BROCADE

DEPLOYMENT GUIDE

SAN Design and Best Practices

A high-level guide focusing on Fibre Channel Storage Area Network (SAN) design and best practices, covering planning, topologies, device sharing in routed topologies, workload monitoring, and detecting server and storage latencies—to help with decisions required for successful SAN design.

53-1004781-01
23 November 2016

Contents

About Brocade

Brocade® (NASDAQ: BRCD) networking solutions help the world's leading organizations transition smoothly to a world where applications and information reside anywhere. This vision is designed to deliver key business benefits such as unmatched simplicity, non-stop networking, application optimization, and investment protection.

Innovative Ethernet and storage networking solutions for data center, campus, and service provider networks help reduce complexity and cost while enabling virtualization and cloud computing to increase business agility.

To help ensure a complete solution, Brocade partners with world-class IT companies and provides comprehensive education, support, and professional services offerings (www.brocade.com).

Introduction

This document is a high-level design and best-practices guide based on Brocade products and features, focusing on Fibre Channel SAN design. Covered topics include the early planning phase, understanding possible operational challenges, and monitoring and improving an existing SAN infrastructure.

Emphasis is given to in-depth information on how the data flows between devices affect a SAN infrastructure at a design and operational level.

The guidelines in this document do not apply to every environment, but they will help guide you through the decisions that you need to make for successful SAN design. Consult your Brocade representative or refer to the documents in Appendix E for details about the hardware and software products, as well as the interoperability features mentioned in text and illustrations.

> NOTE
> This is a "living" document that is continuously being expanded, so be sure to frequently check MyBrocade for the latest updates to this and other best-practice documents.

Audience and Scope

This guide is for technical IT architects who are directly or indirectly responsible for SAN design based on the Brocade® Gen 5 and Gen 6 Fibre Channel SAN platforms. It describes many of the challenges that face SAN designers today in both greenfield and legacy storage environments. While not intended as a definitive design document, this guide introduces concepts and guidelines to help you avoid potential issues that can result from poor design practices. This document describes best-practice guidelines in the following areas:

- Architecting a core data center infrastructure
- Capacity planning
- SAN topology
- Inter-switch connectivity
- Data flows
- Predeployment infrastructure testing
- Device connections
- Workloads/virtualization
- Distance extension
- Fibre Channel routing

NOTE
A solid understanding of SAN concepts and Brocade Fibre Channel technology is assumed. Please refer to Appendix E for recommended additional publications.

Approach

While some advanced features and specialized SAN applications are discussed, these topics are covered in greater detail in separate documents. The primary objective of this guide is to provide a solid foundation to facilitate successful SAN designs—designs that effectively meet current and future requirements. This document addresses basic administration and maintenance, including capabilities to identify early warning signs for end-device (initiator or target) latency, which can cause congestion in the SAN fabric. However, you should consult product documentation and documents in Appendix E for more details. Comprehensive discussions of SAN fabric administration, storage network cabling, and Fibre Channel security best practices are covered in separate documents.

Overview

Although Brocade SAN fabrics are plug-and-play and can function properly even if left in a default state, Fibre Channel networks clearly benefit from a well-thought-out design and deployment strategy. In order to provide reliable and efficient delivery of data, your SAN topology should follow best-practice guidelines based on SAN industry standards and considerations specific to Brocade.

This document does not consider physical environment factors such as power, cooling, and rack layout. Rather, the focus is on network connectivity (both inter-switch and edge device) and software configurations.

NOTE
The scope of this document is switch-centric and does not discuss end-device setup, configuration, or maintenance. Fabric monitoring, management, diagnostics, cabling, and migration are covered in separate documents.

Architecting a SAN

The SAN planning process is similar to any type of project planning and includes the following phases:

- Phase I: Gathering requirements
- Phase II: Developing technical specifications
- Phase III: Estimating project costs
- Phase IV: Analyzing Return on Investment (ROI) or Total Cost of Ownership (TCO) (if necessary)
- Phase V: Creating a detailed SAN design and implementation plan

When selecting which criteria to meet, you should engage users, server and storage subject matter experts (SMEs), and other relevant experts to understand the role of the fabric. Since most SANs tend to operate for a long time before they are renewed, you should take future growth into account as SANs are difficult to re-architect. Deploying new SANs or expanding existing ones to meet additional workloads in the fabrics requires a critical assessment of business and technology requirements. Proper focus on planning will ensure that the SAN, once deployed, meets all current and future business objectives, including availability, deployment simplicity, performance, future business growth, and cost. Tables in Appendix B are provided as a reference for documenting assets and metrics for SAN projects.

A critical aspect for successful implementation that is often overlooked is the ongoing management of the fabric. Identifying systems-level SMEs for all components that make up the SAN, as well as adequate and up-to-date training on those components, is critical for efficient design and operational management of the fabric. When designing a new SAN or expanding an existing SAN, you should take into account the following parameters:

Application Virtualization

- Which applications will run under a virtual machine (VM) environment?
- How many VMs will run on a physical server?
- Migration of VMs under what conditions (business and nonbusiness hours; is additional CPU or memory needed to maintain response times)?
- Is there a need for solid-state storage media to improve read response times?

Homogenous/Heterogeneous Server and Storage Platforms

- Are blade servers or rack servers used?
- Is auto-tiering in place?
- Which Brocade Fabric OS® (FOS) versions are supported in a multivendor storage environment?
- What is the planned refresh cycle of servers and storage platforms (2 years/3 years)?

Scalability

- How many user ports are needed now?
- How many devices will connect through an access gateway?
- How many inter-switch links (ISLs)/Brocade UltraScale inter-chassis links (ICLs) are required to minimize congestion in the fabric?
- What distances for ISL/ICL connections need to be supported?
- Does the fabric scale out at the edge or the core?

Backup and Disaster Tolerance

- Is there a centralized backup? (This will determine the number of ISLs needed to minimize congestion at peak loads.)
- What is the impact of backup on latency-sensitive applications?
- Is the disaster solution based on a metro Fibre Channel (FC) or Fibre Channel over IP (FCIP) solution?

Diagnostics and Manageability

- What is the primary management interface to the SAN (command-line interface [CLI], Brocade Network Advisor, or third-party tool)?
- How often will Brocade FOS and Brocade Network Advisor be updated?
- How is cable and optics integrity validated?

Investment Protection

- Is support needed for adding Gen 6 switches into a Gen 5 fabric?
- Is support needed for storage technologies like NVMe over Fabrics?
- What device interoperability support is required?
- Is interoperability required for other technologies such as UCS?

Operational Considerations

While Brocade fabrics scale in terms of port density and performance, the design goal should be to ensure simplicity for easier management, future expansion, and serviceability. Examples of this simplicity may include using a two-tier edge-core topology; avoiding the use of both Inter-Fabric Routing (IFR) and Virtual Fabrics (VF) where not required; and turning on port monitoring parameters for critical applications.

NOTE

Refer to the *FOS Scalability Matrix* for currently tested and supported scalability limits. Any requirements beyond the tested scalability limits should be pretested in a nonproduction environment, or system resources like CPU and memory utilization should be actively monitored to minimize fabric anomalies.

Be the Pilot

Whether building a new SAN or connecting to an existing SAN, prestaging and validating a fabric/application before putting it into production ensures that there are baseline metrics in terms of rated throughput, latency, and expected Cyclic Redundancy Check (CRC) errors based on patch panel and physical cable infrastructure.

Predeployment Cabling and Optics Validation

For SANs built with Brocade Gen 5 and Gen 6 Fibre Channel switches equipped with 16-Gb or higher optics, Brocade ClearLink Diagnostics enables the use of predeployment testing to validate the integrity of the physical network infrastructure before operational deployment. Part of Brocade Fabric Vision technology, ClearLink is an offline diagnostics tool that allows users to perform an automated battery of tests to measure and validate maximum throughput speeds as well as latency and distance across fabric links. ClearLink Diagnostics can also be used to verify the health and integrity of all 16- and 32-Gbps transceivers in the fabric on a one-by-one basis. It is recommended that diagnostics be conducted before deployment or when there are CRC errors that could be caused by physical-layer issues.

A ClearLink Diagnostics port (D_Port) requires that only the individual ports attached to the tested link go offline, allowing the remainder of the ports to stay online in isolation from the link. It can also be used to test links to a new fabric switch without allowing the new switch to join or even be aware of the current fabric, providing an opportunity to measure and test ISLs before they are put into production. This fabric-based physical-layer validation enables the following:

- Transceiver health check
- Transceiver uptime
- Local and long-distance measurements (5-meter granularity for 16-Gbps and 32-Gbps Small Form-factor Pluggable [SFP] optics and 50 meters for 10-Gbps SFP optics)
- Link latency measurements between D_Ports
- Link power (dB) loss
- Link performance

The following is an overview of Brocade ClearLink Diagnostics guidelines and restrictions. Refer to the *Brocade Fabric OS Administration Guide* for a more detailed discussion of diagnostic port usage.

- ClearLink Diagnostics is supported only on Gen 5 Fibre Channel platforms running FOS 7.0 and later and Gen 6 Fibre Channel platforms running FOS 8.0 and later.
- Complete ClearLink support, including optical and electrical loopback testing, requires Brocade branded 16-Gbps or 32-Gbps SWL transceivers, 16/32-Gbps LWL transceivers, or 16-Gbps ELWL transceivers.
- Brocade FOS 7.1 and later support connections between Brocade switches and between switches and devices equipped with ClearLink Diagnostics-enabled host or target adapters. Check with your adapter vendor or refer to the *Brocade Fabric OS Administration Guide* for details on supported models.
- ClearLink is supported on ISL links with ports on both ends of the ISL in ClearLink D Port mode.
- Running ClearLink tests between a switch and a non-Brocade HBA requires that a Fabric Vision license be installed or the combination of Fabric Watch and Advanced Performance Monitoring on a Gen 5 switch (FOS 7.3.2 or later).
- Brocade FOS 7.1 provides ClearLink Diagnostics support on UltraScale ICLs on the Brocade DCX® 8510 Backbone.

- ClearLink Diagnostics on UltraScale ICLs, FC16-64 blades, 8-Gbps LWL/ELWL SFPs, and 10-Gbps SFPs skips the electrical and optical loopback tests, because the 8-Gb and 10-Gbps SFP and Quad Small Form-factor Pluggable (QSFP) transceivers used do not support it.
- If Brocade inter-switch link (ISL) Trunking is deployed, use a minimum of two ports for the trunk. This enables the user to take down one of the links for diagnostic testing without disrupting the traffic on the remaining trunk members.
- Make sure that there are at least two ISLs before taking a port offline for diagnostic testing. This ensures redundancy and prevents fabric segmentation in case a link is taken down for diagnostics.

Refer to "Appendix A: ClearLink Diagnostics" in the *SAN Fabric Resiliency and Administration Best Practices* guide for details on enhancements in each FOS release.

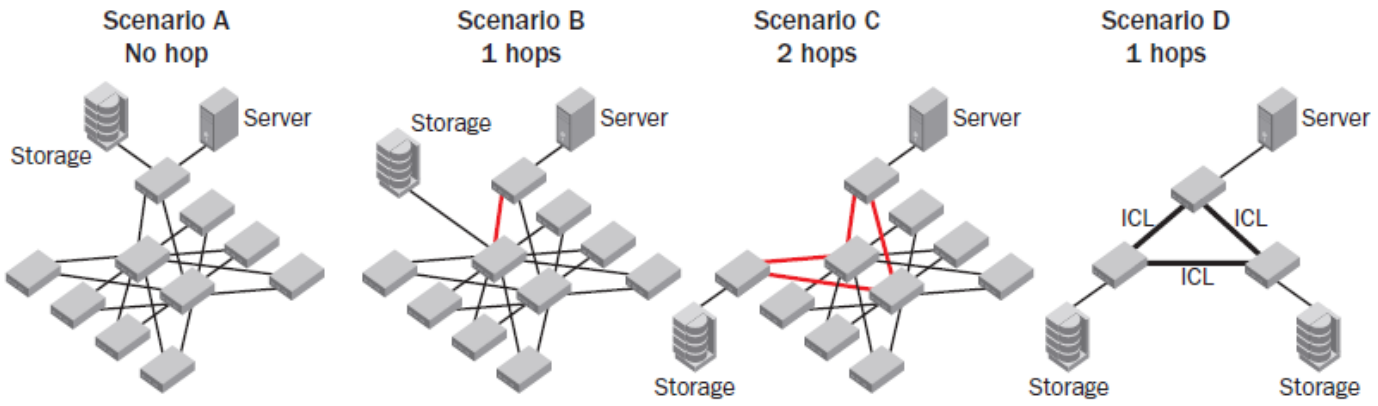
SAN Design Basics

This section provides high-level guidelines for installing a typical SAN. The focus is on best practices for core-edge or edge-core-edge fabrics. The discussion starts at the highest level, the data center, and works down to the port level, providing recommendations at each point along the way.

Topologies

A typical SAN design comprises devices on the edge of the network, switches in the core of the network, and the cabling that connects all the devices together. Topology is usually described in terms of how the switches are interconnected, such as ring, core-edge, and edge-core-edge or fully meshed. At this point the focus is on switch topology with SLs —device connectivity is discussed in later sections. The recommended SAN topology to optimize performance, management, and scalability is a tiered, core-edge topology (sometimes called core-edge or tiered core edge). This approach provides good performance without unnecessary interconnections. At a high level, the tiered topology has a large number of edge switches used for device connectivity and a smaller number of core switches used for routing traffic between the edge switches, as shown in Figure 1.

FIGURE 1 Four Scenarios of Tiered Network Topologies (Hops Shown in Heavier, Red Connections)



The difference between these four scenarios is device placement (where devices are attached to the network) and the associated traffic flow, which is discussed further in the "Data Flow Considerations" section later in this document.

- Scenario A has localized traffic, which can have small performance advantages for performance-optimized workloads but does not provide ease of scalability or manageability.

- Scenario B, also called edge-core, separates the storage and servers, thus providing ease of management and higher scalability. An edge-core topology has only one hop from server to storage, providing similar performance benefits as full mesh while allowing higher scalability.
- Scenario C also known as edge-core-edge, has both storage and servers on edge switches, which provides ease of management and the most flexible scalability.
- Scenario D is a full-mesh topology, and server to storage is no more than one hop. Designing fabrics with UltraScale ICLs is an efficient way to save front-end ports, and users can easily build a large (for example, 1536-port or larger) fabric with minimal SAN design considerations.

Edge-Core Topology

The edge-core topology (Figure 1 on page 10) places initiators (servers) on the edge tier and storage (targets) on the core tier. Since the servers and storage are on different switches, this topology provides ease of management as well as good performance with minimal fabric latency, with most traffic traversing only one hop from the edge to the core. (Storage-to-storage traffic is two hops if the second storage is on another core switch, but the two cores can be connected if fabrics are redundant.) The disadvantage to this design is that the storage and core connections are in contention for expansion as they scale; however, using modular platforms allows for flexibility while allowing the use of ICL ports for intra-switch connectivity to free up device ports.

Edge-Core-Edge Topology

The edge-core-edge topology (Figure 1 on page 10) places initiators on one edge tier and storage on another edge tier, leaving the core for switch interconnections or connecting devices with network-wide scope, such as Dense Wavelength Division Multiplexers (DWDMs), interfabric routers, storage virtualizers, tape libraries, and encryption engines. Since servers and storage are on different switches, this design enables independent scaling of compute and storage resources and ease of management with traffic traversing two hops from the edge through the core to the other edge. The principal benefit to edge-core-edge is that it provides an easy path for expansion since ports and/or switches can readily be added to the appropriate tier as needed.

Full-Mesh Topology

A full-mesh topology (Figure 1 on page 10) allows you to place servers and storage anywhere, since communication between source and destination is no more than one hop. Using director-class switches with UltraScale ICL ports for interconnectivity is essential to this design in order to ensure maximum device port availability and utilization.

Scalability

Regardless of the chosen topology, UltraScale ICLs on the Brocade DCX 8510 and X6 directors allow architects to build fabrics that are scalable and cost-effective compared to the previous generation of SAN products. UltraScale ICLs enable support for up to 12 chassis using a 4x8 core/edge design, allowing connectivity for up to 6,144 device ports, Using a full-mesh topology, ICLs enable connectivity for up to 9 directors, or 4,608 device ports.

NOTE

From a resiliency perspective, hop count is not a concern if the total switching latency is less than the disk I/O timeout value. Flash storage users who want to minimize fabric latency may consider lower hop counts to maximize performance.

High Performance Workloads

Over the last few years, enterprises have come to leverage low-latency, high-throughput flash arrays for demanding, performance-sensitive workloads. Brocade's Gen 6 Fibre Channel with IO Insight enhancements is perfectly suited to these types of workloads due to the submicrosecond latency through the switch and the increased bandwidth offered by 32-Gb throughput speeds while providing accurate IO latency instrumentation. Performance testing has shown that 16Gb- and even 8Gb-attach all-flash arrays can realize dramatic benefits by connecting to a Gen 6 SAN and host adapter, offering gains up to 2x over Gen 5 SANs. The Gen 6 standard also includes the use of Forward Error Correction (FEC) to ensure transmission reliability and a highly deterministic data flow. FEC corrects up to 140 corrupt bits per 5280-bit frame at the receiving end of the link, avoiding the need to retransmit frames when bit errors are detected.

For these demanding workloads, a no-hop fabric connection through a single ASIC switch like the Brocade G620 or locally switched on a single director port blade will minimize SAN fabric latency to submicrosecond speeds. Local switching is the ability to switch data traffic through a single ASIC by using both ingress and egress switching ports in a common port group. When using switches that contain multiple switching ASICs like the Brocade 6520 or port blades in a DCX 8510/X6 director, configuring host and target connections on the ports that share a common ASIC will minimize latency by avoiding the need to move data across multiple ASICs/port groups or across a director backplane to a different blade. To find details on port groups and local switching configuration, refer to the *Brocade Fabric OS Administration Guide* and the hardware installation guide for the appropriate product.

Because Gen 6 FC is backwards compatible with Gen 5 networking, a Gen 6 edge switch or director can be added into an existing Gen 5 fabric. This allows for local all-flash connectivity to a Gen 6 switch to gain the performance advantages of Gen 6 while preserving the investment in Gen 5 networks.

Redundancy and Resiliency

An important aspect of SAN topology is the resiliency and redundancy of the fabric. The main objective is to remove any single point of failure. Resiliency is the ability of the network to continue to function and/or recover from a failure, while redundancy describes duplication of components, even an entire fabric, to eliminate a single point of failure in the network. Brocade fabrics have resiliency built into Brocade Fabric OS (FOS), the software that runs on all Brocade B-Series switches, that can quickly "repair" the network to overcome most failures. For example, when a link between switches fails, FSPF quickly recalculates all traffic flows. Of course this assumes that there is a second route, which is when redundancy in the fabric becomes important.

The key to high availability and enterprise-class installation is redundancy. By eliminating a single point of failure, business continuance can be provided through most foreseeable and even unforeseeable events. At the highest level of fabric design, the complete network should be redundant, with two completely separate fabrics that do not share any network equipment (routers or switches).

Servers and storage devices should be connected to both networks utilizing some form of Multi-Path I/O (MPIO) solution, such that data can flow across both networks seamlessly in either an active/active or active/passive mode. MPIO ensures that if one path fails, an alternative is readily available. Ideally, the networks would be identical, but at a minimum they should be based on the same switch architecture to ensure consistency of performance. In some cases, these networks are in the same location. However, in order to provide for Disaster Recovery (DR), two separate locations are often used, either for each complete network or for sections of each network. Regardless of the physical geography, there are two separate networks for complete redundancy.

In summary, recommendations for the SAN design are to ensure application availability and resiliency via the following:

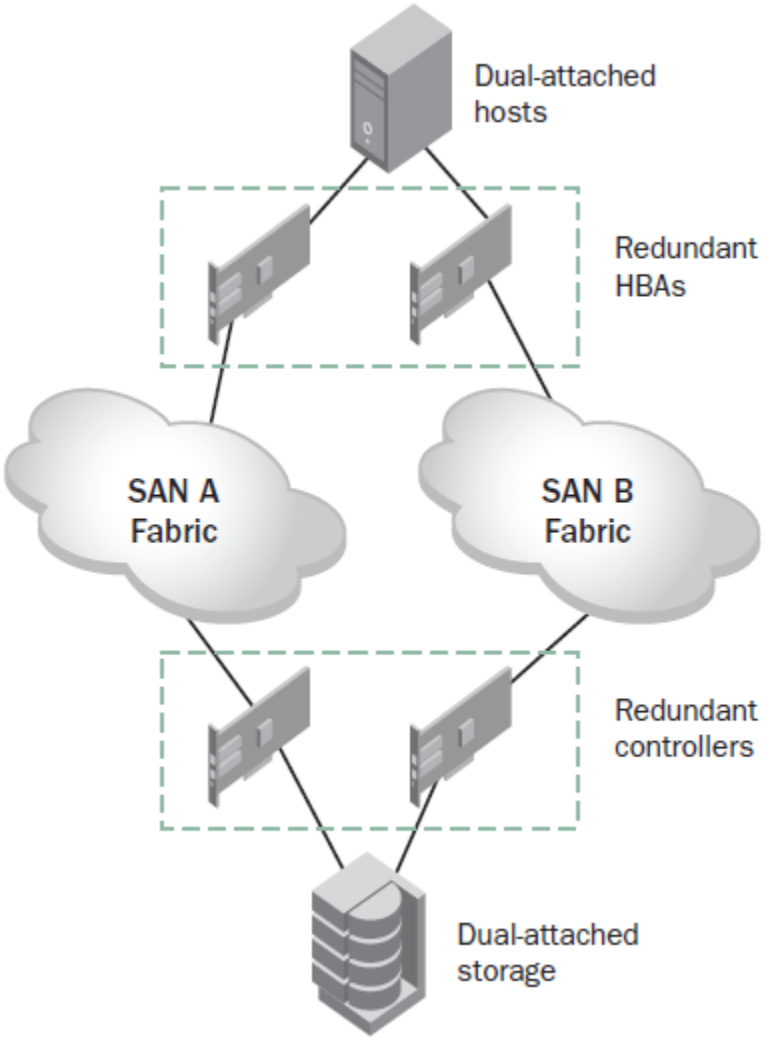
- Redundancy built into fabrics to avoid a single point of failure
- Servers connected to storage via redundant fabrics
- MPIO-based failover from server to storage
- Redundant fabrics based on similar architectures
- Redundant ISLs/ICLs for interswitch connectivity
- Separate storage and server tiers for independent expansion

- Core switches of equal or higher performance compared to the edges
- The highest performance switch in the fabric defined to be the principal switch

Switch Interconnections

As mentioned previously, there should be at least two of every element in the SAN to provide redundancy and improve resiliency. The number of available ports and device locality (server/storage tiered design) determines the number of ISLs needed to meet performance requirements. This means that there should be a minimum of two trunks, with at least two ISLs per trunk. Each source switch should be connected to at least two other switches, and so on. In Figure 2, each of the connection lines represents at least two physical cable connections.

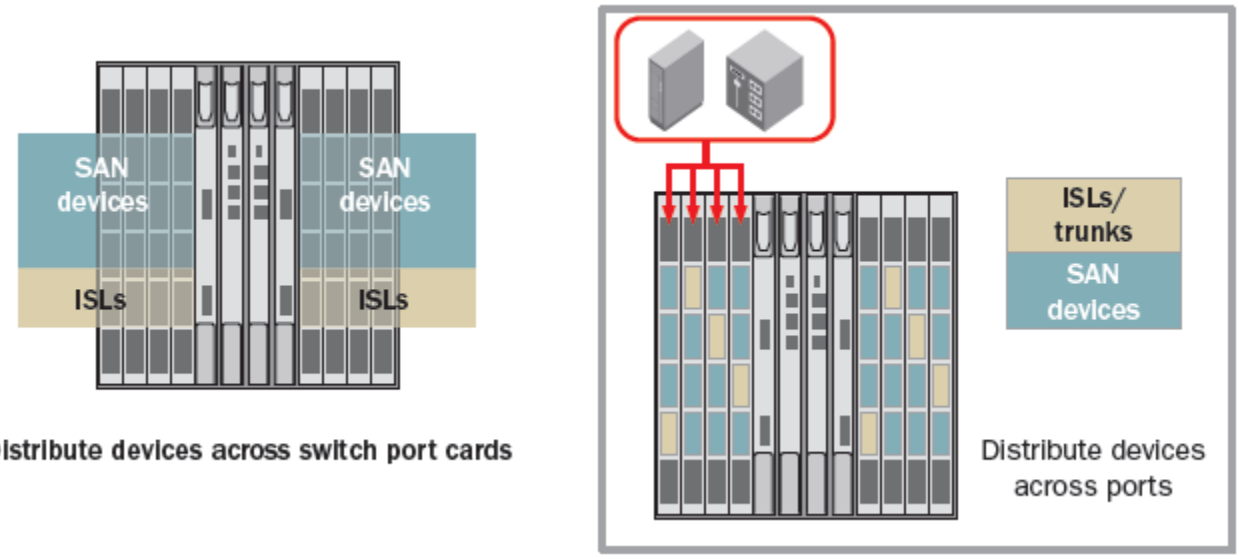
FIGURE 2 Connecting Devices Through Redundant Fabrics



In addition to redundant fabrics, redundant links should be placed on different blades, different ASICs, or at least different port groups whenever possible, as shown in Figure 3. (Refer to "Appendix C" to determine trunk groups for various port blades. For more details, see the *Brocade Fabric OS Administration Guide*.) Whatever method is used, it is important to be consistent across the fabric. For example,

do not place ISLs on lower port numbers in one chassis (as shown in the left diagram in Figure 3) and stagger them in another chassis (as shown in the right diagram in Figure 3). Doing so would be mismatched ISL placement.

