of the SAN design approaches discussed in this document are similar to techniques used in other areas of computer and network design. For example, the core/edge topology described in Section 3 is reminiscent of the star topology common in Ethernet networks.

The third part briefly discusses several solutions that can be built using SANs and how these particular solutions influence a SAN design. These solutions are described in further detail in Brocade's SOLUTIONware series of technical briefs, available from *http://www.brocade.com/san/solutionware.jhtml.*

The Brocade Product Line

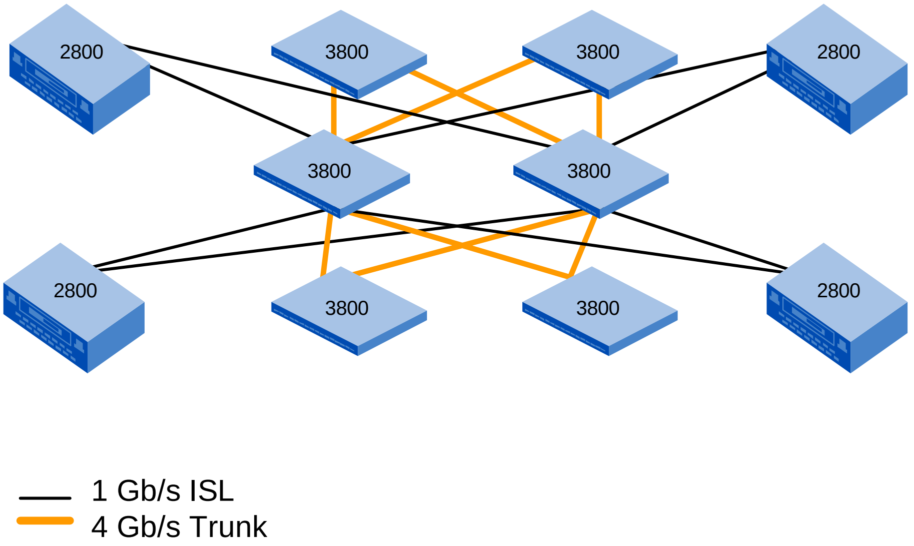
SilkWorm 2000

The SilkWorm 22x0 and the SilkWorm 20x0 switches have a single power supply and do not have hot swappable fans. The SilkWorm 2400 and the SilkWorm 2800 implement hot swap and redundant power supplies, delivering increased availability levels. The SilkWorm 20x0 and SilkWorm 2400 switches are 8-port switches and the SilkWorm 22x0 and the SilkWorm 2800 are 16-port switches. The SilkWorm 2000 series of switch utilize a v2.x version of Fabric OS. To ensure maximal functionality and scalability, v2.6.x or higher should be used. .

SilkWorm 3000

The SilkWorm 3000 series of switches utilize an advanced ASIC technology that enables advanced SAN features such as trunking and frame filtering. Frame filtering enables more granular levels of zoning and end-to-end performance monitoring. The SilkWorm 3000 series of switches are capable of 2 Gbit/sec speeds. The SilkWorm 3800 is a 16-port switch configured with redundant power supplies and fans. The impacts 2 Gbit/sec, port count, and trunking upon a SAN design are discussed earlier in this section and it is suggested to place these switches adjacent to each other or in the core of a core/edge topology, as shown in Figure 32. The SilkWorm 3000 series of switch utilize a v3.x version of Fabric OS, which is backwards compatible with v2.x versions of firmware. The size of fabric built should take into consideration the support limits of a particular Fabric OS version, which are identified in Section 5. The SilkWorm 3800 enables the creation of very dense fabrics in relatively small space due to its 1U height.

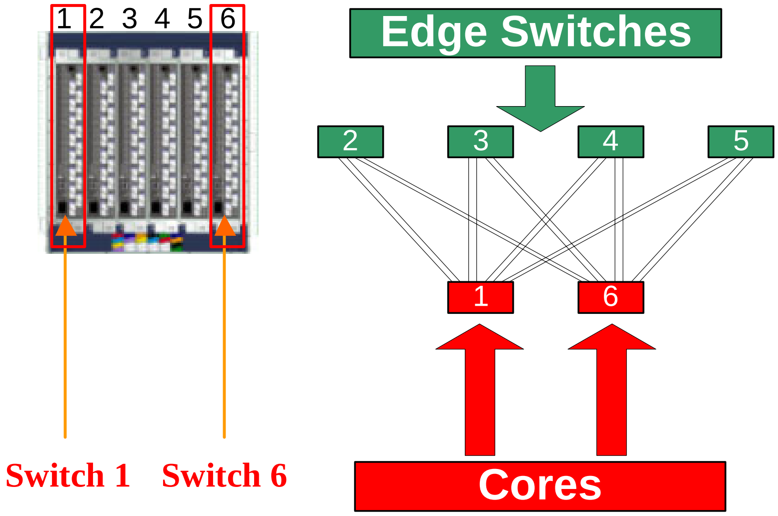
Figure 1. Place the SilkWorm 3800 in the Core of a Core/Topology



SilkWorm 6400

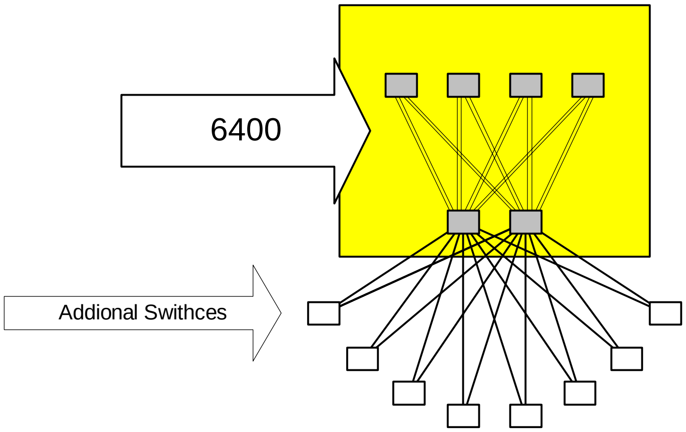
The SilkWorm 6400 is a SilkWorm 2250 based integrated fabric. The switches are connected in a core/edge topology. The SilkWorm 6400 uses six domains – one for each switch in the integrated fabric. The SilkWorm 6400 utilizes a v2.4.1x or greater version of Fabric OS. The size of fabric built should take into consideration the support limits of a particular Fabric OS version, which are identified in Section 5. When connecting other switches to the SilkWorm 6400 or interconnecting multiple SilkWorm 6400s, there are some guidelines for doing so. The switches numbers 1 and 6 are core switches (see Figure 2). When adding devices to the SilkWorm 6400, it is recommended to place the devices on edge switches 2, 3, 4, and 5 first. When devices are placed on the core switches (switches 1 and 2), it limits scalability and performance (see Device Attachment Strategies in Section 2 for the reasons why).

Figure 2. SilkWorm 6400 Switch Functions and Locations



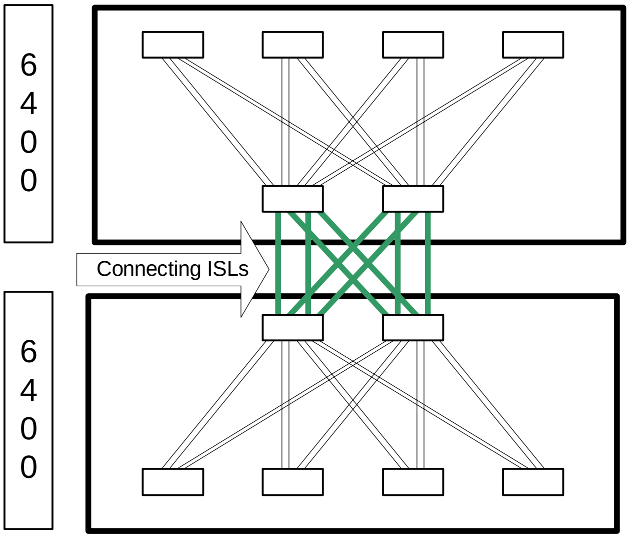
When adding switches to expand the fabric, connect those switches to the SilkWorm 6400 core switches. Doing so preserves the balanced performance characteristics of the core/edge topology. It is not necessary to attach each new edge switch with two ISLs to each core switch unless performance requirements dictate. If a single ISL is used to attach each edge switch to each core of the SilkWorm 6400, it is possible to create a 160-port fabric as shown in Figure 3.

Figure 3. Expanding the SilkWorm 6400 by Adding Switches to the Core



When connecting multiple SilkWorm 6400s together, it is most efficient to connect the SilkWorm 6400s at the cores, as shown in Figure 4. The topology shown in Figure 4 is an effective one for joining two SilkWorm 6400s that are separated by a distance or to build a larger fabric.

Figure 4. Connecting SilkWorm 6400s Together



SilkWorm 12000

The SilkWorm 12000 has a bladed architecture. Each chassis contains up to 128 ports, each of which runs at either 1 or 2 Gbit/sec. These ports are divided into two logical 64-port switches. The same advanced ASIC technology introduced in the SilkWorm 3800 is used in the SilkWorm 12000, with the same advanced features supported, and high availability/failover features have been added as well. A new Fabric OS (version 4.0) is used on the SilkWorm 12000.

There are quite a few impacts to SAN design when using a SilkWorm 12000. Some are obvious: if a 64-port solution is desired, no network is required. For large fabrics, fewer switches are required. Some implications are less obvious. In order to attach a SilkWorm 2000 or 3000 series switch to a SilkWorm 12000 network, new firmware might be required on those switches. In addition, some configuration parameters (e.g. Core Switch PID Format) may need to be changed. For a comprehensive discussion of the impacts to SAN design of using the SilkWorm 12000, please see the *Brocade SilkWorm 12000 Design, Deployment, and Management Guide* (53-0000251-xx).

Note: When deploying the SilkWorm 12000 into existing fabrics that also include SilkWorm 2000 and 3000 series switches, it is necessary to change the Core Switch PID format setting on those switches. Doing so may have an impact on existing applications, as enabling this setting will change the 24-bit address. Using a redundant fabric architecture can mitigate or eliminate this impact.

Design Considerations

SANs are built in order to solve business problems. The problem statement could be: *"Our nightly backups don't complete within the allowable window, so we need them to run faster"* or *"We need to optimize our storage utilization since we are becoming short on data center floor space."* When evaluating the following list of design considerations, keep in mind that the *overriding* concept is always the same: a SAN solution must solve the business problem that drove the creation of a SAN in the first place. In order to be effective, a SAN solution should:

- Solve the underlying business problem.
- Meet business requirements for availability and reliability[1].
- Provide the appropriate level of performance.
- Be effectively manageable.
- Be scalable and adaptable to meet current and future requirements.
- Be cost effective.
- Improve operational control of storage infrastructure.

What *is* the correct level of performance? It might – and indeed usually does – vary from host to host within an organization. What does "effectively manageable" mean? If an enterprise management environment is already in place, it could mean integrating SAN management into that environment or it might mean that it is necessary to evaluate SAN-specific management packages to get the required tools. These and other requirements must be well thought out before the design is created.

[1] Note that SANs may require higher availability than any individual attached node requires in order to meet the availability goals of the overall organization.

Design Terminology

These terms and definitions are provided to ensure that a consistent language for describing SANs is used throughout the document and so that the reader understands what these terms mean. This section is not intended to be all-inclusive. For example, latency is briefly defined here, but its significance is not discussed until later.

Blocking: The inability of one device to connect to another device. Brocade's Virtual Channel implementation of Fibre Channel does not block. The term blocking is often confused with the term congestion.

Congestion: If two or more sources contend for the same destination, performance for each source may decrease; however, available bandwidth is shared fairly by all sources contending for the same destination. Congestion is the realization of the potential of over-subscription. Congestion may be due to contention for a shared storage port or host port, or an ISL.

Core Switch: Also known as a "core fabric switch." This is one of the switches at the logical center of a core/edge fabric. There are generally at least two core switches per core/edge fabric to enable resiliency within the fabric. Ports on a core switch are normally used for ISLs.

Edge Switch: This is one of the switches on the logical outside edge of a core/edge fabric. There are generally many more edge switches than core switches. Ports on edge switches are often used for node connections.

Fabric: One or more interconnected Fibre Channel switches. The term "Fabric" only refers to the interconnected switches, not to nodes or devices connected to the fabric.

Fabric Topology: A topology is "the logical layout of the components of a computer system or network and their interconnections." A fabric topology is the layout of the switches that form a fabric.

Fabric Port Count: The number of ports available to connect nodes in a fabric. ISLs ports (E-ports) are not included in this count.

Fan-in: The ratio of storage ports to a single host port.

Fan-out: The ratio of host ports to a single storage port.

FSPF: Fabric Shortest Path First protocol. The FSPF protocol was developed by Brocade and subsequently adopted by the Fibre Channel standards community for allowing switches to discover the fabric topology and route frames correctly. It is now the industry standard routing protocol for Fibre Channel networks.

Hop Count: For evaluating SAN designs, the hop count is identical to the number of ISLs that a frame must traverse to reach its destination.

ISL: Inter-Switch Link. ISLs connect two switches via E-ports.

ISL Over-Subscription Ratio: In networks where all ports operate at the same speed, the over-subscription ratio for an ISL is the number of different ports that could contend for the use of its bandwidth. If there are 14 node ports on a switch and 2 ISLs, the ratio is 14:2, or 7:1. When there is a mixture of port speeds, the exact calculation can become unnecessarily complex. This will be discussed later. The rule of thumb is that the lower the ratio is, the better performance is likely to be. However, in most environments, designing for a ratio lower than 7:1 does not provide greater real-world performance; it just adds cost.

Latency: The time it takes for a frame to traverse from its source to its destination is referred to as the latency of the link. Sometimes a frame is switched from source to destination on a single switch and other times a frames must traverse several hops between switches before it reaches its destination.

Locality: The degree that I/O is confined to a particular switch or segment of a fabric. If two devices that need to communicate with each other are located on the same switch or fabric segment, then these two devices are said to have high locality. If these same devices are located on different switches or segments of a fabric and these two devices need to communicate with each other, then these devices are said to have low locality.

Node: Any SAN device – usually either a host or storage device – that attaches to a fabric.

Node Count: The number of nodes attached to a fabric.

Over-Subscription: A condition where more nodes could potentially contend for the use of a resource – such as an ISL – than that resource could simultaneously support, that resource is said to be over-subscribed.

Radius: The greatest "distance" in hops between any edge switch and the center of a fabric can be thought of at that fabric's radius. Low radius networks have lower hop counts and latency than high radius fabrics. The unit of measurement for a fabric radius is hops.

Resilience: The ability of a fabric to adapt to or tolerate a failure of a component.

SAN: A Storage Area Network (SAN) can consist of one or more related fabrics and the connected nodes.

SAN Architecture: The overall design or structure of a storage area network solution. This includes one or more related fabrics, each of which has a topology. Other components may also be included, such as host, storage, and other SAN devices.

SAN Port Count: The number of ports available for connection by nodes in the entire SAN. The SAN Port Count equals the fabric port count in a single fabric SAN and is equal to the sum of each fabric's port count in a multi-fabric SAN.

Scalability: The ease with which a particular design can grow and adapt without requiring a significant change in architecture or requiring a substantial re-layout of existing nodes.

SPOF: A single point of failure. A SPOF in a SAN is any component – either hardware or software – that could cause a fabric or a SAN to fail.

Tiering: The process of grouping particular SAN devices by function and then attaching these devices to particular switches or groups of switches based on that function

SAN Solutions

The adoption of SANs is being driven by a variety of objectives. Some examples are:

- The need for more efficient usage of enterprise storage arrays
- Decreasing size of backup/restore windows
- Increasing size of data set to be backed up
- The need for improved high availability and disaster tolerance solutions
- The need to enhance storage resource management

While many SAN users begin their SAN experience with one particular SAN solution, the SAN quickly becomes the basis for many other applications. For example, a company may start out with SAN-based backup and very quickly integrate storage consolidation and clustering into the existing SAN foundation. In that respect, a SAN decision is a strategic one, and should receive an appropriate level of attention.

Three of the most popular SAN solution categories are Storage Consolidation, LAN-Free Backup, and High Availability. Each of these SAN solutions are generically described below and key attributes of each are discussed in terms of their affect on SAN design. For a more detailed discussion regarding the configuration, design, and implementation of a SAN solution, reference Brocade SOLUTIONware, which is available from: *http://www.brocade.com/san/solutionware.jhtml*.

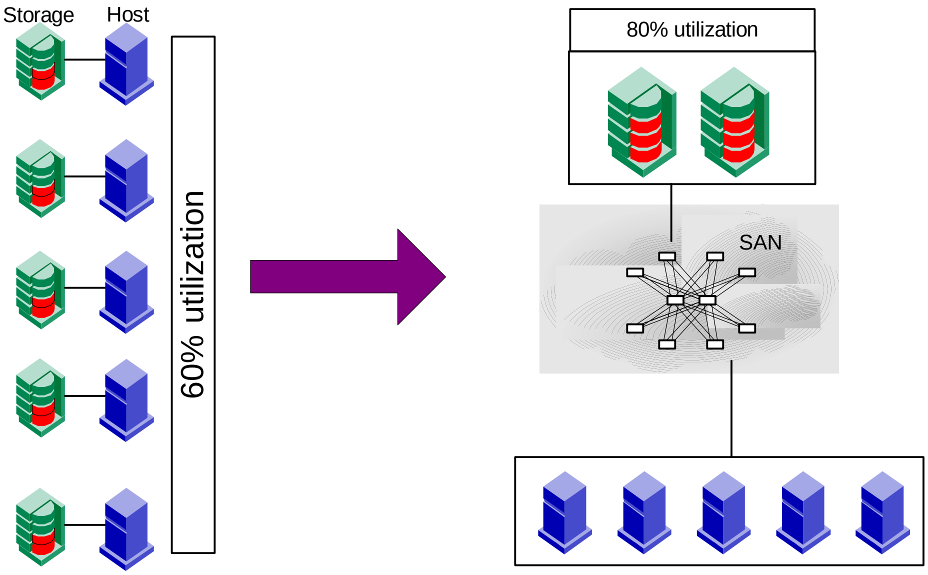
Storage Consolidation

Storage consolidation is a way of optimizing storage resource utilization. It is often the result of migrating directly attached storage (DAS) and hosts to a SAN environment. In a SAN, it is no longer necessary to have a one-to-one correspondence between a host port and a storage port. Instead, many hosts can share a single storage port, and a single host port can access many storage devices. This immediately reduces cost on hosts because fewer HBAs are needed, and on storage because fewer controllers are needed. In addition, savings can accrued by reducing storage management, power, cooling, and, floor space costs. However, the greatest savings comes from improved utilization of free space on enterprise storage subsystems. With the lowering cost of FC HBA and switch infrastructure, the storage consolidation value proposition has never been better.

Assume that 20 hosts each have 100 GB of storage in a direct attach environment, requiring a total 2000 GB of storage. Some space on each system is free. This is known as white space, or headroom. The average utilization of this directly attached storage (DAS) is 50%, leaving 50% white space. The total storage utilized is 1200 GB, which leaves 800 GB of white space.

With the use of a SAN, it is possible to achieve much higher utilization since every host has access to all storage in the SAN. In this example, a modest 10-20% improvement in storage utilization could result in a savings of several hundred GB of storage. In addition, a reduction in associated ownership costs of that surplus storage would occur. In the storage consolidation model, if a host is not using all of its storage, it is possible to rapidly reallocate this extra storage to a different host. It is also possible to add additional storage for all servers to access, rather than having to purchase storage for specific hosts. In a direct attach environment, it is more difficult to do so, forcing the need to have very high white space overhead to allow growth. A conservative 60% utilization scenario with DAS and storage consolidation environments are compared in Figure 5.

