White Paper

Advanced SAN Design Using Cisco MDS 9500 Series Multilayer Directors

Cisco Systems® released the first-generation Cisco® MDS 9000 Family high-density, multiprotocol, intelligent storage area network (SAN) switches in December 2002. The switches offered intelligent features and functionality, with port densities higher than those previously existing in the Fibre Channel switching marketplace.

The second-generation Cisco MDS 9000 Family linecards, supervisors, and chassis add additional intelligence to the SAN switching fabric, more than doubling the port densities offered by other SAN switches, while preserving investment protection through compatibility with existing chassis, linecards, and supervisors.

This white paper provides technical guidance for the factors to consider when designing a SAN.

AUDIENCE

Escalating storage requirements, rising management costs, the requirement to share information, and increasing levels of underutilized disk storage resources are encouraging the consolidation of storage resources and the migration from direct-attached storage (DAS) to SANs. SANs provide the basis for managing increased storage costs by efficiently pooling and scaling storage resources, enabling:

- Efficient utilization of storage resources
- Multiple system access to multiple storage devices, regardless of location
- Consolidation of multiple low-end storage systems into centralized, high-end systems
- Reduced administrative overhead with centralized storage management

The aim of this white paper is to provide SAN architects and storage administrators with the background knowledge to use technical features and innovations within Cisco MDS 9000 Family switches to decrease the overall cost and complexity of SAN deployments while providing high levels of availability and manageability.

SAN DESIGN

SAN design doesn't have to be rocket science. Modern SAN design is about deploying ports and switches in a configuration that provides flexibility and scalability. It is also about making sure the network design and topology look as clean and functional one, two, or five years later as the day they were first deployed.

Problems of Early SAN Designs

The first SAN deployments hardly qualified as networks. Built using fixed-configuration 8-port or 16-port switches, they offered limited port counts, no fault isolation or fabric segmentation functionality, limited switch-to-switch connectivity with minimal traffic load-balancing, and no traffic management capabilities. Management tools focused on element management rather than network management.

Fibre Channel topologies have evolved a long way from the early SAN days. SANs started as simple single-switch topologies to provide additional connectivity to storage devices. Storage vendors bundled small Fibre Channel switches with their storage arrays to improve bandwidth, I/Os per second (IOPS), and array connectivity fan-out capabilities.

As storage requirements continued to grow exponentially, the need to optimize storage usage by way of consolidation and better utilization became necessary as a method to reduce overall total cost of ownership (TCO). This growth has led to the need for more scalable and flexible SAN topologies to meet increasing storage requirements and increasing performance requirements of computing platforms.

Principles of SAN Design

The underlying principles of SAN design are relatively straightforward: plan a network topology that can handle the number of ports necessary now and into the future; design a network topology with a given end-to-end performance and throughput level in mind, taking into account any physical requirements of a design (for example, whether the data center is or will in the future be located on multiple floors of a building or in multiple buildings or locations); and provide the necessary connectivity with remote data centers to handle the business requirements of business continuity and disaster recovery.

These underlying principles fall into five general categories:

- **Port density and topology requirements**—Number of ports required now and in the future
- **Device performance and oversubscription ratios**—Determination of what is acceptable and what is unavoidable
- **Traffic management**—Preferential routing or resource allocation
- **Fault isolation**—Consolidation while maintaining isolation
- **Control plane scalability**—Reduced routing complexity

Port Density and Topology Requirements

The single most important factor in determining the most suitable SAN design is determining the number of end ports—both now and over the anticipated lifespan of the design. As an example, the design for a SAN that will handle a network with 100 end ports will be very different from the design for a SAN that has to handle a network with 1500 end ports.

From a design standpoint, it is typically better to overestimate the port count requirements than to underestimate them. Designing for a 1500-port SAN does not necessarily imply that 1500 ports need to be purchased initially, or even ever at all. It is about helping ensure that a design remains functional if that number of ports is attained, rather than later finding the design is unworkable. As a minimum, the lifespan for any design should encompass the depreciation schedule for equipment, typically three years or more. Preferably, a design should last longer than this, because redesigning and reengineering a network topology become both more time-consuming and more difficult as the number of devices on a SAN expands.