FIGURE 3 Examples of Distributed ISL Placement for Redundancy



NOTE

In Figure 3, ISL trunks are placed on separate application-specific integrated circuits (ASICs) or port groups. It is important to match ISL placement between devices and across fabrics to ensure simplicity in design and assist in problem determination.

ISL Connectivity Using Q-Flex Ports on Brocade G620 Switches

In addition to 48 ports of connectivity using SFP+ transceiver connections, Brocade G620 switches with Gen 6 FC provide an additional 16 ports of 32-Gb connectivity via four Q-Flex QSFP connectors. Using MPO-to-MPO cabling, Q-Flex ports can be configured between G620 switches as ISL connections, preserving the use of the SFP+ ports for device connectivity, and increase of up to 33% more scalability over the Brocade 6510 Gen 5 switch. Connecting all four Q-Flex ports between switches will allow for an oversubscription as low as 3:1 between two G620 switches or 6:1 in a three-switch fabric. G620 Q-Flex ports can also connect to Brocade DCX 8510 directors at 16-Gbps speeds or X6 directors at 16-Gbps or 32-Gbps speeds. Check with your Brocade representative for optics support for your desired configuration. For larger fabrics, additional SFP+ ports can be configured as ISLs to configure oversubscription at the desired ratio.

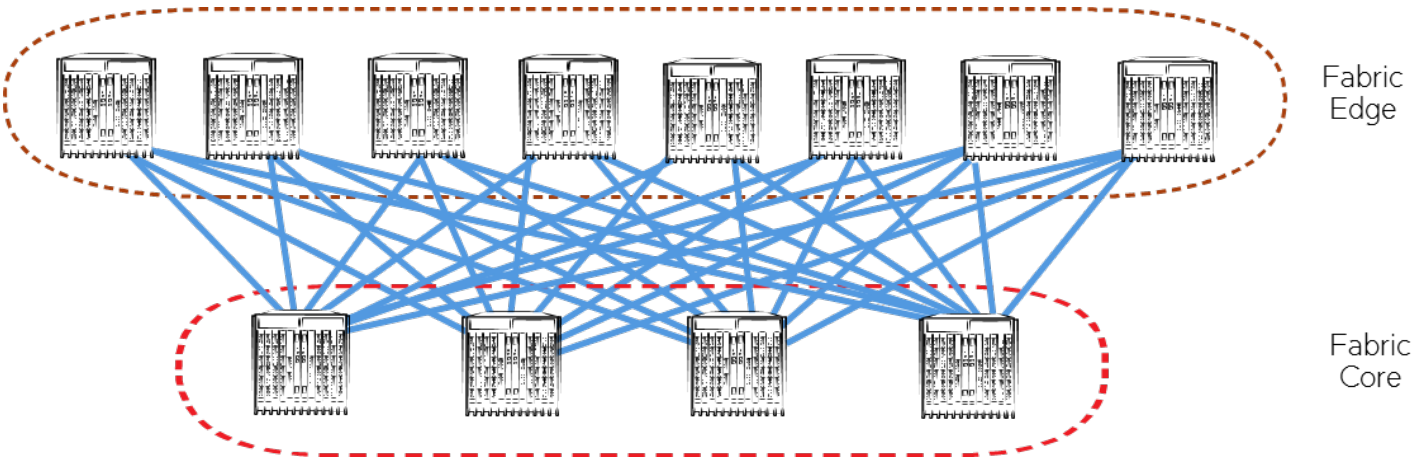
Q-Flex ports are enabled with a separate Q-Flex Ports on Demand (POD) license kit that includes four QSFP optics. Port groups for the Q-Flex ports are as follows:

- QSFP port 0 is grouped with QSFP port 1.
- QSFP port 2 is grouped with QSFP port 3.

UltraScale ICL Connectivity for Gen 5 Brocade DCX 8510-8/8510-4 and Gen 6 X6-8/X6-4 Directors

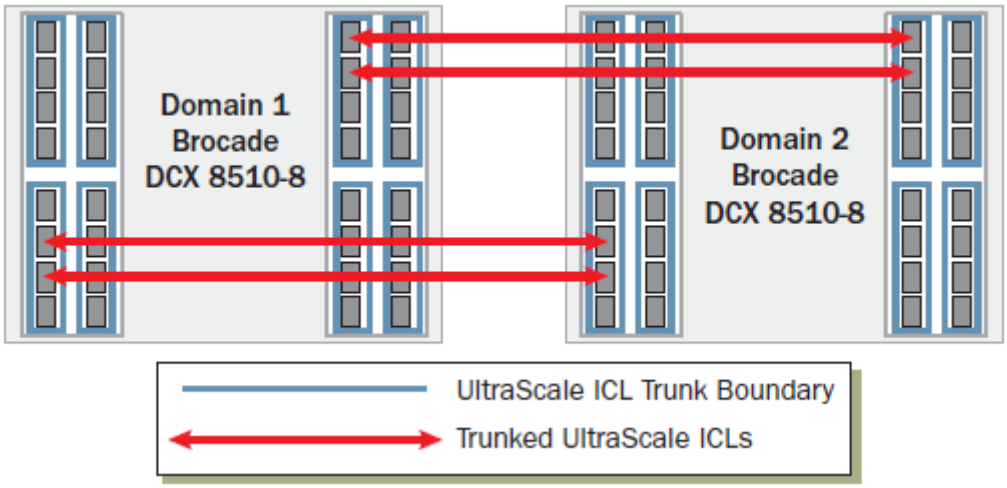
The Brocade DCX 8510 and X6 platforms use second-generation UltraScale ICL technology from Brocade with optical QSFPs. The Brocade DCX 8510-8 and X6-8 support up to 32 QSFP ports per chassis, and the Brocade DCX 8510-4 and X6-4 support up to 16 QSFP ports to help preserve director ports for connections to end devices. Each QSFP port actually has four independent links, each of which terminates on a different ASIC within the core blade. Each core blade in a Brocade DCX 8510-8 or X6-8 director has four ASICs. A pair of connections between two QSFP ports can create 128 Gbps (Gen 5) or 256 Gbps (Gen 6) of bandwidth. Figure 4 shows a core-edge design based on Gen 6 UltraScale ICLs supporting 2304 32-Gbps ports with a minimum of 512 Gbps of bandwidth between the chassis (12:1 oversubscription). As more UltraScale ICLs are added, oversubscription can be reduced to 6:1 with redundant ICL links between director chassis. Note that X6 ICL ports can also connect to DCX 8510 ICL ports at 16-Gb speeds with port speed on the X6 ICL configured to 16 Gb and use of the 16-Gb QSFP on the attached X6 port.

FIGURE 4 12-Chassis UltraScale ICL-Based Core-Edge Design



To connect multiple Brocade DCX 8510 or X6 chassis via UltraScale ICLs, a minimum of four ICL ports (two on each core blade) must be connected between each chassis pair, as shown in Figure 5. With 32 ICL ports available on the Brocade DCX 8510-8 or X6-8 (with both ICL POD licenses installed), this supports ICL connectivity with up to eight other chassis in a full-mesh topology and at least 256 Gbps of bandwidth to each connected Brocade DCX 8510 and 512 Gbps of bandwidth between connected X6-8 chassis. The dual connections on each core blade must reside within the same ICL trunk boundary on the core blades. If more than four ICL connections are required between a pair of Brocade DCX 8510/X6 chassis, additional ICL connections should be added in pairs (one on each core blade).

FIGURE 5 Minimum ICL Connections Needed Between Brocade DCX 8510 Chassis



For ICL connections between DCX 8510 or X6 chassis over 100m apart, 16-Gb optics supporting up to 2 km of distance are available. However, only 10 ICL ports on a DCX 8510 chassis can be configured at the full 2-km distance due to buffer credit limitations. Alternatively, more ports can be supported with less distance and fewer credits, and all 16 ICL ports can be configured with 11 buffers to support 1.375 km of distance. The *Brocade Fabric OS Administration Guide* provides more detailed instructions for assigning buffer credits to ICL ports to achieve the desired combination of port count and distance. On an X6 chassis, all 16 ICL ports can be configured to support the full 2 km of distance at 16-Gb speeds.

Brocade UltraScale ICL Connection Best Practices

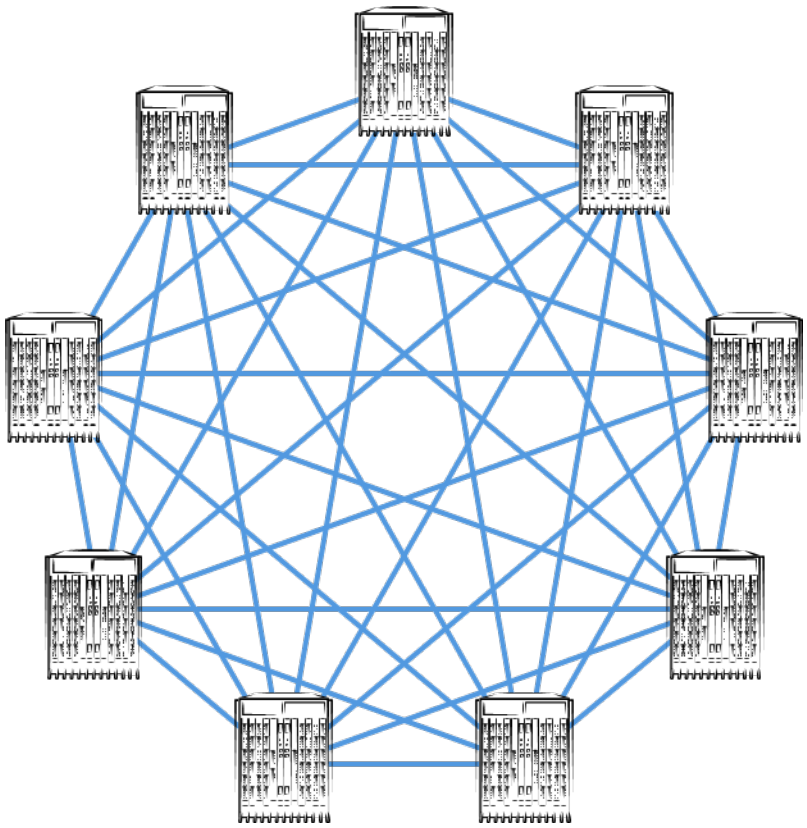
Each core blade in a chassis must be connected to each of the two core blades in the destination chassis to achieve full redundancy.

> NOTE
> For redundancy, use at least one pair of links between two core blades.

Mesh Topology

A mesh design provides a single hop between source and destination. Beginning with FOS 7.3.0x, Brocade supports a 9-chassis mesh design with up to100 meter distances using select QSFPs and OM4 fiber. In the configuration shown in Figure 6, up to 1152 16-Gbps ports are supported using UltraScale ICLs with a 12:1 oversubscription. As more UltraScale ICLs are added, oversubscription can be reduced to 3:1.

FIGURE 6 9-Chassis UltraScale ICL-Based Full-Mesh Topology



NOTE

Refer to the *Scale-out Architectures with Brocade X6 and DCX 8510 UltraScale Inter-Chassis Links* feature brief for details. UltraScale ICL connections are considered a "hop of no concern" in a FICON fabric.

When using core-edge or edge-core-edge SAN design methodologies, edge switches should connect to at least two core switches with trunks of at least two ISLs each. Each of those trunks should be attached to a different blade/port group. In order to be completely redundant, there would be a completely mirrored second fabric and devices must be connected to both fabrics utilizing MPIO.

Recommendations for switch ISL/UltraScale ICL connectivity are:

- There should be at least two core switches.
- Every edge switch should have at least two trunks to each core switch.
- Select small trunk groups (keep trunks to two ISLs) unless you anticipate very high traffic volumes. This ensures that you can lose a trunk member without losing ISL connectivity.
- Place redundant links on separate blades.
- Trunks should be in a port group (ports within an ASIC boundary).
- Allow no more than 30m in cable difference for optimal performance for ISL trunks.
- Use the same cable length for all UltraScale ICL connections.
- Use either ISL or UltraScale ICL connectivity into the same domain. Mixing the two types of connections is not supported.
- Use the same type of optics on both sides of the trunks: Short Wavelength (SWL), Long Wavelength (LWL), or Extended Long Wavelength (ELWL).

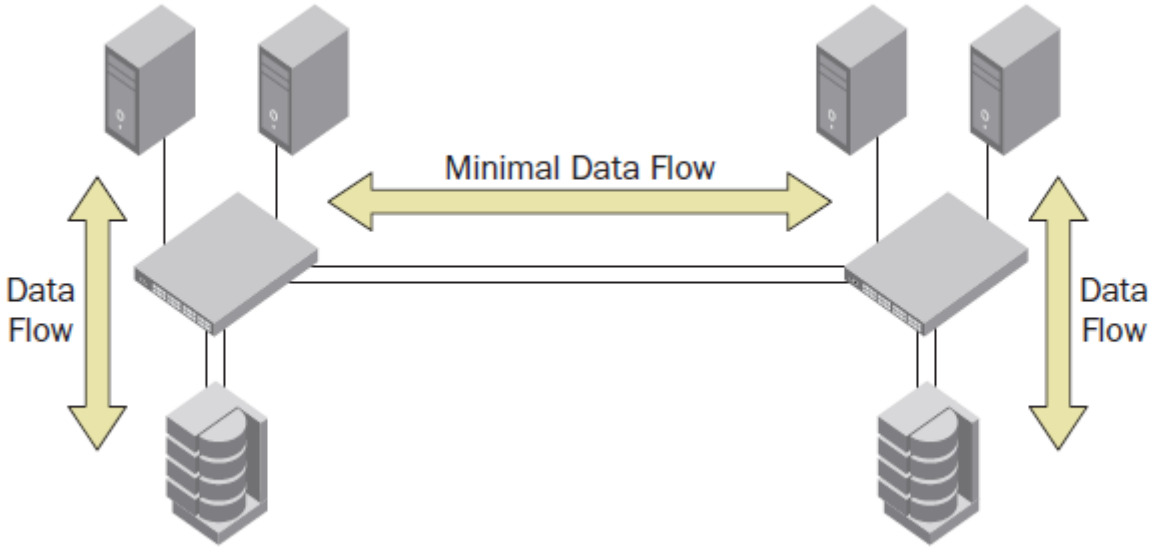
Device Placement

Device placement is a balance between traffic isolation, scalability, manageability, and serviceability. With the growth of virtualization and multinode clustering on the UNIX platform, frame congestion can become a serious concern in the fabric if there are interoperability issues with the end devices.

Traffic Locality

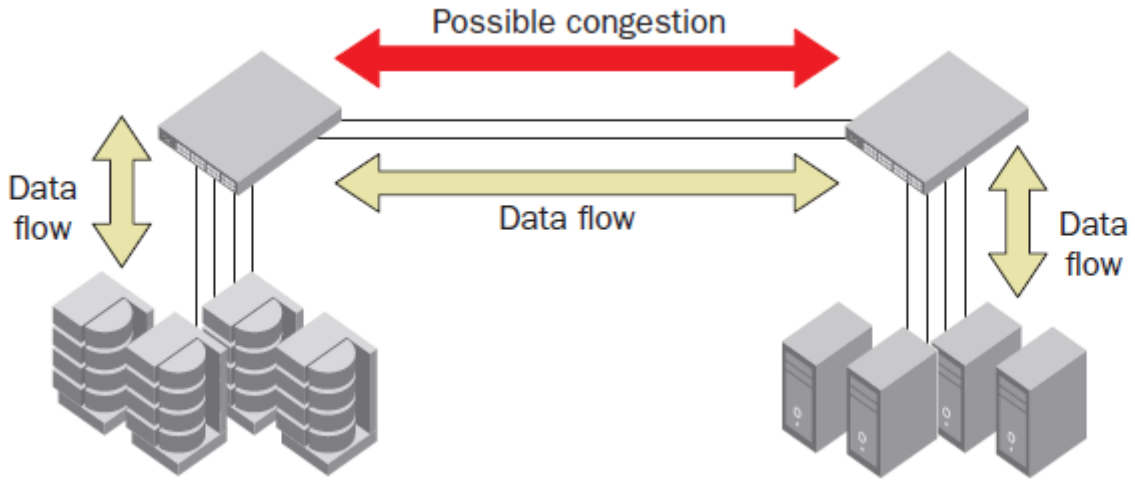
Designing device connectivity depends a great deal on the expected data flow between devices. For simplicity, communicating hosts and targets can be attached to the same switch.

FIGURE 7 Hosts and Targets Attached to the Same Switch to Maximize Locality of Data Flow



However, this approach does not scale well. Given the high-speed, low-latency nature of Fibre Channel, attaching these host-target pairs on different switches does not mean that performance is adversely impacted for common workloads. Though traffic congestion is possible, it can be mitigated with proper provisioning of ISLs/UltraScale ICLs. With current generation switches, locality is not required for performance or to reduce latencies. For mission-critical applications that depend on extremely fast response times, architects may want to localize the traffic when using flash storage or in very exceptional cases, particularly if the number of ISLs available is restricted or there is a concern for resiliency in a multihop environment.

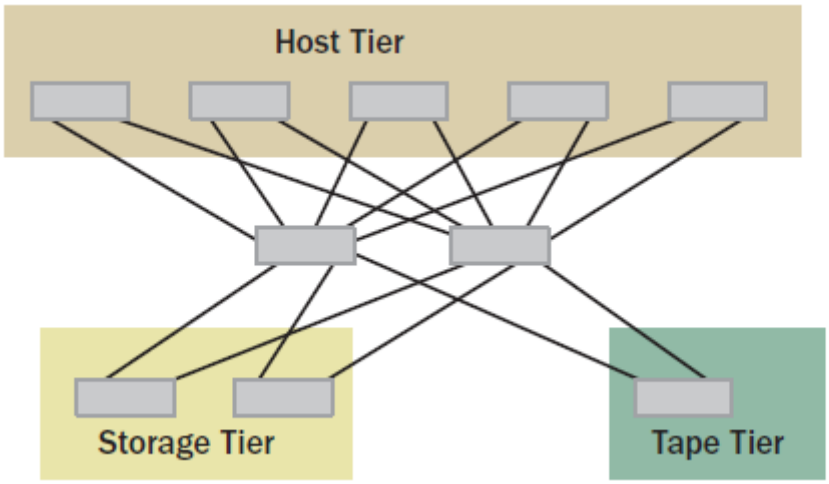
FIGURE 8 Hosts and Targets Attached to Different Switches for Ease of Management and Expansion



One common scheme for scaling a core-edge topology is dividing the edge switches into a storage tier and a host/initiator tier. This approach lends itself to ease of management as well as ease of expansion. In addition, host and storage devices generally have different performance requirements, cost structures, and other factors that can be readily accommodated by placing initiators and targets in different tiers.

The following topology provides a clearer distinction between the functional tiers.

FIGURE 9 Device-Type-Based Edge-Core-Edge Tiered Topology



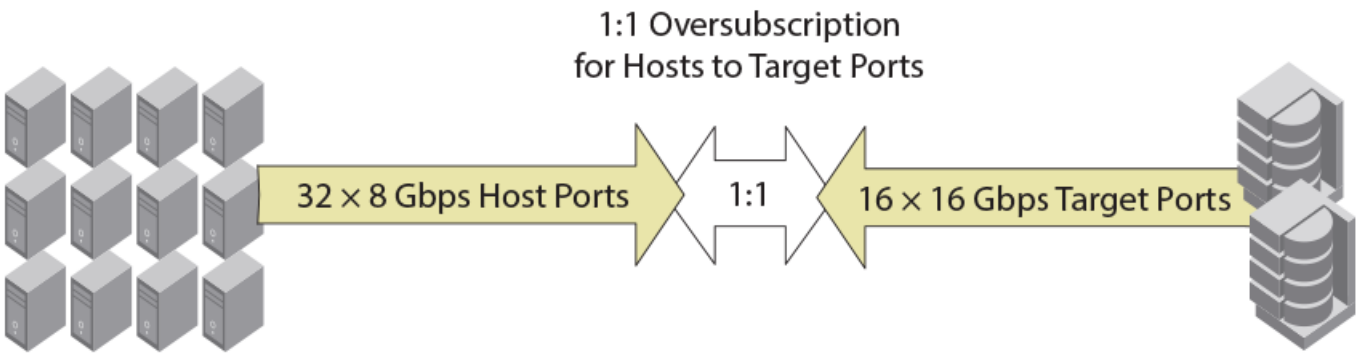
Data Flow Considerations

Fan-In Ratios and Oversubscription

A critical aspect of data flow is the *fan-in-ratio* or *oversubscription*, in terms of source ports to target ports and devices to ISLs. This is also referred to as the *fan-out-ratio* if viewed from the storage array perspective. The ratio is the number of device ports that share a single port, whether ISL, UltraScale ICL, or target. This is always expressed from the single entity point of view, such as 7:1 for seven hosts utilizing a single ISL or storage port.

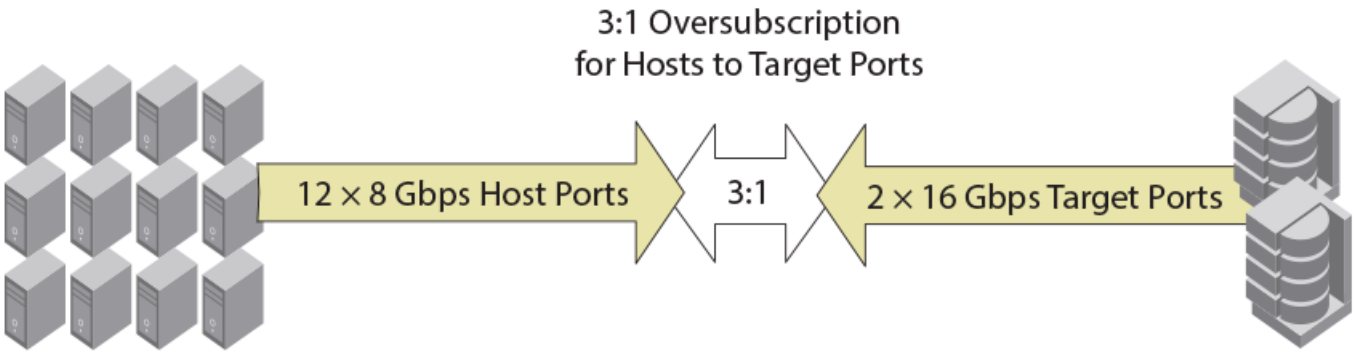
What is the optimum number of hosts that should connect per to a storage port? This seems like a fairly simple question. However, once you take into consideration clustered hosts, VMs, workload characteristics, and number of Logical Unit Numbers (LUNs) (storage) per server, the situation can quickly become much more complex. Determining how many hosts to connect to a particular storage port can be narrowed down to three considerations: port queue depth, I/O per second (IOPS), and throughput. Of these three, throughput is the only network component. Thus, a simple calculation is to add up the expected peak bandwidth usage for each host accessing the storage port. The total should not exceed the supported bandwidth of the target port, as shown in Figure 10.

FIGURE 10 Example of One-to-One Oversubscription



In practice, however, it is highly unlikely that all hosts perform at their maximum level at any one time. With a traditional application-per-server deployment, the Host Bus Adapter (HBA) bandwidth is overprovisioned. However, with virtual servers (KVM, Xen, Hyper-V, proprietary UNIX OSs, and VMware) the game can change radically. Network oversubscription is built into the virtual server concept. To the extent that servers leverage virtualization technologies, you should reduce network-based oversubscription proportionally. It may therefore be prudent to oversubscribe ports to ensure a balance between cost and performance. An example of 3-to-1 oversubscription is shown in Figure 11.