Figure 5. Direct Attach / Storage Consolidation Comparison



Since many hosts depend upon continuous access to their storage in a storage consolidation solution, designing a highly available SAN to ensure this continuous access is critical. Resilient and redundant fabric designs are highly recommended, especially in large storage consolidation solutions. These topics are discussed further in Section 2 under Availability on page 16.

In a storage consolidation solution, many devices contend for a shared storage port. The performance-limiting factor is often the over-subscription or fan-out ratio of that port, and not the network. Because of this, it is possible to design SANs with a certain amount of over-subscription without adversely affecting application performance. The relationship between applications, SAN design, and performance is explored further in Section 2.

Because the benefits of storage consolidation grow proportionally with the number of hosts and storage, the capability for a SAN to scale is important. You can choose a SAN architecture that can grow from tens of ports to hundreds, and in some cases, thousands of ports, while minimizing or eliminating downtime. Topologies such as the core/edge are optimal for enabling this type of scaling. A discussion of fabric topologies and multi-fabric architectures is presented in Section 3.

LAN-Free Backup

A SAN-based backup is, in some respects, a form of storage consolidation in that an I/O device (the tape drive) is available to be shared by many hosts. The difference is that the shared device is tape, rather than disk. This distinction can affect SAN design in several ways:

- Currently, tape libraries tend to be single-attach, so the multi-pathing approaches used in storage consolidation will usually not work.
- Backup devices tend to be more sensitive to I/O disruption than disk arrays. Arrays can recover from small glitches; tape solutions sometimes do not recover as easily. This is a known issue in the industry and something being addressed with the emergence and adoption of the FC-TAPE standard.

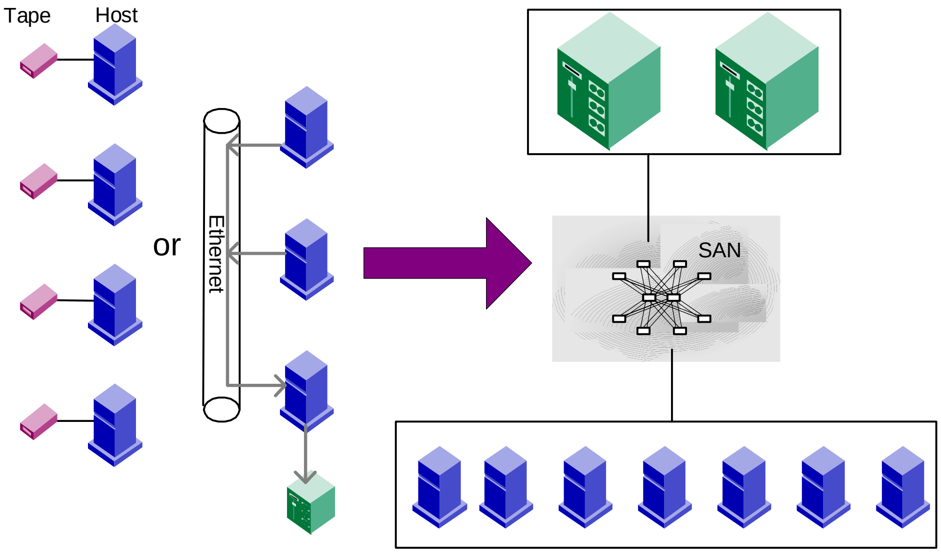
- The availability of tape drives is usually not as critical as that of disk arrays.
- Per-port performance requirements are usually lower for tape than for disk.

Non-SAN based backups take the form of direct attach tape drives, or backup over IP networks. IP backups contend with the normal day-to-day traffic already on the Local Area Network (LAN). Using direct attach tape on each host is costly because of the number of tape devices, tape management, and increased infrastructure cost for floor space, power, cooling, etc.

High-speed SAN-enabled backups reduce backup and restore windows and can enable disaster tolerance by locating libraries at remote sites. SAN based backup improves on traditional backup by enabling the sharing of fewer, larger tape libraries and by minimizing or eliminating the performance issues associated with traditional backup architectures, as depicted in

Figure 6. It is also effective to leverage the performance capabilities of Fibre Channel by running backups in more traditional mode by backing up clients via IP over Fibre Channel (IPFC) to backup server, which in turn then writes the data to tape via SCSI over Fibre Channel (FCP).

Figure 6. Direct attach and LAN based backup compared to a SAN based backup



A disruption in a backup SAN is usually not as critical as a disruption in a storage consolidation SAN. Mission critical applications require continuous access to storage, while a tape backup normally can be restarted without end users seeing the effect. Therefore, a SAN architecture solely used for backups may not require the highest availability enabled by a dual fabric architecture, and a single resilient fabric may provide sufficient availability. Core/edge, mesh, and ring topologies are all candidates for a backup SAN.

Clustering

High-availability (HA) clusters are used to support critical business applications. They provide a redundant, fail-safe installation that can tolerate equipment, software, and/or network failures, and continue running with as little impact upon business as possible.

HA clusters have been in use for some time now. However, until the advent of Fibre Channel, they were very limited in size and reliability. This is because clusters require shared storage, and sharing SCSI storage subsystems is difficult and unreliable. In fact, sharing a SCSI device between more than two initiators is *completely* impractical due to SCSI cabling limitations, and SCSI's poor support for multiple initiators.

Clustering technology has therefore been greatly enhanced by the network architecture of SANs. SANs provide ease of connectivity, and the ability to interconnect an arbitrarily large number of devices. SANs can support as few as two hosts in a failover configuration, and can be expanded to support "many-to-one" configurations. The primary advantages that a SAN affords a cluster are connectivity, scalability, and reliability.

Summary

SAN design can appear to be a challenging task, due to the large number of variables involved in picking an appropriate design strategy. The key to a successful SAN design is to thoroughly define the requirements of the *solution(s),* which the SAN will support. This understanding will make it possible to choose appropriate design from those discussed in the following sections.

Section 2: SAN Design Concepts

Several SAN design concepts are discussed in depth in this section to provide a foundation for describing a fabric topology or SAN architecture in more detail. These key concepts are: availability, scalability, and performance. To effectively describe these SAN design concepts, it is necessary to refer to certain topologies, which are detailed further in Section 3.

Differing levels of detail can be given on any networking topic. This guide takes a middle-of-the-road approach. Most of the important concepts are given, and topologies are listed at the end of the guide for quick reference. For a *more* detailed discussion of SAN design, see chapters 5 and 7 in *Building SANs with Brocade Fabric Switches* from Syngress Press. For a *less* detailed approach, see the *Sales Guide to SAN Design* (part number 53-0000253-xx) on Brocade's web site.

Availability

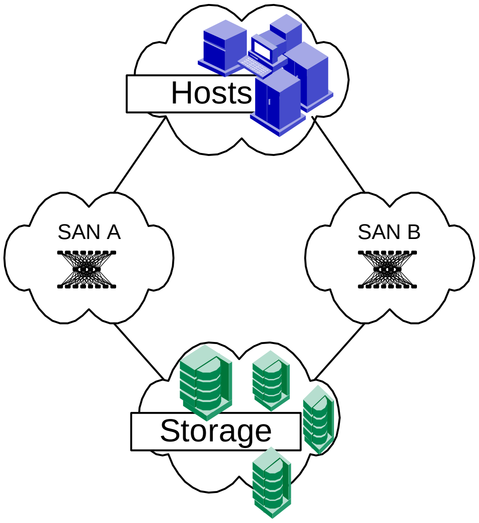
A computer system or application is only as available as the weakest link. To build a highly available computer system it is not sufficient to only have a highly available SAN. It is necessary to account for availability throughout the entire computer system: dual HBAs, multi-pathing software, highly available and multi-ported storage subsystems, and clustering software are some of the components that may make up such a system.

When building a highly available computer system, use *redundant* components. One of anything is not HA. Webster's dictionary provides the following definition for the word redundant:

> *Redundant: serving as a duplicate for preventing failure of an entire system (as a spacecraft) upon failure of a single component*

As the prevalence of SANs increases and businesses integrate SANs into their entire computing infrastructure, there is just too much risk in relying on any single entity, even a single fabric. A single fabric, no matter how resiliently designed, is not immune to all failures: human error, disaster, software failure, or a combination of unforeseen events can cause the failure of up to the entire fabric. Using redundant fabrics increases the level of availability significantly, as redundant fabrics mitigate all of these possible causes of failure. Figure 7 depicts a dual fabric redundant SAN.

Figure 7. A Resilient/Redundant SAN Architecture



Availability Classifications

Devices attached to a fabric may require highly reliable access to support applications such as storage consolidation, server clustering, high availability, or business continuance operations. There are four primary categories of availability in SAN architecture. In order of increasing availability, they are:

Single fabric, non-resilient: All switches are connected to form a single fabric, which contains at least one single point of failure. The Cascade topology is an example of this category of SAN.

Single fabric, resilient: All switches are connected to form a single fabric, but there is no single point of failure that could cause the fabric to segment. Topologies such as ring, full mesh, and core/edge topologies are examples of single, resilient fabrics.

Multi-fabric, non-resilient: The most common multi-fabric SAN is the dual fabric SAN. In a dual fabric non-resilient SAN, half of the switches are connected to form one fabric, and the other half form a separate fabric. This model can be extended to more than two fabrics if desired. Within each fabric, at least one single point of failure exists. This design can be used in combination with dual-attached hosts and storage devices to keep a solution running even if one fabric fails, or if a rolling upgrade is needed.

Multi-fabric, resilient: The most common multi-fabric SAN is the dual fabric SAN. In a dual fabric resilient SAN, half of the switches are connected to form one fabric, and the other half form a separate fabric. This model can be extended to more than two fabrics if desired[2]. No fabric has a single point of failure that could cause the fabric to segment. This design can be used in combination with dual-attached hosts and storage devices to keep an application running even if one entire fabric fails due to operator error, catastrophe, or quality issues. This is the best design approach for high-availability environments. Another key benefit of this design is the

[2] This is frequently done in storage consolidation solutions. To be most effective, the arrays must have many ports and be able to present the same LUN out all ports simultaneously.

ability to take part of the SAN offline for rolling upgrades or maintenance without affecting production operations on the remaining fabric(s). Thus, upgrades can be performed without *path* downtime.

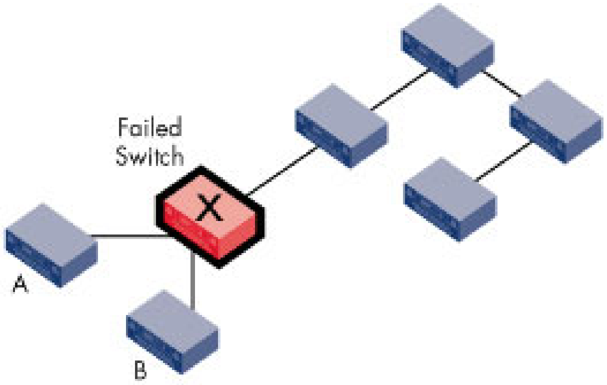
Both resilient and non-resilient dual fabrics can be referred to as “redundant fabric SANs.” Redundant designs are always recommended for HA systems, and any large SAN deployment where downtime for the entire SAN could affect hundreds of servers.

Resilient Fabrics

Many fabric topologies are available that that provide at least two internal fabric routes between all switches that comprise that fabric. These topologies are considered resilient because each topology can withstand a switch or ISL failure while the remaining switches and overall fabric remain operational. This self-healing capability is enabled by the Brocade-authored Fabric Shortest Path First (FSPF) protocol[3].

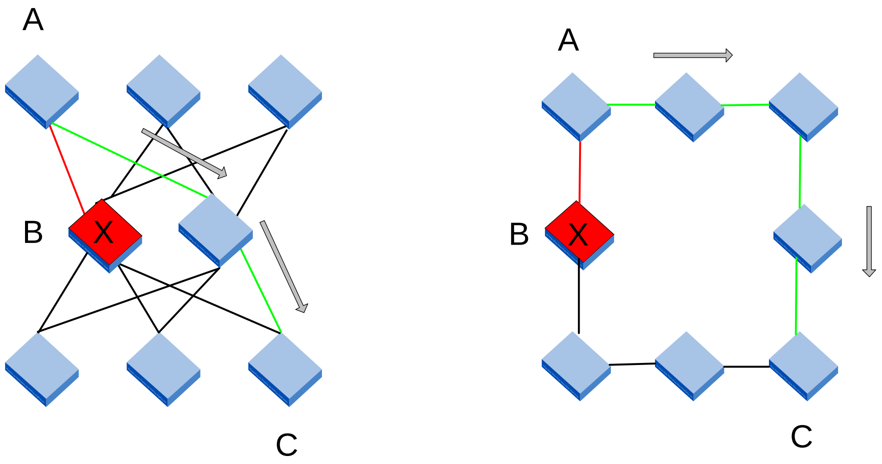
Figure 8 depicts the failure of a switch in a Cascade topology. Switches A and B are unable to communicate with the remaining switches when the switch marked with the “X” fails, resulting in the fabric segmenting into three separate fabrics. However, a switch failure in a Ring, core/edge, or other resilient topology fabric does not cause a loss of communication with the remaining switches, as shown in Figure 9. If switch B fails, switch A can still communicate with switch C through the alternate path indicated by the arrows. The fail over to alternate paths is effectively transparent to the attached devices. This fail over is performed by FSPF, which automatically reroutes the data around the failure.

Figure 8. Resilience in a Cascade topology



[3] While originally a Brocade-only protocol, FSPF has been accepted by the standards bodies as the standard protocol for Fibre Channel fabric routing.

Figure 9. Resilience in a Core/Edge and Ring topologies



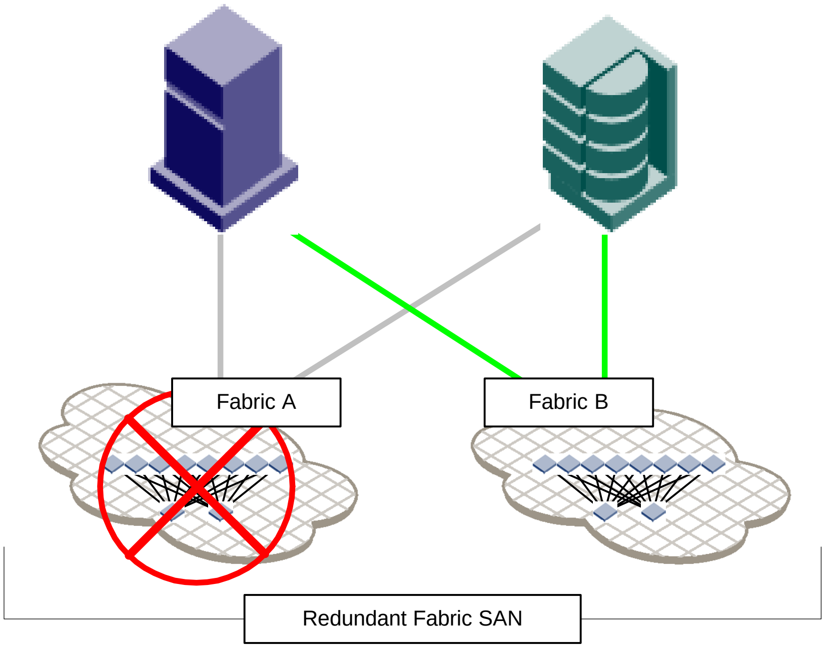
Redundant Fabrics

Resilient fabrics and the fault tolerant components that comprise them are very reliable. However, no single fabric can ever truly be an HA solution. Because all switches in a single resilient fabric have common software components, the fabric itself is still potentially subject to failures caused by things like disaster, operator error, and software malfunctions. To account for those categories of error, another level of availability must be used: The redundant fabric SAN. This is sometimes known as a multi-fabric SAN.

Redundancy in SAN design is the duplication of components up to and including the entire fabric to prevent the failure of the SAN solution. Even though an airplane navigation system (e.g. a GPS) is resilient to failures, most jumbo jets also have a redundant navigation system (e.g. a magnetic compass and a map) so that the jet will not get lost even if the resiliency fails to keep the primary navigation system up.

Using a fully redundant fabric makes it possible to have an entire fabric fail as a unit or be taken offline for maintenance without causing downtime for the attached nodes. When describing availability characteristics, what we are concerned with is *path* availability. If a particular link fails, but the path to the data is still there, no downtime is experienced by the users of the system. It is possible that a performance impact may occur, but this is a very small event compared to one or many crashed servers. Dual fabrics must be used in conjunction with multiple HBAs, multiple RAID controllers, and path switchover software to be effective. Figure 10 illustrates the ability of redundant fabrics to withstand large-scale failures.

Figure 10 Failure of an entire fabric



In a redundant SAN architecture, there must be at least two *completely separate* fabrics – just as a high-availability server solution requires at least two completely separate servers. Duplicating components and providing switchover software is well established as the most effective way to build HA systems. By extension, multi-fabric SAN architectures are the best way to achieve HA in a SAN.

In addition to enhancing availability, using redundant fabrics also enhances scalability. Using dual redundant fabrics essentially doubles the maximum size of a SAN. If a fabric is limited by vendor support levels to 20 switches / 200 ports and a single fabric solution with dual attach devices is utilized, then the SAN is limited to 200 ports. Two hundred dual attach ports is equivalent to 100 devices. However, if a dual fabric with dual attach device solution is utilized, the SAN is capable of supporting 400 ports or 200 devices[4].

Any devices that are dual attached and are capable of supporting an active-active dual-path essentially double the potential bandwidth. An active-active dual path means that I/O is capable of using both paths in normal operation. Some devices only support active-passive dual-pathing. With active-passive dual-pathing, the passive path is utilized only when the primary path fails.

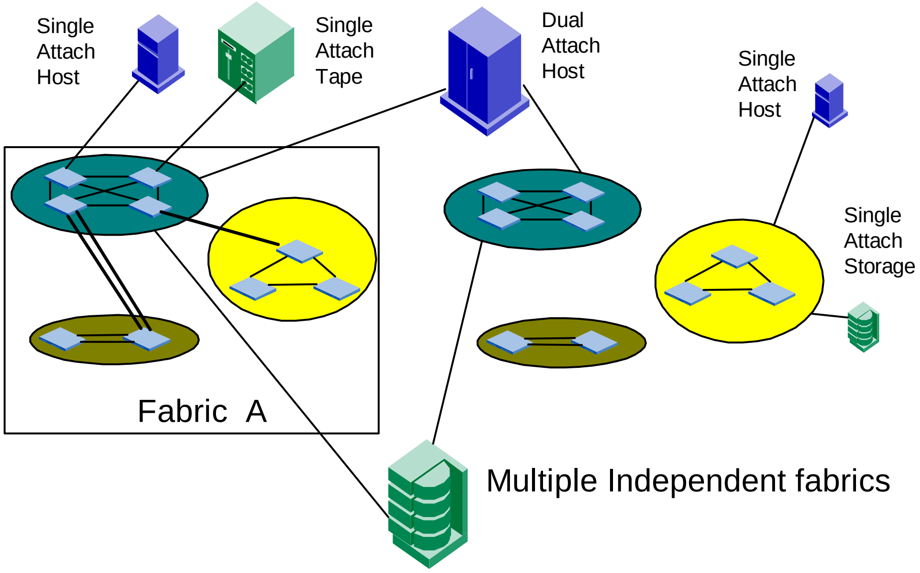
Some devices, such as tape drives, are not currently capable of supporting multiple paths. It is possible to address this issue by equally distributing tape devices between the redundant fabrics and configuring the backup applications to use an alternate tape drive should an outage on one of the fabrics occur.