Where existing SAN infrastructure is present, determining the approximate port count requirements is not difficult. You can use the current number of end-port devices and the increase in number of devices during the previous 6, 12, and 18 months as rough guidelines for the projected growth in number of end-port devices in the future.

For new environments, it is more difficult to determine future port-count growth requirements, but once again, it is not difficult to plan based on an estimate of the immediate server connectivity requirements, coupled with an estimated growth rate of 30 percent per year.

A design should also consider physical space requirements. For example, is the data center all on one floor? Is it all in one building? Is there a desire to use lower-cost connectivity options such as iSCSI for servers with minimal I/O requirements? Do you want to use IP SAN extension for disaster recovery connectivity? Any design should also consider increases in future port speeds, protocols, and densities. Although it is difficult to predict future requirements and capabilities, unused module slots in switches that have a proven investment protection record open the possibility to future expansion.

Device Performance and Oversubscription Ratios

Oversubscription is a necessity of any networked infrastructure and directly relates to the major benefit of a network—to share common resources among numerous clients. The higher the rate of oversubscription, the lower the cost of the underlying network infrastructure and shared resources. Because storage subsystem I/O resources are not commonly consumed at 100 percent all the time by a single client, a fan-out ratio of storage subsystem ports can be achieved based on the I/O demands of various applications and server platforms. Most major disk subsystem vendors provide guidelines as to the recommended fan-out ratio of subsystem client-side ports to server connections. These recommendations are often in the range of 7:1 to 15:1.

When considering all the performance characteristics of the SAN infrastructure and the servers and storage devices, two oversubscription metrics must be managed: IOPS and network bandwidth capacity of the SAN. The two metrics are closely related, although they pertain to different elements of the SAN. IOPS performance relates only to the servers and storage devices and their ability to handle high numbers of I/O operations, whereas bandwidth capacity relates to all devices in the SAN, including the SAN infrastructure itself. On the server side, the required bandwidth is strictly derived from the I/O load, which is derived from factors including I/O size, percentage of reads versus writes, CPU capacity, application I/O requests, and I/O service time from the target device. On the storage side, the supported bandwidth is again strictly derived from the IOPS capacity of the disk subsystem itself, including the system architecture, cache, disk controllers, and actual disks.

In most cases, neither application server host bus adapters (HBAs) nor disk subsystem client-side controllers are able to handle full wire-rate sustained bandwidth. Although ideal scenario tests can be contrived using larger I/Os, large CPUs, and sequential I/O operations to show wire-rate performance, this is far from a practical real-world implementation. In more common scenarios, I/O composition, server-side resources, and application I/O patterns do not result in sustained full-bandwidth utilization. Because of this fact, oversubscription can be safely factored into SAN design. However, you must account for burst I/O traffic, which might temporarily require high-rate I/O service. The general principle in optimizing design oversubscription is to group applications or servers that burst high I/O rates at different time slots within the daily production cycle. This grouping can examine either complementary application I/O profiles or careful scheduling of I/O-intensive activities such as backups and batch jobs. In this case, peak time I/O traffic contention is minimized, and the SAN design oversubscription has little effect on I/O contention.

Best-practice would be to build a SAN design using a topology that derives a relatively conservative oversubscription ratio (for example, 8:1) coupled with monitoring of the traffic on the switch ports connected to storage arrays and Inter-Switch Links (ISLs) to see if bandwidth is a limiting factor. If bandwidth is not the limited factor, application server performance is acceptable, and application performance can be monitored closely, the oversubscription ratio can be increased gradually to a level that is both maximizing performance while minimizing cost.

Traffic Management

Are there any differing performance requirements for different application servers? Should bandwidth be reserved or preference be given to traffic in the case of congestion? Given two alternate traffic paths between data centers with differing distances, should traffic use one path in preference to the other?

For some SAN designs it makes sense to implement traffic management policies that influence traffic flow and relative traffic priorities.