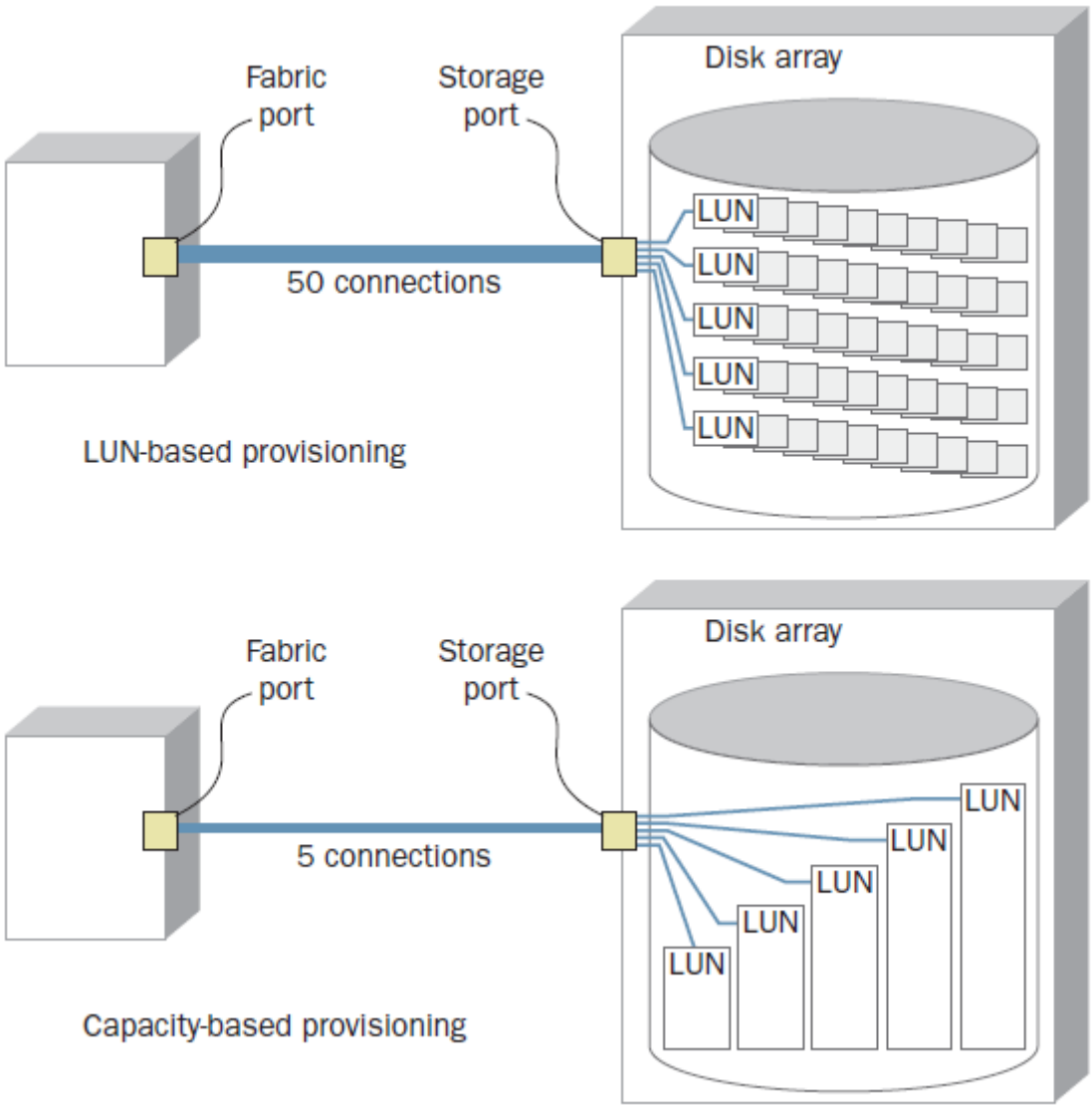
FIGURE 11 Example of Three-to-One Oversubscription



Another method is to assign host ports to storage ports based on the I/O capacity requirements of the host servers. The intended result is a small number of high-capacity servers or a larger number of low-capacity virtual servers assigned to each storage port, thus distributing the load across multiple storage ports.

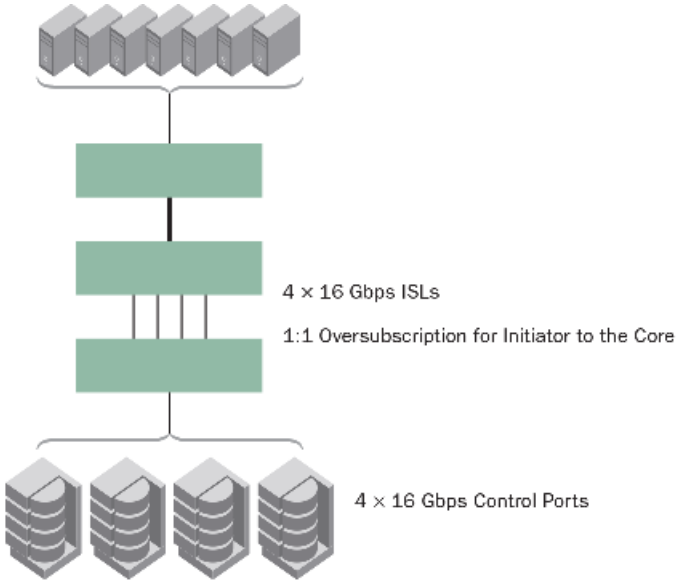
Figure 12 shows the impact of the two different LUN provisioning strategies described above. Notice that there is a huge difference between the fan-in to the storage port, based on the number of LUNs provisioned behind the port.

FIGURE 12 Two Different LUN Provisioning Strategies



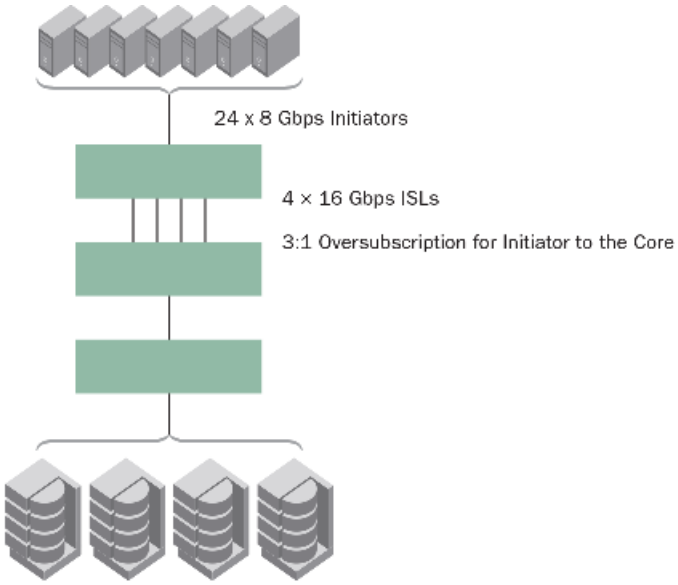
Regardless of the method used to determine the fan-in/fan-out ratios, port monitoring should be used to determine actual utilization and what adjustments, if any, should be made. In addition, ongoing monitoring provides useful heuristic data for effective expansion and efficient assignment of existing storage ports. For determining the device-to-ISL fan-in ratio, a simple calculation method works best: the storage port should not be oversubscribed into the core (for example, a 16-Gbps storage port should have a 16-Gbps pipe into the core).

FIGURE 13 One-to-One Oversubscription for Targets into the Core



The realized oversubscription ratio of host-to-ISL should be roughly the same as the host-to-target ratio, taking into account the bandwidth (that is, if there are four hosts accessing a single 8-Gbps storage port, then those four hosts should have an 8-Gbps pipe into the core.) In other words, match device utilization and speeds with ISL speeds, as shown in Figure 14.

FIGURE 14 Three-to-One Oversubscription for Hosts Coming into the Core



Recommendations for avoiding frame congestion (when the number of frames is the issue rather than bandwidth utilization) include:

- Use more and smaller trunks.
- Storage ports should follow the array-vendor-suggested fan-in ratio for ISLs into the core. Follow vendor-suggested recommendations when implementing a large number of low-capacity LUNs.

- Bandwidth through the core (path from source/host to destination/target) should exceed storage requirements.
- Host-to-core subscription ratios should be based on both the application needs and the importance of the application.
- Plan for workload peaks, not average usage.

For mission-critical applications, the ratio should exceed peak load enough such that path failures do not adversely impact the application. In other words, have enough extra bandwidth to avoid congestion if a link fails.

Resource Contention

While there may sometimes be a temptation to maximize the utility of a server by running a high number of virtualized hosts on a physical server, this can also create delays in responses that may at first be interpreted as network congestion. Increasing the number of workloads sharing common storage resources means that a larger number will be performing read and write operations simultaneously, increasing the probability of a resource availability conflict. A SCSI reserve command issued to prevent another initiator from accessing the data during a read or write operation in process will result in a blocked request, requiring that the initiator retry the reserve following a predetermined interval. As this happens with increased frequency due to contention among applications, increasingly long delays in accessing target devices will slow down application performance. To avoid this situation, the number of virtualized hosts per server may need to be reduced or the response times decreased by deploying faster flash-enhanced storage.

Congestion in the Fabric

Congestion is a major source of poor performance in a fabric. Sufficiently impeded traffic translates directly into poor application performance.

There are two major types of congestion: traffic-based and frame-based. Traffic-based congestion occurs when link throughput capacity is reached or exceeded and the link is no longer able to pass more frames. Frame-based congestion occurs when a link has run out of buffer credits and is waiting for buffers to free up to continue transmitting frames.

Traffic Versus Frame Congestion

At higher link throughput speeds, the emphasis on fabric and application performance shifts from traffic-level issues to frame congestion. With current link speeds and Brocade features, such as Brocade ISL Trunking and UltraScale ICLs, it is very difficult to consistently saturate a link. Most infrastructures today rarely see even two-member trunks reaching a sustained 100-percent utilization. Frame congestion can occur when the buffers available on a Fibre Channel port are not sufficient to support the number of frames the connected devices wish to transmit. This may be the case if there are a high number of VMs running on a physical server generating command frames, even while the server's CPU utilization, IOPS, and data throughput levels stay relatively low. The deployment of flash-enhanced storage can further amplify this effect by accelerating the response times from storage and increasing the rate of exchanges between VMs and storage. This situation can result in credit starvation backing up across the fabric. This condition is called back pressure, and it can cause severe performance problems.

The sources and mitigation for traffic congestion are well known and are discussed at length in other parts of this document. The remainder of this section focuses on the sources and mitigation of frame-based congestion.

Sources of Frame Congestion

Frame congestion is primarily caused by latencies somewhere in the SAN—usually hosts and occasionally storage devices. These latencies cause frames to be held in ASICs and reduce the number of buffer credits available to all flows traversing that ASIC. The congestion backs up from the source of the latency to the other side of the connection and starts clogging up the fabric. Back pressure can be created from the original source of the latency to the other side and all the way back (through other possible paths across the fabric) to the original source again. Once this situation arises, the fabric is very vulnerable to severe performance problems.

Sources of high latencies include:

- Storage devices that are not optimized or where performance has deteriorated over time.
- Distance links where the number of allocated buffers has been miscalculated or where the average frame sizes of the flows traversing the links have changed over time.
- Hosts where the application performance has deteriorated to the point that the host can no longer respond to incoming frames in a sufficiently timely manner.
- Incorrectly configured HBAs.
- Massive oversubscription on target ports and ISLs.
- Tape devices.

Other contributors to frame congestion include behaviors where short data frames or high levels of command frames are generated in large numbers, such as:

- Extremely high VM-to-host consolidation ratios generating a high level of SCSI RESERVE/RELEASE commands.
- Clustering software that verifies the integrity of attached storage.
- Clustering software that uses control techniques such as SCSI RESERVE/RELEASE to serialize access to shared file systems.
- Host-based mirroring software that routinely sends SCSI control frames for mirror integrity checks.
- Virtualized environments, both workload and storage, that use in-band Fibre Channel for other control purposes.

Mitigating Congestion

Frame congestion cannot be corrected in the fabric. Devices that exhibit high latencies, whether servers or storage arrays, must be examined and the source of poor performance eliminated. Since these are the major sources of frame congestion, eliminating them typically addresses the vast majority of cases of frame congestion in fabrics.

Brocade has introduced a new control mechanism in an attempt to minimize the effect of some latencies in the fabric. Edge Hold Time (EHT) is a new timeout value that can cause some blocked frames to be discarded earlier by an ASIC in an edge switch where the devices typically are provisioned. EHT is available from Brocade FOS 6.3.1b and later and allows for frame drops for shorter timeout intervals than the 500 milliseconds typically defined in the Fibre Channel standard. EHT accepts values from 500milliseconds all the way down to 80 milliseconds. The EHT default setting for F_Ports is 220 milliseconds, and the default EHT setting for E_Ports is 500 milliseconds. Note that an I/O retry is required for each of the dropped frames, so this solution does not completely address high-latency device issues.

EHT applies to all the F_Ports on a switch and all the E_Ports that share the same ASIC as F_Ports. It is a good practice to place servers and ISLs on different ASICs since the EHT value applies to the entire ASIC, and it is recommended that the ISL EHT stay at 500 ms.

> NOTE
> EHT applies to the switch and is activated on any ASIC that contains an F_Port. For example, if EHT is set to 250 ms and the ASIC contains F_Ports and E_Ports, the timeout value for all the ports is 250 ms.

Behaviors that generate frequent large numbers of short frames cannot typically be changed—they are part of the standard behavior of some fabric-based applications or products. As long as the major latencies are controlled, fabrics tolerate this behavior well.

IO Insight

IO Insight, supported on Gen 6 Fibre Channel switches and directors with FOS 8.0.1 and later, adds flow-level device latency and IO metrics to monitor storage performance. IO Insight metrics can be included in MAPS policies, allowing admin notification of performance degradation. This enables the ability to baseline performance from MAPS to alert the administrator of developing congestion issues that impact IO and storage response latency times.

Port Monitoring

One side effect of frame congestion can be very large buffer credit zero counts on ISLs and F_Ports. This is not necessarily a concern, unless counts increase rapidly in a very short period of time. Brocade's Fabric Performance Impact (FPI) monitoring feature (previously referred to as Bottleneck Detection in Fabric OS 7.x and earlier) can help to more accurately assess the impact of a lack of buffer credits. FPI monitoring does not require a license, but it is disabled by default. The recommended best practice is to enable FPI monitoring on all switches in the fabric and to leave it on to continuously gather statistics.

FPI monitoring, when applied to F_Ports (devices), detects high-latency devices and provides notification on the nature and duration of the latency. This is a huge advantage to the storage administrator, because there is now a centralized facility that can potentially detect storage latencies while they are still intermittent.

FPI monitoring can also serve as a confirmation to host information when storage latencies are suspected in poor host performance. The reverse (eliminating the storage as the source of poor performance) is also true. FPI monitoring can also be applied to ISLs (E_Ports) and will highlight issues on those links.

The sampling interval and number of notifications are configurable, as well as the alerting mechanisms. With Brocade FOS 6.4, notifications can be configured for Reliability, Availability, and Serviceability (RAS) log and Simple Network Management Protocol (SNMP). Brocade Network Advisor can be configured to automatically monitor and detect bottlenecks in the fabric. You can easily pinpoint areas of network congestion with visual connectivity maps and product trees.

Design Guidelines

Edge Hold Time (EHT)

- EHT is recommended primarily for initiators (hosts). Extreme care must be taken if you choose to apply EHT to target ports because a target port can service a large number of initiators. A large number of frame drops on a target port can potentially affect a very large number of running applications. Those applications may be more tolerant to poor performance than to a large number of I/O retries.
- There is no calculation for determining the best value for EHT. EHT can be set from 100 to 500 milliseconds. The lower the value, the more frame drops you can expect. Brocade recommends that you take a value of approximately 250 milliseconds and observe the results.
- EHT is less effective when initiators and targets share the same switch, because the timeout value will apply equally to both storage and host ports.
- EHT applies to the entire ASIC. If possible, ISLs should be placed on a different ASIC than the servers.

Fabric Performance Impact Monitoring

A phased approach to deploying FPI monitoring works best. Given the potential for a large number of alerts early on in the process, Brocade recommends starting with a limited number of storage ports and incrementally increasing the number of ports monitored over time. Once the storage latencies are dealt with, you should move on to the host (initiator) ports and ISLs. You can increase the number of ports monitored once the chronic latency problems have been resolved.

Slow Drain Device Quarantine (SDDQ)

SDDQ is an action associated with FPI monitoring rules supported in FOS 7.4 and beyond. SDDQ is triggered when FPI detects an F_Port in either the IO_PERF_IMPACT state or the IO_FRAME_LOSS state. SDDQ sends the PID associated with the slow draining device to all switches in a fabric, instructing them to move the traffic destined for the slow draining device into a low-priority virtual channel (VC) and to free up buffer credits on the regular medium-priority VC for traffic destined to other devices. This effectively removes the impact of the slow drain device from the fabric performance without disruption to the workload traffic by isolating the slow drain device in a quarantine state while remaining online. Additional detail on SDDQ is available in the *SAN Fabric Resiliency and Administration Best Practices*.

Availability and Health Monitoring

With Brocade Fabric Vision, IT organizations can monitor fabrics on both a real-time and historical basis. This allows users to address performance issues proactively and rapidly diagnose the underlying causes, then quickly resolve the issues before the SAN becomes the bottleneck for critical applications. An overview of the major components is provided below. A complete guide to health monitoring is beyond the scope of this document. For more detailed information, refer to the *SAN Fabric Resiliency and Administration Best Practices*, the *Brocade Fabric OS Troubleshooting and Diagnostics Guide*, the appropriate *Brocade SAN Health* and *MAPS* guides, and the *Brocade Network Advisor SAN User Manual*.

Health and Threshold Monitoring

Brocade Monitoring and Alerting Policy Suite (MAPS)

MAPS is a health and threshold monitoring tool that allows you to constantly monitor all directors and switches for developing fault situations and that automatically alerts you to problems long before they become costly outages. MAPS was introduced as an upgrade to Fabric Watch starting with Brocade FOS 7.2.0.

MAPS tracks a variety of SAN fabric elements and events. Monitoring fabric-wide events, ports, bit errors, and environmental parameters enables early fault detection and isolation as well as performance measuring. You can configure fabric elements and alert thresholds on an individual port or on a port group, and you can also easily integrate MAPS with enterprise system management solutions.

MAPS provides predefined monitoring thresholds based on the administrative philosophy you select: conservative, moderate, or aggressive. Thresholds can also be customized on an individual level. MAPS can provide notifications before problems arise, such as reporting when network traffic through a port is approaching the bandwidth limit. This information enables the fabric administrator to perform pre-emptive network maintenance, such as trunking or zoning, and avoid potential network failures.

MAPS lets you define how often to measure each switch and fabric element and specify notification thresholds. Whenever fabric elements exceed these thresholds, MAPS automatically takes action, such as administrative notification using e-mail messages, SNMP traps, and log entries, or automated actions, such as port fencing that prevents a slow drain device from impacting the rest of the network.

Brocade MAPS Recommendations

Brocade MAPS is an optional feature that provides monitoring of various switch elements. Brocade MAPS monitors ports based on the port type, for example, F_Port and E_Port classes, without distinguishing between initiators and targets. Since the monitoring thresholds and desired actions are generally different for initiators, targets and ISL ports, it is recommended that these ports be combined into a common group so that Brocade MAPS rules can be applied using common thresholds and actions.

> NOTE
> For additional details, see the *Brocade Monitoring and Alerting Policy Suite Configuration Guide* for your version of Fabric OS.

RAS Log

RAS log is the Brocade FOS error message log. Messages are organized by Brocade FOS component, and each one has a unique identifier as well as severity, source and platform information and a text message.

RAS log is available from each switch and director via the "errdump" command. RAS log messages can be forwarded to a syslog server for centralized collection and analysis or viewed within Brocade Network Advisor via the Master Log.

Audit Log

The Audit log is a collection of information created when specific events are identified on a Brocade platform. The log can be dumped via the **auditdump** command, and audit data can also be forwarded to a syslog server for centralized collection.

Information is collected on many different events associated with zoning, security, trunking, FCIP, FICON, and others. Each release of Brocade FOS provides more audit information.

Brocade SAN Health

Brocade SAN Health provides snapshots of fabrics showing information such as switch models and firmware levels, connected device information, performance information, zone analysis, and ISL fan-in ratios.

Design Guidelines

Brocade strongly recommends implementing some form of monitoring of each switch. Often issues start out relatively benign and gradually degrade into more serious problems. Monitoring the logs for warning, critical and error severity messages will go a long way in avoiding many problems.

- Plan for a centralized collection of RAS log and perhaps Audit log via syslog. You can optionally filter these messages relatively easily through some simple scripting programs or perform advanced correlation using an event management engine.
- Brocade platforms are capable of generating SNMP traps for most error conditions. Consider implementing some sort of alerting mechanism via SNMP or email notifications.

Monitoring and Notifications

Error logs should be looked at regularly. Many end users use combinations of syslog and SNMP in combination with MAPS and the logs to maintain a very close eye on the health of their fabrics. You can troubleshoot network-related issues such as syslog events and SNMP traps through the Event Manager within Brocade Network Advisor.

Brocade Network Advisor also collects, monitors, and graphically displays real-time and historical performance data, so you can proactively manage your SAN network.

Brocade Professional Services can be engaged to assist with implementing these and other advanced features.

Available Paths

It is recommended that the SAN be deployed with at least two paths between source and destination. Often, there are more than two paths and the utilization of these paths is dependent on the routing policy configuration.

- Port-Based Routing (PBR) assigns a single route between source port and destination port. Although this minimizes disruption caused by changes in the fabric, it represents a less efficient use of available bandwidth.
- Exchange-Based Routing (EBR), Brocade's default routing protocol, uses all available (equal-cost) routes between source port and destination port, with individual exchanges assigned a single route. Although it represents a more efficient use of available bandwidth, it is potentially more disruptive unless Dynamic Load Sharing (DLS) is implemented with the lossless feature.

The number of available paths can be adjusted by changing the size of trunk groups. While a trunk can have two to eight members, it may prove beneficial to have more trunks with fewer members. Spreading ISLs across multiple trunks uses more of the fabric bandwidth by spreading traffic across more paths. Keep at least two members in each trunk to avoid unnecessary frame loss if a trunk member fails.

Latencies

There are many causes of latencies:

- Slow devices such as disk-based storage arrays
- Oversubscribed devices
- Long-distance links
- Servers that are not responding rapidly enough to I/O requests they have previously made
- Degraded cables and SFPs causing many retried I/Os

There is very little that can be done in the fabric to accommodate end-device latencies: they typically must be addressed through other means. Array latencies can be dealt with by array or LUN reconfiguration or data migration. Long-distance problems might require more long-distance bandwidth or reconfiguration of the distance setting on the switch. Applications might require tuning to improve their performance, and failing links and SFPs must be identified and replaced. At best, the fabric can help identify the source of the problem. Brocade has been working hard to enhance RAS features in Brocade FOS in line with changing customer requirements. Some of these features are described briefly in the sections that follow.

Misbehaving Devices

All fabrics, regardless of the equipment vendor, are vulnerable to the effects of badly behaving devices, that is, a server or storage device that for some reason stops functioning or starts flooding the fabric with data or control frames. The effects of such behavior can be very severe, causing other applications to failover or even stop completely. There is nothing that the fabric can do to anticipate this behavior. Brocade has implemented several new features that are designed to rapidly detect a misbehaving device and isolate it from the rest of the fabric.

Isolating a single server has much less impact on applications than disabling a storage array port. Typically, a storage port services many applications, and the loss of that storage can severely impact all the applications connected to it. One of the advantages of a core-edge design is that it is very simple to isolate servers from their storage and ensure that any action applied to host port for a given behavior can be very different than the action applied to a storage port for the same behavior.

Detailed guidance on monitoring for misbehaving devices and configuring fabrics to respond to developing issues can be found in the *SAN Fabric Resiliency and Administration Best Practices* guide.

Design Guidelines

- Transaction-based systems: Make sure that ISL/UltraScale ICLs traversed by these systems to access their storage do not contain too many flows. The fan-in from the hosts/initiators should not exceed a ratio of 10 to 1. Also ensure that there is as little interference from other applications as possible, to ensure that latencies and congestion from other sources do not affect the overall performance of the applications.
- I/O-intensive applications: Bandwidth is the most common constraint for these systems. Modern fabrics typically provide more bandwidth than is needed except for the most powerful hosts. Take care to ensure that these high performing systems do not interfere with other applications, particularly if utilization spikes at specific times or if batch runs are scheduled. When in doubt, add more paths (ISLs or trunks) through the fabric.
- Clusters: Clusters often have behavioral side effects that must be considered. This is particularly true during storage provisioning. It is possible, for example, for a cluster to inundate the fabric and storage arrays with LUN status queried and other short frame requests. This behavior can cause frame congestion in the fabric and can stress the control processors of the arrays. Make sure that you spread out the LUNs accessed by the hosts in the cluster across as many arrays as possible.
- Congestion: Traffic congestion (total link capacity regularly consumed) is remedied by adding more links or more members to a trunk. Frame congestion is typically addressed by dealing with the nodes causing the congestion.

- Misbehaving devices: As stated earlier, there is little that can be done in the fabric to mitigate the effects of a badly behaving device other than to remove it from the fabric. Brocade supports a Brocade FOS capability called Port Fencing, which is designed to isolate rogue devices from the network. Port Fencing works with Brocade MAPS to disable a port when a specific threshold has been reached. Port Fencing, in combination with FPI monitoring, can be used for detecting and isolating high-latency devices from impacting the rest of the devices in the fabric.
- Initiator and targets: If possible, isolate host and storage ports on separate switches for much greater control over the types of controls that you can apply to misbehaving and high-latency devices. The effect on applications is typically much less severe if a host is disabled versus disabling a storage port, which may be servicing flows from many servers.

Monitoring

- Brocade Network Advisor is a powerful proactive monitoring and management tool that offers customizable health and performance dashboards to provide all critical information in a single screen. With Brocade Network Advisor, you can manage your entire network infrastructure.
- Use Brocade MAPS to monitor switch and director resource consumption, port utilization, and port errors. Brocade MAPS is also used to trigger Port Fencing.
- IO Insight metrics can be included in MAPS policies for Gen 6 devices, allowing admin notification of performance degradation. For workloads that depend on low latency response times from flash-enhanced storage, this enables the ability to baseline performance to ensure consistency of performance and fast diagnosis of IO-related operational issues.
- FPI monitoring is very useful in detecting latencies in devices and across links. It can help clarify whether high buffer credit zero counts are actually a problem. Once device latencies have been addressed, it is often useful to apply other controls, such as Port Fencing or Slow Drain Device Quarantining to improve the resiliency of the fabric by isolating new misbehaving devices or future high latencies.
- Brocade SAN Health is a free utility that provides a lot of useful information to the storage or SAN administrator. SAN Health provides visibility into ISL fan-in ratios, creates Visio diagrams of fabrics, verifies firmware levels on switches, and can provide a host of other valuable information.

IOPS and VMs

Another method for determining bandwidth and/or oversubscription is to use the IOPS between host and storage devices. This method is greatly simplified with the use of IO Insight on a Gen 6 switch; otherwise estimation is required of the typical number of IOPS and I/O size to calculate both average and estimated peak loads in terms of Megabytes per second (MB/sec). Next, look at the paths through the network for these I/Os, along with I/Os from other devices using the same network paths. Then use these data points to calculate bandwidth utilization and/or oversubscription ratios for devices and ISLs.