[4] Note that the per-port cost of the single 400-port fabric vs. the dual 200-port fabric is the same.

Any single attached devices, such as a tape drive, non-critical storage and hosts can be single-attached, by alternately assigning them between the fabrics. When implementing a logical group of single-attached devices, ensure that all devices required by them reside on the same fabric.

When deploying redundant fabrics, it is not always necessary to deploy symmetrical fabrics. For example, when using a dual fabric, the first fabric could consist of several interconnected SAN islands, while the second fabric consists of isolated islands, as shown in Figure 11. Redundancy is still maintained for dual attach devices.

Figure 11. Asymmetric Redundant SAN



Scalability

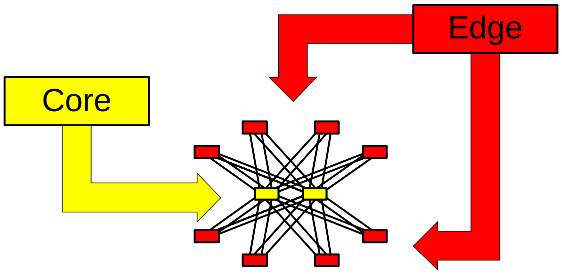
The scalability of a SAN is the size to which that SAN could be expanded without fundamental restructuring. Scalability is so important to SAN design that it is frequently the first criteria used in deciding how to approach the SAN architecture: the designer starts with asking "how many ports does the SAN need now, and how many *will* it need in the near future" and then designs a solution to meet the port count requirement.

SANs should be designed to scale to the largest size that they could reasonably be expected to need to achieve in a reasonable time frame, rather than merely using the requirements at the time of implementation as a target. This will prevent the SAN from being "painted into a corner," and needing to be fundamentally restructured after entering production.

Investment protection is another area that relates to scalability. If an existing switch is replaced with a newer or higher port count switch to increase scalability, it is valuable to reuse the existing switch elsewhere in the fabric. Proper initial planning facilitates this as well.

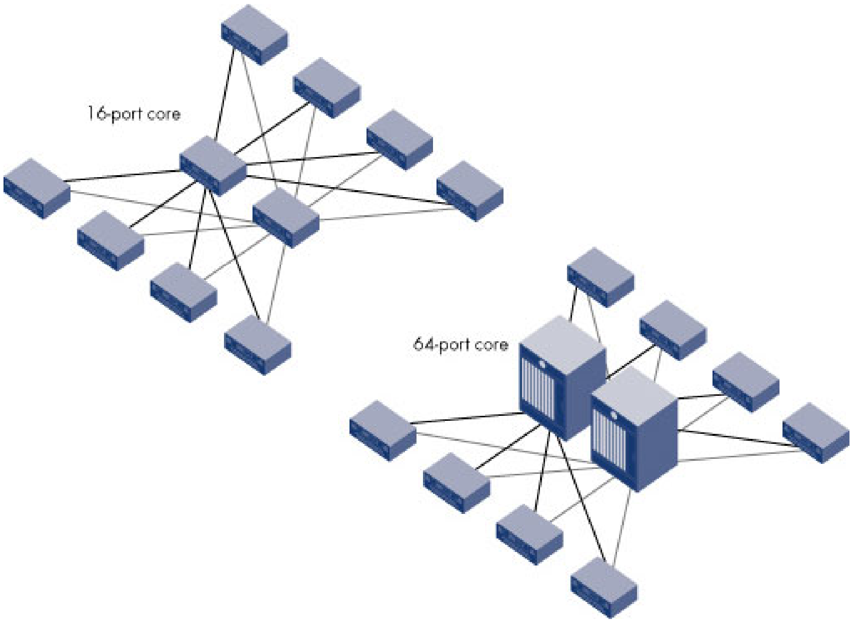
The core/edge fabric topology is the most frequently deployed topology in cases where scalability needs are great. It is derived from the star topology, which is common in traditional data networks. With a star topology, each network device is connected to a common central network, frequently known as the backbone. The edge network devices may possibly have several hops separating them from the backbone. The core/edge fabric topology (see Figure 12) is a similar design, except that the core is redundant, and there is typically only one level of edge switches (few hops).

Figure 12. Core/Edge fabric topology



A core/edge topology is scalable from many perspectives. It is possible to use variable size switches in the cores and the edges. The larger the core switch, the larger the fabric can grow. If large cores and edges are utilized, it is possible to build very large fabrics. “Large” is a relative term. If using 64-port core and edge switches, it is possible to build a core/edge fabric that yields 3,968 or more fabric ports using the same architecture[5]. This concept of scaling a fabric by using variable port count switches is shown in Figure 13.

Figure 13. Core/Edge topology with superior scalability

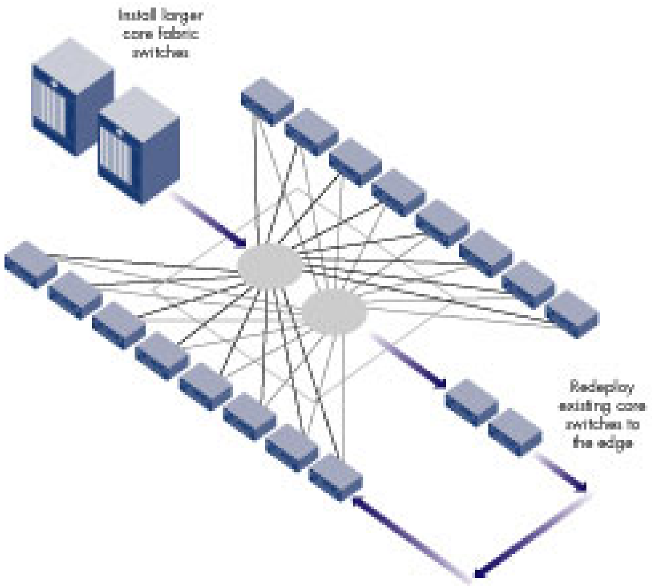


[5] This number represents the theoretical maximum scalability of the topology. Currently, no fabrics of this size have actually been built and tested

A reasonable network progression might start with 16-port core switches and migrate to 64-port cores when the scalability limits of the smaller cores are reached. See the *SilkWorm 12000 Core Migration Users Guide* (part number 53-0000477-xx) for detailed information on how such a migration would be performed, as well as tips on how to configure the network initially to facilitate this migration.

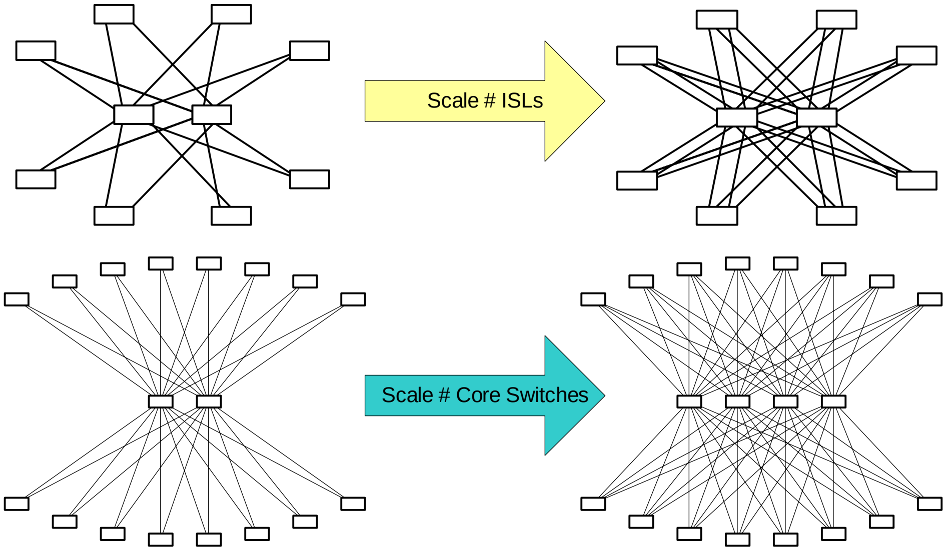
As shown in Figure 14, when additional ports are required, the 16-port switches in the core can be replaced with the higher density 64-port switches. The 16-port switches can then be redeployed at the edge.

Figure 14. **Upgrading the core to a higher port-count switch and redeploying the former core to the edge**


Install larger core fabric switches
Redeploy existing core switches to the edge

If greater bandwidth is required between the edge switches, it is possible to scale the number of ISLs from the edge switches to the core, as shown in Figure 15. It is also possible to scale the bandwidth between any two-edge switches by increasing the number of core switches.

Figure 15. Increased bandwidth between edge switches through the addition of ISLs or the addition of core switches



Performance

Fibre Channel currently performs at 1 Gbit/sec and 2 Gbit/sec with plans for 10 Gbit/sec currently in the standards bodies. Surprisingly, few applications are capable of sustaining even 1 Gbit/sec in bandwidth. For many SAN users, performance is a secondary concern – unless poor performance is inhibiting a SAN solution from functioning properly. The key to avoiding performance issues is identifying the performance requirements of the SAN during the design phase of the SAN life cycle. Then balance these requirements with the other factors that can affect a SAN into the overall SAN design. If a SAN is designed effectively, it can accommodate changes in performance without requiring a redesign. An effective SAN architecture is an architecture that can accommodate changes in performance requirements and incorporate additional switches, ISLs, and higher speed links with minimal impact to production SAN operations.

Blocking and Congestion

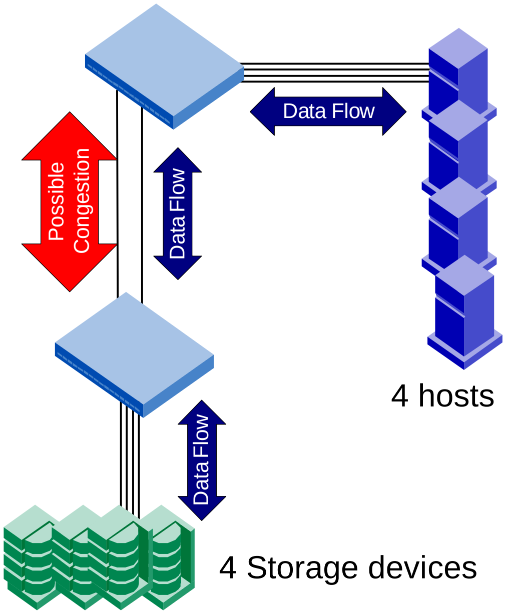
Due to the nature of the Brocade advanced virtual channel architecture, all SilkWorm switches are non-blocking. Any two ports on the switch—for example, ports A and B—can communicate with each other at full bandwidth as long as no other traffic is competing for ports A and B. However, when two or more switches are interconnected by an ISL, there is a potential for congestion. Congestion is the result of less available bandwidth than what the traffic patterns demand. Frequently, the term blocking is incorrectly used to describe congestion.

Locality

If devices that communicate with each other are connected to the same switch or groups of switches then these devices have high locality. If two devices must cross an ISL to communicate, then these devices have low locality.

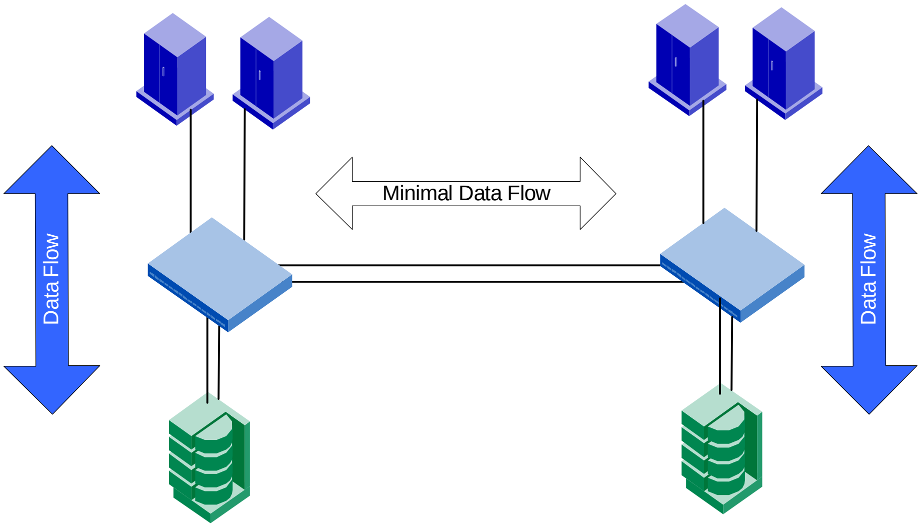
Figure 16 depicts the scenario of zero traffic localization. When host and storage devices need to communicate in a zero localization scenario, all traffic must traverse through ISLs. If four 1 Gbit/sec hosts in the figure need to concurrently communicate with four 1 Gbit/sec storage devices/connection at full bandwidth, congestion occurs in the ISLs. This is because eight devices (four hosts, four storage devices) that could potentially generate 800 MB/sec of I/O, must share only 400 MB/sec of bandwidth. Of course, in reality, most devices cannot sustain full throughput and they would not all peak at the same time. This is why many hosts can share a single storage port, and why many devices can share a single ISL. If all eight devices were connected to the same switch, they could communicate with each other at a potential aggregate bandwidth of 800 MB/sec without congestion. When a single switch is not large enough to support the number of required devices, a network of switches is needed.

Figure 16. Zero locality SAN



With a little planning, it is usually possible to design a SAN with a significant degree of locality, as shown in Figure 17. While higher levels of locality are desirable, it is still possible to build very effective SANs with minimal to no locality. In fact, some SANs are deliberately designed with zero locality to maximize the administrative simplicity that a zero locality design provides. It is a straightforward process to design a tiered SAN that delivers sufficient bandwidth in a zero locality environment. The value in doing so is that tiered SANs require no planning or management to add hosts or storage – just attach hosts to host-designated switches and storage to storage-designated switches. This topic is discussed later in this section (see Device Attachment Strategies on page 34).

Figure 17. 100 percent locality SAN



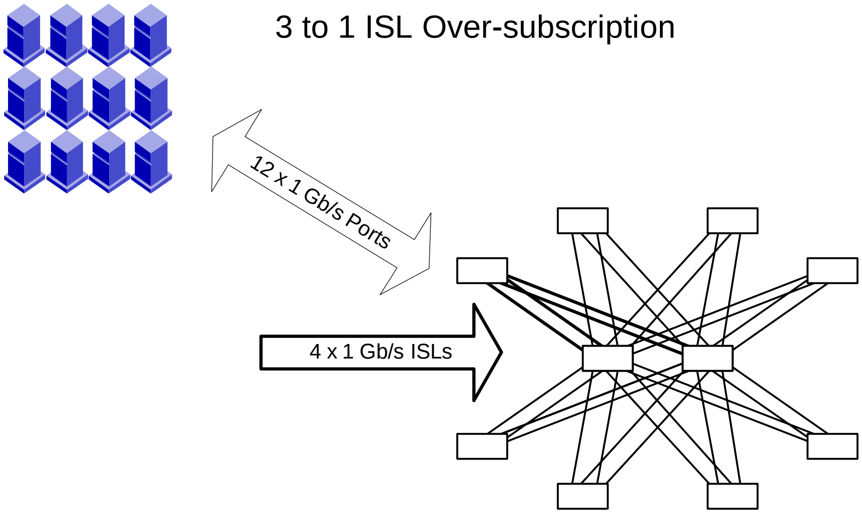
ISL Over-Subscription

In many cases, ISL over-subscription is not a performance-limiting factor in SAN design. Storage port fan-out, low application I/O requirements, and performance limits on edge devices are much more likely to be the areas to focus on for maximum performance improvement. It is usually sufficient to apply a rule of thumb and use the same ISL over-subscription ratio used for storage port fan-out. (This is usually around 7:1.) However, sometimes it is beneficial to understand ISL over-subscription at a detailed level so that it can be analyzed in performance models.

When all ports operate at the same speed, ISL over-subscription is the ratio of node, or data input ports that might drive I/O between switches to the number of ISLs over which the traffic could cross. In Figure 18, the over-subscription ratio on the leftmost switch is three node ports to one ISL. This is usually abbreviated as 3:1. There are twelve hosts connected to the upper left edge switch and only four ISLs to the core. Thus, there are three hosts for each ISL. If all of these hosts tried to simultaneously use the ISLs at full speed—even if the hosts were accessing different storage devices—each would receive only about one-third of the potential bandwidth available.

The simple over-subscription formula is “ISL Over-Subscription = Number of Nodes : Number of ISLs”, or $I_o=N_n:N_i$. This is reduced as a fraction so that $N_i=1$.

Figure 18. ISL over-subscription with 1 Gbit/sec devices

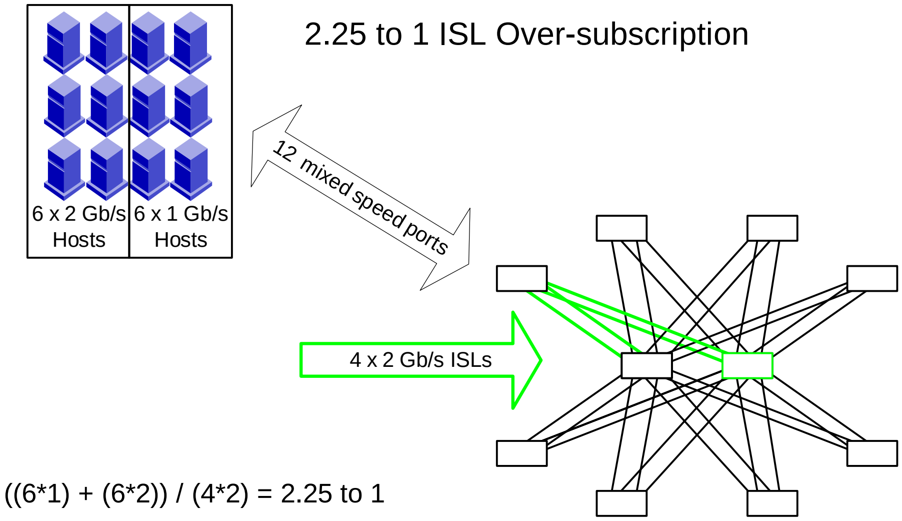


With the advent of 2 Gbit/sec devices today and 10 Gbit/sec speeds to follow, it is necessary to put some additional thought into calculating ISL over-subscription with variable speed hosts, storage, and ISLs. In Figure 19, six 1 Gbit/sec hosts and six 2 Gbit/sec hosts are depicted. These share access to four 2 Gbit/sec ISLs. To calculate the ISL over-subscription ratio, average the speed of the input ports and divide this result by the speed of the output ports. Multiply the node portion of the ratio by that number. For Figure 19, the ISL over-subscription ratio is 2.25:1.

If it is rare to need to consider over-subscription beyond the “rule of thumb” level, it is virtually unheard of to need the following formula. However, in the interest of completeness, the mixed-speed over-subscription formula is “ISL Over-Subscription = ((Average of Node Speeds / ISL Speed) x Number of Nodes) : Number of ISLs”, or $I_o=((A_{ns}/I_s)N_n):N_i$. For Figure 8, the ISL over-subscription ratio is 2.25:1. $A_{ns}=((6*1)+(6*2))/12)=1.5$; $I_s=2$; $N_n=12$ so $I_o=((1.5/2)12):4$, which reduces to 2.25:1.

There are other ways to organize the formula, as shown in Figure 19. In that calculation, N_n is reduced out of the formula.