Fault Isolation

Consolidating multiple areas of storage into a single physical fabric both increases storage utilization and reduces the administrative overhead associated with centralized storage management. The major drawback is that faults are no longer isolated within individual storage areas. Many organizations would like to consolidate their storage infrastructure into a single physical fabric, but both technical and business challenges make this difficult.

Technology such as virtual SANs (VSANs) enables this consolidation while increasing the security and stability of Fibre Channel fabrics by logically isolating devices that are physically connected to the same set of switches. Faults within one fabric are contained within a single fabric (VSAN) and are not propagated to other fabrics.

Control Plane Scalability

A SAN switch can be logically divided into two parts: a data plane, which handles the forwarding of data frames within the SAN; and a control plane, which handles switch management functions, routing protocols, Fibre Channel frames destined for the switch itself such as Fabric Shortest Path First (FSPF) routing updates and keepalives, name server and domain-controller queries, and other Fibre Channel fabric services.

Control plane scalability is the primary reason storage vendors set limits on the number of switches and devices they have certified and qualified for operation in a single fabric. Because the control plane is critical to network operations, any service disruption to the control plane can result in business-impacting network outages. Control plane service disruptions (perpetrated either inadvertently or maliciously) are possible, typically through a high rate of traffic destined to the switch itself. These result in excessive CPU utilization and/or deprive the switch of CPU resources for normal processing. Control plane CPU deprivation can also occur when there is insufficient control plane CPU relative to the size of the network topology and a network-wide event (for example, loss of a major switch or significant change in topology) occurs.

FSPF is the standard routing protocol used in Fibre Channel fabrics. FSPF automatically calculates the best path between any two devices in a fabric through dynamically computing routes, establishing the shortest and quickest path between any two devices. It also selects an alternative path in the event of failure of the primary path. Although FSPF itself provides for optimal routing between nodes, the Dijkstra algorithm on which it is commonly based has a worst-case running time that is the square of the number of nodes in the fabric. That is, doubling the number of devices in a SAN can result in a quadrupling of the CPU processing required to maintain that routing.

A goal of SAN design should be to try to minimize the processing required with a given SAN topology. Attention should be paid to the CPU and memory resources available for control plane functionality and to port aggregation features such as Cisco PortChannels, which provide all the benefits of multiple parallel ISLs between switches (higher throughput and resiliency) but only appear in the topology as a single logical link rather than multiple parallel links.

ADVANCED CISCO MDS FEATURES USED TO SIMPLIFY SAN DESIGN AND DEPLOYMENT

A common practice in many organizations embarking on SAN deployment is to build numerous separate SANs rather than one larger consolidated SAN. This practice has generally been followed for many reasons relating to both products and procedures.

From a product perspective, many organizations have experienced instability caused by limited scalability of existing SAN switches and simply could not build larger consolidated fabrics. This is still true today in companies where many switches have inadequate CPU and memory resources, rendering them unable to participate in larger fabrics requiring larger control plane processing. For this reason, many organizations built numerous smaller fabrics. This practice is a costly one because of the wasteful nature of dedicating entire physical fabrics and their ISLs to a subset of applications, along with the inability to move free ports from one fabric to another application in need on another fabric.

The procedural reasons for building separate SANs are more complicated to overcome. Many organizations deployed separate applications on separate SANs to enforce security and management domain requirements, so that a given SAN administrator could be assigned to manage a given fabric independently from the other fabrics. Separate SANs also meant that changes to the storage configuration in one application area did not require coordinating downtime across multiple applications. Any inadvertent misconfiguration won’t result in downtime for other applications. Interoperability concerns still exist with SANs today, and for this reason many applications, storage types, and operating systems are often isolated on different fabrics, again for stability reasons. These are difficult problems to overcome in traditional fabrics.

Although there are numerous product and procedural reasons for building separate SANs, doing so does not come without additional costs. Some SANs might be full (no unused ports left), meaning that additional capital investment is required for expansion, even if there are other SANs with free (unused) ports that could otherwise be used. As such, SANs typically end up with stranded ports—unused ports in one fabric that cannot be used where they are required.