The use of VMs and the mobility of these VMs can make such IOPS calculations a challenge, as loads can shift when VMs move. Thus, the administrator needs to be aware of the potential VM loads on each physical server and their associated application loads for VMs.

While these calculations can certainly produce an accurate picture of bandwidth requirements for the storage network, they can be complicated even in a small network topology. This is why the simple approach discussed above is generally recommended.

Routed Topologies—MetaSANs

The FC-FC routing service enables Fibre Channel SANs to share devices between two or more fabrics without merging those fabrics. The advantages for a routed topology are a reduced number of switch domains and zones for management within an edge fabric, fault isolation to a smaller edge fabric, interoperability with legacy fabrics, and increased security. In general, edge fabrics with Fibre Channel Routing (FCR) topologies follow the same basic best practice design guidelines as traditional fabrics, core-edge architectures for example. The FCR feature can be used for local or between fabrics across dark fiber or Wide-Area Networks (WANs) using FCIP.

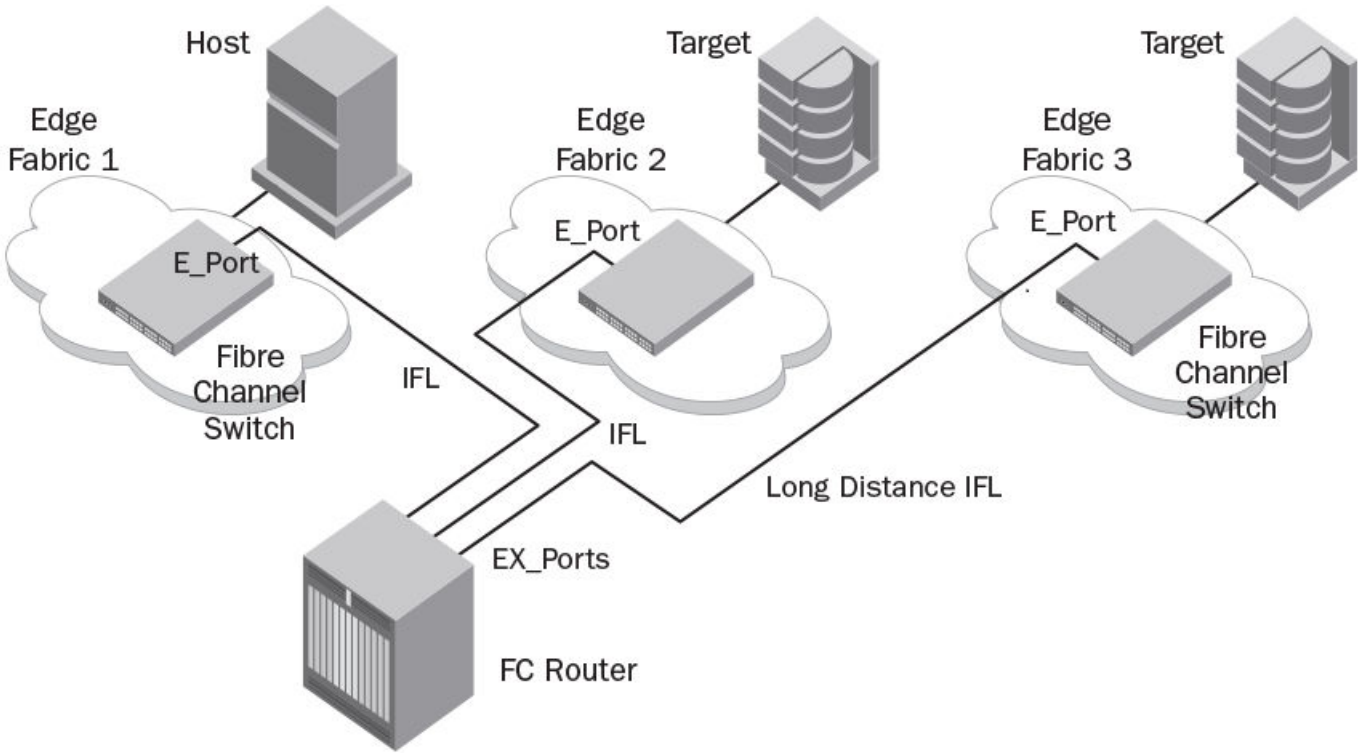
> NOTE
> Refer to *Brocade SAN Scalability Guidelines* for FCR scalability limits.

The primary considerations for using FCR are as follows:

- A limited number of LUNs shared between fabrics
- A limited number of servers that need to share LUNs between fabrics
- Share archiving devices like tape libraries
- The migration of legacy Brocade FOS fabrics to current Brocade FOS-based platforms
- OEM support
- Security separation in a managed services environment

There should be redundancy at the fabric, switch, and Inter-Fabric Link (IFL) levels (see Figure 15–Figure 17). A routed SAN, or MetaSAN, environment consists of multiple edge fabrics interconnected through one or more backbone fabrics. Multiple backbone fabrics are in parallel and belong to only the A or B fabric, not both. A core-edge topology can be used at both the backbone level and at the edge fabric level, such that edge fabrics and backbone fabrics are both deployed in a core-edge fashion.

FIGURE 15 Typical MetaSAN Topology



The implementation and configuration of ISLs (and IFLs, in the case of FCR) should be based on the expected data flow between the switches and/or fabrics in question and the desired level of redundancy between edge switches and across the routed SAN. Below are some architectural examples of MetaSAN topologies.

FIGURE 16 Example of a Simple Core-To-Core Attached Backbone in a Redundant Routed Fabric Topology

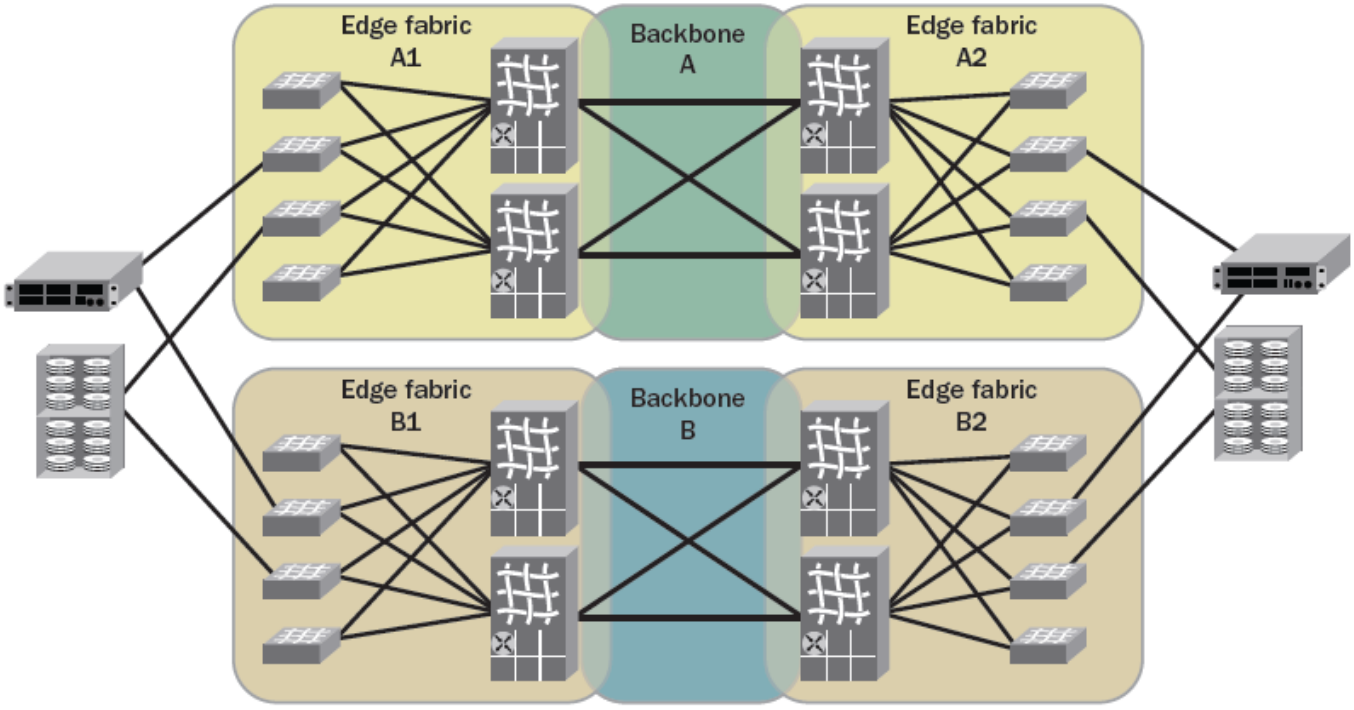
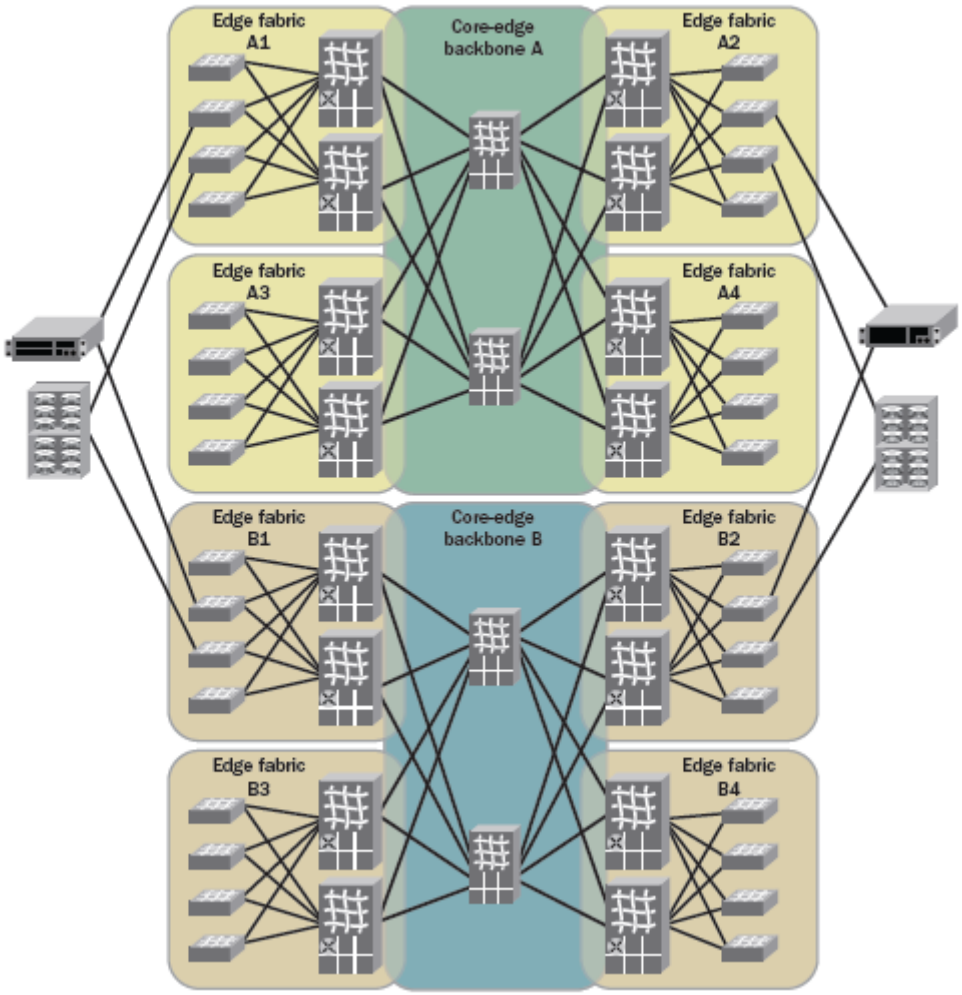


FIGURE 17 Example of an Edge-Core-Edge Backbone in a Redundant Routed Fabric Topology



Backbone Considerations

There are many factors to consider when designing backbone fabrics. As mentioned above, the general SAN topology recommendations are applicable to backbone fabrics. There should be redundant fabrics, switches, and paths between the end-points (source and destination). Consider the following factors when identifying the best switch platforms and backbone topology, including switch interconnections:

- The number of edge fabrics impacts the backbone topology, as well as the manner in which edge fabrics are attached to the backbone. Brocade 8, 16 and 32 Gbps platforms can support FCR functionality on all standard FC ports, and they provide a much more flexible solution when compared to legacy FCR platforms.
- Composition of edge fabrics:
  - Legacy switches: The presence of legacy Brocade or McDATA switches anywhere in the SAN environment impacts the features that are supported and, depending on the platform and firmware version, may have other impacts as well.
  - Advanced SAN application/features: If you are considering implementing advanced SAN applications and/or features, the key factor is support (or compatibility) of the application on the SAN switch platforms being considered, as well as the ability to support those features across FCR.

- Projected inter-fabric traffic patterns:
  - Quantity (bandwidth utilization): You should provision a sufficient number of IFLs between each edge and the backbone to accommodate the projected traffic (at peak load) to and from each edge fabric. In addition, you should provision enough ISLs within the backbone to accommodate the projected traffic (at peak load) that will traverse the backbone.
  - Bursty versus continuous traffic: Bursty traffic is more forgiving than continuous traffic, since it generally handles temporary spikes in latency (unavailability of bandwidth). If the traffic pattern is largely made up of continuous streams of data, then provision extra bandwidth.
- Small versus large frame size: Fibre Channel is a high-speed, low-latency protocol. It relies, however, on buffer-to-buffer credits to handle flow control. This mechanism is a fundamental part of the Fibre Channel standard and ensures lossless connections. Thus, a series of 100 small frames uses the same number of buffers as a series of 100 large frames. Large frames, on the other hand, use more bandwidth. In other words, a large amount of small-frame traffic can fully utilize available buffers, while consuming only a very small amount of available bandwidth. Therefore, you need to consider not only bandwidth, but also the typical frame size. If the bulk of frames are expected to be smaller in size, then additional links and/or buffers should be allocated to the paths that will be handling those smaller frame I/O patterns. Pay extra attention to this type of congestion, because backbones can become congested and adversely impact the performance of all connected edge fabrics. When in doubt, overprovision IFLs.
- Distance (location of fabrics): The distance between the end-points of the data transmission is an issue of providing adequate buffers for the projected traffic, and all of the potential traffic flows that might traverse the long-distance link(s) need to be considered. Given that long-distance solutions generally already have increased latency (simple physics of time to cover distance), it is important that long-distance links be overprovisioned for capacity, such that unexpected spikes do not adversely impact the data flow or, potentially, the entire network.
- Virtual Fabrics (VF): If VF is enabled, the base switch is like a backbone switch, and a base fabric is like a backbone fabric. All switches in a backbone fabric must have the same backbone fabric ID, which must be unique from the edge fabric.

  NOTE
  The VF Fabric ID is also the backbone, and Fabric ID and EX_Ports and VEX_Ports can be configured only on the base switch.

- Zoning: Traffic Isolation (TI) zones and FCR: Some VE_Port-based features, such as tape pipelining, require the request and corresponding response traffic to traverse the same VE_Port tunnel across the MetaSAN. Use TI Zones to ensure that the request and response traverse the same VE_Port tunnel; you must set up TI Zones in the edge and backbone fabrics. In addition to setting up TI Zones, you must also ensure that the devices are in an LSAN zone so that they can communicate with each other. If failover is enabled and the TI path is not available, an alternate path is used. If failover is disabled and the TI path is not available, then devices are not imported.
- **A potential for growth exists in the following:**
  - Number of fabrics: If the number of fabrics is likely to increase, then deploy backbone fabrics such that they can readily accommodate additional edge fabrics and additional traffic loads.
  - Size of fabrics: If the size of edge fabrics is likely to grow, and the inter-fabric traffic is expected to grow accordingly, provision additional IFLs and ISLs such that the capacity of available paths stays well ahead of current usage. That way, incremental growth on the edge can be accommodated without the need to immediately upgrade the backbone.
  - Amount of traffic between fabrics: If the inter-fabric traffic is expected to grow even without growth in the individual edge fabrics, then provision additional IFLs and ISLs such that the capacity of available paths stays ahead of current usage. That way, incremental increases in data flow across the backbone can be accommodated without the need to immediately upgrade the backbone. Make sure that you allow for plenty of room for backbone expansion.

Avoiding Congestion

Just as with a flat Layer 2 fabric, a routed SAN needs to be evaluated for traffic bandwidth and potential bandwidth utilization between all end-points. For routed topologies, this means calculating traffic flowing in and out of every edge fabric and providing enough links into and across the backbone to accommodate that traffic. Use the same guidelines that apply to ISLs when connecting fabrics through IFLs for improved utilization and resiliency. As often happens as fabrics evolve, an edge fabric can be of higher performance versus the backbone, resulting in a completely oversubscribed backbone. This can lead to congestion at peak loads and high latency due to slow device response. Prior to upgrading the edge fabric, consider increasing the number of ISLs or upgrading the backbone to avoid congestion.

Available Paths

The best approach is to have multiple trunked paths between edge fabrics so that traffic can be spread across available resources; however, it is never good practice to attach both A and B fabrics to the same backbone router. From the perspective of FC, you should adhere to the concept of an "air gap" all the way from host to storage. A common device connected to both A and B fabrics can cause a SAN-wide outage. If an air gap is implemented, faults on one fabric cannot affect the other fabric. These faults can manifest from defects in host, fabric, or storage hardware and software, as well as human error. It is not relevant that FCR keeps the fabrics separate, because these types of faults can transcend FCR and cause the entire SAN to fail.

Design Guidelines and Constraints for Routed SANs

Some of the key metrics and rules of thumb for routed SAN topologies are:

- Keep A and B fabrics separated all the way from host to storage from a FC perspective. This is referred to as the "air gap." This does not include IP networks passing FCIP traffic, although FCIP ISL end points should never cross-connect A and B fabrics, as this is the same as a traditional ISL cross connecting A and B fabrics.
- Localize traffic within an edge fabric as much as possible.
- Have a plan for predefining the domains in the fabric (for example, edge switches with a certain range, translate domains in a certain range that connect to the backbone fabric, and unique backbone fabric IDs to avoid domain overlap).
- Consider upgrading the backbone to higher performance prior to upgrading the edge fabric.
- No more than one long-distance hop between source and destination.
- Place long-distance links within the backbone (as opposed to between edge and backbone), as edge fabrics can then be isolated from disruption on the long-distance links. An edge fabric that contains a long-distance link is referred to as a remote edge. Remote edges can be the product of VEX_Ports and EX_Ports that connect to FC-based DWDM. Remote edges are not considered best practice.
- Use Logical SAN (LSAN) zones only for devices that will actually be communicating across the backbone. In other words, do not make every zone an LSAN zone for ease.
- As edge fabrics and the routed network grow, the use of "filters" such as LSAN zone binding and LSAN tagging can improve topology convergence time and efficient usage of FCR resources.
- Make the backbone fabrics redundant to improve resiliency. This means redundancy for each fabric; therefore, fabric A would be redundant and so would fabric B. Fabric B would never be used as the redundancy for fabric A, and vice-versa.
- Backbone fabrics that share connections to the same edge fabrics must have unique backbone fabric IDs. This statement is referring to the case in which there are multiple "A" fabrics and multiple "B" fabrics. This does not refer to sharing connections between A and B fabrics.
- TI Zones within the backbone fabric cannot contain more than one Destination Router Port (DRP) per each fabric.
- TI over FCR is supported only from edge fabric to edge fabric. Traffic isolation from backbone to edge is not supported.

- UltraScale ICLs on the core blade cannot be used for FCR.

Virtual Fabrics Topologies

The Brocade FOS Virtual Fabrics (VF) feature provides a mechanism for partitioning and sharing hardware resources, with the intention of providing more efficient use, deterministic paths for FCIP, increased fault isolation, and improved scalability. Virtual Fabrics use hardware-level fabric isolation between Logical Switches (LSs) and fabrics. Logical Fabrics consist of one or more Logical Switches across multiple physical switches (non-partitioned).

Hardware-level fabric isolation is accomplished through the concept of a Logical Switch, which provides the ability to partition physical switch ports into one or more "logical" switches. Logical Switches are then connected to form Logical Fabrics. As the number of available ports on a switch continues to grow, partitioning switches gives storage administrators the ability to take advantage of high-port-count switches by dividing physical switches into different Logical Switches. Without VF, an FC switch is limited to 512 ports. A storage administrator can then connect Logical Switches through various types of ISLs to create one or more Logical Fabrics.

There are three ways to connect Logical Switches: a traditional ISL, IFL (EX_Port used by FCR), and Extended ISL (XISL). An ISL can only be used for normal L2 traffic between the connected Logical Switches, carrying only data traffic within the Logical Fabric of which the ISL is a member. One advantage of Virtual Fabrics is that Logical Switches can share a common physical connection, and each LS does not require a dedicated ISL. In order for multiple Logical Switches, in multiple Logical Fabrics, to share an ISL, Virtual Fabrics supports an XISL connection, which is a physical connection between two base switches. Base switches are a special type of Logical Switch that are specifically intended for intra- and inter-fabric communication. As mentioned, base switches are connected via XISLs and form the base fabric.

Once a base fabric is formed, the Virtual Fabric determines all of the Logical Switches and Logical Fabrics that are physically associated via the base fabric, as well as the possible routes between them. For each local Logical Switch, a Logical ISL (LISL) is created for every destination Logical Switch in the same Virtual Fabric that is reachable via the base fabric. Thus, an XISL comprises the physical link between base switches and all of the virtual connections associated with that link. In addition to XISL support, the base fabric also supports IFLs via EX_Port connections for communication between Virtual Fabrics. Base switches also interoperate with FC router switches, either in the base fabric or in separate backbone fabrics.

VF Guidelines

If no local switching is used, any set of ports in the chassis/fabric can be used to create a Virtual Fabric. If local switching is used, ports for the VF fabric should be from the same port groups.

Use Case: FICON and Open Systems (Intermix)

Virtual Fabrics enable customers to share FICON and FCP traffic on the same physical platform. As chassis densities increase, this is a viable option for improved hardware utilization while maintaining director class availability. The primary reasons for moving to an Intermix environment are the following:

- Array-to-array RDR of FICON volumes (uses FCP)
- ESCON-FICON migration
- Sharing of infrastructure in a non-production environment
- Reduced TCO
- Growth of zLinux on the mainframe

From a SAN design perspective, the following guidelines are recommended when considering FICON Intermix:

- Connect devices across port blades (connectivity from the same device should be spread over multiple blades).

- One-hop count still applies (there are "Hops of No Concern" in some cases)

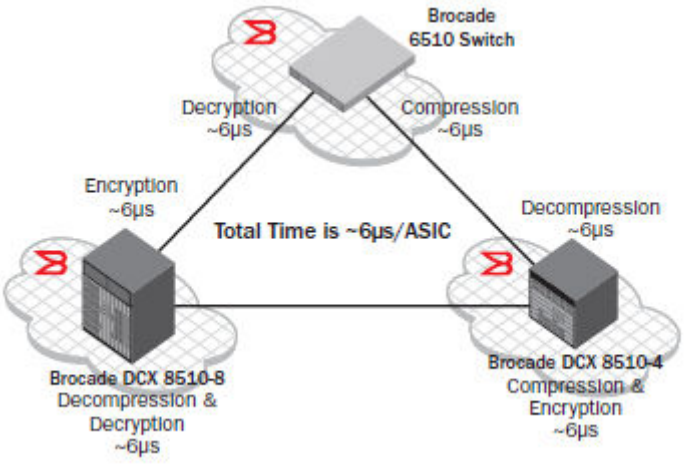
For details, see the *Brocade FICON/FCP Intermix Best Practices Guide*.

Intelligent Services

In-Flight Encryption and Compression

Brocade Gen 5 & Gen 6 Fibre Channel platforms support both in-flight compression and/or encryption at a port level for both local and long-distance ISL links. In-flight data compression is a useful tool for saving money when either bandwidth caps or bandwidth usage charges are in place for transferring data between fabrics. Similarly, in-flight encryption enables a further layer of security with no key management overhead when transferring data between local and long-distance data centers besides the initial setup.

FIGURE 18 Latency for Encryption and Compression



Enabling in-flight ISL data compression and/or encryption increases the latency as the ASIC processes the frame compression and/or encryption. Approximate latency at each stage, including encryption, compression and local switching) is 6.2 microseconds (see Figure 18). For example, compressing and then encrypting a 2KB frame incurs approximately 5.5 microseconds of latency on the sending Condor3-based switch and incurs approximately 5.5 microseconds of latency at the receiving Condor3-based switch in order to decrypt and uncompress the frame with an additional 700ns for local switching. This results in a total latency time of 12.4 microseconds, again not counting the link transit time. The use of FEC, required in Gen 6, adds an additional 400ns per stage.

Virtual Fabric Considerations (Encryption and Compression)

The E_Ports in the user-created Logical Switch, base switch, or default switch can support encryption and compression. Both encryption and compression are supported on XISL ports, but not on LISL ports. If encryption or compression is enabled and ports are being moved from one LS to another, it must be disabled prior to moving from one LS to another.

In-Flight Encryption and Compression Guidelines

- Supported on E_Ports and EX_Ports.

- No more than two ports on one ASIC can be configured with encryption, compression, or both when running a Gen 5 switch at 16 Gbps speed. With Brocade FOS v7.1, additional ports can be utilized for data encryption, data compression, or both if running at lower than 16 Gbps speeds.
- Gen 6 switches and blades support up to four ports with compression at any line speed. ISL encryption is not currently supported on Gen 6 devices.
- ISL ports must be set to Long-Distance (LD) mode when compression is used.
- Twice the number of buffers should be allocated if compression is enabled for long distance, as frame sizes may be half the size.
- If both compression and encryption are used, enable compression first.
- When implementing ISL encryption, using multiple ISLs between the same switch pair requires that all ISLs be configured for encryption-or none at all.
- ISL link encryption is not currently compliant with Federal Information Processing Standards (FIPS) 140-2.