Figure 19. ISL over-subscription with mixed-speed devices



The worst-case scenario of meaningful over-subscription for an ISL on a 16-port edge switch is 15:1.[6] This ratio means that fifteen devices could be contending for the use of one ISL. That obviously is not a recommended configuration, as it would not allow for any redundancy or improvement if congestion were encountered; however, this is not a unique property of Brocade switches. It is a mathematical property of "networks built with 16-port switches where all ports operate at the same speed."

One could argue that more than fifteen nodes outside a switch could contend for access to it. However, this is not a meaningful definition of *ISL* over-subscription, since the nodes would be subject to performance limitations of *node* over-subscription. If two hosts are trying to access one storage port, it does not matter how well the network is built - the over-subscribed storage port will be the limiting factor. See the definitions for *fan-in* and *fan-out* in Section 1.

Bandwidth Consumption and Congestion

An over-subscribed link is one on which multiple devices *might* contend for bandwidth. A congested link is one on which multiple devices *actually are* contending for bandwidth. Traditional data networks have been built with very high levels of over-subscription on links for years. The Internet is probably the best-known example of this, and has links that are over-subscribed at a rate of millions to one.

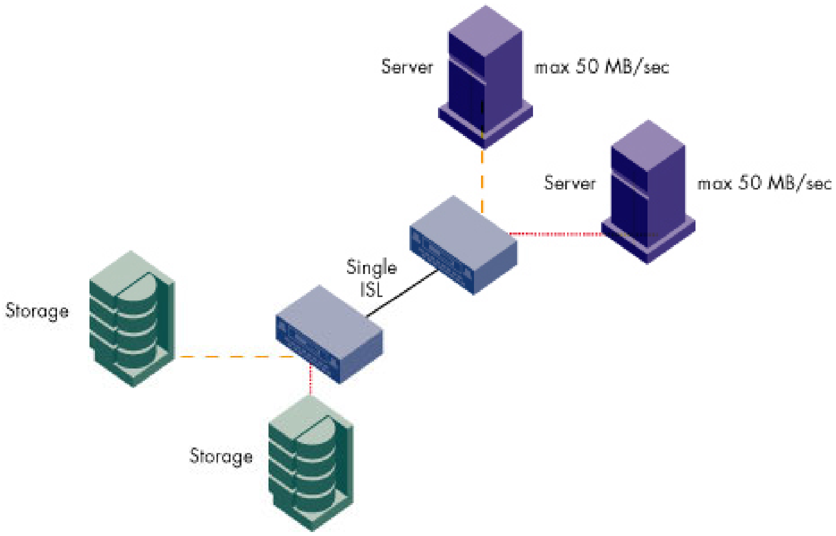
While not capable of supporting Internet-like over-subscription ratios, real-world SANs can be expected to have several characteristics that enable them to function well even with over-subscribed links. These characteristics include bursty traffic, shared resources, low peak usage by devices, good locality, and devices that can generate only a small fraction of the I/O as compared to the available bandwidth. Most networks have all of these characteristics to some degree. Moreover, organizations can often realize substantial cost savings by deliberately designing a SAN with a certain amount of over-subscription.

When performance service levels are critical and the bandwidth requirements are high, lower over-subscription levels or traffic localization should be targeted.

Today, many devices attached to a SAN are not capable of generating traffic at the full Fibre Channel bandwidth of 100 MB/sec or 200 MB/sec. Figure 20, Figure 21, and Figure 22 detail a simplified scenario using a handful of devices to explain SAN bandwidth consumption.

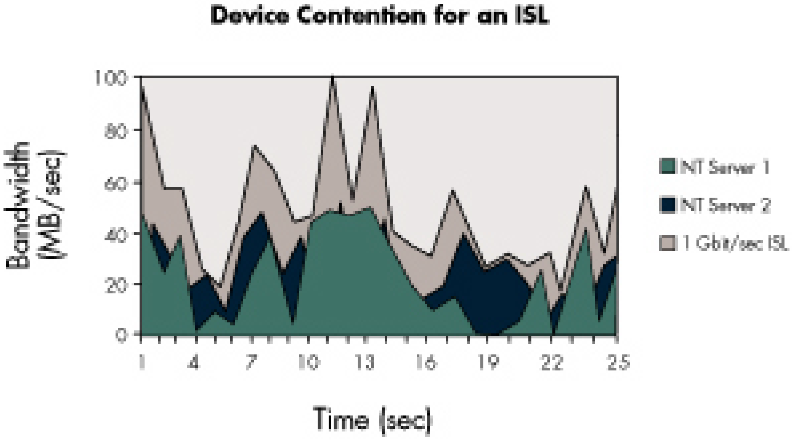
[6] This is extendable: for 32-port switches the theoretical maximum is 31:1; for 64-port switches it is 63:1.

Figure 20. Figure. Low server I/O utilization



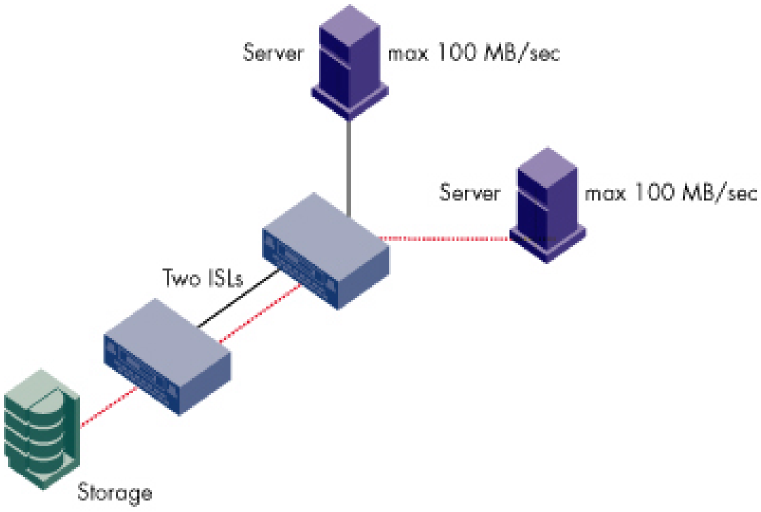
Note that in the ISL graph in Figure 21, the total amount of traffic that is intended to cross between switches never exceeds the 100 MB/sec capacity of the link.

Figure 21. Low bandwidth consumption I/O



Even if the servers in Figure 22 are running at their theoretical maximum performance, there still might be performance bottlenecks with other edge devices. In this example, the two servers are accessing a single storage port, so the 2:1 fan-out of the storage port becomes the limiting factor.

Figure 22. High-bandwidth consumption with two ISLs



The key to managing bandwidth is capturing or estimating performance requirements and matching these requirements to an appropriately designed SAN. If the servers are capable of generating 100 MB/sec of traffic and there are two ISLs, the network routes the traffic over both of them, and congestion does not occur. The SAN can provide 200 MB/sec of bandwidth between the two switches (400 MB/sec using the SilkWorm 12000 or SilkWorm 3800). The storage port can operate at only 50 MB/sec; therefore, each server can average only 25 MB/sec. This scenario is common in storage consolidation environments where many servers need to share a single storage port. However, the I/O requirements for most servers can be surprisingly low (1 or 2 MB/sec) and a single storage port can sustain many hosts without overwhelming its I/O capability.

I/O Profiles

Understanding an application's I/O requirements is essential to the SAN design process.

An individual I/O can be classified as either a read or a write operation. Although I/O is usually a mixture of reads and writes some applications are strongly biased. For example, video *server* I/O activity is normally almost 100 percent reads, while video *editing* cluster I/O may by mostly writes.

I/O can further be classified as random or sequential. Examples of random I/O include an e-mail server or an OLTP server. Sequential I/O is characteristic of decision support (such as data warehousing) or scientific modeling applications.

The third characteristic of I/O is size, which typically ranges from 2 KB to over 1 MB. Typically, user file systems have smaller I/O sizes, whereas video servers or backups may have very large sizes. Table 1 illustrates the application I/O profiles that establish the typical magnitude of application bandwidth consumption.

For SAN design performance purposes, I/O is classified by bandwidth utilization: *light*, *medium*, and *heavy*. It is very important to support test assumptions by gathering actual data when possible. You can gauge the type of I/O activity in your existing environment by using I/O measurement tools such as **iostat** and **sar** (UNIX) or **diskperf** (Microsoft).

Table 1. **Application I/O profiles**

Application	**Bandwidth Utilization**	**Read/Write Max**	**Typical Access**	**Typical I/O Size**
OLTP, e-mail, UFS e-commerce, CIFS	Light	80% read 20% write	Random	8 KB
OLTP (raw)	Light	80% read 20% write	Random	2 KB to 4 KB
Decision support, HPC, seismic, imaging	Medium to Heavy	90% read 10% write (except during “builds”)	Sequential	16 KB to 128 K
Video Server	Heavy	98% read 2% write	Sequential	> 64 KB
SAN applications: LAN-Free backup, snapshots, third-party copy	Medium to Heavy	Variable	Sequential	> 64 KB

Fabric Latency

The time it takes a frame to traverse from its source to its destination is referred to as the latency of the link. Sometimes a frame is switched from source to destination on a single switch and other times a frames must traverse one or more hops between switches before it reaches its destination.

A common misconception is that the hop counts introduce unacceptable latency. For the vast majority of Fibre Channel devices, the latency associated with traversing one or more ISLs is inconsequential. I/O for disk devices is measured in milliseconds. Every hop in the Brocade SAN fabric adds no more than two microseconds of latency (typically 1 microsecond). In a large fabric designed with seven hops between two devices (the Brocade-supported maximum), the latency could be up to fourteen microseconds. The distance between switches also introduces latency, especially for long-distance solutions spread over larger metropolitan areas. The speed of light in optics is approximately five microseconds per kilometer. Brocade addresses the need for longer distance performance with Brocade Extended Fabrics™. This product enables full-bandwidth performance across long distances spanning up to 120 km, with greater distances possible at lower speeds. (This document does not address the performance of SANs where the distances between switches are large enough to add significant latency.)

For most I/O profiles, hop-count latency is inconsequential, from both a switch latency and optical latency standpoint. This is because the millisecond disk I/O latency is several orders of magnitude greater than the microsecond latency of a Fibre Channel fabric. Because it is so small, virtually no applications will be affected by the added latency.

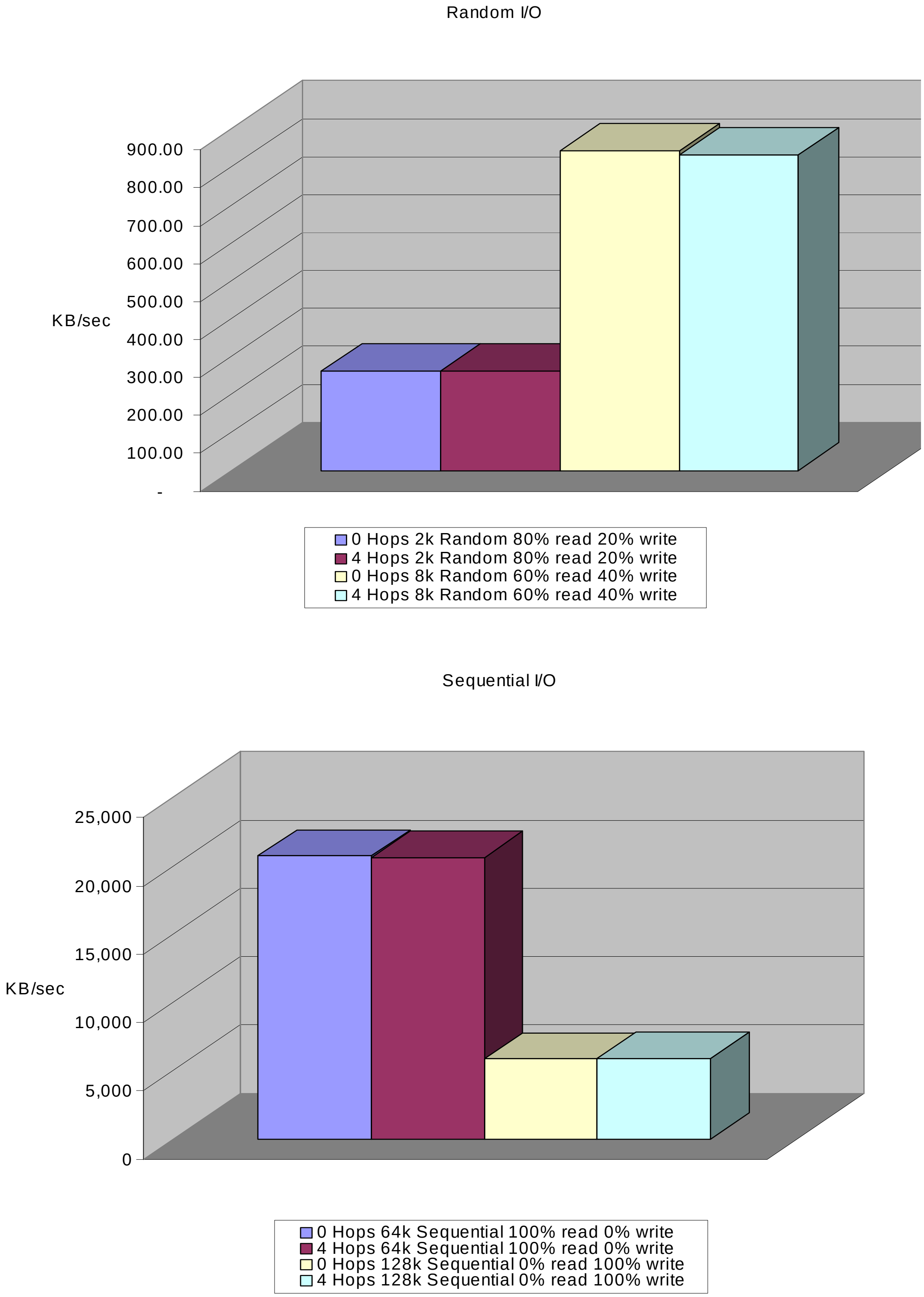
As a result, hop latency is not a reason to keep hop counts low in a SAN design. A more pertinent reason to do so involves over-subscription: the more ISLs a frame has to traverse, the more likely it is to cross a congested ISL. The best hop count for reducing over-subscription is, of course, zero hops (localized traffic). In some cases, however, the second-best-performing scenario is actually two hops, rather than the more intuitive one hop. This is because a two-hop design enables FSPF to perform better load sharing across multiple ISLs. This subject is explained further in this section (see Device Attachment Strategies on page 34).

Table 2 and Figure 23 show collected I/O performance data representing random read/write operations (typical of e-commerce or OLTP applications) and large, sequential reads (typical of decision-support, backup, or video applications). Tests were run on a 1 Gbit/sec host and storage connected to the same switch and across a SAN where the host and storage were separated by four 2 Gbit/sec hops. There was no significant difference in performance between I/Os run locally or across the SAN, so I/O latency was not an issue in either configuration.

Table 2. **I/O latency test results**

HOPS	**I/O Size**	**Read %**	**Write %**	**Access**	**Response Time** (milliseconds)	**Throughput** (KB/sec)
0	2 KB	80%	20%	Random	8	262
4	2 KB	80%	20%	Random	8	261
0	8 KB	60%	40%	Random	9	843
4	8 KB	60%	40%	Random	10	833
0	64 KB	100%	0%	Sequential	3	20,741
4	64 KB	100%	0%	Sequential	3	20,652
0	128 KB	0%	100%	Sequential	22	5890
4	128 KB	0%	100%	Sequential	22	5899

Figure 23. Hop, speed, and optic latency are inconsequential to various I/O patterns



Device Attachment Strategies

While device placement does not *constitute* fabric topology, it may *affect* and be affected by topology. The example in Figure 24 illustrates how a device's placement in a fabric can impact performance and scalability. Where a device attaches can also impact the management of a SAN.

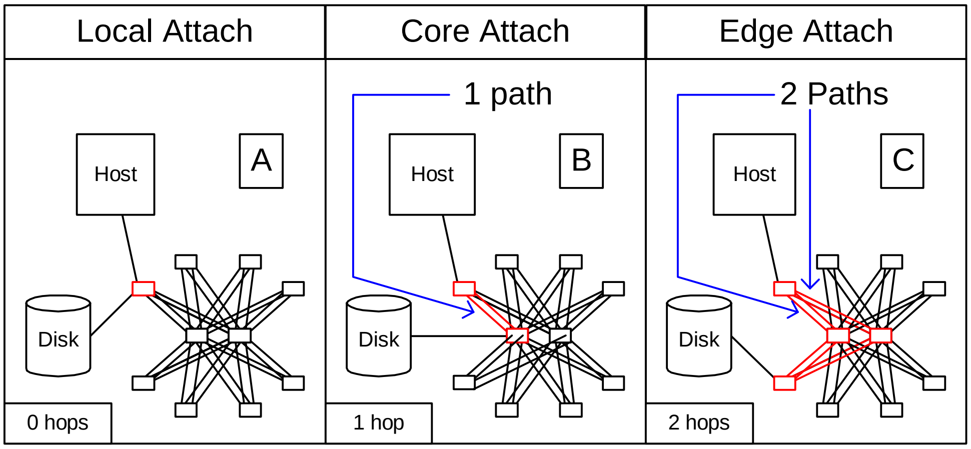
Performance and Scalability Effects

Scenario "A" (Local Attach) in Figure 24 depicts a disk system attached to the same switch as the host that needs to access it. This is a very effective configuration, and is frequently used in high-performance applications because it not only offers zero hop count but also eliminates the need to manage ISL over-subscription. This configuration is useful when most traffic can be localized and congestion is a concern.

Scenario "B" (Core Attach) depicts the case where not all ports on the core are being used by ISLs, and the storage device is directly attached to the core. While this means that only one hop is needed end-to-end, this configuration has two impacts. First, the number of available ports in the SAN is significantly reduced because core ports are no longer available for connecting additional switches. Second, there is only one direct path between the disk switch and the host switch.

Scenario "C" (Edge Attach) is the most typical case. The number of available paths between the host and storage is two. The core switch ports are available for increasing the size of the SAN by adding new edge switches. The impact of connecting nodes to a Core switch is more acute for smaller switches, as there are fewer ports available for scaling. For example, the size of a fabric is reduced by up to 64-ports when connecting a node to a 16-port core switch. This is the case since the same port used to connect a node could be used to expand the fabric by attaching a 64-port switch, such as the SilkWorm 12000. For larger switches, there are more ports for scaling and therefore, the impact of connecting devices to the core is not as significant. However, the impact is still the same: every device connected to a core switch theoretically reduces the size of the fabric by up to 64-ports.

Figure 24. **How Device Placement Can Impact Performance and Scalability In A Core/Edge Fabric**



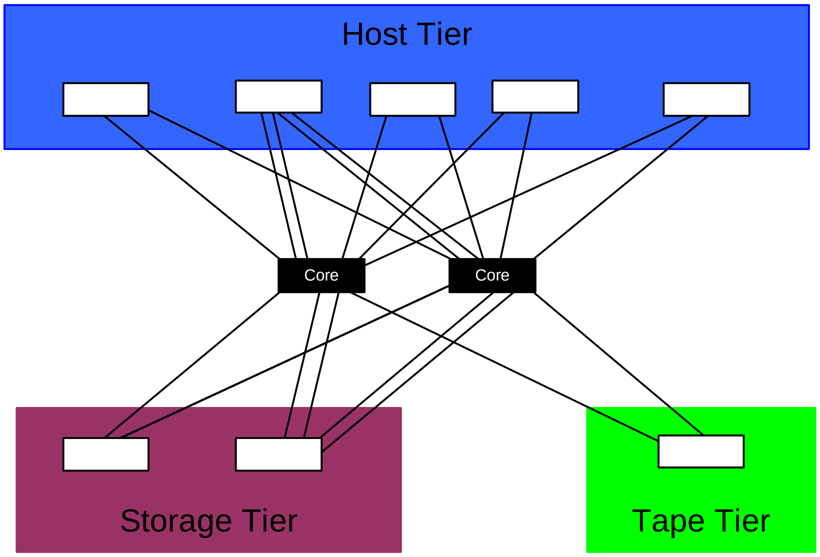
Tiered Fabrics

Tiering is the process of grouping particular devices by function and then attaching these devices to particular switches or groups of switches based on that function. Tiering is the opposite of locality: in a localized SAN, hosts are attached to the same switches as their storage devices; in a tiered SAN, hosts are never attached to the same switches as storage arrays.