The solution to the costly deployment problem of numerous SANs is not solved by simply merging all environments together into one logical and physical fabric. The ideal model is to be able to logically replicate the isolation, stability, and multiple management scopes offered by the SANs atop a common physical fabric in which ports can be easily deployed within fabrics in need. This solution is best termed as 'virtualizing the fabric' and involves an end-to-end architecture supported by integrated hardware and software mechanisms within the intelligent SAN switches themselves.

VSANs

Fabric virtualization is used to provide hardware-based traffic segregation and control plane fabric services on a virtual fabric basis. Cisco fabric virtualization is available on a per-port basis and uses VSANs, a technology that is now part of the ANSI T11 standard for virtual fabrics. Cisco's VSAN technology works by prepending a virtual fabric header tag onto each frame as it enters the switch, with the tag indicating on which VSAN the frame arrived. This tag is subsequently used when the switch makes distributed hardware forwarding decisions.

There are fundamental differences between the Cisco approach to fabric virtualization and those available from other vendors: Cisco frame forwarding application-specific integrated circuits (ASICs) are virtual fabric aware, using the industry-standard Virtual Fabric Trunking (VFT) extended headers (FC-FS-2 standard, section 10.2) in forwarding decisions. Other switches either do not offer any fabric virtualization capabilities or populate different forwarding tables into different linecards, essentially building a per-line-card partitioning feature out of inconsistently programming forwarding tables.

VSANs can be used to provide any form of logical separation: production, development, test, open-systems hosts, tape, replication, cross-site connectivity, segregation by operating system, and so on.

VSAN Trunking

VSAN Trunking is the ability to trunk multiple VSANs across a single ISL or group of ISLs, enabling a common group of ISLs to be used as a pool for connectivity for multiple fabrics between switches. VSAN Trunking uses industry-standard VFT Extended Headers to provide traffic segregation across common trunked ISLs.

The primary benefit of VSAN Trunking is in consolidating and reducing the number of distinct ISLs required between switches. For example, for organizations that have multiple fabrics between data centers (for example, individual fabrics for synchronous replication, IBM Fiber Connection [FICON], and open-systems tape), VSAN Trunking enables a common pool of ISLs to be used, reducing the number of individual ISLs. This typically results in substantial cost savings through a reduction in the number of dense wavelength-division multiplexing (DWDM) transponders or dark fibre pairs necessary to transport all the fabrics between sites. Furthermore, because individual fabrics often have very different load profiles, grouping them together can result in a higher overall throughput because individual fabrics can burst at the rate of all of the ISLs. Where priority needs to be given to specific traffic or devices, quality of service (QoS) can be combined with VSAN Trunking to provider a bandwidth allocation guarantee for specific devices or VSANs.

Inter-VSAN Routing

The Cisco approach to SAN routing is called Inter-VSAN Routing (IVR). IVR provides the ability to selectively route traffic between different VSANs without merging the fabrics together. IVR can be used to simplify SAN designs and provide resiliency for many types of deployments: isolated fabrics for change control; primary data center and secondary data center connectivity; Storage Service Provider (SSP) or management segregation; consolidated tape; and separation of production, development, and test environments with selected connectivity to centralized storage devices.

There are fundamental difference between the Cisco approach to SAN routing and those available from other vendors. IVR makes use of Fibre Channel Network Address Translation (FC-NAT) functionality built into the distributed frame forwarding silicon on every linecard, and can be deployed with no loss of performance or additional latency. Other SAN switch vendors' implementations require external gateway fabric switches, which both consume switch ports (for ISLs to and from partitions and switches) and offer lower performance and higher latency because of traffic having to loop through the gateway device. External gateway devices also typically have their own (non-integrated) proprietary management applications and interfaces which increases management complexity. Furthermore, because the external gateway device is commonly a fabric switch that is not highly available, multiple gateway devices need to be deployed to achieve a highly available SAN routing solution.