Distance Extension Topologies

For a complete DR solution, SANs are typically connected over metro or long-distance networks. In both cases, path latency is critical for mirroring and replication solutions. For native Fibre Channel links, the amount of time that a frame spends on the cable between two ports is negligible, since that aspect of the connection speed is limited only by the speed of light. The speed of light in optics amounts to approximately 5 microseconds per kilometer, which is negligible compared to typical disk latency of 5 to 10 milliseconds. The Brocade Extended Fabrics feature enables full-bandwidth performance across distances spanning up to hundreds of kilometers. It extends the distance ISLs can reach over an extended fiber by providing enough buffer credits on each side of the link to compensate for latency introduced by the extended distance.

Buffer Allocation

Buffer credits are a measure of frame counts and are not dependent on the data size (a 64 byte and a 2KB frame both consume a single buffer). Standard 16-Gb transceivers support up to 125 meters over OM4 cable. (Refer to "Appendix A" for data rates and distances.) Users should consider the following parameters when allocating buffers for long-distance links connected via dark fiber or through a D/CWDM in a pass-thru mode:

- Round-Trip Time (RTT)-in other words, the distance
- Frame processing time
- Frame transmission time

Some good general guidelines are:

- Number of credits = 6 + ((link speed Gb/s * Distance in KM) / frame size in KB)
- Example: 100 KM @2k frame size = 6 + ((16 Gb/s * 100) / 2) = 806
- Buffer model should be based on the average frame size
- If compression is used, number of buffer credits needed is 2x the number of credits without compression.

On the Brocade DCX 8510 or X6 Fibre Channel platforms, 4K unreserved buffers are available per ASIC to support long distance connectivity.

Since FOS v7.1, Brocade provides users additional control when configuring a port of an LD or LS link, allowing users to specify the buffers required or the average frame size for a long-distance port. Using the frame size option, the number of buffer credits required for a port is automatically calculated. These options give users additional flexibility to optimize performance on long-distance links.

In addition, Brocade FOS provides users better insight into long-distance link traffic patterns by displaying the average buffer usage and average frame size via CLI. Brocade FOS also provides the **portBufferCalc** CLI command, which automatically calculates the number of buffers required per port given the distance, speed, and frame size. The number of buffers calculated by this command can be used when configuring the **portCfgLongDistance** command. If no options are specified, the current port's configuration is considered to calculate the number of buffers required.

NOTE
The ClearLink D_Port mode can also be used to measure the cable distance to a granularity of 5 meters between two 16-Gbps platforms; however, ports must be offline.

Fabric Interconnectivity over Fibre Channel at Longer Distances

SANs spanning data centers in different physical locations can be connected via dark fiber connections using Extended Fabrics, a Brocade FOS optionally licensed feature, with wave division multiplexing, such as: Dense Wave Division Multiplexing (DWDM), Coarse Wave Division Multiplexing (CWDM), and Time Division Multiplexing (TDM). This is similar to connecting switches in the data center with one exception: additional buffers are allocated to E_Ports connecting over distance. The Extended Fabrics feature extends the distance the ISLs can reach over an extended fiber. This is accomplished by providing enough buffer credits on each side of the link to compensate for latency introduced by the extended distance. Use the buffer credit calculation above or the CLI tools with Brocade FOS to determine the number of buffers needed to support the required performance.

Any of the first 8 ports on the 16 Gbps port blade and any port on the 32 Gbps port blade can be set to 10 Gbps FC for connecting to a 10 Gbps line card D/CWDM without the need for a specialty line card. If connecting to DWDMs in a pass-thru mode where the switch is providing all the buffering, a 16 Gbps line rate can be used for higher performance.

Recommendations include the following:

- Connect the cores of each fabric to the DWDM.
- If using trunks, use smaller and more trunks on separate port blades for redundancy and to provide more paths. Determine the optimal number of trunk groups between each set of linked switches, depending on traffic patterns and port availability.
- Configure the switches connected to the D/CWDM in 'R_RDY' mode with the D/CWDM configured to passive mode for highest assurances of interoperability.
- Check with your D/CWDM vendor to confirm support for VC_RDY mode and any advanced features such as trunking.

Workloads

Many different kinds of traffic traverse a SAN fabric. The mix of traffic is typically based on the workload on the servers and the effect that behavior has on the fabric and the connected storage. Examples of different types of workload include:

- I/O-intensive, transaction-based applications: These systems typically do high volumes of short block I/O and do not consume a lot of network bandwidth. These applications usually have very high-performance service levels to ensure low response times. Care must be taken to ensure that there are a sufficient number of paths between the storage and hosts to ensure that other traffic does not interfere with the performance of the applications. These applications are also very sensitive to latencies.
- I/O-intensive applications: These applications tend to do a lot of long block or sequential I/O and typically generate much higher traffic levels than transaction-based applications (data mining). Depending on the type of storage, these applications can consume bandwidth and generate latencies in both storage and hosts that can negatively impact the performance of other applications sharing their storage.
- Host High Availability (HA) clustering: These clusters often treat storage very differently from standalone systems. They may, for example, continuously check their connected storage for data integrity reasons and put a strain on both the fabric and the

storage arrays to which they are attached. This can result in frame congestion in the fabric and can cause performance problems in storage arrays.

- Host-based replication: Host-based replication causes traffic levels to increase significantly across a fabric and can put considerable pressure on ISLs. Replicating to poorer-performing storage (such as tier 1-to-tier 2 storage) can cause application performance issues that are difficult to identify. Latencies in the slower storage can also cause "back pressure," which can extend back into the fabric and slow down other applications that use the same ISLs.
- Array-based replication: Data can be replicated between storage arrays as well.

Workload Virtualization

The past several years have witnessed a huge growth in virtualized workloads. Available on IBM mainframes for decades, workload virtualization was initially popularized on Intel-based platforms by VMware ESX Server (now vSphere). Windows, UNIX, and Linux server virtualization are now ubiquitous in enterprise infrastructures.

Intel-Based Virtualization Storage Access

Intel-based VMs typically access storage in two separate ways:

- They use some sort of distributed file system that is typically controlled by the hypervisor (the control program that manages VMs). This method puts the onus on the hypervisor to manage the integrity of VM data. All VM I/O passes through an I/O abstraction layer in the hypervisor, which adds extra overhead to every I/O that a VM issues. The advantage to this approach is that many VMs can share the same LUN (storage), making storage provisioning and management a relatively easy task. Today the vast majority of VMware deployments use this approach, deploying a file system called Shared VMFS.
- They create separate LUNs for each data store and allow VMs to access data directly through N_Port ID Virtualization (NPIV). The advantage of this approach is that VMs can access data more or less directly through a virtual HBA. The disadvantage is that there are many more LUNs to provision and manage.

Most VMs today tend to do very little I/O-typically no more than a few MB/sec per VM via very few IOPS. This allows many VMs to be placed on a single hypervisor platform without regard to the amount of I/O that they generate. Storage access is not a significant factor when considering converting a physical server to a virtual one. More important factors are typically memory usage and IP network usage.

The main storage-related issue when deploying virtualized PC applications is VM migration. If VMs share a LUN, and a VM is migrated from one hypervisor to another, the integrity of the LUN must be maintained. That means that both hypervisors must serialize access to the same LUN. Normally this is done through mechanisms such as SCSI reservations. The more the VMs migrate, the potentially larger the serialization problem becomes. SCSI reservations can contribute to frame congestion and generally slow down VMs that are accessing the same LUN from several different hypervisor platforms.

Design Guidelines

- If possible, try to deploy VMs to minimize VM migrations if you are using shared LUNs.
- Use individual LUNs for any I/O-intensive applications such as SQL Server, Oracle databases, and Microsoft Exchange.

Monitoring

- Use MAPS to monitor Flow Vision flows to alert you to excessive levels of SCSI reservations. These notifications can save you a lot of time by identifying VMs and hypervisors that are vying for access to the same LUN.
- For Gen 6 networks, use IO Insight to monitor storage response latency and IO levels for critical and performance-sensitive workloads.

- For larger environments, Brocade Analytics and Monitoring Platform can monitor traffic across multiple switches and fabrics and alert you to developing contention issues that can degrade performance over time.

UNIX Virtualization

Virtualized UNIX environments differ from virtualized Windows deployments in a few significant ways.

First, the UNIX VMs and hypervisor platforms tend to be more carefully architected than equivalent Windows environments, because more mission-critical applications have traditionally run on UNIX. Frequently the performance and resource capacity requirement of the applications are well understood, because of their history of running on discrete platforms. Historical performance and capacity data will likely be available from the UNIX performance management systems, allowing application architects and administrators to size the hypervisor platforms for organic growth and headroom for peak processing periods.

Second, VM mobility is not common for workload management in UNIX deployments. VMs are moved for maintenance or recovery reasons only. IBM clearly states, for example, that moving VMs is limited to maintenance only. Carefully architected hypervisor/application deployments contain a mix of I/O-intensive, memory-intensive, and processor-intensive workloads. Moving these workloads around disturbs that balance and potentially leads to performance problems. Problem determination also becomes more difficult once VM migrations have to be tracked.

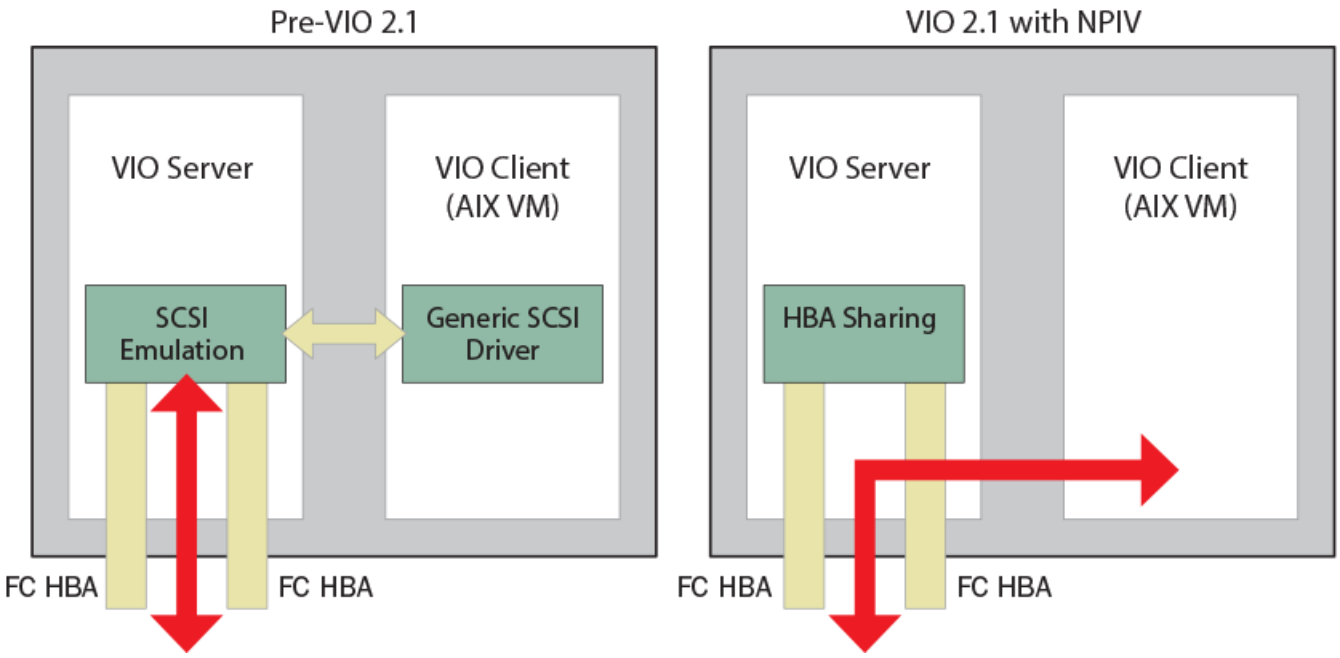
Third, virtualized mission-critical UNIX applications such as large SQL Server database engines typically use much more block I/O than their Windows counterparts, both in volume and in transaction rates. Each hypervisor platform now produces the aggregate I/O of all those mission-critical applications. Backups, especially if they are host-based through backup clients, are also a serious architectural concern.

Recent Changes

Two technical advances create profound changes to storage deployments for mission-critical UNIX applications: NPIV and storage virtualization.

Consider the IBM AIX VIO platform as an example to explain UNIX workload virtualization. (Other vendor systems such as Oracle/Sun Solaris and HP HP-UX behave somewhat differently.) NPIV came later to UNIX, with IBM adopting NPIV in AIX VIO 2.1 to improve traffic through the SCSI I/O abstraction layer. The difference is illustrated in Figure 19.

FIGURE 19 Before and After IBM AIX/VOI 2.1



Pre-NPIV implementations of VIO, shown on the left in Figure 19, performed SCSI I/O through generic SCSI drivers in the VM (the VIO client) in an AIX Logical Partition (LPAR). The VIO server in another LPAR has actual control of the Fibre Channel adapters and provides SCSI emulation to all VIO clients. With VIO 2.1 and later versions, the VIO client performs I/O directly via NPIV to the Fibre Channel HBA through a virtual HBA, and the VIO server simply controls access to HBAs installed in the system, shown on the right.

The use of NPIV significantly reduces the complexity of the I/O abstraction layer. I/O is therefore less of a bottleneck and allows for more LPARs on each AIX hypervisor platform. More LPARs (VMs or VIO clients) means better consolidation ratios and the potential to save capital expenses on hypervisor platforms. I/O utilization per Fibre Channel HBA increases, perhaps necessitating the addition of more FC adapters to accommodate the increased workload. This in turn translates to higher traffic levels and more IOPS per HBA.

As consolidation of UNIX hosts progresses, expect to see much higher activity at the edge of the fabric. As a result you will need to monitor the fabric much more carefully to avoid both traffic and frame congestion. It is also much more likely that the hypervisors themselves will become substantial bottlenecks.

Design Guidelines

- With the higher levels of I/O potentially occurring at each edge port in the fabric, you must ensure that there is sufficient bandwidth and paths across the fabric to accommodate the load. Consider a lot of trunked ISLs and lower subscription ratios on the ISLs, if at all possible. Remember that many flows are partially hidden due to the increased use of NPIV.
- Frame congestion is also a greater possibility. Many of the VMs may still be in clusters and may require careful configuration. Spread out the LUNs across a lot of storage ports.
- Separate the hypervisors on separate directors and, certainly, keep them separate from storage ports. This allows you to very easily apply controls through Brocade MAPS groups without affecting storage.
- Determine what latencies are tolerable to both storage and hosts (VMs and storage), and consider setting Brocade FOS thresholds accordingly.

- Port Fencing is a powerful tool. Once many applications are running in VMs on a single physical platform, take care to ensure that Port Fencing does not disable ports too quickly.

Monitoring

- Fabric Performance Impact Monitoring becomes very important here. Use it to monitor latencies on both the hypervisor and storage ports to identify high latencies as soon as you can. Address the latencies as soon as possible.
- Brocade MAPS is essential in early notification of potential issues in the fabric. Given the much higher concentration of I/O due to the server consolidation, you should closely monitor traffic levels. The tight integration of IO Insight with MAPS and Flow Vision in Gen 6 switches allows for monitoring of I/O levels as well as device-level I/O latency relative to established baseline levels and should be fully utilized.
- Monitor the Class 3 frame discards (C3TX_TO) through Brocade MAPS as well. They are a strong indication of high-latency devices.

Scalability and Performance

Brocade products are designed with scalability in mind, knowing that most installations will continue to expand and that growth is supported with very few restrictions. However, following the same basic principles outlined in previous sections as the network grows will ensure that the levels of performance and availability will continue.

Evaluate the impact on topology, data flow, workload, performance, and perhaps most importantly, redundancy and resiliency of the entire fabric any time one of the following actions is performed:

- Adding or removing initiators:
  - Changes in workload
  - Changes in provisioning
- Adding or removing storage:
  - Changes in provisioning
  - Changes in storage media type (e.g., increased deployment of flash-based storage)
- Adding or removing switches
- Adding or removing ISLs and ICLs
- Change in virtualization (workload and storage) strategies and traffic flow pattern

If these design best practices are followed when the network is deployed, then small incremental changes should not adversely impact the availability and performance of the network. However, if changes are ongoing and the fabric is not properly evaluated and updated, then performance and availability can be jeopardized. Some key points to cover when looking at the current status of a production FC network include:

Reviewing redundancy and resiliency:

- Are there at least two physically independent paths between each source and destination pair?
- Are there two redundant fabrics?
- Does each host connect to two different edge switches?
- Are edge switches connected to at least two different core switches?
- Are inter-switch connections composed of two trunks of at least two ISLs?
- Does each storage device connect to at least two different edge switches or separate port blades?
- Are storage ports provisioned such that every host has at least two ports through which it can access LUNs?

- Are redundant power supplies attached to different power sources?
- Are zoning and security policies configured to allow for patch/device failover?

Reviewing performance requirements:

- Host-to-storage port fan-in/out ratios
- Oversubscription ratios:
  - Host to ISL
  - Edge switch to core switch
  - Storage to ISL
- Size of trunks
- Routing policy and currently assigned routes; evaluate actual utilization for potential imbalances
- Use of FEC for all ISLs and connections to Gen 5 (if supported) and Gen 6 devices

Watching for latencies such as these:

- Poor storage performance
- Overloaded hosts or applications
- Distance issues over constrained long distance links resulting from changes in usage, such as adding mirroring, or too many workloads)
- Deteriorating optics resulting in declining signal strength and increased error rate

In Gen 6 networks, storage response latency can be baselined and monitored continuously using IO Insight in conjunction with MAPS. Deal with latencies immediately; they can have a profound impact on the fabric.

In summary, although Brocade SANs are designed to allow for any-to-any connectivity, and they support provision-anywhere implementations, these practices can have an adverse impact on the performance and availability of the SAN if left unchecked. As detailed above, the network needs to be monitored for changes and routinely evaluated for how well it meets desired redundancy and resiliency requirements.

Supportability

Supportability is a critical part of deploying a SAN. Follow the guidelines below to ensure that the data needed to diagnose fabric behavior or problems have been collected. While not all of these items are necessary, they are all pieces in the puzzle. You can never know which piece will be needed, so having all of the pieces available is best.

- Configure Brocade MAPS monitoring: Leverage Brocade MAPS to implement proactive monitoring of errors and warnings such as CRC errors, loss of synchronization, and high-bandwidth utilization.
- Configure syslog forwarding: By keeping historical log messages and having all switch messages sent to one centralized syslog server, troubleshooting can be expedited and simplified. Forwarding switch error messages to one centralized syslog server and keeping historical log messages enables faster and more effective troubleshooting and provides simple monitoring functionality.
- Create a switch configuration template using the Configuration and Operational Monitoring Policy Automation Services Suite (COMPASS) feature in Brocade Network Advisor to avoid configuration drift from occurring over time. COMPASS can also adopt existing configurations as a template for deploying new switches in the fabric, ensuring consistency across the data center. Brocade Network Advisor policy violation monitoring should also be used to identify switches that are no longer in compliance. COMPASS can also be run on a regular basis to validate ongoing compliance with organizational policy.
- Follow Brocade best practices in the LAN infrastructure for management interfaces: Brocade best practices in the LAN infrastructure recommend a setup of different physical LAN broadcast segments, for example, by placing IP routers between segments or configuring different VLANs for the management interfaces of two fabric switches.

- Enable audit functionality: To provide audit functionality for the SAN, keep track of which administrator made which changes, usage of multiple user accounts (or RADIUS), and configuration of change tracking or audit functionality (along with use of errorlog/syslog forwarding).
- Configure multiple user accounts (LDAP/OpenLDAP or RADIUS): Make mandatory use of personalized user accounts part of the IT/SAN security policy, so that user actions can be tracked. Also, restrict access by assigning specific user roles to individual users.
- Establish a test bed: Set up a test bed to test new applications, firmware upgrades, driver functionality, and scripts to avoid missteps in a production environment. Validate functionality and stability with rigorous testing in a test environment before deploying into the production environment.
- Implement serial console server: Implement serial remote access so that switches can be managed even when there are network issues or problems during switch boot or firmware upgrades.
- Use aliases: Using aliases to give switch ports and devices meaningful names can lead to faster troubleshooting.
- Configure **supportftp:** Configure **supportftp** for automatic file transfers. The parameters set by this command are used by **supportSave** and **traceDump**.
- Configure an NTP server: To keep a consistent and accurate date and time on all the switches, configure switches to use an external time server.

Firmware Upgrade Considerations

Both fixed-port and modular switches support hot code load for firmware upgrades.

- Disruptive versus non-disruptive upgrades:
  - Simultaneous upgrades on neighboring switches
  - Standard FC ports versus application and special-feature ports
- Review the Brocade Fabric OS Release Notes for the following:
  - Upgrade path
  - Changes to feature support
  - Changes to backward compatibility
  - Known issues and defects
- Consider a separate AG firmware upgrade strategy. Brocade Access Gateways have no fundamental requirement to be at the same firmware release level as Brocade FOS. Upgrading only directors and switches minimizes the infrastructure changes required during an upgrade cycle.

NPIV and the Brocade Access Gateway

One of the main limits to Fibre Channel scalability is the maximum number of domains (individual physical or virtual switches) in a fabric. Keeping the number of domains low reduces much of the overhead typically attributed to SAN fabrics. Small-domain-count fabrics are more reliable, perform better, and are easier to manage. You can reduce overhead by doing the following:

- Reducing inter-switch zone transfers
- Reducing name server synchronization
- Reducing RSCN processing

The main reason for using Access Gateway (AG) mode is scalability. Given that embedded switches are smaller-port-count switches, an environment with a lot of blade servers with embedded switches can easily start to encroach on the stated limits on total domain count. Putting these switches into AG mode means they will not be consuming domain. The downside to AG mode has been the functionality

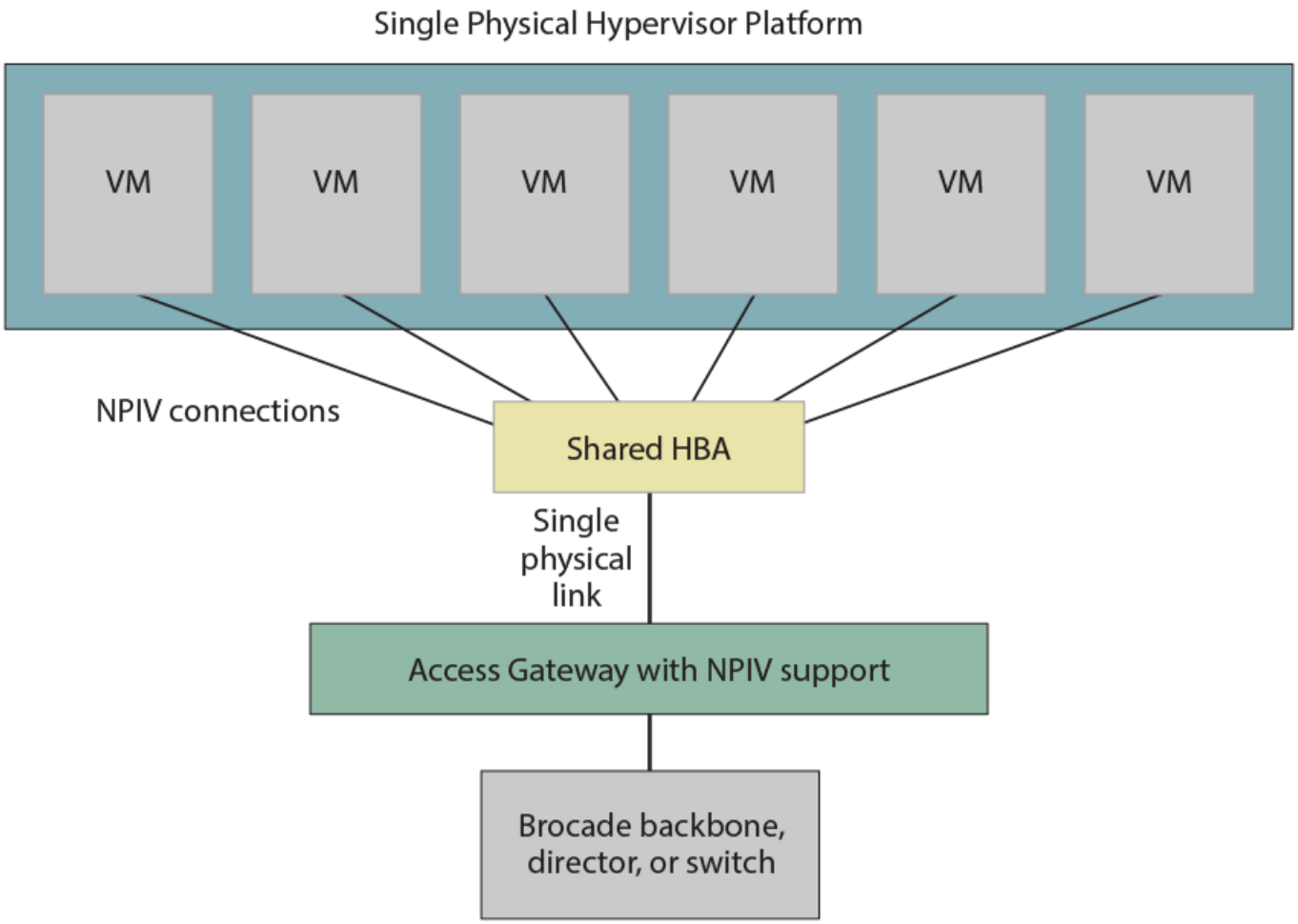
(or feature set) available, although AG continues to expand its functionality with each release. Though there are some scenarios with a clear-cut answer for AG mode, generally it is an evaluation of the SAN environment and the desired functionality that determines if AG is a design option for the environment. In a fabric with lots of legacy devices, identifying and isolating misbehaving devices is easier to do in a full-fabric environment.