It requires some level of effort to plan and manage the layout of a fabric for optimal locality. Sometimes this effort is not necessary if there is a sufficient level of available ISL bandwidth. For example, if it is known that the peak bandwidth that a host generates is 10 MB/sec and there are fourteen hosts on a switch, it is sufficient to only have two ISLs connecting that switch to the remainder of the fabric and tiering is a viable design option. However, if those hosts generate 50 MB/sec concurrently, it is probably more appropriate to adopt a device attachment strategy that involves a high degree of locality, or to use more ISLs.

From a cabling and maintenance perspective, tiering is quite effective. In Figure 25, a group of switches is designated as the storage switch group, another group designated as the tape group, and a final group is designated as the host group. When it becomes necessary to expand backup, storage, or hosts, it becomes a straightforward effort to attach the new devices to an open port on the appropriate tier and to then enable access (i.e. zoning, configure hosts). If a particular tier requires expansion, add a new switch to that group.

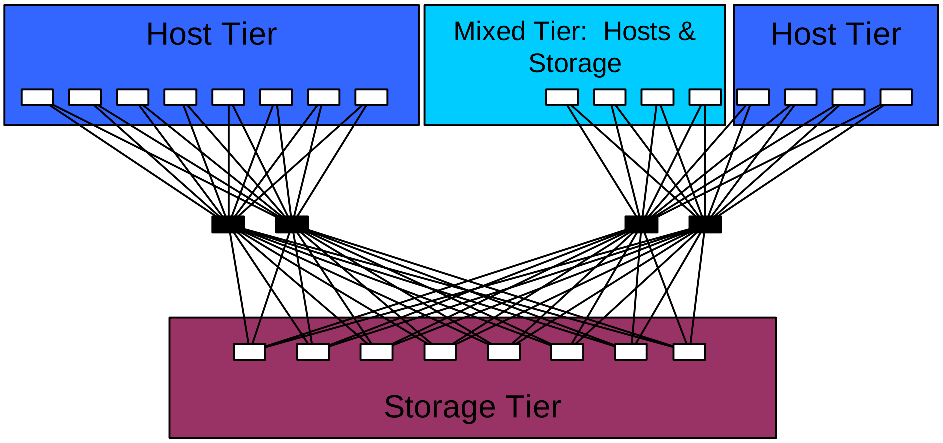
Figure 25. A Tiered SAN



The performance characteristics of a core/edge fabric make this topology an excellent candidate for tiering (see the discussion in Section 3 for the reasons why). Also, note the flexibility to increase bandwidth between devices by adding ISLs to account for varying performance requirements. It is not required to deploy an *entirely* tiered architecture. For performance reasons, it may be desirable to establish a hybrid of tiered switches and some switches that are not tiered. For example, it may be appropriate to connect a high performance host and storage device on the same switch while maintaining a tiered approach for the other devices in the SAN.

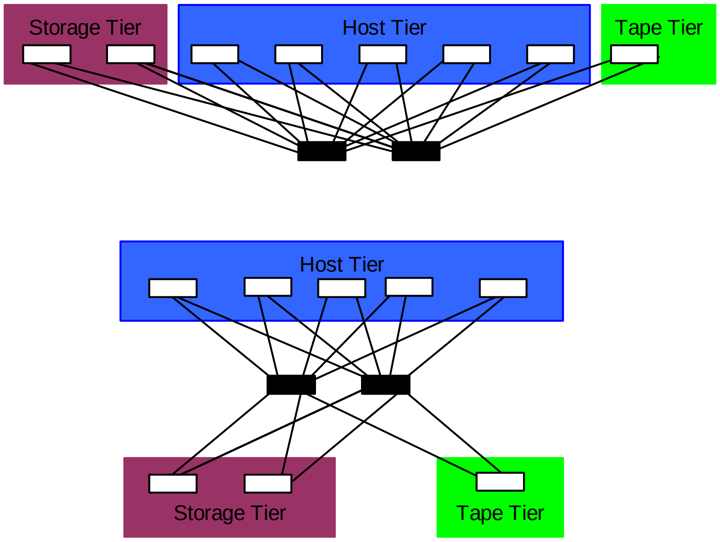
A composite core/edge topology also works well for a tiered SAN architecture, as shown in Figure 26. It is possible to tier by function or "sub-fabrics".

Figure 26. A Tiered Composite Core/Edge Fabric



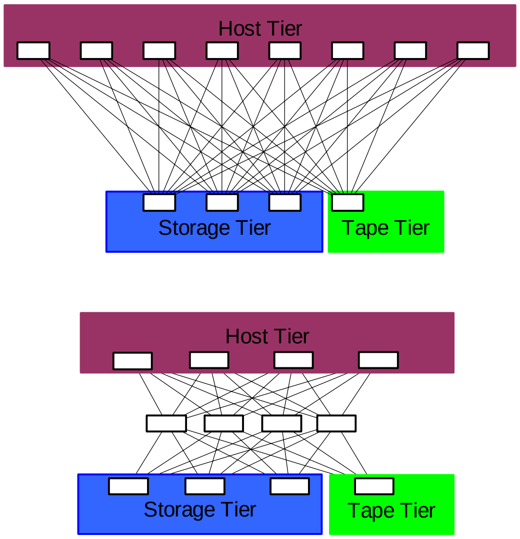
An interesting aspect of a tiered SAN is the visual layout of the switches in the SAN architecture. Note that the two SANs depicted in Figure 27 are identical: each SAN is built with the same number of switches, number of ISLs, ISL connection points, and device attachment points. The only difference is how the switches are laid out in the figure.

Figure 27. Two different graphical representations of the same SAN



In Figure 28, the SANs have the same number of switches, number of ISLs, and ISL connection points; however, the device connection points are different, as the core switches are utilized for device attachment. These two SANs are topologically identical, but functionally different. The scalability and performance caveats apply, as discussed earlier in this section, when attaching devices to the core scalability is diminished. The top SAN in Figure 28 is sometimes called a two-tier SAN and the bottom SAN is sometimes called a three-tier SAN. The device attachment points, not the layout of the switches, differentiate a two-tier SAN from a three-tier SAN.

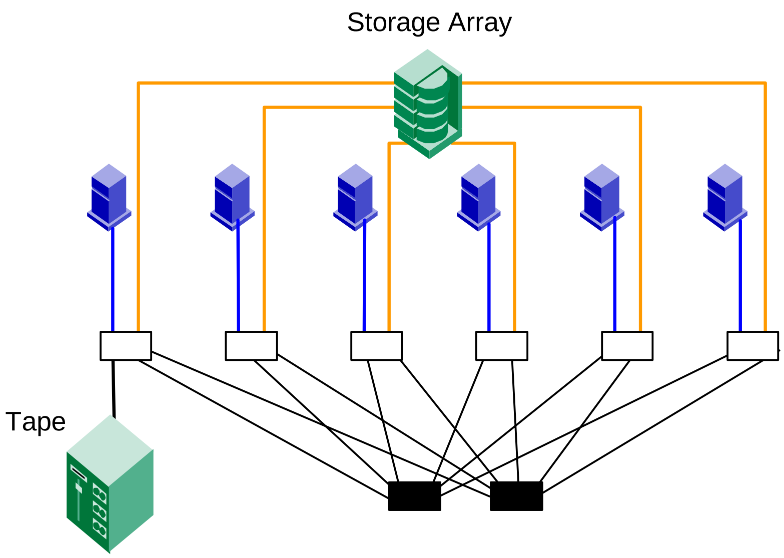
Figure 28. Each SAN is similar in design, but functionally different due to device attachment points



High Locality Device Attachment

For high performance devices, it is desirable to attach devices based on the principle of locality: those devices that communicate with each other most frequently should be placed close together. As mentioned, an architecture that employs locality is the opposite of a tiered SAN architecture. While opposite, these approaches are not mutually exclusive. Frequently, a large SAN architecture will incorporate aspects of both locality and the use of tiers. This approach is depicted in Figure 29. Note that the hosts and storage ports are localized, but the tape library is shared by all hosts, and is not localized.

Figure 29. Achieving High Performance With Device Placement



Section 3: Fabric Topologies

There are multitudes of options for how to go about building a fabric infrastructure. A number of factors influence a choice of fabric topology, such as scalability, performance, and availability. This section discusses some of these factors, and how they can affect a design choice. It also describes some of the most effective fabric topologies, and gives guidance on when to use each. To simplify the SAN design process, consider the use of any of the recommended reference topologies that are detailed in Table 8. These topologies are effective for a wide variety of applications, are tested extensively within Brocade, are in wide deployment in customer environments, are high performance, and scalable.

Brocade's flexible fabric architecture allows arbitrarily complex fabrics to be built when it is necessary to solve complex problems, but also allows simple solutions to be built to solve simple problems. The philosophy is this: "Make easy things *easy*, and hard things *possible*."

Simple Topologies

A simple topology is one that is geometrically simple. A ring of six switches is easy to recognize as a ring. There is no question that it is a ring, and its performance and reliability characteristics are easy to predict. A ring of six switches, each of which has some variable number of switches attached to it is not as easy to define, and would be considered a hybrid of a ring and some number of other topologies. Hybrid topologies are discussed later.

Cascade

A cascaded fabric, which is shown in Figure 30, is like a bus topology: it is a line of switches with one connection between each switch and the switch next to it. The switches on the ends are not connected.

Cascaded fabrics are very inexpensive, easy to deploy, and easy to expand. However, they have the lowest reliability and limited scalability. They are most appropriate in situations where most if not all traffic can be localized onto individual switches, and the ISLs are used primarily for management traffic or low bandwidth SAN applications. See the locality discussion in Section 2. for more information on the principal of locality.

There are cascade variations that use more than one ISL between switches. This will eliminate ISLs as a single point of failure, and greatly increase the reliability of the solution. However, this also increases the cost of the solution, and each switch can still be a single point of failure. Table 3 charts the properties of a cascade topology.

Figure 30. A Cascaded Fabric

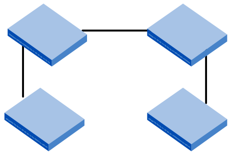


Table 3. Properties of the Cascade Topology

Ratings indicate how well the topology meets the ideal requirements of that property (1 = Not well, 2 = Moderately well, 3 = Very well).	
Properties	**Ratings**
Limit of scalability (edge port count)	114 ports / 8 switches[7]
Ease of scalability	2
Performance	1
Ease of deployment	3
Reliability	1
Cost (per edge port)	3

Ring

A Ring (see Figure 31) is like a cascaded fabric, but with the ends connected. The ring has superior reliability to the cascade because traffic can route around an ISL failure *or* a switch failure. It does cost more than a cascade but only slightly so. The ring is usually preferable to the cascade for that reason. Like the cascade, the ring is most suitable when locality is used to optimize traffic patterns in the fabric. This design is effective for configurations that start small and stay small. It can also be used when implementing SAN over MAN or WAN, where the topology of the MAN/WAN might dictate the topology of the Fibre Channel network -- Rings are common MAN/WAN topologies. Finally, a Ring topology is a good choice when the ISLs are mostly used for management or low bandwidth SAN applications. Table 4 charts the properties of a ring topology.

Figure 31. A Ring Topology

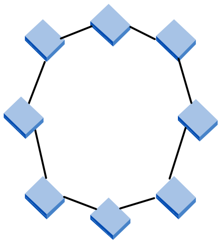


Table 4. Properties of a Ring Topology

Ratings indicate how well the topology meets the ideal requirements of that property (1 = Not well, 2 = Moderately well, 3 = Very well).	
Properties	**Ratings**
Limit of scalability (edge port count)	112 ports / 8 switches[8]
Ease of scalability	2
Performance	1
Ease of deployment	3
Reliability	3
Cost (per edge port)	3

[7] This is based on 16-port switches. It is currently possible to cascade up to four 64-port switches, to achieve 250 available ports, with very high over-subscription ratios. Support for such a configuration is dependent on your support provider.

[8] This is based on 16-port switches. Higher port counts are possible using 64-port switches.

Full Mesh

In a full-mesh topology (see Figure 32 and Figure 33), every switch is connected directly to every other switch. Using 16-port switches, the largest useful full mesh consists of eight switches, each of which has nine available ports. This provides 72 available ports. Adding more than eight switches will actually reduce the number of available ports. Full meshes are best used when the fabric is not expected to grow beyond four or five switches, since the cost of the ISLs becomes prohibitive after that. They can also form effective backbones to which other SAN islands are connected. These networks are best used when any-to-any connectivity is needed. In addition, traffic patterns should be evenly distributed, but overall bandwidth consumption low. Otherwise, a core/edge SAN is a better fit. The full mesh is also a good fit for building elements of hybrid networks, which are discussed later in this section. It is particularly well suited for use in complex core/edge networks due to its low radius. Technically, almost any topology could be described as some sort of mesh. Since this is not a very useful definition, working definitions for two meshes are provided: the full mesh and the partial mesh. There are two special cases for a full mesh:

- A 2-switch full mesh is identical to a 2-switch cascade.
- A 3-switch full mesh is identical to a 3-switch ring.

Scaling a full mesh can require unplugging edge devices. If using a 4-switch full mesh (52 edge ports) and all the ports with edge devices are in use, it will be necessary to unplug one device from each switch in the mesh in order to add another switch. Because of this, full meshes do not have a high rating for ease of scalability. Table 5 charts the properties of a full-mesh topology.

Figure 32. A Full-Mesh Topology

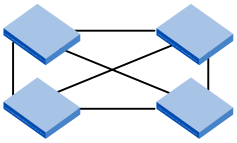


Figure 33. Maximum Size of a Full-Mesh Topology with 16-port switches

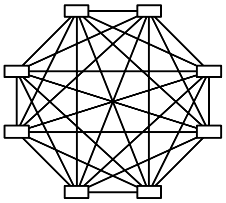


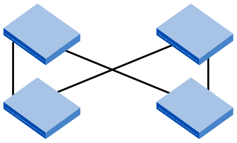
Table 5. Properties of a Full-Mesh Topology

Ratings indicate how well the topology meets the ideal requirements of that property (1 = Not well, 2 = Moderately well, 3 = Very well).	
Properties	**Ratings**
Limit of scalability (edge port count)	72 ports / 8 switches[9]
Ease of scalability	1
Performance	2
Ease of deployment	3
Reliability	3
Cost (per edge port)	1

Partial Mesh

A partial mesh is similar to a full mesh, but with some of the ISLs removed. In most cases, this is done in a structured pattern. Partial meshes are useful when designing a SAN backbone, in which traffic patterns between islands connected to the backbone are well known. For example, this is a viable MAN/WAN technique. The common definition for a partial mesh (see Figure 34) is broad enough to encompass almost all fabrics that are not full meshes. In most cases, this will be done in a structured pattern. For example, each switch will directly connect to its neighbor and to every *other* switch across from it. While this definition is not in general use outside of Brocade, it describes a desirable variant on the full mesh. A core/edge topology is considered a partial mesh topology.

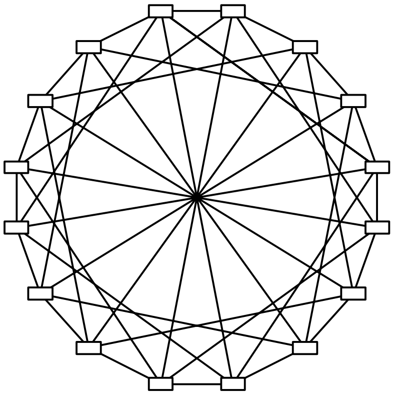
Figure 34. A Partial-Mesh Topology



The network in Figure 34 might be useful if minimal traffic is expected flow horizontally (i.e. from left to right) and that the majority of traffic will flow vertically (i.e. top to bottom). For example, hosts would be connected to the top switches and storage connected to the bottom switches. The network still is fully resilient to failure, and there is no price premium for an ISL that will not be used. Partial meshes also scale farther than full meshes. Figure 35 shows a partial mesh that has 176 free ports. Remember that the largest full mesh has 72 ports. Each switch is connected to its neighbor. Two switches are skipped before the next connection. The worst-case hop count between switches in the event of an ISL failure is three hops.

[9] This is based on 16-port switches. It is possible to full-mesh three or more 64-port switches. Support of such configurations are dependent upon you support provider.

Figure 35. Maximum Size of a Partial-Mesh Topology



While these networks can be scaled to produce a large number of edge ports, they still have less than ideal performance characteristics. None of the networks listed thus far will benefit much from FSPF load sharing capabilities, for example. Since bandwidth is more frequently a concern than is hop count, the ability of a topology to load share across ISLs is key to its performance.

In addition, partial meshes can be difficult to scale without downtime. The procedure for moving from the full-mesh fabric in Figure 33 to the partial-mesh fabric in Figure 35 would require not only adding new switches and potentially disconnecting nodes, but also actually *disconnecting ISLs* that were already in place. The same is true for scaling between many partial-mesh designs. This is disruptive to many production SANs, especially if redundant fabrics are not used. As a result, meshes – either full or partial – are recommended only for networks that will change infrequently and where the traffic patterns of the connected devices are known. They might also be used as a static component of a network. For example, a full mesh could be used in an environment where the “SAN islands” architecture was employed, or as the core of a complex core/edge design, which is discussed later in this section. Table 6 charts the properties of a partial-mesh topology.

Table 6. Properties of a Partial-Mesh Topology

Ratings indicate how well the topology meets the ideal requirements of that property (1 = Not well, 2 = Moderately well, 3 = Very well).	
Properties	**Ratings**
Limit of scalability (edge port count)	176+ ports / 16+ switches[10]
Ease of scalability	1
Performance	1
Ease of deployment	2
Reliability	3
Cost (per edge port)	2 to 3

[10] This is based on 16-port switches. It is currently possible to build a partial mesh of three or more 64-port switches, depending upon your support provider’s support guidelines..

Core/Edge

With a core/edge topology, it is easier to satisfy SAN functional requirements. Given a diverse set of requirements: performance, locality, data integrity issues, connectivity, scalability, and security, the core/edge topology provides the most flexible architecture to address these overall requirements.

The core/edge fabric is a variation on the well-established "star" topology popular in Ethernet LANs. There are a few differences, however. Because Fibre Channel uses a routing protocol (i.e. FSPF) with load sharing, Fibre Channel fabrics can take full advantage of multiple core switches. In an Ethernet network, multiple switches at the center of a star would usually act in an active/passive backup relationship, using a Spanning Tree Protocol or a variation on the Cisco proprietary Hot Standby Router Protocol.

These differences make multi-core fabrics very popular, since it is possible to easily scale the fabric's bandwidth by adding core elements. In addition, the requirements of a core fabric switch are more stringent than those of the center switch in an Ethernet star. Because of Fibre Channel's channel-like properties, the acceptable performance and reliability characteristics are very high.