Port Bandwidth Reservations

The Cisco MDS 9000 Family first-generation line-card modules uniquely offered the choice of both performance-optimized (non over-subscribed, non-blocking) and host-optimized (over-subscribed, non-blocking) Fibre Channel line-card modules. This provides a choice in switching module density, performance, and flexible price per port and enables SAN designers to build larger SAN fabrics with fewer switches while still maintaining reasonable host-to-storage oversubscription ratios.

Cisco second-generation line-card modules continue to provide this flexibility, but rather than hardcoding performance characteristics of individual modules in the line-card module itself, switch forwarding resources can be dynamically configured to provide dedicated switch forwarding resources to individual ports that need it. This enables the mixing and matching of high-performance devices and devices with lesser performance requirements within port groups on the same linecard.

Each of 12-port 1/2/4 Gbps, 24-port 1/2/4 Gbps and 48-port 1/2/4 Gbps line-cards has four port groups with 12.8 Gbps of forwarding resource shared among all front-panel ports that make up that port group. The 12-port linecard has 3 front-panel ports per port group, the 24-port linecard has 6 front-panel ports per port group, and the 48-port linecard has 12 front-panel ports per port group (Figure 1).

Figure 1. Port Group on a 48-port 1/21/2/4 Gbps Linecard Module

By default, a 12-port linecard offers line-rate throughput at all of 1/2/4 Gbps. The 24-port linecard offers line-rate throughput at 1/2 Gbps but is 2:1 over-subscribed at 4 Gbps if all ports in the same port group are attempting to push line-rate traffic simultaneously. The 48-port linecard offers line-rate throughput at 1 Gbps but is 2:1 and 4:1 over-subscribed respectively for 2 Gbps and 4 Gbps, if all ports in the same port group attempt to push line-rate traffic simultaneously (Table 1).

Table 1. Forwarding Performance for Cisco MDS 9000 Family Second-Generation 1/21/2/4 Gbps Linecards

Cisco Part Number	Line-Card Description	Number of Front-Panel Ports per 12.8-Gbps Port Group	Default Data Link Layer Bandwidth Allocation per Port	Data Link Layer Bandwidth Requirements for Line-Rate 1-Gbps Fibre Channel		Data Link Layer Bandwidth Requirements for Line-Rate 2-Gbps Fibre Channel		Data Link Layer Bandwidth Requirements for Line-Rate 4-Gbps Fibre Channel	
				Bandwidth	Ratio	Bandwidth	Ratio	Bandwidth	Ratio
DS-X9124	12-port 1/21/2/4-Gbps Fibre Channel module	3	4.25 Gbps	0.85 Gbps*	1:1	1.70 Gbps*	1:1	3.40 Gbps*	1:1
DS-X9124	24-port 1/21/2/4-Gbps Fibre Channel module	6	2.125 Gbps						2:1
DS-X9148	48-port 1/21/2/4-Gbps Fibre Channel module	12	1.0625 Gbps				2:1		4:1

* 1-Gbps, 2-Gbps, and 4-Gbps Fibre Channel are 1.0625 Gbps, 2.125 Gbps, and 4.25G bps at the physical link layer. They are all encoded at 8b/10, and thus actual data-link-layer bandwidth requirements are 0.85 Gbps, 1.7 Gbps, and 3.4 Gbps, respectively.