Last, for configurations with hosts and targets on the same AG, the traffic does need to go through the fabric switch, but it is handled within the local switch and does not need to traverse to another switch in the fabric and then back again. The theoretical domain limit in a single fabric is 239, but most fabrics are typically limited to a much smaller number (56 is recommended in Brocade fabrics). The domain count limit typically comes into play when a large number of small-port-count switches are deployed. Large-bladed server deployments, for example, can easily push the domain count up over recommended limits when embedded blade switches are part of the implementation. FC switches in blade server enclosures typically represent fewer than 32 ports.

NPIV was originally developed to provide access to Fibre Channel devices from IBM mainframes and to improve the efficiency of mainframe I/O for virtualized environments. NPIV is part of the Fibre Channel standard and has been put to use in many open systems storage deployments. Brocade switches and directors as well as the Brocade Access Gateway support NPIV.

NPIV allows for many flows (connections) to share a single physical link. Figure 20 illustrates a single platform that supports flows from separate VMs through a single upstream link to a fabric via a shared HBA.

FIGURE 20 VMs Supported on a Single Link to a Fabric via NPIV



A device or switch connecting to another switch via an NPIV-enabled port does not require a domain ID, does not do any zoning, and behaves much more like an end device (or group of devices) than a switch. The Brocade Access Gateway was originally designed to reduce domain ID proliferation with the introduction of embedded blade switches, which use low-port-count switches that reside in blade server chassis. In most environments, these embedded switches are deployed in large quantities, which not only lead to high-domain-count fabrics, but also increases switch administration overhead. The Brocade Access Gateway eliminates or reduces both of these issues and is supported on all Brocade embedded switches and some fixed-port switch platforms. The Brocade Access Gateway connects initiators such as host HBAs on its "downstream" F_Ports to one or more fabrics via "upstream" N_Ports.

Benefits of the Brocade Access Gateway

- Scalability: You can add many Access Gateways to a fabric without increasing the domain count. A major scalability constraint is avoided when small-port-count switches or embedded switches are part of an infrastructure. Registered State Change Notifications (RSCNs) are also greatly reduced-only those that are related to the initiators on the downstream Access Gateway ports are passed on through to the fabric. Since it is essentially a device, the Access Gateway can connect to more than one fabric from its upstream ports. Brocade Access Gateways can be cascaded to reduce the number of fabric connections required to support a given workload or traffic level from the attached hosts.
- Error isolation and management: Most initiator errors are not propagated through to the fabric. Disconnecting an upstream port, for example, does not cause a fabric rebuild. Most management activities on the Brocade Access Gateway are also isolated from the fabric. One possible scenario is server administrators managing the Access Gateways and storage administrators simply providing LUNs and zoning support for the servers using NPIV.
- Increased resiliency: The Brocade Access Gateway supports F_Port Trunking, which increases the resiliency of connections into the fabric. Losing a trunk member simply reduces the bandwidth of the upstream trunk. While a few frames may be lost, no host connections are affected.
- Other: Hosts or HBAs can be configured to automatically fail over to another upstream link, should the one they are using fail. The Brocade Access Gateway also implements many advanced features such as Adaptive Networking services, Trunking, hot code load, Brocade MAPS, Brocade ClearLink, Credit Recovery, and Forward Error Correction.

Constraints

The advantages of the Brocade Access Gateway are compelling, but there are constraints:

- Although benefits are much more obvious for servers, the Brocade Access Gateway supports storage devices, but the traffic must flow through the fabric, which has its own limitations.
- There are a maximum number of 254 NPIV connections per upstream port.
- The number of Brocade Access Gateways per switch is limited only by what the fabric switches can support.

The primary factors are:

- The total number of devices that attach to the fabric through the Access Gateways
- The number of devices per Access Gateway N_Port
- The total number of devices attached to the switch and fabric

See the Brocade Scalability Guidelines for details.

- The number of fabrics to which a single Brocade Access Gateway can be connected is limited to the number of N_Ports on that Access Gateway. In general, most deployments require a single Access Gateway connection to only one or two fabrics. Note that the ability to connect different upstream ports to different fabrics does not reduce the requirement for redundancy. All attached servers should have dual paths to their storage through different fabrics via separate Access Gateways.

Design Guidelines

Use the Brocade Access Gateway when you deploy bladed environments or have a lot of low port-count switches and when you need to connect different servers in different fabrics from a single bladed enclosure. The Access Gateway can be very valuable when you want to separate the management of blade enclosures so that the enclosure is completely managed by server administrators, and the fabric is handled by storage administrators. Management separation is provided through the NPIV connection, which allows the Access Gateway to be managed separately by tools such as integrated blade server enclosure management tools without any adverse effects on the fabric.

Monitoring

Monitoring is somewhat difficult for NPIV flows. Traditional SAN monitoring has been based at the port level where hosts are connected. Multiple flows across ISLs and IFLs and into storage ports are common, but multiple host behaviors into initiators are a relatively new concept. The Brocade Access Gateway has been enhanced to include many features found in the standard version of Brocade FOS, such as Port Fencing, device security policies, and FPI monitoring.

Maintenance

There is usually no need to keep the Brocade Access Gateway firmware levels synchronized with the firmware levels deployed in the fabrics to which it is connected (and Brocade supports connections from other vendors' NPIV-enabled devices, where firmware synchronization is impossible). This can be significant for very large fabrics with many devices, including many Access Gateways. The version of Brocade FOS running on fabric switches can be upgraded at one time and the Access Gateways at another time, which greatly reduces the amount of change required to the infrastructure during a single maintenance window.

See the *Brocade Fabric OS Release Notes* to determine if a synchronized Brocade FOS upgrade of Brocade Access Gateway devices is required.

Backup and Restore

Backup and restore is part of an overall Disaster Recovery strategy, which itself is dependent on the criticality of data being backed up. In addition to storage consolidation, data backups are still a primary driver for a SAN-based infrastructure. This is commonly known as LAN-free backup, leveraging high-speed Fibre Channel for transport.

> NOTE
> Since tape drives are streaming devices, it is important to determine and maintain the optimal transfer rate. Contact your tape drive vendor if this information is not available.

The key factors for backup and restore include the following:

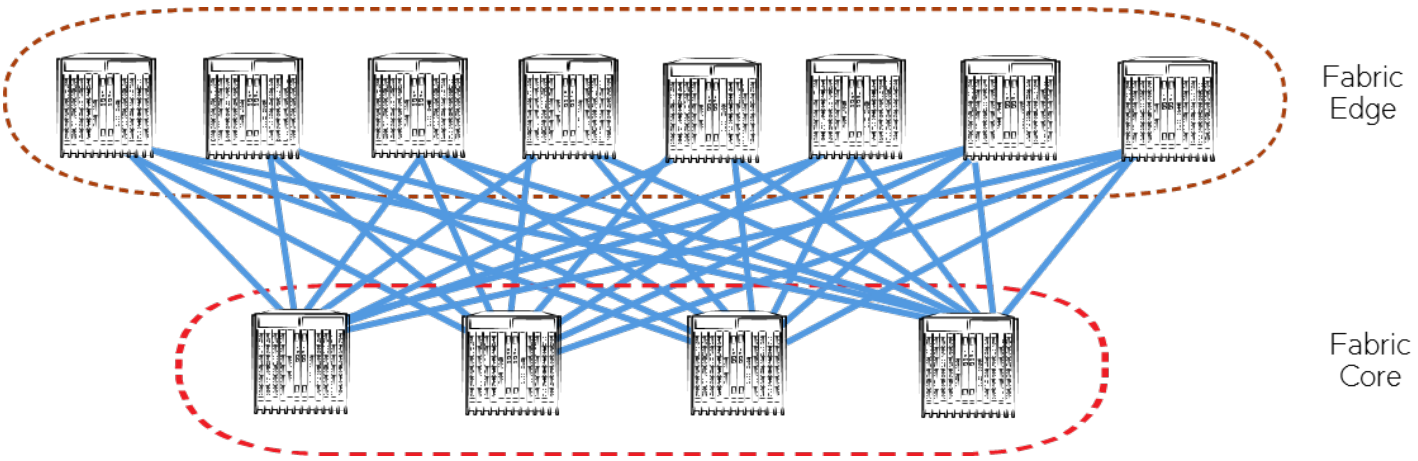
- Restoring backup data successfully is the most critical aspect of the backup/recovery process. In addition to ensuring business continuity in the event of a man-made or natural disaster, it is also a regulatory compliance requirement.
- Backups must be completed most, if not all, of the time.
- You should leverage backup reports so that administrators can keep track of tape media utilization and drive statistics as well as errors.
- If tapes are kept offsite for storage and Disaster Recovery, encrypt the data for security purposes.

Create a process and document procedures to validate backups periodically. Back up not only application data, but also include switch configurations to ensure that in the event of a switch failure a new switch can quickly be configured. Use COMPASS, Brocade SAN Health, Brocade Network Advisor, or the Brocade FOS CLI to capture switch configurations.

Determining SAN Bandwidth for Backups

At a minimum, available bandwidth in the fabric should be able to support applications and backup throughput at the same time. For example, in a core-edge topology, the ISL or ICL paths from the storage to the host should be able to support total throughput of all active backups and all applications without congestion. As shown in Figure 21, these paths should be redundant so that the failure of an ISL or ICL will not cause congestion in the fabric, impacting application or backup performance.

FIGURE 21 The Same Core-Edge Tiered Topology



The key drivers for data recovery include the following:

- How quickly access to data is restored, called the Recovery Time Objective (RTO), determined by the potential for lost revenue during the recovery period.
- The point in time in which the last valid data transaction was captured, called the Recovery Point Objective (RPO) which determines the needed frequency of the backups required to not exceed the acceptable data loss.
- Where the recovered data is located.

Improving the Backup Infrastructure

Determine if the existing backup infrastructure can support expanding SANs driven by data growth:

- Look at the backup schedule and how long it takes to complete the backup, to see if there are better time periods to run the job, or schedule to a different library for faster completion.
- Use tape multiplexing or compression.

If budgets permit, other options to improve backups to meet business objectives include the following:

- Add additional drives or libraries.
- Deploy a deduplication appliance.
- Use Virtual Tape Libraries (VTLs).

From a SAN perspective, consider the following:

- Add additional ISLs, or break down existing trunks into no more than two ports in the trunk to create TI Zones. This minimizes the impact of backup traffic on other application traffic.
- Make sure that there are redundant paths to the backup tier (see the section on Device Placement for details).
- Utilize MAPS to monitor for overutilization of ISLs during backup windows.

- For Brocade DCX 8510 director chassis with open slots in the core, add a high-density port blade such as the Brocade FC16-64 to expand the backup tier and add additional backup devices.

To reduce the time to recover from a backup, implement a two-tier disk-tape system with incremental backup to disk and migration to tape in off-hours and full backups only during downtime, such as on weekends. Another option is to implement a Continuous Data Protection (CDP) system, in which after a full backup only changed files or disk blocks are backed up. This provides the ability to restore at a granular level.

For a detailed discussion of backup and recovery concepts and issues, see *Strategies for Data Protection*, by Tom Clark, on Brocade Bookshelf (www.brocade.com/bookshelf).

Storage

Storage arrays have evolved dramatically over the last few years. Performance has increased exponentially, capacities have exploded, and more LUNs are supported than ever before. The performance and capacity of low-end arrays has also improved. New features include the following:

- Tiers comprising all-flash or flash-enhanced storage media have become commonplace for mid-range and enterprise-class environments.
- Some arrays time out and reset their ports if they do not receive acknowledgements from the connected host after specific intervals.
- Use of in-band Fibre Channel for control purposes has increased, putting extra stress on FC port buffer usage.

The use of flash-enhanced storage not only reduces the response times and increases the I/O level by up to 10x, is also increases the efficiency of the host platforms, allowing them to run at a higher level of utilization and execute more I/O commands rather than waiting on responses from slow disk media. Adding a layer of flash can therefore open up bottlenecks and relieve congestion resulting from long wait times.

Still, storage array performance can degrade over time, which can be attributed to factors that include:

- Proliferation of VMs and the "IO blender" effect, whereby IO processes become increasingly non-sequential and increase reads and writes to disk.
- Insufficient controller cache.
- Misaligned LUN configurations where OS block sizes do not match well to those on block storage.
- Misconfigured control values such as edge hold time and queue depths.
- Provisioning strategies can favor capacity over usage. An example of this might be a policy that dictates the number of terabytes allocated per storage port. Applications accessing the LUNs can overload the array capacity in order to service the requests.

Fixing degraded array performance is never easy. It usually involves some data migration or array reconfiguration. FPI monitoring can be used to detect these conditions early, and the origin of the congestion identified with Flow Vision. Fast diagnosis and early action allows changes to be implemented before performance degradation becomes chronic.

Design Guidelines

- Ensure that network throughput capacity matches or exceeds the capabilities of storage devices.
- Be careful if you deploy mixed arrays with different performance characteristics. Experience has shown that it is very easy for a Tier 3 storage array, depending on how it is used, to impact the performance of arrays in the same fabric. Troubleshooting in these situations is very difficult.

- Network switch and optic speeds should be limited to two generations at most, with a best practice of uniform speed to avoid target ports from slowing to accommodate lower-speed ports and creating a back-pressure situation.
- Enable Forward Error Correction for ISL connections as well as connectivity to Gen 5 devices that support it. FEC is on by default for Gen 6 networks and should always be used to ensure consistency of network transit times.
- Control the number of LUNs behind each storage port based on the type of usage they will receive.
- Check on any special short-frame traffic to avoid frame congestion at array ports. It may be necessary to increase the number of buffers at the array port to accommodate the extra control traffic.
- Use advance Brocade FOS threshold timers to monitor hosts and storage arrays to ensure that array ports do not reset due to a high-latency host, and thus do not adversely impact other connected hosts.

Monitoring

- FPI monitoring is indispensable; many high-latency array ports can be identified and their performance problems addressed before issues come to the attention of the server administrator.
- Use Brocade MAPS to monitor Class 3 frame discards due to TX timeout so that severe latencies on storage array ports can be identified.
- Use Flow Vision to monitor traffic at the flow level and identify hot spots that can cause congestion and impact performance across the fabric.
- Use IO Insight on Gen 6 fabrics to baseline and monitor storage response latency and queue depths.

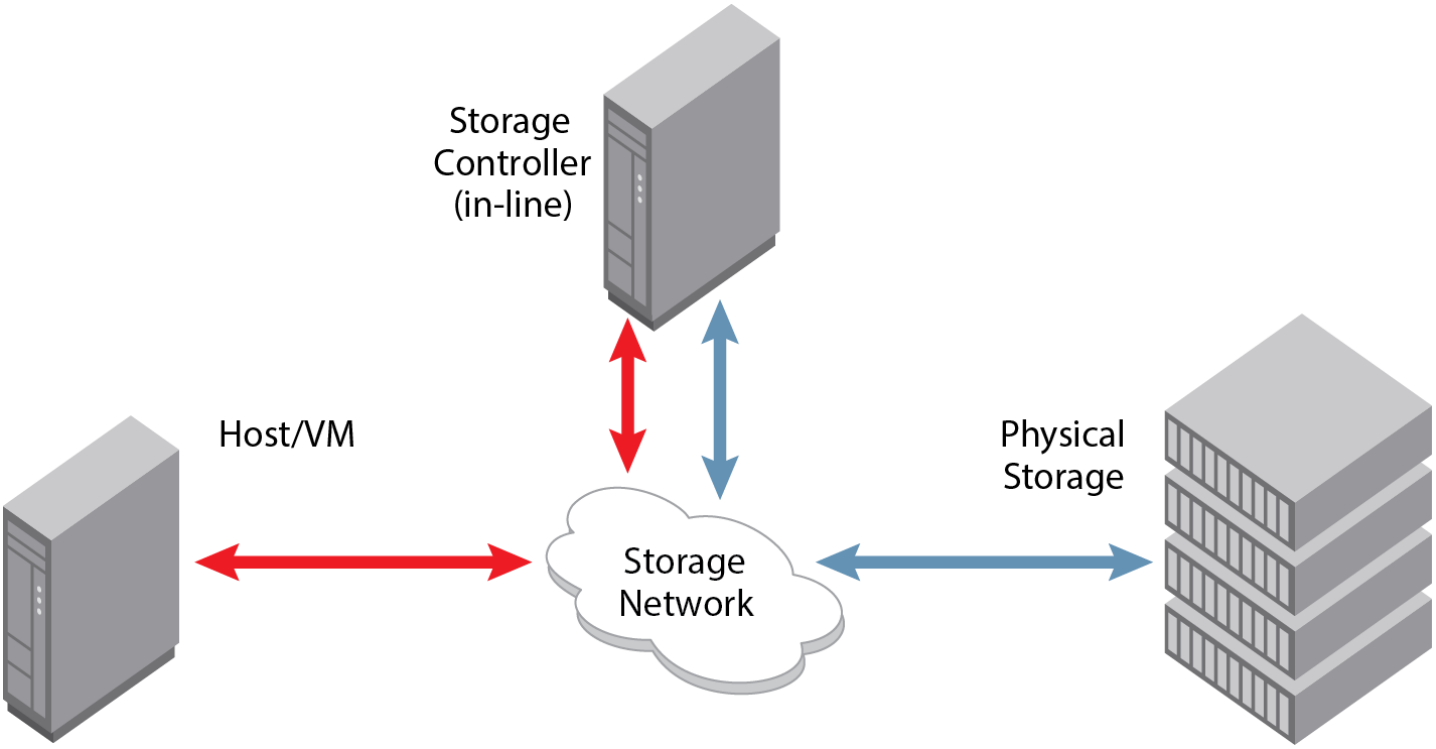
Storage Virtualization

Storage virtualization enables LUNs accessed by servers to be abstracted from the physical storage (typically storage arrays) on which they actually reside. (These are not the same as traditional storage array LUN allocations, which can also be viewed as a form of virtualization.) Virtualized LUNs that are disassociated from their actual storage allow for more flexible storage provisioning processes. Performance may also improve, as the virtual LUNs can be striped across multiple storage arrays.

There are two general types of storage virtualization: one uses an external controller (called in-line virtualization), and in the other, the virtualization occurs inside a storage array. In-line solutions are slightly more flexible, because they can use physical storage from a variety of sources and vendors.

Figure 22 shows a typical implementation of an in-line virtualized storage solution. The host or VM accesses storage via a storage controller (shown on top) through the storage network. The red arrows indicate data access to and from the storage controller. The storage controller typically controls all access to the physical storage, shown on the right (and indicated by the blue arrows). This creates a very flexible storage solution, because logical LUNs can be striped across several physical arrays to improve performance, and logical LUNs can be manipulated completely transparently to the host or VM.

FIGURE 22 Typical Implementation of an In-Line Virtualization Storage Solution



The major benefit of this type of storage virtualization is that storage can now be provisioned in units of capacity (500 gigabytes or a terabyte) rather than physical LUNs. This is a first step toward viewing storage as a service instead of as physical units. VM provisioning now becomes less complex and easier to automate. Look into products such as IBM SAN Volume Controller, Hitachi Data Systems Virtual Storage Platform, EMC VPLEX, and HP StoreVirtual virtual storage appliance (VSA) for information about how these products work.

Design Guidelines

- Each storage controller in an in-line solution serves as both an initiator and a target.
- ISL utilization increases with in-line virtualized storage. Make sure that you have enough ISL bandwidth to handle the increased load.
- There is also the possibility that the in-line storage heads will communicate through Fibre Channel or generate many more SCSI control frames to manage their attached storage, which can contribute to frame congestion. You may need to increase the number of buffers at the ports that connect to the storage controller to accommodate this behavior.
- It is much more difficult to determine initiators and targets with in-line virtualized storage. Since they are on the same switch, be careful about deploying tools such as Port Fencing.

Monitoring

- Utilize MAPS to track thresholds of critical parameters that indicate developing issues such as frame discards, overutilization of links and CRC errors.
- FPI monitoring is very useful in determining latencies associated with virtualized storage.
- Use Flow Vision to look at high-traffic-usage flows and identify the source of bottlenecks and congestion.

- Consider running Flow Vision constantly for the most mission-critical applications to keep a historical record of application performance profile and intermittent irregularities.
- For "frequent caller" applications, run Flow Vision on a regular basis where time permits to verify good health.

Security

There are many components to SAN security in relation to SAN design, and the decision to use them is greatly dependent on installation requirements rather than network functionality or performance. One clear exception is the zoning feature used to control device communication. The proper use of zoning is key to fabric functionality, performance, and stability, especially in larger networks. Other security-related features are largely mechanisms for limiting access and preventing attacks on the network (and are mandated by regulatory requirements), and they are not required for normal fabric operation.

Zoning: Controlling Device Communication

The SAN is primarily responsible for the flow of data between devices. Managing this device communication is of utmost importance for the effective, efficient, and also secure use of the storage network. Brocade Zoning plays a key role in the management of device communication. Zoning is used to specify the devices in the fabric that should be allowed to communicate with each other. If zoning is enforced, then devices that are not in the same zone cannot communicate.

In addition, zoning provides protection from disruption in the fabric. Changes in the fabric result in notifications (RSCNs) being sent to switches and devices in the fabric. Zoning puts bounds on the scope of RSCN delivery by limiting their delivery to devices when there is a change within their zone. (This also reduces the processing overhead on the switch by reducing the number of RSCNs being delivered.) Thus, only devices in the zones impacted by the change are disrupted. Based on this fact, the best practice is to create zones with one initiator and one target with which it communicates ("Single-Initiator Zoning"), so that changes to initiators do not impact other initiators or other targets, and disruptions are minimized (one initiator and one target device per zone). In addition, the default zone setting (what happens when zoning is disabled) should be set to No Access, which means that devices are isolated when zoning is disabled.

Zones can be defined by either switch port or device World Wide Name (WWN). While it takes a bit more effort to use WWNs in zones, it provides greater flexibility; if necessary, a device can be moved to anywhere in the fabric and maintain valid zone membership.

Peer Zoning

As the number of zones increase, it may become difficult to configure and maintain the zones using the single-initiator zoning practice. In addition, the storage requirements of defining unique zones for each host and target may exceed the zone database size limits. "One-to-Many Zoning" defines a zone having one target and other members as initiators. This approach has the advantages of being easier to manage and using less storage, but zoning all the initiators together in this manner results in less effective use of hardware resources and greater RSCN traffic.

"Peer Zoning" allows one or more "principal" devices to communicate with the rest of the devices in the zone and manage a Peer Zone. "Non-principal" devices in the zone can communicate with the principal device only, but they cannot communicate with each other. This approach establishes zoning connections that provide the efficiency of Single-Initiator Zoning with the simplicity and lower memory characteristics of one-to-many zoning.

Target Driven Zoning

When the principal device is a storage device, this is referred to as Target Driven Zoning. Support for target driven zoning requires support for a third party management interface and enabled at the switch F_Port connected to the storage device service as the principal device. Refer to Fabric OS Administrator's Guide 7.4 or later and your storage vendor for additional limitations and considerations.

Zone Management: Dynamic Fabric Provisioning (DFP)

Brocade Gen 5 and Gen 6 Fibre Channel SAN platforms support dynamically provisioned, switch-generated virtual WWNs, enabling SAN admins to create a fabric-wide zone database prior to acquiring and connecting any HBAs to the switch. DFP enables SAN administrators to pre-provision services like zoning, QoS, Device Connection Control (DCC), or any services that require port-level authentication prior to servers arriving in the fabric. This enables a more secure and flexible zoning scheme, since the fabric assigns the WWN to use. The FA-WWN can be user-generated or fabric-assigned (FA-WWN). When a supported HBA (currently available from QLogic) is replaced or a server is upgraded, zoning and LUN mapping does not have to be changed, since the new HBA is assigned the same FA-WWN as before. DFP is supported on both switches with or without the Brocade Access Gateway support. The switch automatically prevents assignment of duplicate WWNs by cross-referencing the Name Server database, but the SAN Administrator has the ultimate responsibility to prevent duplicates from being created when it is user-assigned.