The introduction of trunking (see ISL Trunking^{TM} on page 50) further increases the effectiveness of a core/edge fabric due to more efficient utilization of the ISLs and lessened management requirements. In a resilient core/edge fabric, which is shown in Figure 36, two or more switches reside in the center of the fabric (the core) and interconnect a number of other switches (the edge). Switches that reside in the middle of the fabric are referred to as core switches. The switches that are interconnected by the core switches are referred to as edge switches. The simple form of the core/edge fabric has two or more core elements, each of which consists of a single switch. In a simple core, the core switches do not connect with each other. Edge switches in a core/edge fabric also do not connect to each other. They only connect to the core switch. Several variants of the core/edge network do not meet this definition. These are discussed later in this section (see Hybrid Topologies later in this section). Devices such as hosts and storage are attached to free ports on the edge switches. These ports are referred to as edge ports. Free ports on the core switches should usually be reserved for additional edge switches. The scalability of a core/edge fabric is reduced when a device is attached to a core switch. (see Device Attachment Strategies in Section 2).

A key benefit of the core/edge topology is the use of FSPF, which automatically distributes the load across all paths equally. In fact, all edge-to-edge paths are equal in a true core/edge topology. There are two or more paths between any two edge switches in a resilient core/edge topology. Because of this, core/edge fabrics have very good performance under varying to zero locality conditions. This concept is depicted in Figure 37.

Figure 36. Equal load distribution and access in a Core/Edge topology

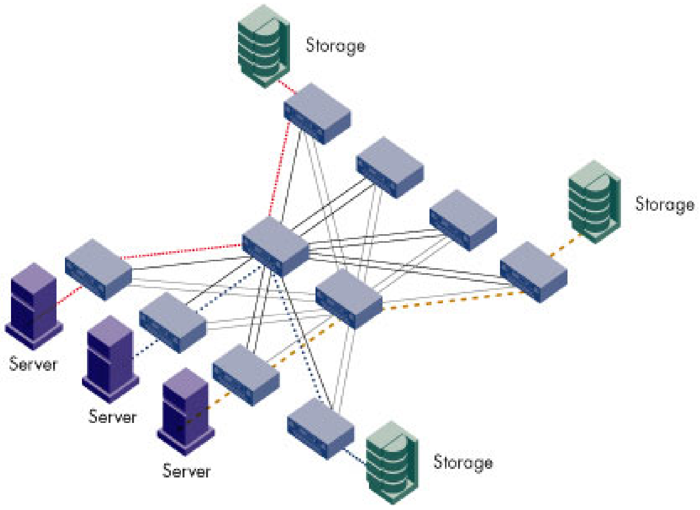
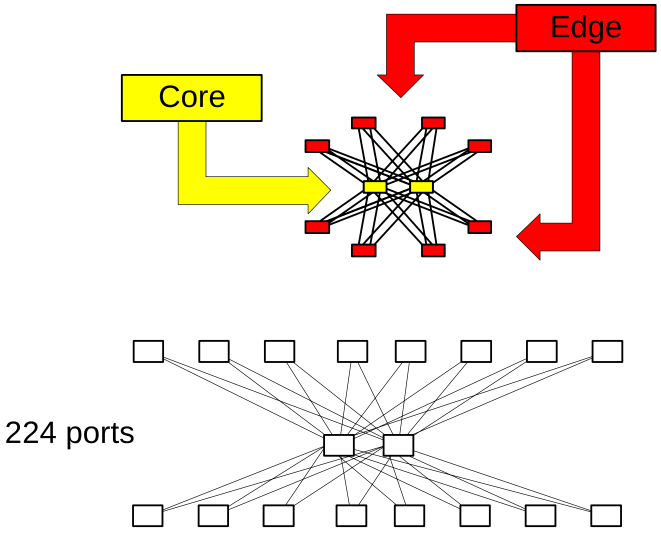


Figure 37. A Core/Edge Topology



The core/edge topology is preferred for scalable, available, and high performance fabrics for a number of reasons. The core/edge topology is:

- Easy to grow without downtime or disconnection of links and devices
- Pay as you grow
- Flexible
- Easy to transition to future large core fabric switches
- Investment protection as the smaller core switches are redeployed to the edge
- Simple and easy to understand
- Well-tested and reliable
- Widely deployed in production environments

- Capable of exhibiting stellar performance, with full utilization of FSPF load sharing and redundancy features
- Conducive to performance analysis. Because the core/edge topology is symmetrical, it is a straightforward process to identify performance issues. Every device has an equivalent path to any other device and the same available bandwidth between any two devices. To identify a performance issue it is only necessary to monitor the core switches. With other topologies, this is not the case.
- Currently scalable to hundreds of ports (using 16-port switches) with the potential to scale to thousands of ports (using 64 or higher port switches)
- Able to solve most design problems, fits wells with many SAN solutions, and is an effective choice when design requirements are not well known

Table 7. Properties of a Simple Core/Edge Topology When Using

Ratings indicate how well the topology meets the ideal requirements of that property (1 = Not well, 2 = Moderately well, 3 = Very well).	
Properties	**Ratings**
Limit of scalability (edge port count)	224 ports / 20 switches[11]
Ease of scalability	3
Performance	3
Ease of deployment	3
Reliability	3
Cost (per edge port)	2

Hybrid Topologies

In some cases, it may be desirable to use a topology that is a composite of simple topologies, which is known as a hybrid topology. For example, if a SAN is spread across a MAN or WAN link, it can be an effective strategy to “glue” two core/edge SANs together using a complex core approach. There are several reasons to build a hybrid SAN:

- Geographically distributed nodes
- Cable plant restrictions
- Scalability requirements

It is unrealistic to expect all SAN users to deploy a core/edge topology or for that matter even a simple topology. Hybrid topologies are more complex and can be more expensive to build and maintain – especially for high port count fabrics. However, in certain cases it is necessary to expand beyond several hundred ports using 16-port switches or (theoretically) several thousand ports using 64-port switches. It is also possible that geographic/cable restrictions need to be designed into the solution. In these cases, hybrid topologies are an option. The elegance of Brocade SilkWorm switches and Fabric OS is the freedom to design and implement any number of topologies to suit the specified requirements. The key to designing hybrid topologies is not to exceed the recommended number of switches and hop count per fabric. Two core/edge based hybrid topologies that have a moderate level of utility and are reasonably simple follow. It is certainly possible to build other types of hybrid topologies.

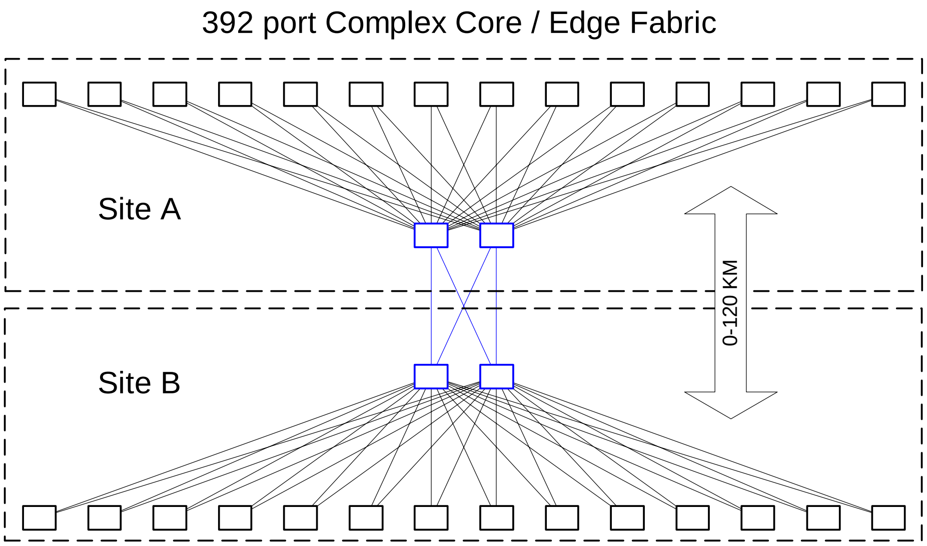
[11] This is based on 16-port switches. With 64-port switches, supported topologies with a greater number of ports are possible.

It is possible to build a fabric consisting of thousands of ports using 64-port switches in a simple core/edge topology. Even with higher port count switches, the need to design fabrics that take into account massive fabric sizes, cable plant restrictions or geographically distributed fabrics will not go away. The same hybrid topologies that work for 16-port switches can also be deployed using higher port count switches to create mega-SANs. What seems large today will appear small tomorrow.

Complex Core

A resilient complex core/edge fabric has two or more core elements, each of which consists of multiple interconnected switches. In contrast, a simple core consists of two or more core elements, each of which consists of a single switch. The core elements can be constructed out of any of the simple topologies discussed previously. However, it is generally best to construct them out of high performance topologies, like a full mesh or simple core/edge if there is to be significant edge-to-edge traffic. It is also advisable to keep the core elements "low radius," in order to insure that the overall fabric hop count never exceeds the 7-hops, which is the Brocade supported upper limit. In practical terms, this means that you should keep the hop count within each core as low as possible. Figure 38 is an example of a 392-port (using 16-port switches) resilient complex core/edge fabric and is an effective design for spanning a fabric between two separate sites. The complex core is actually a partial mesh.

Figure 38. Resilient Complex Core/Edge Fabric

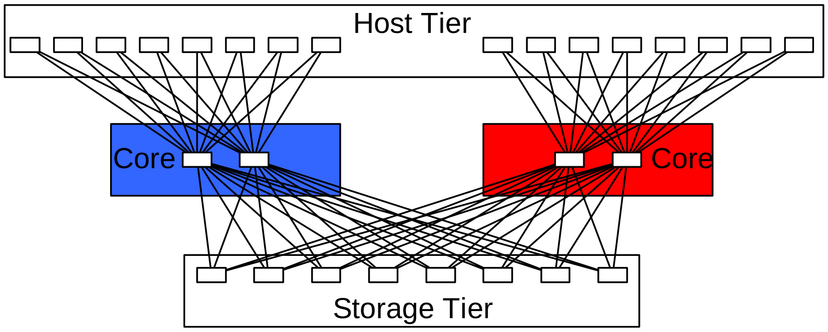


Composite Core/Edge

A composite resilient core/edge SAN has two or more cores, as shown in Figure 25. Each core consists of two or more single-switch elements. It could be viewed as two core/edge fabrics "glued together at the edge." This type of topology is useful to achieve a higher port count than available from a simple topology and is a topology that works well with a tiered approach to performance optimization. Generally, any host needs access to any storage and there are usually more hosts than storage devices in a fabric. In the configuration shown in Figure 39, all hosts have equal access (2-hops) to any storage device; however it takes 4-hops should it be necessary for a host on the left tier to communicate with a host on the right tier, which is acceptable since hosts normally do not communicate with each other.

Figure 39. A Composite Resilient Core/Edge Topology

320-Port Composite Core/Edge Fabric



- Two cores (Blue, Red)
- Each core composed of two switches
- Edge switches asymetrically connected

Summary

Table 8 summarizes various topologies that can be built using Brocade SilkWorm 16-port switches. Table 9 lists the currently supported port counts for networks built with SilkWorm 12000 switches. (Properties other than limit of scalability are the same for both 16-port and 64-port switches.)

The information provided is intended to help identify the effective topologies available and the advantages/disadvantages of these topologies. Many more topologies can be built than are discussed in this section. This is one benefit of using Brocade SilkWorm switches. The choice to build any particular topology is up to the designer. The core/edge topology is a general-purpose topology that delivers scalability, performance, and availability. To simplify the SAN design process, consider the use of any of the recommended reference topologies that are detailed in Table 8.

Table 8. Topology Attribute Summary

Ratings indicate how well the topology meets the ideal requirements of that property (1 = Not well, 2 = Moderately well, 3 = Very well).		
Properties	**Topology**	**Ratings**
Limit of scalability (edge port count)		
	Cascade	114 ports / 8 switches
	Ring	112 ports / 8 switches
	Full Mesh	72 ports / 8 switches
	Partial Mesh	176+ ports / 16+ switches
	Core/Edge	224 ports / 20 switches
	Hybrid	300+ ports / 32+ switches
Ease of scalability		
	Cascade	3
	Ring	2
	Full Mesh	1
	Partial Mesh	1

	Core/Edge	3
	Hybrid	2
Performance		
	Cascade	1
	Ring	1
	Full Mesh	2
	Partial Mesh	1
	Core/Edge	3
	Hybrid	2
Ease of deployment		
	Cascade	3
	Ring	3
	Full Mesh	3
	Partial Mesh	2
	Core/Edge	3
	Hybrid	1
Reliability		
	Cascade	1
	Ring	3
	Full Mesh	3
	Partial Mesh	3
	Core/Edge	3
	Hybrid	3
Cost (per edge port)		
	Cascade	3
	Ring	3
	Full Mesh	1
	Partial Mesh	2 to 3
	Core/Edge	2
	Hybrid	1

Because the SilkWorm 12000 has four times as many ports per switch as Brocade’s other products, networks built using that platform have fundamentally different scalability characteristics. The second column in the following table reflects the current tested and production supportable configurations for the SilkWorm 12000. Much larger configurations are being tested as of this writing, and these will be available through the SE channels as they become finalized. The third column reflects the *theoretical* maximum scalability of the topology.

Table 9. SilkWorm 12000 Scalability Summary

Topology	**Scalability (current)**	**Scalability (theoretical)**
Cascade	374	498
Ring	372	496
Full Mesh	354	1056
Partial Mesh	350+	Thousands
Core/Edge	**512**	**Thousands**

Section 4: Switch Platform Considerations

The range of Brocade SilkWorm switches enable the designer to build a variety of topologies. The platforms utilized in a SAN topology do have bearing on the overall SAN architecture. This section discusses how particular platform features, such as ISL Trunking or port speed, can impact a SAN design.

Port Count

When designing a SAN, it is possible to use a mixture of 8-, 16-, and 64-port switches. When designing a core/edge topology, using the higher port count switches in the core enables greater scalability.

ISL Trunking^{TM}

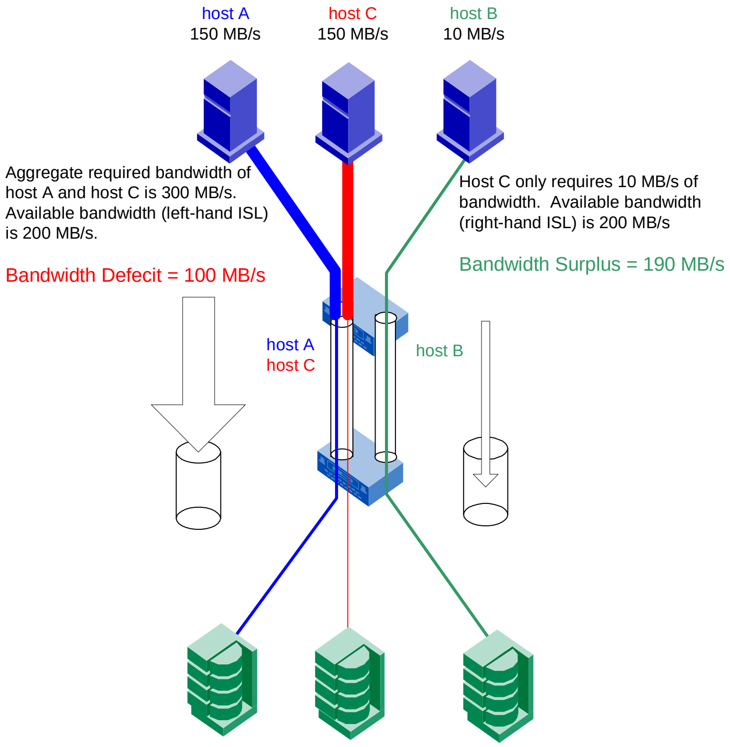
Trunking is a feature that enables traffic to be evenly distributed across available inter-switch links (ISLs) while preserving in-order delivery. A trunk logically joins two, three, or four ISLs into one logical ISL – up to 8 MB/sec. Trunking capable core switches out-perform other cores due in large part to trunking. Use of trunking can minimize or eliminate congestion in the SAN because trunking optimizes ISL utilization. Additionally, the use of trunking minimizes the effort of managing a SAN since ISLs are now managed as a group instead of individually. Trunking also increases availability, since no interruption of I/O occurs if a non-Master ISL fails. As long as at least one ISL link remains, I/O continues if an ISL failure occurs -- although at a lower bandwidth.

The SilkWorm 3200, 3800, and 12000 support trunking. For trunking to work, it is necessary to interconnect two trunking capable switches, for example by connecting a SilkWorm 3800 to a 12000. It is not possible to trunk by for example connecting a SilkWorm 3000 series of switch to a SilkWorm 2000 series of switch.

What Is Trunking?

As a practical matter, due to limitations of some Fibre Channel devices, frame traffic between a source device and a destination device must be delivered in-order within an exchange. This restriction forces the current Brocade frame routing method to fix a routing path within a fabric. Therefore, certain traffic patterns may become unevenly distributed, leaving some paths congested and other paths underutilized. As all other switch vendors use this same methodology, they all suffer this same limitation. This situation is depicted in Figure 40. Hosts A and C require an aggregate of 300 MB/sec of bandwidth while host B requires 10 MB/sec of bandwidth. The nature of Fabric Shortest Path First (FSPF) is to route traffic in a round robin fashion while preserving in-order delivery. In Figure 40, hosts A & C, which require 150 MB/sec each of bandwidth, will share the left-hand ISL. While host C will utilize the right-hand ISL. Note that the aggregate bandwidth required is 310 MB/sec. Hosts A & C will experience congestion, as only 200 MB/sec is available and the bandwidth requirement is 300 MB/sec. Host B will experience no congestion, as there is sufficient bandwidth to support the 10 MB/sec requirement. The two ISLs provide an aggregate bandwidth of 400 MB/sec. If the routing initialization happened in a different order, there would not be any congestion. While the scenario depicted in Figure 40 is a trivial scenario, the probability of encountering such a scenario increases with the number of nodes present in the fabric and the size of the fabric itself.

Figure 40 Uneven distribution of traffic without trunking



Aggregate required bandwidth of host A and host C is 300 MB/s. Available bandwidth (left-hand ISL) is 200 MB/s.

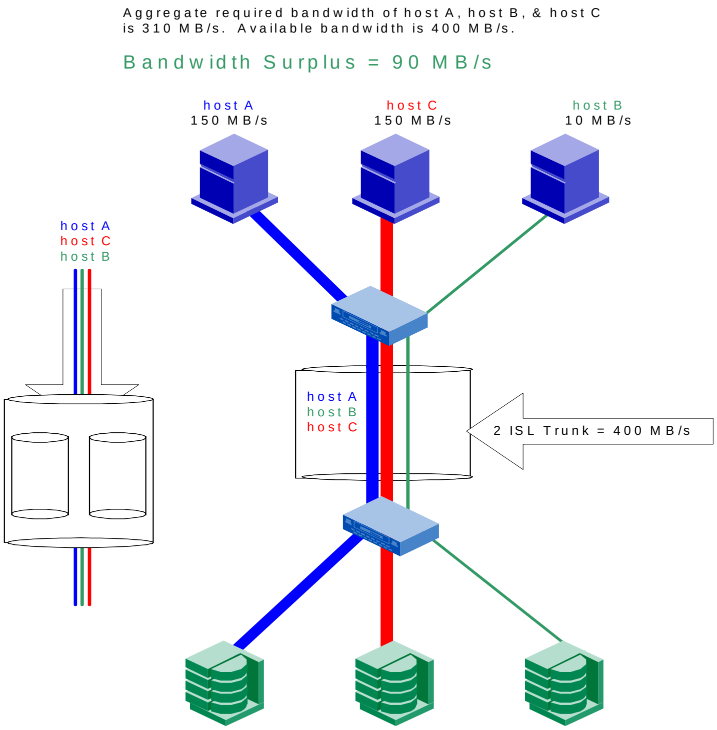
Bandwidth Defecit = 100 MB/s

Host C only requires 10 MB/s of bandwidth. Available bandwidth (right-hand ISL) is 200 MB/s

Bandwidth Surplus = 190 MB/s

Figure 41 depicts a similar scenario as shown in the Figure 40 configuration. Notice that all variables are the same, except that trunking is utilized. The trunk is a logical aggregation of the two 200 MB/sec ISLs into one logical 400 MB/sec trunk. Note that trunking allows the grouping of up to four 2 Gbit/sec ISLs into one logical 8 Gbit/sec trunk. The trunk provides sufficient bandwidth for all traffic streams.