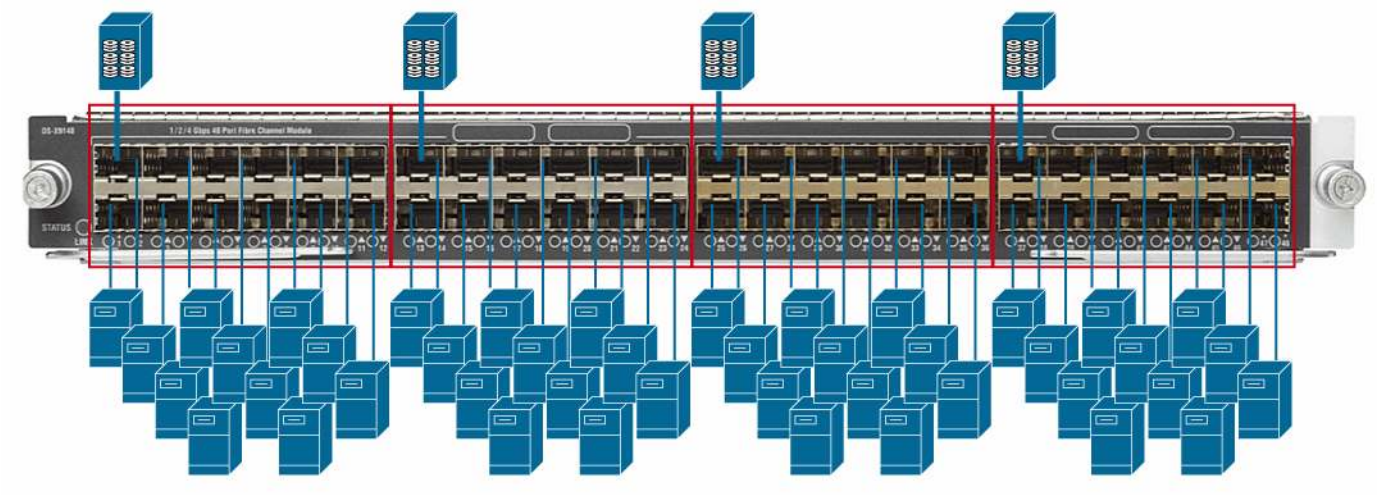
Port bandwidth reservations enable bandwidth to be dedicated to individual ports such that they are capable of sustaining line-rate 4 Gbps on 24-port linecards and line-rate 2 Gbps and 4 Gbps on 48-port linecards. That is, both the 24-port linecard and the 48-port linecard can be configured such that they reserve capacity for up to 12 ports of line-rate traffic at 4 Gbps.

This flexibility allows mixing and matching of devices that require high performance (for example, ports attached to storage arrays and ISLs) alongside devices that do not require sustained high throughput (for example, ports attached to hosts). In reality, even the highest performance SAN-attached devices do not require dedicated 100 percent line-rate capacity, but rather only require the ability to burst to 100 percent line rate.

As an example, let us say that we are provisioning storage resources at a ratio of one storage port for every 11 host ports (11:1 oversubscription). Using 48-port linecards for connectivity, you could configure the linecard to dedicate 4 Gbps of capacity to each port attached to a storage array and share the remaining ~9.4 Gbps across the remaining 11 ports in a 12-port port group (Figure 2).

On an end-to-end basis, overall oversubscription is 11:1—one storage-attached port for every 11 host-attached ports. Within a single port group, bandwidth is dedicated for traffic to and from the storage array (4 Gbps dedicated, helping ensure no oversubscription), and the remaining 11 ports for hosts share ~9.4 Gbps of forwarding resource capacity. This equates to an oversubscription ratio of ~2.2:1 at 2 Gbps or ~4.4:1 at 4 Gbps—a lower oversubscription ratio than what is desired end-to-end (11:1), thus not affecting the overall design.

Figure 2. Port Bandwidth Reservations Enable Dedicated Switch Forwarding Resources for Devices that Require Sustained High Performance (e.g. Storage and ISLs) and Sharing of Forwarding Resources between Devices that Don't Require Sustained Line-Rate



As shown in Figure 2, port bandwidth reservations are a very important SAN design feature that enables high-density, high-performance SAN designs to be built.

Cisco PortChannels

PortChannels are an ISL aggregation feature that can be used to construct a single logical ISL between switches from up to 16 physical ISLs. From a SAN design perspective, this is useful in terms of providing both higher-throughput connectivity between switches and a high-resiliency connection. PortChannels can be built using ports from any linecard on any module. Traffic across a PortChannel bundle is automatically load-balanced according to a per-VSAN load-balancing policy, providing very granular traffic distribution across physical interfaces within a PortChannel bundle while preserving in-order delivery.

With PortChannels, physical interfaces can be both added to and removed from a PortChannel bundle without interrupting the flow of traffic, enabling nondisruptive configuration changes to be made to production environments.