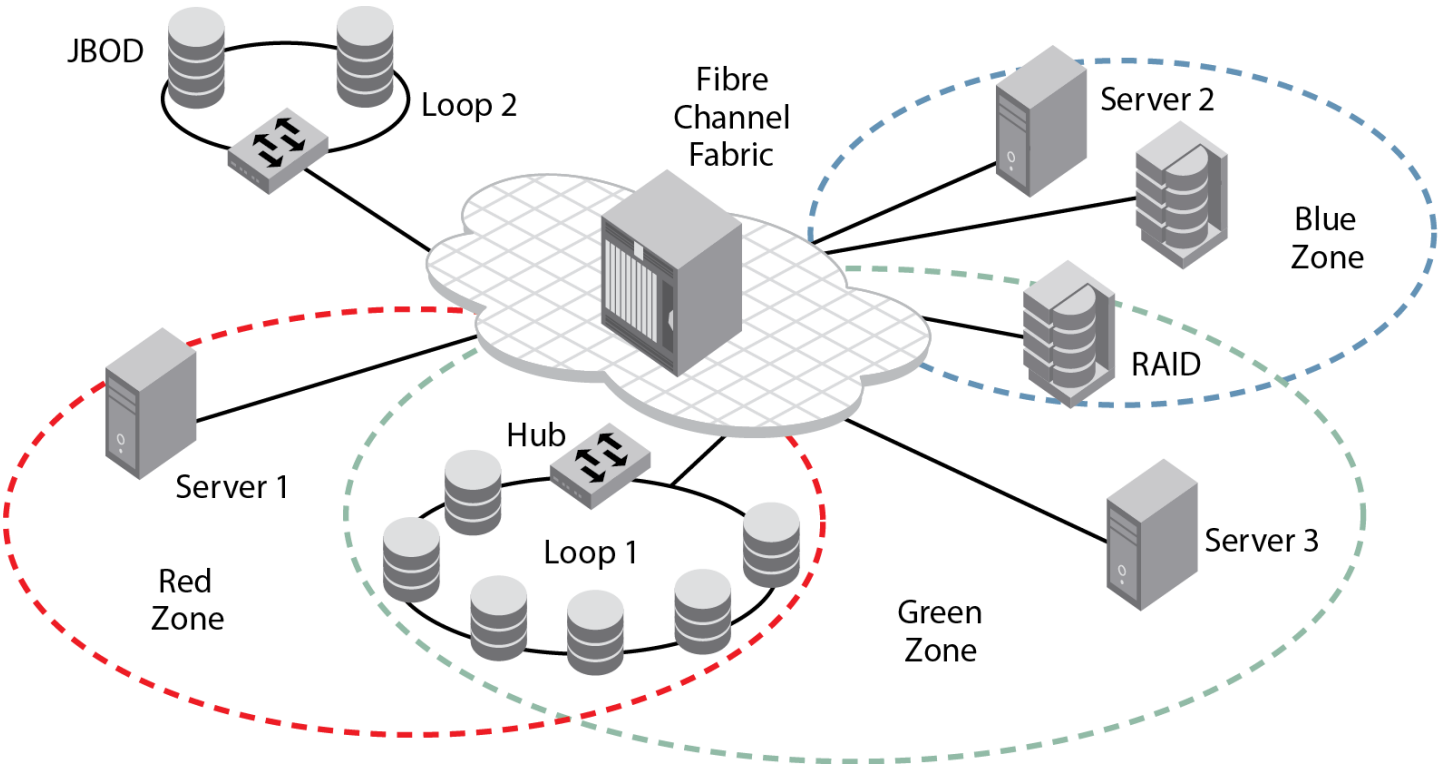
Zone Management: Duplicate WWNs

In a virtual environment like VMware or HPs Virtual Connect, it is possible to encounter duplicate WWNs in the fabric, most often as a transient condition. This impacts the switch response to fabric services requests like "get port WWN," resulting in unpredictable behavior, and represents a security risk by enabling spoofing of the intended target. The fabric's handling of duplicate WWNs is not meant to be an intrusion detection tool but a recovery mechanism. Prior to Brocade FOS v7.0, when a duplicate entry is detected, a warning message is sent to the RAS log, but no effort is made to prevent the login of the second entry.

Starting with Brocade FOS v7.0, handling of duplicate WWNs is as follows:

- Same switch: The choice of which device stays in the fabric is configurable (default is to retain existing device)
- Local and remote switches: Remove both entries
- Zoning recommendations include the following:
  - Always enable zoning.
  - Create zones with only one initiator (shown in Figure 42) and target, if possible.
  - Define zones using device WWPNs (World Wide Port Names).
  - Default zoning should be set to No Access.
- Use FA-WWN if supported by the HBA and HBA driver.
- Delete all FA-PWWNs (Fabric-Assigned Port World Wide Names) from the switch whose configuration is being replaced before you upload or download a modified configuration.
- Follow vendor guidelines for preventing the generation of duplicate WWNs in a virtual environment.

FIGURE 23 Example of Single Initiator Zones



Role-Based Access Controls (RBACs)

One way to provide limited accessibility to the fabric is through user roles. Brocade FOS has predefined user roles, each of which has access to a subset of the CLI commands. These are known as Role-Based Access Controls (RBAC), and they are associated with the user login credentials.

Access Control Lists (ACLs)

Access Control Lists are used to provide network security via policy sets. Brocade FOS provides several ACL policies including a Switch Connection Control (SCC) policy, a Device Connection Control (DCC) policy, a Fabric Configuration Server (FCS) policy, an IP Filter, and others. The following subsections briefly describe each policy and provide basic guidelines. A more in-depth discussion of ACLs can be found in the Brocade Fabric OS Administrator's Guide.

SCC Policy

The SCC policy restricts the fabric elements (FC switches) that can join the fabric. Only switches specified in the policy are allowed to join the fabric. All other switches will fail authentication if they attempt to connect to the fabric, resulting in the respective E_Ports being segmented due to the security violation.

Use the SCC policy in environments where there is a need for strict control of fabric members. Since the SCC policy can prevent switches from participating in a fabric, it is important to regularly review and properly maintain the SCC ACL.

DCC Policy

The DCC policy restricts the devices that can attach to a single FC Port. The policy specifies the FC port and one or more WWNs allowed to connect to the port. The DCC policy set comprises all of the DCC policies defined for individual FC ports. (Note that not every FC port has to have a DCC policy, and only ports with a DCC policy in the active policy set enforce access controls.) A port that is present in the active DCC policy set will allow only WWNs in its respective DCC policy to connect and join the fabric. All other devices will fail authentication when attempting to connect to the fabric, resulting in the respective F_Ports being disabled due to the security violation.

Use the DCC policy in environments where there is a need for strict control of fabric members. Since the DCC policy can prevent devices from participating in a fabric, it is important to regularly review and properly maintain the DCC policy set.

FCS Policy

Use the FCS policy to restrict the source of fabric-wide settings to one FC switch. The policy contains the WWN of one or more switches, and the first WWN (that is online) in the list is the primary FCS. If the FCS policy is active, then only the primary FCS is allowed to make and/or propagate fabric-wide parameters. These parameters include zoning, security (ACL) policies databases, and other settings.

Use the FCS policy in environments where there is a need for strict control of fabric settings. As with other ACL policies, it is important to regularly review and properly maintain the FCS policy.

IP Filter

The IP Filter policy is used to restrict access through the Ethernet management ports of a switch. Only the IP addresses listed in the IP Filter policy are permitted to perform the specified type of activity via the management ports.

The IP Filter policy should be used in environments where there is a need for strict control of fabric access. As with other ACL policies, it is important to regularly review and properly maintain the IP Filter policy.

Authentication Protocols

Brocade FOS supports both Fibre Channel Authentication Protocols (FCAPs) and Diffie-Hellman Challenge Handshake Authentication Protocols (DH-CHAPs) on E_Ports and F_Ports. Authentication protocols provide additional security during link initialization by assuring that only the desired device/device type is connecting to a given port.

Policy Database Distribution

Security Policy Database Distribution provides a mechanism for controlling the distribution of each policy on a per-switch basis. Switches can individually configure policies to either accept or reject a policy distribution from another switch in the fabric. In addition, a fabric-wide distribution policy can be defined for the SCC and DCC policies with support for strict, tolerant, and absent modes. This can be used to enforce whether or not the SCC and/or DCC policy needs to be consistent throughout the fabric.

- Strict mode: All updated and new policies of the type specified (SCC, DCC, or both) must be distributed to all switches in the fabric, and all switches must accept the policy distribution.
- Tolerant mode: All updated and new policies of the type specified (SCC, DCC, or both) are distributed to all switches (Brocade FOS v6.2.0 or later) in the fabric, but the policy does not need to be accepted.
- Absent mode: Updated and new policies of the type specified (SCC, DCC, or both) are not automatically distributed to other switches in the fabric; policies can still be manually distributed.

Together, the policy distribution and fabric-wide consistency settings provide a range of control on the security policies from little or no control to very strict control.

For a detailed discussion of SAN security concepts and issues, see the *Brocade Fibre Channel Security Best Practices* guide or the Brocade handbook Securing Fibre Channel Fabrics, by Roger Bouchard.

Capacity Planning

Gathering Requirements

The SAN project team should interview all stakeholders (IT application owners, finance, corporate facilities, IT lab administrators, storage and network administrators, and end users) who have a vested interest in the project-and this applies equally to planning for both new and updated SANs.

Application Owners

As critical stakeholders, application owners care because everyone is measured on application uptime. Application outages are something that users notice, and they can have severe financial impact for a business. With a redundant or a resilient infrastructure, hardware outages are transparent to the user, and only SAN administrators need to pay attention. To better understand their requirements, questions to ask the application owners include:

- What is the business goal for this application? (Is it a database that multiple applications rely on for business transactions?)
- What are the availability requirements?
- Is the application latency sensitive?
- Are there peak periods of utilization or other traffic patterns?
- What are the IOPS requirements in terms of read/writes?
- What is the worst-case response time before an outage?
- Is the application running on a cluster?
- Has the application been benchmarked to determine the CPU and memory resources required?
- Is there application downtime that can be used for applying patches, software upgrades, and maintenance?
- Can the application run on a VM? If so, how many other VMs can co-exist on the same physical hardware?

The business criticality of the application will determine the SAN design and the DR strategy, including backup and recovery. If the application is mission critical, the infrastructure must be fully redundant, with no single point of failure for both mainframe or distributed open systems architectures.

Server and Storage Administrators

Once the application requirements have been defined, identify physical server and storage on which the application and data will reside to determine the overall high-level architecture of the SAN, especially if this includes existing equipment as well as new equipment.

- Gather information about the server(s) on which the applications are running (blade or rack, CPU, memory, HBA/embedded FC switch, OS level, OS patch level, HBA driver version)?
- How many HBAs are in the rack servers?
- Does each server have single or multiple port HBAs?
- What is the primary storage for the application? Is there enough storage capacity to support this application and data?
- What is the desired storage response time (latency)? Will longer latency times resulting from a disk hit severely impact performance?

- For hybrid storage, how much storage is allocated to SSD?
- For HDD-based storage, what is the current cache utilization? Is there enough cache to meet required response times? What is the average drive utilization (the greater the utilization, the longer the response times) for HDDs and SSDs? Contact your drive vendor to identify response times based on utilization for sizing workloads.

Utilization	25%	50%	75%
Drive response time (ms or µs)			

- How much storage capacity is allocated to flash (more cache allows for more consistent response time) if using a hybrid array?
- Are storage tiers used in the environment? What is the policy used for migrating data? Are different tiers used for online storage? What is the impact?
- What is the raid level used? This will determine available disk space and performance for the application.
- How many FC ports are there in the array?
- Are the arrays front-ended by a storage virtualization controller? If so, what is the additional latency?
- What are the recommended fan-in and fan-out ratios for the arrays used for this application? What are the limits?
- Is there a Disaster Recovery (DR) site? If so, how is it connected (dark fiber, FCIP)?
- What is the available/required bandwidth between the intra-site for DR? Can the existing storage infrastructure support DR with the additional load?
- What tools are used for mirroring and replication (host-based or array-based)? If host-based, was the failover tested? If so, was there any impact in application uptime? If storage-based, was the failover tested? Did the LUNs appear on the active ports? Was there an impact to application uptime?

SAN Administrator: General

A SAN Administrator is responsible for the day-to-day operation of the network. The SAN design must be easy to monitor, manage, and maintain. If the current SAN is being expanded, adequate performance metrics should be collected to ensure that the existing design can be expanded to address new workloads.

- Are there performance (bandwidth) or latency issues in the existing SAN?
- Are procedures in place to address redistribution of capacity when switch port utilization exceeds 75 percent?
- Is the current design two-tier (core-edge) or three-tier (edge-core-edge)?
- Is the SAN centrally managed by a tool such as IBM Tivoli Netcool or HP OpenView?
- If there is an existing SAN, how is it managed (CLI, Brocade DCFM)? Is there a separate network for SAN management?
- Are access control policies in place for change management (zoning)? Is there a zoning policy? Are there devices in the zone database that no longer exist? What type of zoning is used (port or WWN)?
- Is the current SAN a redundant configuration?
- Is there an identified server to capture logs from the fabric?
- Is the traffic equally distributed across the ISLs or the trunks?
- Is historical performance data available for initiators, targets, and ISLs?
- How many unused ports are available per switch?

SAN Administrator: Backup and Restore

Backup and restore continue to be the primary drivers for SANs. As data growth continues to increase, backup windows continue to shrink. What is often overlooked is the restore time, which for some customers can take days.

Some topics to consider for backup and restore as you plan for SAN expansion or a new design are these:

- If the backup site is local, what is the window to complete the backup? If the backup site is remote, what is the window to complete the backup? How much of the bandwidth pipe is available?
- Is there a dedicated backup server, or do other applications share the server? Is the backup SAN on a separate SAN or a shared network?
- How often are full backups completed, and how long does it take? How often are backups checked for the integrity of the backup? How often do the backups fail to complete? What are the primary reasons (link down, tape drive failure, low throughput, other)? What is the restore time for Tier 1 and 2 applications?
- In a VM environment, is there a centralized proxy backup management, or does each VM have its own backup agent?
- Is a tiered backup implemented (disk, VTL, tape)?
- Is backup validation a regulatory requirement? If so, what processes are in place to ensure compliance?

NOTE
Brocade offers certification courses in *Open Systems and Mainframe SAN Design and Management*.

Facilities

Facility requirements are often overlooked as SANs grow due to business expansion or data center consolidation after mergers. Even when a SAN design meets application requirements, if physical plant, power, cooling, and cable infrastructure are not available, a logically designed SAN may have to be physically distributed, which can impact application performance and ongoing servicing.

Consider the following:

- Is there existing space for new SAN devices (servers, switches, and storage)? What is the physical real estate (floor space, number of racks, rack dimensions), and do the racks have internal fans for cooling?
- What is the available power (AC 120/240), and what is the in-cabinet power and plug type? Is it the same as existing types, or do you need new power supplies?
- What method of cooling is available (hot/cool aisle, other), and what is the worst-case temperature that the data center can tolerate?
- What is the cable infrastructure (OM-4, OM-3, other), and are cables already installed? Is an upgrade needed to achieve the desired cable run length at the required data rate?
- Is there a structured cable plant with patch panels, and so forth? If so, how many patch panels will the data traverse? Is the patch panel connector type (LC or MPO/MTP) appropriate to the desired cabling?
- Has the cabling infrastructure been tested for optical and signal integrity to ensure optimal performance?

Finance

Once the technical specifications have been determined, a reasonable cost estimate can be calculated based on available equipment, new purchases required, manpower, and training. Financial metrics for a total cost analysis should include the following:

- Lease versus buy
- Budget for equipment
- Budget for service and support (is 24x7 required?)
- Budget for daily operation
- Available administrative support

Tools for Gathering Data

Brocade Network Advisor

Brocade Network Advisor greatly simplifies daily operations while improving the performance and reliability of the overall SAN. This software management tool offers customizable dashboards, visibility into up to two years of historical data, and alert notifications to help you proactively monitor and manage your SAN network. As a result, you can optimize storage resources, maximize performance, and enhance the security of storage network infrastructures.

Brocade Network Advisor provides comprehensive management of data center fabrics, including configuration, monitoring, and management of the Brocade DCX 8510 and X6 director families, switches and routers. Brocade Network Advisor also integrates with leading storage partner data center automation solutions to provide end-to-end network visibility through frameworks such as the Storage Management Initiative-Specification (SMI-S).

Additional suggestions for Brocade Network Advisor usage to get maximum value include:

- Use dashboards for fast identification of potentially developing problems caused by faulty media or misbehaving devices, and use troubleshooting drill-down menus to identify root causes and remediation actions.
- Use integrations and free plugins where applicable between Brocade Network Advisor and VMware tool sets including vCenter, vRealize Operations Manager, and Log Insight to pull data from those applications to assist in performance visibility and troubleshooting. The vCenter integration point also enables a VMID to be retrieved by Brocade Network Advisor to enable the mapping of a location of that virtual machine to a physical server as a first step in addressing any operational issues.
- Use the modeling button that simulates fabric operation before committing a change and its potential impact on zoning and fabric merges that prevent potential mishaps like loss of connectivity to applications and storage devices.
- Save Brocade Network Advisor data out to an Open Database Connectivity (ODBC)-compliant data set every 5 days to be able to go back and identify short-term trends that are event driven or that occur seasonally and may not be readily identifiable across long-term data sets.
- Consider granting read-only access to certain frequent caller application owners to provide them insight into reporting tools and identify problems that may be developing within their domain, allowing them to address them up-front.

Brocade SAN Health

Brocade SAN Health is a free tool that allows SAN administrators to securely capture, analyze, and report comprehensive information about Brocade fabrics with switches running Brocade FOS and M-EOS operating systems and Cisco MDS fabrics running SANOS/NXOS. It can perform tasks such as:

- Taking inventory of devices, switches, firmware versions, and SAN fabrics
- Capturing and displaying historical performance data
- Comparing zoning and switch configurations against best practices
- Assessing performance statistics and error conditions
- Alerting the administrator of relevant Technical Support Bulletins (TSBs)
- Producing detailed reports (in Microsoft Excel) and diagrams (in Microsoft Visio)

NOTE
In mainframe FICON environments, collect the Input/Output Configuration Program (IOCP) in plain-text format (build I/O configuration statements from HCD), and upload the data. Brocade SAN Health matches the IOCP against the RNID data.

Download Brocade SAN Health and find details and instructions on how to use it at:

www.brocade.com/services-support/drivers-downloads/san-health-diagnostics/index.page

MAPS

Brocade Monitoring and Alerting Policy Suite (MAPS) provides policy-based threshold monitoring and alerting, and may be the single most valuable tool available to enable storage admins to meet their SLAs. MAPS monitors over 250 parameters across SAN fabrics as well as extended links using FCIP across a WAN for multiple categories. Some various examples of these include:

- Port Health
- Back-end Health
- FRU Health
- Security Violations
- Fabric State Changes
- Traffic Performance
- FCIP Health
- Fabric Performance Impact

MAPS policies can be configured using a set of rules that apply to groups of objects, and can be simply and easily distributed and updated across an environment. As conditions are monitored, exceeding a threshold will trigger a pre-defined action to take ranging from alerts to port fencing or device quarantine. Thresholds can be defined to respond using an aggressive, moderate or conservative parameter levels and can take different actions as threshold levels are exceeded, with increasingly strong actions taken as higher threshold levels are exceeded.

For configuration and usage details on MAPS, refer to the *Brocade Monitoring and Alerting Policy Suite Configuration Guide*.

Flow Vision

Flow Vision was designed as a diagnostics tool and is supported on all Brocade SAN platforms running Fabric OS 7.2 and later. Flow Vision provides the SAN administrator with visibility into fabric traffic flows and with the ability to copy traffic flows for later analysis. Flow Vision also allows for test-flow generation at line-rate speeds to pre-validate SAN hardware performance and connectivity. It is recommended to use the flow generation capability before operational deployment where possible to confirm optimal health and the ability to support spikes in throughput.

For the most mission-critical applications, consider running Flow Vision constantly to keep a historical record of the application performance profile and intermittent irregularities. For frequent callers such as critical application owners, run Flow Vision on a regular basis where time permits to verify good health.

IO Insight

IO Insight is a capability supported by Brocade's Gen 6 Fibre Channel switch products that provides even deeper flow-level IO statistics. These include storage device latency and IOPS metrics such as first IO response time, IO completion time, and number of pending IOs for a specific host and target or target/LUN, providing IO workload monitoring and early detection of storage performance degradation.

These IO Insight metrics should be added into MAPS policies and dashboards for notification of storage response time and performance degradation. This reporting is of tremendous value for performance-sensitive workloads, enabling administrators to meet their critical SLAs. IO Insight should be monitored for storage devices that support those critical apps to provide feedback to application and storage administrators on performance over time on device reliability and performance optimization. For example, pending IOs will measure current queue depth of an HBA and can be used to fine tune the server queue depth configuration.

For configuration and usage details on Flow Vision and IO Insight, refer to the *Brocade Flow Vision Configuration Guide*.

Analytics and Monitoring Platform

To attain the highest level of insight for large and performance-sensitive environments, Brocade offers the Analytics Monitoring Platform (AMP), an appliance-based engine that connects to the network non-disruptively via an Analytics Switch Link (ASL) to an analytics port connection. AMP monitors application traffic from across the entire fabric, such as read/write latency and IOPS transfer rate, IO queue depth, protocol errors, SCSI reserves and releases, and fabric latency. For larger and performance sensitive network infrastructures, the investment in a storage monitoring and analytics platform may be an investment that provides a high return.

Storage Traffic Patterns

Most storage arrays have tools for gathering port and LUN level performance data (contact the array vendor for the appropriate tool). It is recommended that gathering a week's worth of data will help in determining if the there are enough resources to accommodate the new application requirements.

The data should reflect both normal and high utilization, such as data that reflects the end of a quarter.

The metrics to collect are as follows:

- Percent of reads
- MB/s reads
- Percent of writes
- MB/s writes
- Worst-case latency (ms)
- Number of SCSI commands/second
- Cache hits (disk-based storage)
- Flash hits (hybrid architectures)
- Queue depth

Server Traffic Patterns

On the server side, there are Windows and UNIX tools for collecting CPU, memory, and network utilization built into the OS. HBA vendors also provide tools to gather the following on a per-port basis:

- Percent of reads
- MB/s reads
- Percent of writes
- MB/s writes
- Worst-case latency (ms)
- HBA queue depth

This is an example of a guideline for determining the queue depth for HBAs attached to an EMC array:

Queue depth value = 8*n/h

(where n= number of members in a metavolume group of disks, where within in the disk contiguous blocks are allocated; h = number of HBAs that can see the metavolume).

If there is an embedded switch in the server, the following information should be gathered:

- Tx frames
- Rx frames

- Total throughput

If the server hosts virtual machines, similar metrics should be collected per VM. As in the storage data collection, a week's worth of data should be collected during normal and highest utilization periods.

Backup Traffic Patterns

To understand the utilization of existing backup infrastructure, collect one week's worth of data, including when full backups are conducted. A table in Appendix B provides a template for capturing the physical infrastructure for backup.

Tape Library

If an existing SAN is used for backup, run CLI commands such as **portPerfShow** and **portStatsShow** for ports connected to the tape library, and use the library management utilities to collect traffic statistics to create a profile of the current environment, to determine the following:

- Low and high utilization periods
- Drives used most often
- Tape cartridges used most often
- Tape drive volume in MB/h

Backup Media Server

On the backup media server, collect CPU, memory, FC port, and Ethernet network utilization. This helps validate that the existing backup infrastructure is working as designed to meet the backup window. It can also help determine if media server performance is impacted in a VM environment. If backup performance is impacted by non-backup traffic in the fabric, use Traffic Isolation zones or increase the number of ISLs to improve performance.

Summary

Once the initial discussions with key stakeholders are complete, data should be analyzed to support an optimized SAN design given business drivers, funding, and available resources. Sometimes it can be difficult to analyze the requirements from various organizations, and creating a radar chart may help to visually analyze competing requirements from internal groups (see "Appendix B"). If edge switch count is increasing, consider consolidating to high-density core enterprise-level platforms, which increase port density while reducing power consumption and the number of domains to manage.

Appendix A: Important Tables

The following table shows the support distance based on cable type and data rates.

Speed Name	OM1 Link Distance 62.5-µm core and 200 MHZ*km	OM2 Link Distance 50-µm core and 500 MHz*km	OM3 Link Distance 50-µm core and 2000 MHz*km	OM4 Link Distance 50-µm core and 4700 MHz*km	OS1 Link Distance 9-µm core and ~infinite MHz*km
1GFC	300	500	860	*	10,000
2GFC	150	300	500	*	10,000
4GFC	50	150	380	400	10,000

Speed Name	OM1 Link Distance 62.5-µm core and 200 MHZ*km	OM2 Link Distance 50-µm core and 500 MHz*km	OM3 Link Distance 50-µm core and 2000 MHz*km	OM4 Link Distance 50-µm core and 4700 MHz*km	OS1 Link Distance 9-µm core and ~infinite MHz*km
8GFC	21	50	150	190	10,000
10GFC	33	82	300	*	10,000
16GFC	15	35	100	125	10,000
32GFC	-	20	70	100	10,000

TABLE 1 LWL Optics Support (SFP+)

Transceiver Data Rate (Gbps)	Distance (KM)
4	4, 10, & 30
8	10, 25
10	10
16	10
32	10

Appendix B: Matrices

This section provides example checklists and tables that you can use to identify dominant factors, including facilities that will have an impact on the SAN design.

TABLE 2 Current Fabrics

SAN/Fabric	# of Switches	Type of Switches	Total Ports	Domains	# of Servers	# of Storage Devices	Location	Notes
Fabric 1								
Fabric 2								
Fabric 3								
Fabric 4								
Fabric 5								

TABLE 3 Individual Fabric Details

SAN/Fabric	Domain Number	Serial Number	Model	Speed	WWN	IP Addresses	Brocade FOS/M-EOS Version	Notes
Switch 1								
Switch 2								
Switch 3								
Switch 4								
Switch 5								

TABLE 4 Device Details

Servers & Storage	Vendor	Model	WWN	Alias	Zone	OS Version	Application	Fabric/ Switches	Notes
Server 1									
Server 2									

TABLE 4 Device Details (continued)

Servers & Storage	Vendor	Model	WWN	Alias	Zone	OS Version	Application	Fabric/ Switches	Notes
Server 3									
Storage 1									
Storage 2									
Storage 3									

The following table details the metrics that need to be collected and their impact on SAN design and performance.