Figure 41 Performance and management benefits of trunking


Aggregate required bandwidth of host A, host B, & host C is 310 MB/s. Available bandwidth is 400 MB/s.
Bandwidth Surplus = 90 MB/s
host A
150 MB/s
host C
150 MB/s
host B
10 MB/s
host A
host C
host B
host A
host B
host C
2 ISL Trunk = 400 MB/s

How Does Trunking Impact SAN Design?

Trunking optimizes the utilization of ISLs and reduces the SAN administration effort. These benefits enhance the utility of ISLs and so enhance the utility of designs that make use of ISLs. Given two SANs of equivalent functionality, the SAN that is easier to manage has the advantage. Instead of monitoring multiple ISLs, a single trunk is now monitored. The high performance capabilities of a trunking capable switch make these types of switches ideal for placement in the core of a core/edge fabric. Because the trunk is efficient, it is less likely that congestion will be experienced and it is possible that fewer ISL are required in the fabric topology, yielding additional ports for attaching SAN devices.

When designing a SAN with trunking capable switches or introducing trunking capable switches to an existing SAN, it is important to place these switches adjacent to each other when working with Mesh, Ring, or Cascade topologies.

When working with a core/edge topology, it is recommended to place the SilkWorm 3000 series switches in the core. As new trunking capable switches are added to a core/edge topology, it will allow the added switches to connect with trunking capable switches. Migrating existing SilkWorm 2000 series switches from the core to the edge and inserting trunking capable switches into the core is an effective strategy.

The ISLs that comprise a trunk must all reside in the same contiguous four-port groups, which are termed quads. For example, on the SilkWorm 3800, the following ports are grouped into quads: 0-3, 4-7, 8-11, and 12-15. Therefore, when connecting two switches in a fabric with one, two or three ISLs, consider leaving open the other ports on the quad for future trunk growth.

How 2 Gbit/sec Switches Impact a SAN Design

A switch that supports auto-sensing of both 1 Gbit/sec and 2 Gbit/sec device connections, such as the SilkWorm 3800 or SilkWorm 12000, introduces many benefits and choices to the SAN designer. As SAN devices evolve from 1 Gbit/sec to 2 Gbit/sec capable, the role of a 2 Gbit/sec switch becomes very important for connecting SAN devices. Designing in such a capability “future-proofs” a SAN and extends the life span of the initial design. As an interconnect between switches, 2 Gbit/sec ISLs deliver high performance. Devices that are not 2 Gbit/sec capable can still benefit from a switch’s 2 Gbit/sec capabilities, as it is possible to combine multiple 1 Gbit/sec connections over a 2 Gbit/sec ISL or trunk. Many 1 Gbit/sec devices today barely utilize the full bandwidth of a 1 Gbit/sec connection. This should not be a surprise and is why it is possible to design a SAN with over-subscription. The advent of 2 Gbit/sec ISLs essentially doubles the performance of a similarly designed SAN built with 1 Gbit/sec ISLs and nodes. This means that twice the performance is available, if required, or it is possible to scale back the number of ISLs to yield additional ports for device connections. Trunking amplifies this performance benefit, as the ISLs are now faster **and** used more efficiently.

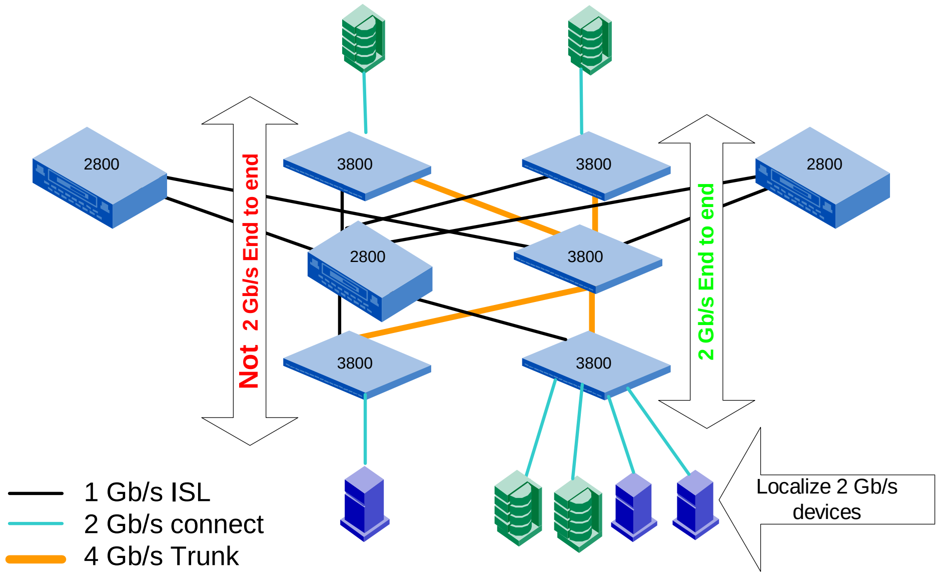
2 Gbit/sec Switch Placement

When designing a SAN with 2 Gbit/sec switches, the same guidelines that apply to trunking apply to 2 Gbit/sec capabilities. Place these switches adjacent to each other to take advantage of 2 Gbit/sec ISLs. Of course, it is also possible to connect a SilkWorm 2000 series switch to a trunking capable switch, as Brocade trunking capable switches are backwards compatible and will negotiate a 1 Gbit/sec ISL.

For core/edge topologies, place trunking capable switches in the core. If 2 Gbit/sec connectivity is required, it is acceptable to attach these devices to the 2 Gbit/sec cores if 2 Gbit/sec edge switches are not yet implemented. By placing 2 Gbit/sec switches in the core, it ensures that a 2 Gbit/sec path exists end to end. If a significant number of 2 Gbit/sec devices are required and the performance requirements are high, an effective strategy is to localize the 2 Gbit/sec devices on the same switch or group of switches. Figure 42 depicts 2 Gbit/sec devices with and without 2 Gbit/sec end-to-end paths as well as the localization of some 2 Gbit/sec devices.

(Note: This example is presented to establish the benefits of placing a trunking capable switch in the core and the issues that can arise by keeping a SilkWorm 2000 series switch in the core. This example is a poor design by choice and is not something that a designer should implement unless there are compelling reasons to do so.)

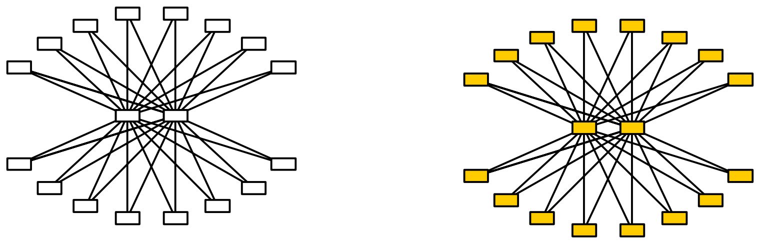
Figure 42. End to end 2 Gbit/sec path and 2 Gbit/sec device locality



More Ports or Higher Performance?

With 2 Gbit/sec switches, the designer has a choice: more ports or higher performance. In Figure 43, two 224-port core/edge topologies are shown. Note that with the same number of switches and ISLs, the SilkWorm 3800 based topology delivers better performance, with an ISL subscription ratio of 3.5:1 as compared to 7:1 with a SilkWorm 2800 based fabric.

Figure 43. With the same number of switches and ISLs, the SilkWorm 3800 based topology delivers better performance.



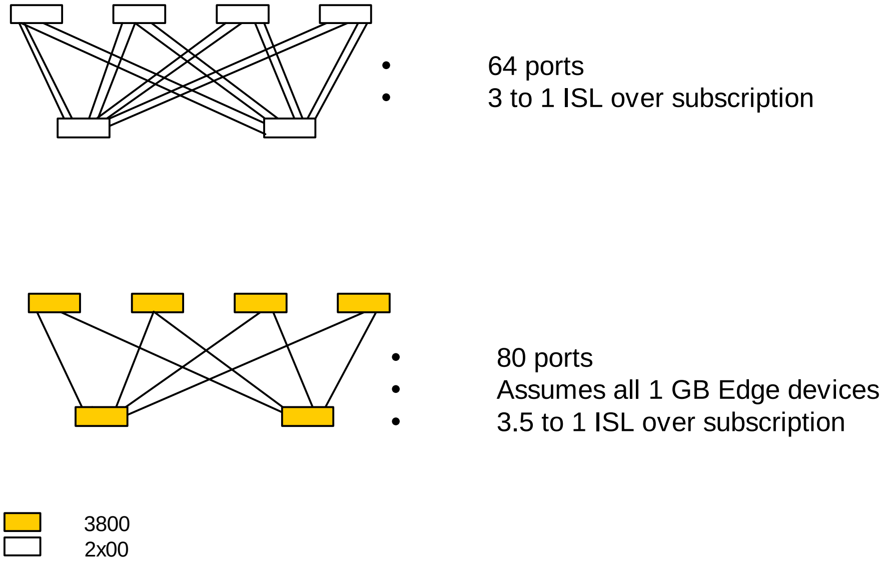
- 18 switch resilient star = 224 ports @ 7 to 1 ISL oversubsciption ratio

- **Assumes all 1 GB Edge devices**
- 18 switch resilient star = 224 ports @ 3.5:1 ISL oversubscription ratio

🟨 3800
□ 2x00

An alternative is to utilize fewer ISLs in a topology to yield similar performance and more device ports. A core/edge topology with six switches is shown in Figure 44. The top topology is designed with two ISLs from each edge switch to each core switch and yields 64-ports with a 3:1 ISL subscription ratio. The bottom topology is designed with one ISL from each edge switch to each core switch and yields 80-ports with a 3.5-to 1-ISL subscription ratio.

Figure 44. With the same number of switches and fewer ISLs, the SilkWorm 3800 based topology delivers similar performance and a higher port count than a SilkWorm 2800 based topology



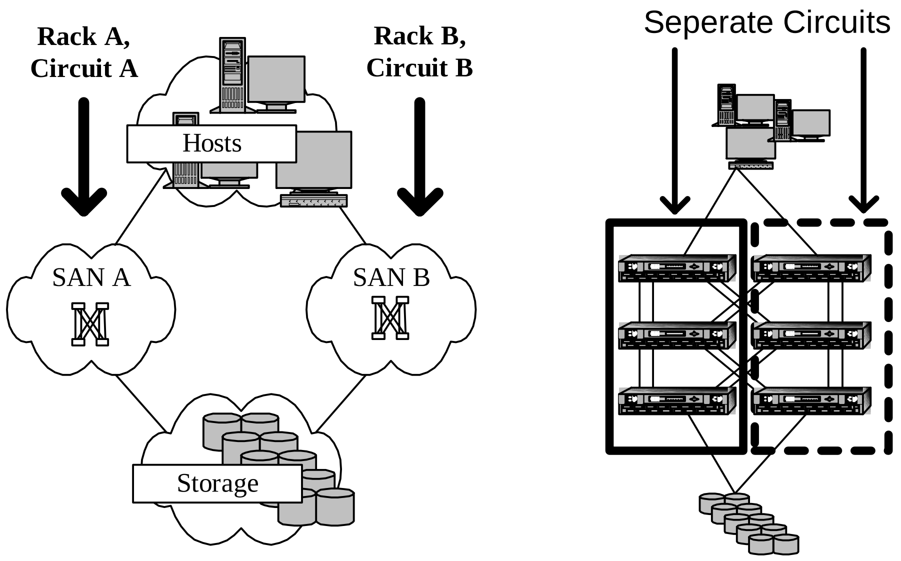
- 64 ports
- 3 to 1 ISL over subscription

- 80 ports
- Assumes all 1 GB Edge devices
- 3.5 to 1 ISL over subscription

Cabling For High Availability

When cabling SAN, be careful that a power outage does not take down the entire fabric or SAN. This means placing each fabric of a redundant fabric on different power circuits or cabling single fabrics in such a way that the fabric can remain functional if a power circuit fails, as shown in Figure 45.

Figure 45. Cabling A SAN for Availability



SilkWorm 12000 Device Attachment Strategies

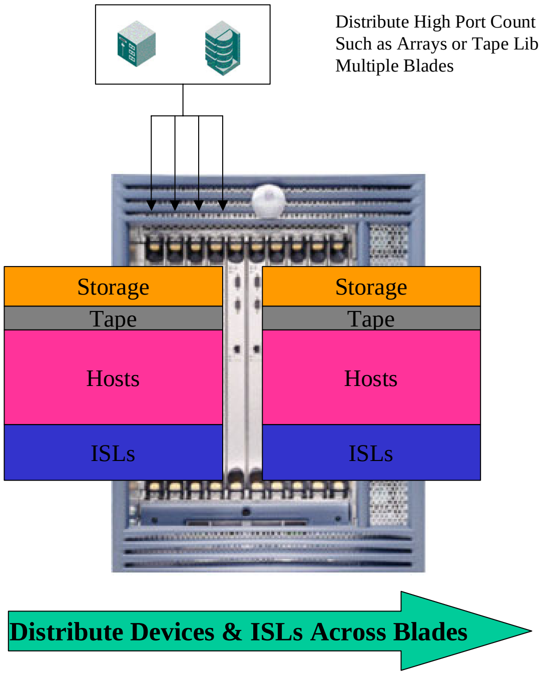
You must take availability, scalability, and performance into account when attaching devices to the SilkWorm 12000. Due to the high port density characteristics of the SilkWorm 12000, it is frequently easy to localize devices that communicate with each other onto the same switch. Localizing traffic enhances performance, as fewer ISLs are utilized and higher scalability since more ports are available for nodes.

Attaching Nodes for Availability

To maximize availability, distribute devices and ISLs across cards. This will minimize the impact to the SAN in the unlikely event of a 16-port card failure. To effectively distribute the connections, it is important to understand the connection types and relationships. For example, a large storage array may have sixteen ports. If these connections were evenly distributed across the cards of a SilkWorm 12000 switch, the failure of a 16-port card would only affect a few of the array ports. Similarly, when connecting devices by type (i.e. host, storage), distribute these connections across the SilkWorm 12000 16-port cards. Figure 46 depicts the attaching of devices across 16-port cards for availability. While it is not necessary to attach devices in groups, as shown, it does make it easier to manage the device connections.

Note: Distribute devices across 16-port cards from left to right for optimal availability; not from top to bottom.

Figure 46. Attaching devices to a 16-port card

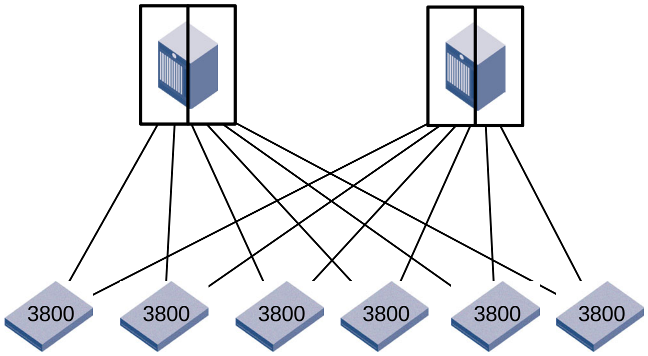

Distribute High Port Count Devices, Such as Arrays or Tape Libraries Across Multiple Blades
Storage
Tape
Hosts
ISLs
Storage
Tape
Hosts
ISLs
Distribute Devices & ISLs Across Blades

SilkWorm 12000 Availability Considerations

The dual switch – single chassis implementation of the SilkWorm 12000 ideally suits this switch for the role of a core switch. Recall that for a resilient core/edge fabric, two is the minimum number of core switches. The SilkWorm 12000 is a single chassis that houses two logical switches. The two logical switches are powered by an active Control Processor (CP) with a failover of both switches to the standby CP card occurring should the active CP card fail. During the failover, there is a brief disruption of I/O for both switches. Additionally, any environmental problem or operator error that could take out the whole chassis would then disrupt both fabrics simultaneously. For this reason we recommend one fabric per chassis. This means either connecting the two logical switches together or to other switches in the same fabric. Some designers may opt for a two chassis core for optimal availability and performance.

Note: For the highest availability and to avoid any environmental problem that could take out the whole chassis, consider using two chassis' for the core of a core/edge fabric also, it is suggested to only use one fabric per chassis to mitigate environmental problems that could take out the whole chassis.

Figure 47. Use Two SilkWorm 12000 Chassis For The Highest Availability



Section 5: Testing and Fabric Support Levels

Brocade is currently testing and validating fabrics consisting of hundreds of ports and has plans for testing multi-thousand port fabrics. The ultimate limitation in fabric design today and as defined in the Fibre Channel standards is a maximum of 239 physical switches, be they 8, 16, or 64 port versions. As a practical matter, no vendor has yet tested networks of this size due to the expense and complexity of implementing such a network. The current practical switch-count limit is lower than 239 switches, based upon empirical testing. Another limit on SAN design is the hop count and ISLs.

Brocade partners with many OEMs and resellers who supply switches to end-users. Many of these partners provide direct support for the switches they sell. These partners extensively test and support specific configurations, including switches, storage, HBAs, and other SAN components. In some cases, the large fabric configurations supported by Brocade partners will differ from the guidelines Brocade presents in this document. In fact, several Brocade switch partners have developed their own SAN design guidelines, including in-depth support matrixes that specify support configurations and firmware versions.

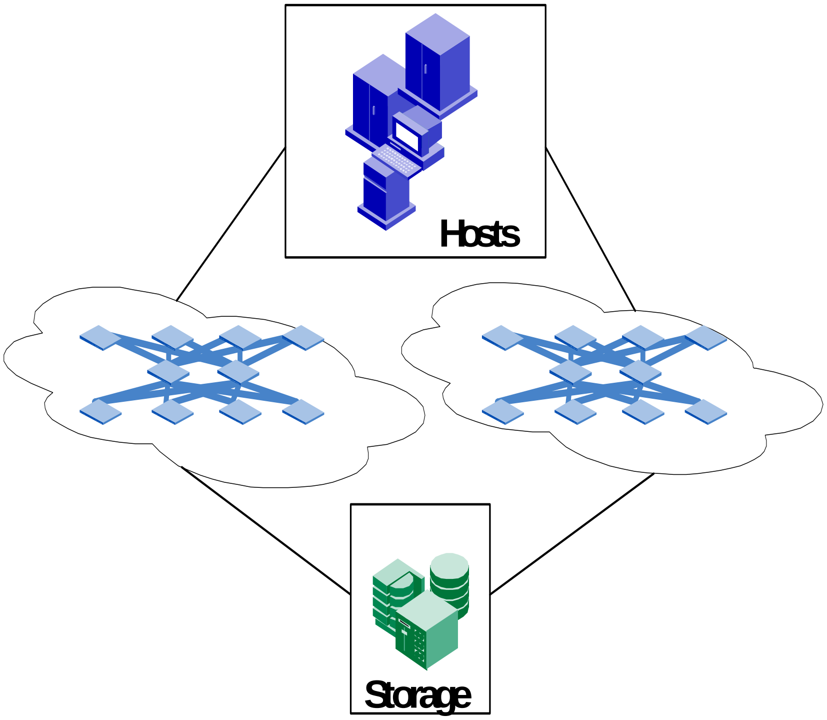
Those partners who do not provide direct switch support sell their customers Brocade Service Plans with the switches and these customers obtain their support from Brocade. For more information on support for SilkWorm switches, you can contact your switch provider. For more information on Brocade Service Plans, visit *www.brocade.com*.