Multiprotocol: SAN Connectivity Using Integrated iSCSI and IP SAN Extension Using Fibre Channel over IP

iSCSI

iSCSI enables hosts to connect to SAN-attached storage at a far lower price point than could be achieved using native Fibre Channel connectivity. iSCSI uses Ethernet and TCP/IP as its transport, with data being passed over existing IP-based host connectivity. Because hosts can use their existing IP and Ethernet network connections to access storage elements, storage consolidation efforts can be extended to the midrange server class at a relatively lower price point compared to Fibre Channel, while improving the utilization and scalability of existing storage devices.

iSCSI gateway capabilities can be enabled on Cisco MDS 9000 Family switches using 8-port IP services modules, 4-port IP services modules, and 2-port multiprotocol services modules.

Fibre Channel over IP

Over short distances, such as within a data center, SANs are typically extended over optical links with multimode optical fiber. As the distance increases, such as within a large data center or campus, single-mode fiber or single-mode fiber with coarse wavelength-division multiplexing (CWDM) is typical. Over metropolitan distances Dense Wave Division Multipexing (DWDM) is preferable. DWDM is also used where higher consolidation density or aggregation of FICON, Enterprise Systems Connection (ESCON), and 10-Gigabit Ethernet data center links is required. In contrast, Fibre Channel over IP (FCIP) can be used to extend a Fibre Channel SAN across any distance. FCIP can be used over metro and campus distances or over intercontinental distances where IP might be the only transport available.

SAN extension using FCIP typically has many cost benefits over other SAN extension technologies. It is relatively common to have existing IP infrastructure between data centers that can be leveraged at no incremental cost. Additionally, IP connectivity is typically available at a better price point for long-distance links compared to pricing for optical Fibre Channel transport services.

FCIP is a means of providing a SAN extension over an IP infrastructure, enabling storage applications such as asynchronous data replication, remote tape vaulting, and host initiator to remote pooled storage to be deployed irrespective of latency and distance. FCIP tunnels Fibre Channel frames over an IP link, using TCP to provide a reliable transport stream with a guarantee of in-order delivery.

Cisco MDS 9000 Family switches offer a number of enhancements on top of standards-based FCIP to improve the functionality and usability enabled by FCIP:

- IVR is included with IP SAN extension functionality. IVR enables IP SAN extension without compromising fabric stability and reliability by allowing for routing between SAN extensions and sites without the need to create a common merged fabric.
- Fibre Channel traffic can be very bursty, and traditional TCP can amplify that burstiness. With traditional TCP, the network must absorb these bursts through buffering in switches and routers. Packet drops occur when there is insufficient buffering at these intermediate points. To reduce the probability of drops, Cisco MDS 9000 switches use traffic shaping to reduce the burstiness of the TCP traffic leaving the Gigabit Ethernet interfaces. This is enabled through the use of a variable-rate, per-flow shaping and by controlling the TCP congestion window size.
- Various enhancements to TCP based on RFC1323, including modifications to TCP slow start and TCP retransmissions, are used within the Cisco MDS to provide higher levels of throughput to FCIP than would otherwise be possible using standard TCP.
- Compression (hardware based on MPS-14/2 module, software based on IPS-4/8 modules) enables higher traffic levels to be sustained through the WAN. Compression utilizes standards-based LZS and deflate algorithms, typically achieving compression ratios of 2:1 to 3:1 with a real-world traffic mix (as measured using industry-standard tests such as Canterbury Corpus) and up to 30:1 compression on very compressible data. Hardware compression supports up to 150 MBps (~1400 Mbps) of compressed throughput per port; software compression is available for up to 300 Mbps of compressed throughput per port. Higher aggregate throughput can be achieved by load-balancing traffic across multiple FCIP tunnels and Gigabit Ethernet interfaces port channeled together.
- FCIP Write Acceleration (FCIP-WA) is a SCSI protocol spoofing mechanism designed to improve application performance by reducing the overall service time for SCSI write I/Os and replicated write I/Os over distance. Most SCSI FCIP write I/O exchanges consist of two or more round trips between the host initiator and the remote target array or tape. With FCIP-WA, multiple round trips are eliminated so that there is one round trip per SCSI FCIP write I/O operation. This means that write operations are typically accelerated with FCIP-WA by a factor of 2:1, halving the I/O service time associated with synchronous and asynchronous replication or approximately doubling the distance between data centers for the same total actual I/O latency.
- FCIP Tape Acceleration (FCIP-TA) is used to improve remote tape backup performance by minimizing the effect of network latency or distance on remote tape applications. With FCIP-TA, a local Cisco MDS 9000 IPS or MPS module proxies as a tape library and a remote IPS or MPS module (where the tape library is located) proxies as a backup server. Similar to FCIP-WA, FCIP-TA recognizes and proxies elements of the upper-level SCSI protocol in order to minimize the number of end-to-end round trips required to transfer a unit of data and optimally make use of the available network bandwidth. FCIP-TA achieves this by proxying the transfer ready and SCSI status responses while maintaining data integrity in the event of a variety of error conditions. The net benefit of FCIP-TA is that tape throughput can be maintained irrespective of latency and distance.