TABLE 5 Metrics and Impact on SAN Design and Performance

Metric	Source	Impact
Servers in the SAN	Estimate/Brocade SAN Health	Normal operations
Host Level Mirroring	Estimate	Distance, ISL congestion, traffic levels
Clusters (MSFT, HACMP, NetApp) Average number of nodes Workload level	Estimate Estimate: High/Med/Low	In-band heartbeat, frame congestion, host fan-in, traffic isolation
Virtualization: VIO Server # of servers Consolidation ratio	Estimate Estimate Estimate	Frame congestion, edge traffic increase/port, server fan-in on target ports, device latencies
Virtualization: VMware # of VMware servers Consolidated ratio Shared VMFS? DRS? RDM? I/O intensive	Estimate Estimate Yes/No Yes (%)/No Yes (%)/No High/Med/Low	Frame congestion, device latencies, and SCSI2 reservations

TABLE 6 Consolidated SAN Snapshot

SAN Requirements Data (Complete for Each SAN)	
Fabric Information	
Target # of user ports per fabric	
Target # of total ports per fabric	
Target # of switches per fabric (# switches/switch type, total switches)	
Number of fabrics	
Number of sites in environment	
Topology (core-edge, ring, mesh, other)	
Maximum hop count	
Expected growth rate (port count)	
Fabric licenses	
SAN Device Information	
Number/types of hosts and OS platforms	
Number/types of storage devices	

TABLE 6 Consolidated SAN Snapshot (continued)

SAN Requirements Data (Complete for Each SAN)	
Number/types of tapes	
Number/types of HBAs	
Other devices (VTL/deduplication appliance)	
Total number of SAN devices per fabric	
Customer requirement for failover/redundancy, reliability of SAN (multipathing software utilized)	
Application Details	
SAN Application (Storage Consolidation, Backup and Restore, Business Continuance)	
Fabric management application(s)	
Performance	
Maximum latency (ms)	
Targeted ISL oversubscription ratio (3:1, 7:1, 15:1, other)	

TABLE 7 Application-Specific Details

Backup/Restore Infrastructure		
Servers		
System	OS Version, Patch Level	HBA Driver Version
Server 1/HBA		
Server 2/HBA		
Server 3/HBA		
Backup Software		
Vendor	Version	Patch
FC Switch		
Vendor	Model	Firmware
Brocade		
Storage		
Vendor	Model	Firmware
Array 1		
Array 2		
Tape Library		
Vendor	Model	Firmware
Library		

NOTE

Keep a similar table for each application.

TABLE 8 Quantitative Analysis: Radar Maps

SAN/Storage Admin Concerns	Rank (1 is low, 10 is high)	Notes
ISL utilization	8	Is traffic balanced across ISLs during peaks?
Switch outage	1	Have there been switch outages? If so what was the cause?
Zoning policy	6	Is the zoning policy defined?
Number of switches in the fabric	10	Is the current number of switches a concern for manageability?

TABLE 8 Quantitative Analysis: Radar Maps (continued)

SAN/Storage Admin Concerns	Rank (1 is low, 10 is high)	Notes
Scalability	6	Can the existing design scale to support additional switches, servers, and storage?
Redundancy	10	Is the existing SAN redundant for supporting a phased migration or firmware update?
Server: High availability	10	Does the cluster software fail over reliably?
Storage: High availability	10	Do the LUNs fail over reliably?
Available disk pool	6	Is there sufficient disk pool to support additional apps?
Management tools for SAN	4	Are the right management tools used for SAN management?
Application response	7	Have there been any instances of slow application response but no outage?

FIGURE 24 SAN Admin Radar Map

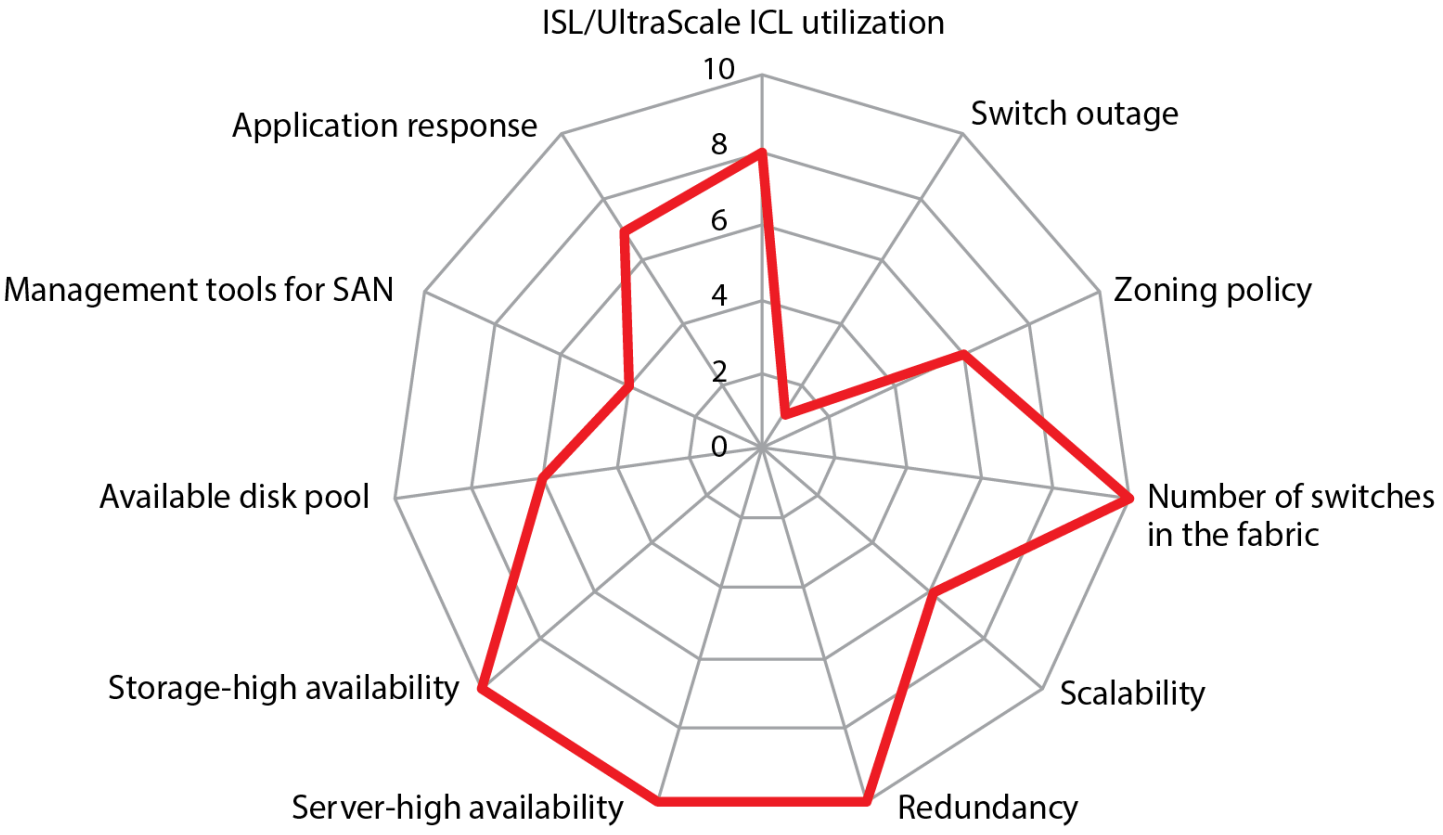


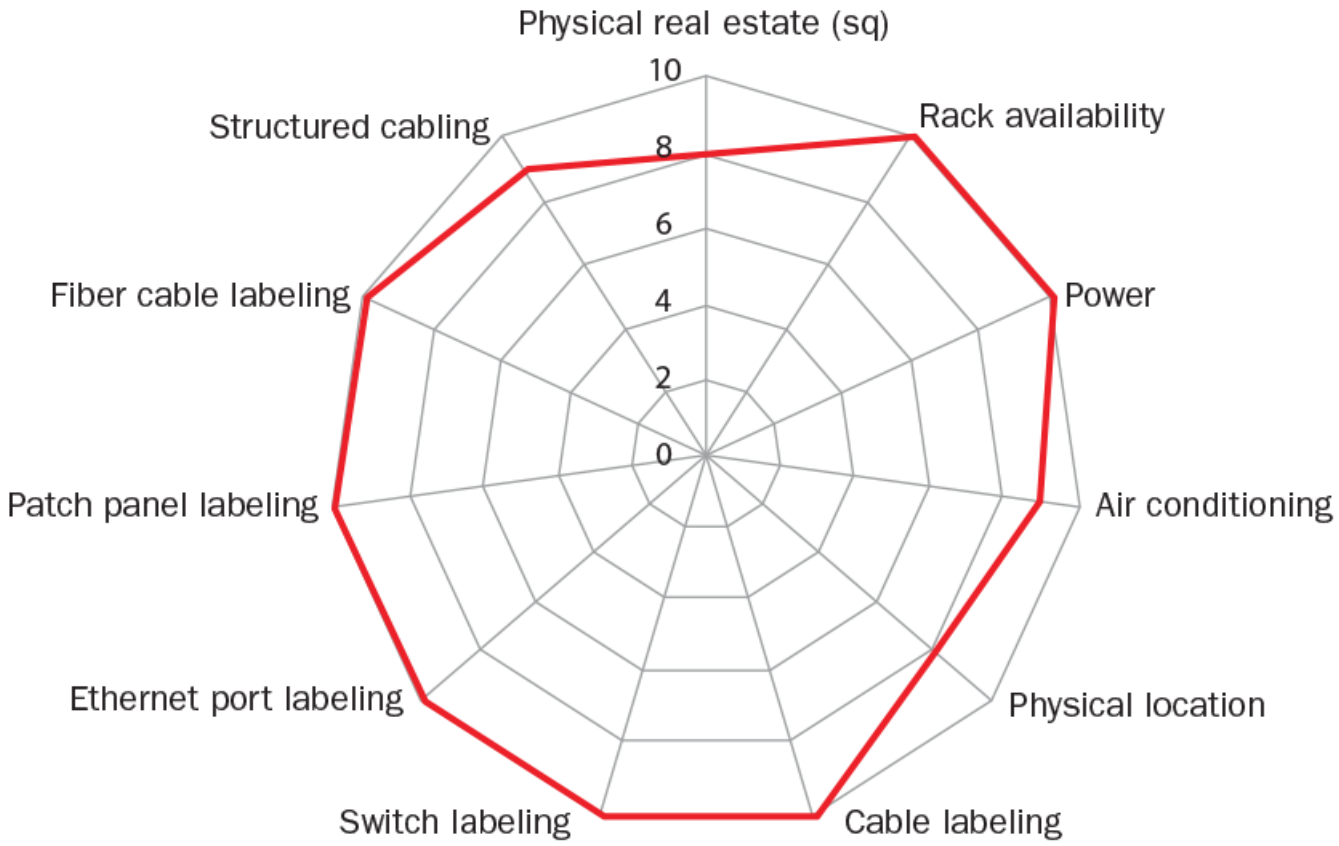
TABLE 9 Facilities Radar Map

Facility	Rank (1 is low, 10 is high)	Notes
Concern for physical real estate	8	What is the total available space for all the hardware?
Support racks	10	How many racks are needed?
Power	10	Is there adequate power?
Air conditioning	9	Is there adequate air conditioning?
Physical location	8	How important is it to have all the equipment in the same physical location or aisle?

TABLE 9 Facilities Radar Map (continued)

Facility	Rank (1 is low, 10 is high)	Notes
Cable labeling	10	Are cables labeled for easy identification?
Switch labeling	10	Are switches labeled for easy identification?
Ethernet port labeling	10	Are Ethernet ports labeled for easy identification?
Patch panel labeling	10	Are patch panels labeled for easy identification?
OM-3 fiber cables used	10	Are OM-3 fiber cables in use?
Structured cabling	9	Is structured cabling in place to support SAN expansion?

FIGURE 25 Facilities Radar Map



Appendix C: Port Groups

A port group is a group of eight ports, based on the user port number, such as 0-7, 8-15, 16-23, and up to the number of ports on the switch or port blade. Ports in a port group are usually contiguous, but they might not be. Refer to the hardware reference manual for your product for information about which ports can be used in the same port group for trunking. The *Brocade Fabric OS Administration Guide* provides guidelines for trunk group configuration. The ports are color-coded to indicate which can be used in the same port group for trunking (trunking port groups can be up to eight ports).

Director Platforms

The Brocade FC16-64, FC16-48, and FC16-32 blades for the Brocade DCX 8510 Backbone and the Brocade FC32-48 blade for the Brocade X6 Director platforms provide trunk groups with a maximum of 8 ports per trunk group. The trunking octet groups are in the following blade port ranges: 0-7, 8-15, 16-23, 24-31, 32-39, and 40-47. (Trunk groups 32-39 and 40-47 are not applicable to FC16-32.) Trunk boundary layout is on the faceplate of the blade, and trunk groups on the FC32-48 are color-coded for easy identification.

FIGURE 26 Brocade FC16-32 Trunk Groups

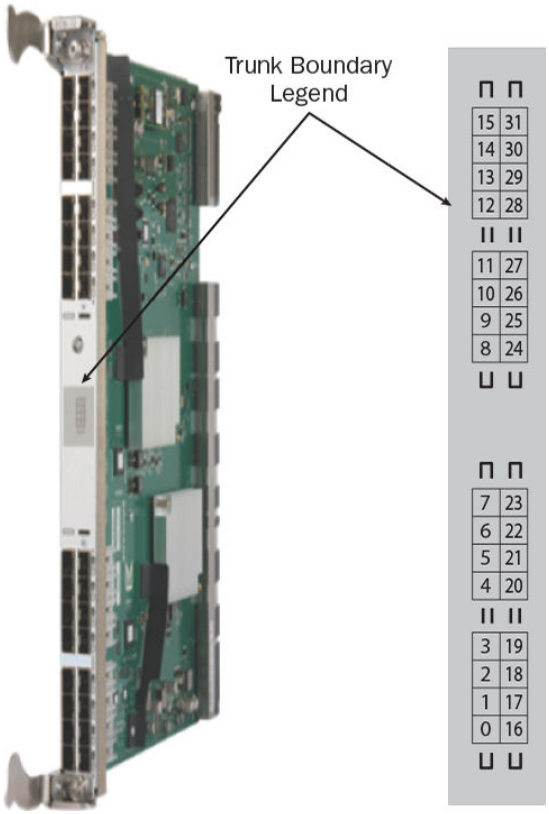
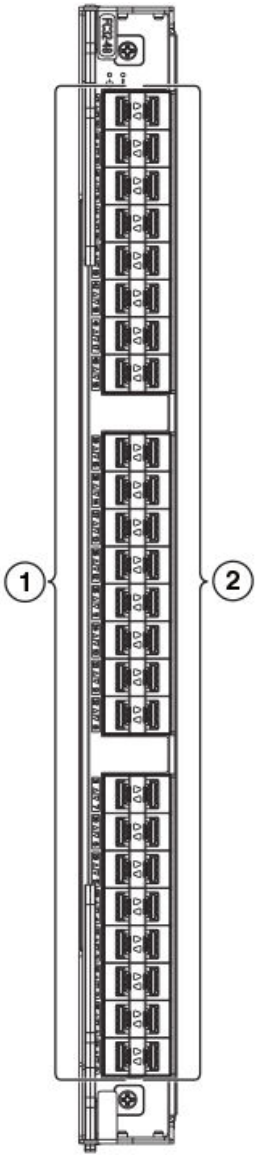


FIGURE 27 FC32-48 Blade Port Numbering



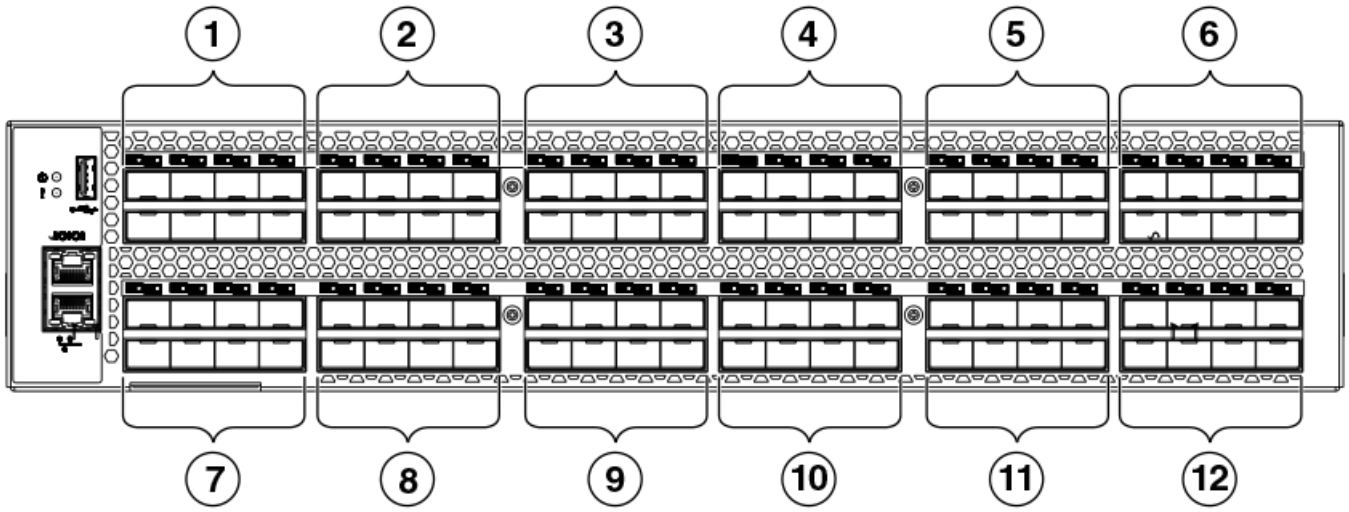
1. FC ports 0-23 (numbered bottom to top)
2. FC ports 24-47 (numbered bottom to top)

Switch Platforms

Trunk grouping on Brocade switches follows the same pattern of octet grouping as the blades based on the user port number, such as 0-7, 8-15, 16-23, and up to the number of ports on the switch.

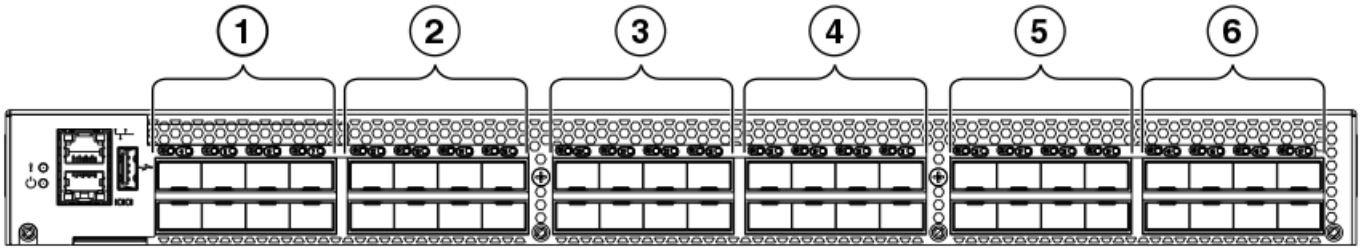
Brocade 6520 Trunk Groups

FIGURE 28 Brocade 6520 Front Port Groups



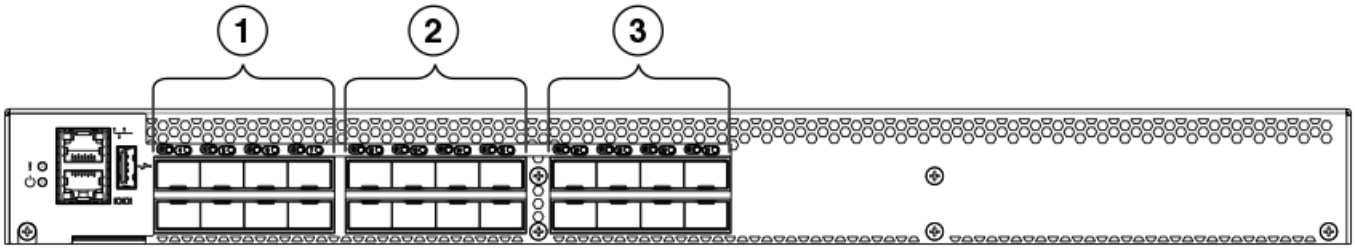
Brocade 6510 Trunk Groups

FIGURE 29 Brocade 6510 Front Port Groups



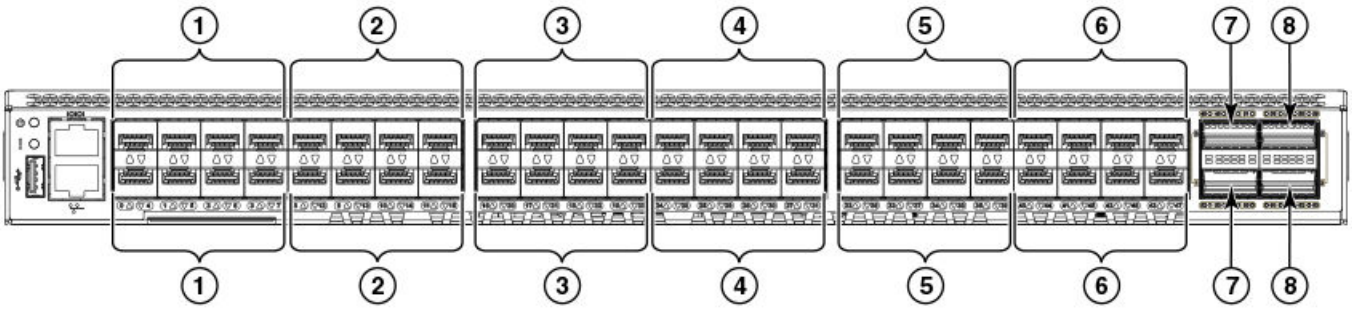
Brocade 6505 Trunk Groups

FIGURE 30 Brocade 6505 Front Port Groups



Brocade G620 Trunk Groups

FIGURE 31 Brocade G620 Trunk Port Groups



Appendix D: Terminology

Term	Brief Description
Base switch	Base switch of an enabled virtual fabric mode switch, providing a common communication infrastructure that aggregates traffic from Logical Switches in the physical switch in a common Base Fabric
ClearLink Diagnostics	Diagnostics tool that allows users to automate a battery of tests to verify the integrity of optical cables and 16 Gbps transceivers in the fabric
D_Port	A fabric port configured in ClearLink diagnostics testing mode for cable and optics integrity testing
Default switch	Default logical switch of an enabled virtual fabric mode switch, automatically created when Virtual Fabrics is enabled on a VF-capable switch
E_Port	A standard Fibre Channel mechanism that enables switches to network with each other
Edge Hold Time	Enables the switch to time out frames for F_Ports sooner than for E_Ports
EX_Port	A type of E_Port that connects a Fibre Channel router to an edge fabric
F_Port	A fabric port to which an N_Port is attached
FCIP	Fibre Channel over IP, which enables Fibre Channel traffic to flow over an IP link
FCR	Fibre Channel Routing, which enables multiple fabrics to share devices without having to merge the fabrics
IFL	Inter-Fabric Link, a link between fabrics in a routed topology
ISL	Inter-Switch Link, used for connecting fixed port and modular switches
Logical Switch	Logical Switch of an enabled virtual fabric mode switch, managed In the same way as physical switches and configurable in any mode
Oversubscription	A condition in which more devices might need to access a resource than that resource can fully support
Port group	A set of sequential ports that are defined (for example, ports 0-3)
QoS	Quality of Service traffic shaping feature that allows the prioritization of data traffic based on the SID/DID of each frame
Redundancy	Duplication of components, including an entire fabric, to avoid a single point of failure in the network (fabrics A & B are identical)
Resilience	Ability of a fabric to recover from failure, could be in a degraded state but functional (for example, ISL failure in a trunk group)
TI Zone	Traffic Isolation Zone, which controls the flow of interswitch traffic by creating a dedicated path for traffic flowing from a specific set of source ports
Trunk	Trunking that allows a group of ISLs to merge into a single logical link, enabling traffic to be distributed dynamically at the frame level

Term	Brief Description
UltraScale ICL	UltraScale Inter-Chassis Link, used for connecting director chassis (Gen 5 and Gen 6) without using front-end device ports
VC	Virtual channels, which create multiple logical data paths across a single physical link or connection
VF	Virtual fabrics, a suite of related features that enable customers to create a Logical Switch, a Logical Fabric, or share devices in a Brocade Fibre Channel SAN

Appendix E: References

Software and Hardware Product Documentation

Refer to the version for your particular Fabric OS release.

- *Brocade Fabric OS Release Notes*
- *Brocade Fabric OS Administration Guide*
- *Brocade Fabric OS Command Reference Manual*
- *Brocade Monitoring and Alerting Policy Suite Administration Guide*
- *Brocade Flow Vision Configuration Guide*
- *Brocade Fabric OS Access Gateway Administration Guide*
- *Brocade Fabric OS Upgrade Guide*
- *Brocade Fabric OS Extension Configuration Guide*
- *Brocade Fabric OS Troubleshooting and Diagnostics Guide*
- *Hardware Installation Guides for Backbones and Directors*
- *Brocade Network Advisor SAN User Manual*

Technical Briefs

- SAN Fabric Resiliency and Administration Best Practices

Brocade Fabric OS v7.x Compatibility Matrix

- Brocade Fabric OS v7.x Compatibility Matrix
- Brocade SAN Scalability Guidelines: Brocade Fabric OS v7.X
- Brocade Fabric OS Target Path Selection Guide

Brocade SAN Health

- www.brocade.com/services-support/drivers-downloads/san-health-diagnostics/index.page

Brocade Bookshelf

- *Principles of SAN Design* (updated in 2007) by Josh Judd
- *Strategies for Data Protection* by Tom Clark

- *Securing Fibre Channel Fabrics* (updated in 2012) by Roger Bouchard
- *The New Data Center* by Tom Clark

Other

- www.snia.org/education/dictionary
- www.vmware.com/pdf/vi3_san_design_deploy.pdf
- www.vmware.com/files/pdf/vcb_best_practices.pdf