To determine whether a SAN is supported, it is necessary to work with your support provider to determine if you SAN design is valid. Important variables that determine the supportability of a particular SAN are the number of switches, version of Fabric OS, the topology, number of ISLs, number of free ports, and the hop count.

Section 6: Reference Topologies

It is always beneficial to learn from others' experience. In addition to extensive internal testing programs, and leveraging testing done by partners, Brocade has made a serious effort to learn from end users what they have built, and how it has worked for them. This section contains some examples of networks that are successfully deployed by many end users. These designs are also extensively tested by Brocade. The foundation topology for the reference topologies that follow is the core/edge topology. Other topologies can and do work very well. If the requirements are scalability, availability, and performance, it is probable that these reference topologies will work for you. As discussed throughout this document, it is possible to tune the core/edge topology for performance and size by adding edge switches, core switches, or ISLs. For the best availability, it is recommended to design a dual fabric SAN, which can be accomplished by deploying any two of the reference topologies, as shown in Figure 48.

Figure 48. Dual, Resilient 96-Port Fabrics



If you have an estimate of your port count and performance requirements, you can pick one of the following designs and "be done with it". You do not have to build the entire design at once. You can build *towards* the target design, and "pay as you grow."

Note that two ISL over subscription ratios are provided in the reference topologies tables since it is necessary to account for 2 Gbit/sec capable switches. The ISL subscription ratios do not change if all devices and ISLs are the same speed. However, if the edge devices are all 1 Gbit/sec and the ISLs are 2 Gbit/sec, the ratios *do* change. If your mix of devices is different, you can calculate the ISL over-subscription ratio as detailed in ISL Over-Subscription on page 26.

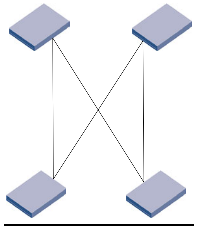
The reference topologies are presented in three sections: topologies built using only 16-port switches, topologies built using a mixture of 16-port and 64-port switches, and topologies built using only 64-port switches. They are presented this way because typically the designs are conceived that way. E.g. a user who already has a set of 16-port switches will not need to refer to designs which only use 64-port switches, whereas an enterprise customer who is deploying a "green field" SAN and wants maximal scalability usually will *only* look at 64-port switches.

16-Port Switch Designs

56-Port Fabric

This is an effective small fabric. It looks similar to a mesh topology. If you plan to grow this fabric, device placement becomes important. Instead of filling up all switches with devices, start adding new switches to the core switches to enable growth. In the fashion, it is possible to grow the fabric to 224-ports, because it becomes a "starter core/edge fabric." If the fabric is targeted to start small and stay below 56-ports, it is acceptable to fill the empty core ports with edge devices. The performance characteristics will differ from a core/edge topology under this scenario due to device placement on the core switches (see Section 2, Device Attachment Strategies) and it may be necessary to employ some level of locality or to insert additional ISLs.

Figure 49. 56-port Core/Edge Fabric

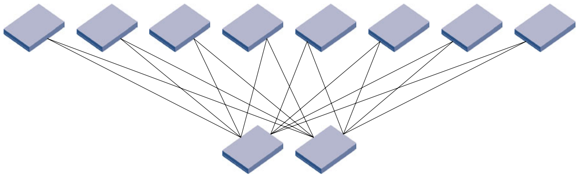


Identification	2 edge by 2 core by 1 ISL (2*e* x 2*c* x 1*i*)	
Edge ports	56	
Switch count	4	
ISL over-subscription		
Same Speed Edge Devices and ISLs	**1 Gb Edge Devices / 2 Gb ISLs**	
7 to 1	**3.5 to 1**	

96-Port Fabric

This fabric is effective if 96 high performance ports are required. To scale this fabric it is necessary to migrate the core switches to the edge and utilize higher port count core switches to fill in the vacancies. Alternatively, if a high performance, high port count fabric is required; it may be more appropriate to use four switches in the core (see 160-Port Fabric reference topology), with a single ISL connecting each Edge switch to each core switch. Notice that the high performance is enabled by the use of two ISLs between each Edge switch and each core switch. It is not necessary to utilize two ISLS between the edges and cores if the performance is not required. If single ISLs are utilized between Edge and core, it creates a fabric that yields 128-ports.

Figure 50. 96-port Core/Edge Fabric

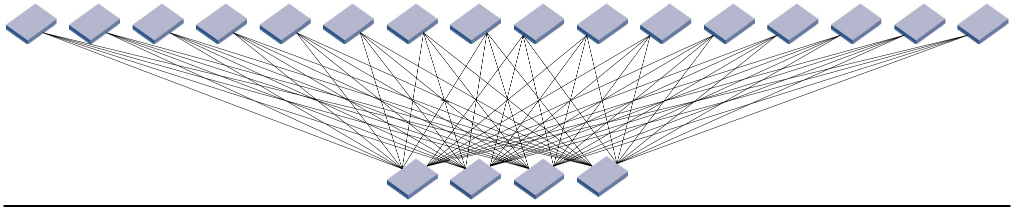


Identification	8 edge by 2 core by 2 ISLs (8*e* x 2*c* x 2*i*)
Edge ports	96
Switch count	10
ISL over-subscription	
Same Speed Edge Devices and ISLs	**1 Gb Edge Devices / 2 Gb ISLs**
3 to 1	**1.5 to 1**

160-Port Fabric

This mid-sized fabric is capable of scaling to 192 ports by adding Edge switches and even larger with higher port count core switches. Note that four switches are utilized in the core to deliver high performance and enable scalability and it is possible to build a similar topology with only two core switches. This topology provides the same performance as the 96-port reference topology.

Figure 51. 160-port Core/Edge Fabric

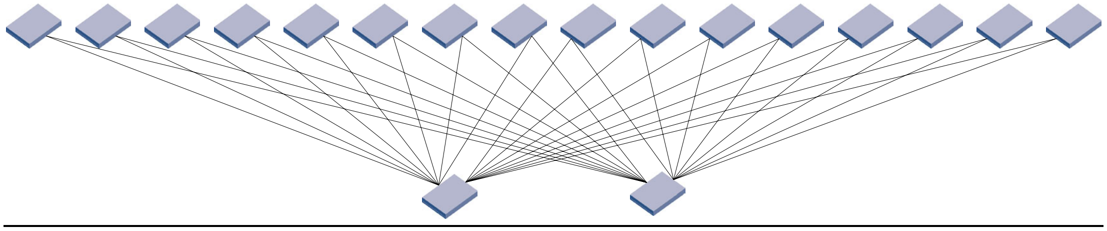


Identification	16 edge by 4 core by 1 ISL (12*e* x 4*c* x 1*i*)
Edge ports	160
Switch count	16
ISL over-subscription	
Same Speed Edge Devices and ISLs	**1 Gb Edge Devices / 2 Gb ISLs**
3 to 1	**1.5 to 1**

224-Port Fabric

This is the largest simple core/edge topology possible using 16-port switches. If a larger SAN is required, use larger core switches or build a Hybrid topology.

Figure 52. 224-port Core/Edge Fabric

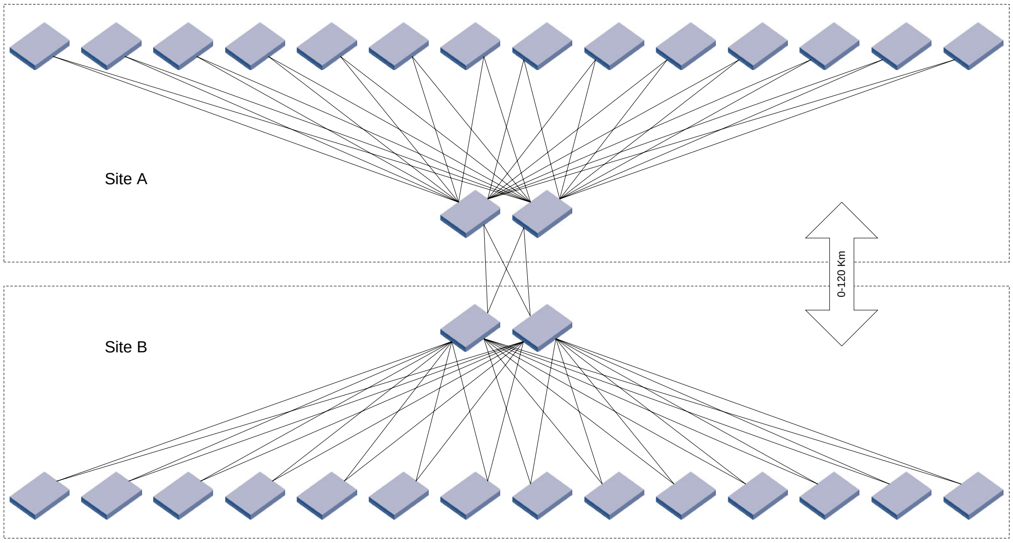


Identification	16 edge by 2 core by 1 ISL (16*e* x 2*c* x 1*i*)
Edge ports	224
Switch count	18
ISL over-subscription	
Same Speed Edge Devices and ISLs	**1 Gb Edge Devices / 2 Gb ISLs**
7 to 1	**3.5 to 1**

Extended Distance Topology

This is the recommended topology to utilize when connecting two geographically separate sites. The fabric maximum size is 392-ports when using 16-port switches. It is possible to build a smaller fabric using this topology. Scaling performance by adding ISLs requires a smaller configuration or the replacement of the existing complex core with larger core switches. To maintain performance, locality within each location is necessary, as the bandwidth between locations is minimal. Note that ISL over-subscription within a location is 7:1.

Figure 53. 392-Port Complex Core/Edge Fabric



Identification	28 edge by 4 core (complex) by 1 ISL (28*e* x 4*c[complex]* x 1*i*)
Edge ports	392
Switch count	32
ISL over-subscription (no locality)	
Same Speed Edge Devices and ISLs	**1 Gb Edge Devices / 2 Gb ISLs**
49 to 1	**24.5 to 1**
ISL over-subscription (locality within local core/edge sub-fabric)	
Same Speed Edge Devices and ISLs	**1 Gb Edge Devices / 2 Gb ISLs**
7 to 1	**3.5 to 1**

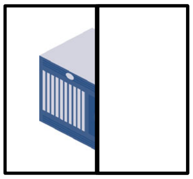
64-Port Switch Designs

Note: New supported configurations are being added constantly; check with your switch support provider for the most current list of supported configurations.

64-Port Fabric

The SilkWorm 12000 can support up to 64 ports in a single logical switch. It is therefore possible to solve design problems of up to 64 ports with a single chassis. It is possible to start with 32 ports and scale to 64 ports in each half of the chassis, for a total of 128 usable ports per chassis.

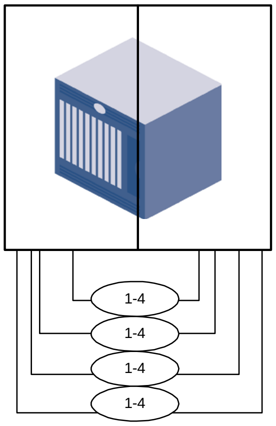
While this design is usually considered undesirable because of the fundamental limit to its scalability, in some environments it may work perfectly well. For example, it is an easy way to use the SilkWorm 12000 to replace legacy director products without making fundamental changes to the SAN architecture.



1 x 12000
64 Available Fabric Ports

124-Port Fabric

The two logical switches in a chassis can be connected to form one fabric. A minimum of two ISLs should be used to ensure high availability. In this configuration, the fabric would have 124 available ports. If greater performance between the switches is desired, more ISLs can be added. A reasonable design target might be to start with a 7:1 over-subscription ratio on the ISLs. In this scenario, each blade in the SilkWorm 12000 would have two ISLs. This would result in a fabric with 112 free ports. For the most performance-critical applications, a ratio of 3:1 could be targeted. Each blade would have four ISLs in that scenario, and the total free port count would be 96 in the fabric.

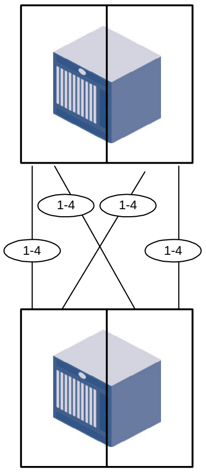


2 x 12000
up to 124 Available Fabric Ports

Identification	2 64-port switch cascade with 4 to 16 ISLs
Edge ports	96 to 124
Switch count	2
ISL over-subscription	
Same Speed Edge Devices and ISLs	**1 Gb Edge Devices / 2 Gb ISLs**
7 to 1, using 8 ISLs	**3.5 to 1, using 8 ISLs**

248-Port Fabric

The same principal used in the 124-port fabric (above) can be extended to interconnect two chassis. Using one ISL per blade on just two blades per switch, the resulting fabric has 248 available ports. Targeting a 7:1 over-subscription ratio with two ISLs per blade one each blade, the number of ports in the fabric is 224. Targeting a 3:1 over-subscription ratio by using four ISLs per blade, the port count is 192 in the fabric.

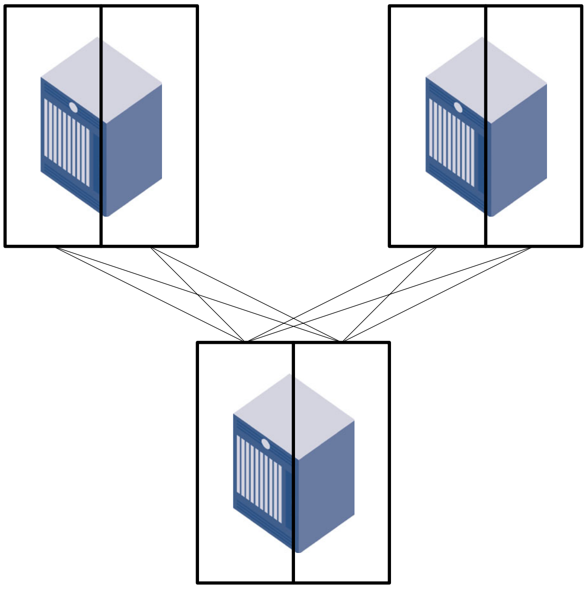


4 x 12000
up to 240 Available Fabric Ports

Identification	2 64-port edge by 2 64-port core by 4 to 8 ISLs
Edge ports	192 to 248
Switch count	4
ISL over-subscription	
Same Speed Edge Devices and ISLs	**1 Gb Edge Devices / 2 Gb ISLs**
7 to 1, using 8 ISLs	**3.5 to 1, using 8 ISLs**

368-Port Fabric

The 248-port design above can be further extended to three chassis, or six 64-port logical switches. If this is done, the total available port can be as high as 368 ports. Targeting a 7:1 over-subscription ratio, the free port count is 320. Targeting a 3:1 over-subscription ratio by using four ISLs per blade, the port count is 256 in the fabric.



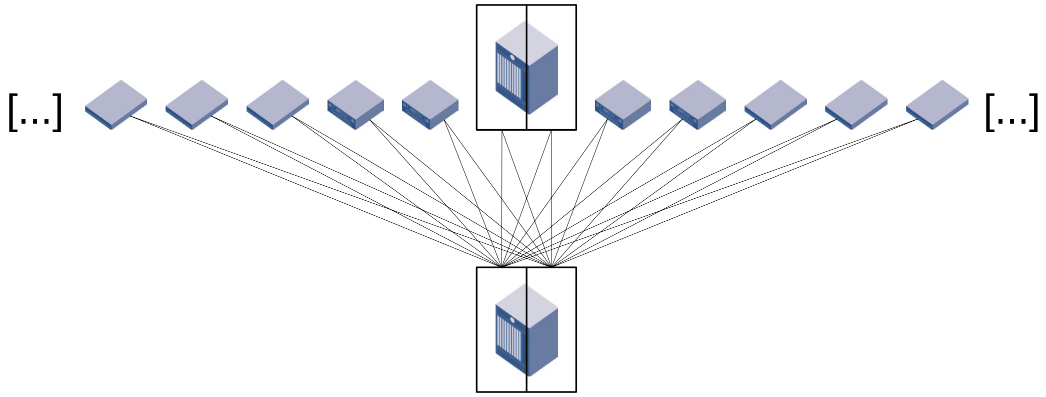
Identification	4 64-port edge by 2 64-port core by 4 to 8 ISLs
Edge ports	256 to 368
Switch count	6
ISL over-subscription	
Same Speed Edge Devices and ISLs	**1 Gb Edge Devices / 2 Gb ISLs**
7 to 1, using 8 ISLs	**3.5 to 1, using 8 ISLs**

16- and 64-Port Switch Mixed Designs

Note: New supported configurations are being added constantly; check with your switch support provider for the most current list of supported configurations.

440-Port and 512-Port Fabrics

It is possible to use a SilkWorm 12000 as a core, and use 16-port switches at the edge. Currently, the total number of supported edge switches is 26. This number is increasing as testing progresses. Using 7:1 as an over-subscription target, it is possible to design a network with 440 available ports. Up to two of the edge switches could be SilkWorm 12000 logical switches. The network could then scale to up to 512 available ports



Identification	2 64-port edge plus 24 16-port edge by 2 64-port core by 2-8 ISLs
Edge ports	512
Switch count	28
ISL over-subscription	
Same Speed Edge Devices and ISLs	**1 Gb Edge Devices / 2 Gb ISLs**
7 to 1	**3.5 to 1**

Copyright

IMPORTANT NOTICE

This document is the property of Brocade. It is intended solely as an aid for installing and configuring Storage Area Networks constructed with Brocade switches. This document does not provide a warranty to any Brocade software, equipment, or service, nor does it imply product availability. Brocade is not responsible for the use of this document and does not guarantee the results of its use. Brocade does not warrant or guarantee that anyone will be able to recreate or achieve the results described in this document. The installation and configuration described in this document made use of third party software and hardware. Brocade does not make any warranties or guarantees concerning such third party software and hardware.

NOTICE: THIS DOCUMENT IS FOR INFORMATIONAL PURPOSES ONLY AND DOES NOT SET FORTH ANY WARRANTY, EXPRESS OR IMPLIED, CONCERNING ANY EQUIPMENT, EQUIPMENT FEATURE, OR SERVICE OFFERED OR TO BE OFFERED BY BROCADE. BROCADE RESERVES THE RIGHT TO MAKE CHANGES TO THIS DOCUMENT AT ANY TIME, WITHOUT NOTICE, AND ASSUMES NO RESPONSIBILITY FOR ITS USE. THIS INFORMATIONAL DOCUMENT DESCRIBES FEATURES THAT MAY NOT BE CURRENTLY AVAILABLE. CONTACT A BROCADE SALES OFFICE FOR INFORMATION ON FEATURE AND PRODUCT AVAILABILITY.

Export of technical data contained in this document may require an export license from the United States Government.

Brocade Communications Systems, Incorporated
1745 Technology Drive
San Jose, CA 95110