- High-performance encryption and decryption are available on MPS-14/2 modules and the Cisco MDS 9216i Fabric Switch through IP Security (IPSec) authentication, data integrity, and hardware-assisted encryption. Encryption and decryption using AES128 is supported at wire rate and can be used to transport sensitive storage traffic over untrusted WAN links.
- Where higher levels of performance or resiliency are necessary, PortChannels can be used to bundle up to 16 physical Gigabit Ethernet interfaces across multiple modules, providing up to 16 Gbps of IP SAN extension throughput. Individual interfaces can be added and removed from a PortChannel bundle in a nondisruptive manner, without affecting the flow of traffic.

Up to three FCIP tunnels can be enabled per Gigabit Ethernet interface, allowing for point-to-multipoint IP SAN extension. FCIP can be enabled on Cisco MDS 9000 Family switches using 8-port IP services modules, 4-port IP services modules, 2-port IP + 14-port Fibre Channel multiprotocol services modules.

SAMPLE SAN DESIGNS

The maximum number of end ports required over the lifespan of a SAN design, desired oversubscription ratio between hosts and storage, and the maximum port count available on a single physical switch are the primary factors that determine the optimal type of topology for a given SAN design.

Generally speaking, SAN designs fit into two distinct categories:

- **SAN designs that can be built using a single physical switch**—SAN designs that can be built using a single physical switch are commonly referred to as collapsed core designs. This terminology refers to the fact that a design is conceptually a core/edge design, making use of core ports (non over-subscribed) and edge ports (over-subscribed) but that it has been collapsed into a single physical switch. Traditionally, a collapsed core design on Cisco MDS 9000 Family switches would utilize both non over-subscribed (storage) and over-subscribed (host-optimized) line-cards. With the introduction of second-generation line-card modules, collapsed core designs can now be built using a single type of line-card (for example, a 48-port linecard), in conjunction with the port bandwidth reservation feature.
- **SAN designs that require multiple physical switches to meet the port count and oversubscription requirements**—Multiple design choices are available for multiswitch fabrics, with the most common design being one of a structured core/edge design. In a core/edge design, storage is always put in the core, and hosts are always attached at the edge. This design is considered structured because SAN traffic flows are typically not peer-to-peer but instead many-to-one (hosts to storage). In addition to core/edge, other common designs available include edge/core/edge (storage edge, core, host edge), used where core/edge provides insufficient scalability and an additional edge tier is necessary to handle the large number of devices; collapsed edge, where the design uses three tiers but the two edge layers are collapsed into the same physical switches (enabled through mixing over-subscribed and non over-subscribed linecards in the edge layer); cascaded; ring; and mesh topologies.

Rather than describing every possible type of topology in detail, this paper only provides examples for collapsed core (single switch) and structured core/edge (multiple switch) designs. While there is nothing inherently wrong in other storage topologies, the high port density offered by the Cisco MDS 9513 Multilayer Director can be used with these two types of designs to address even the largest SAN deployments in the world.

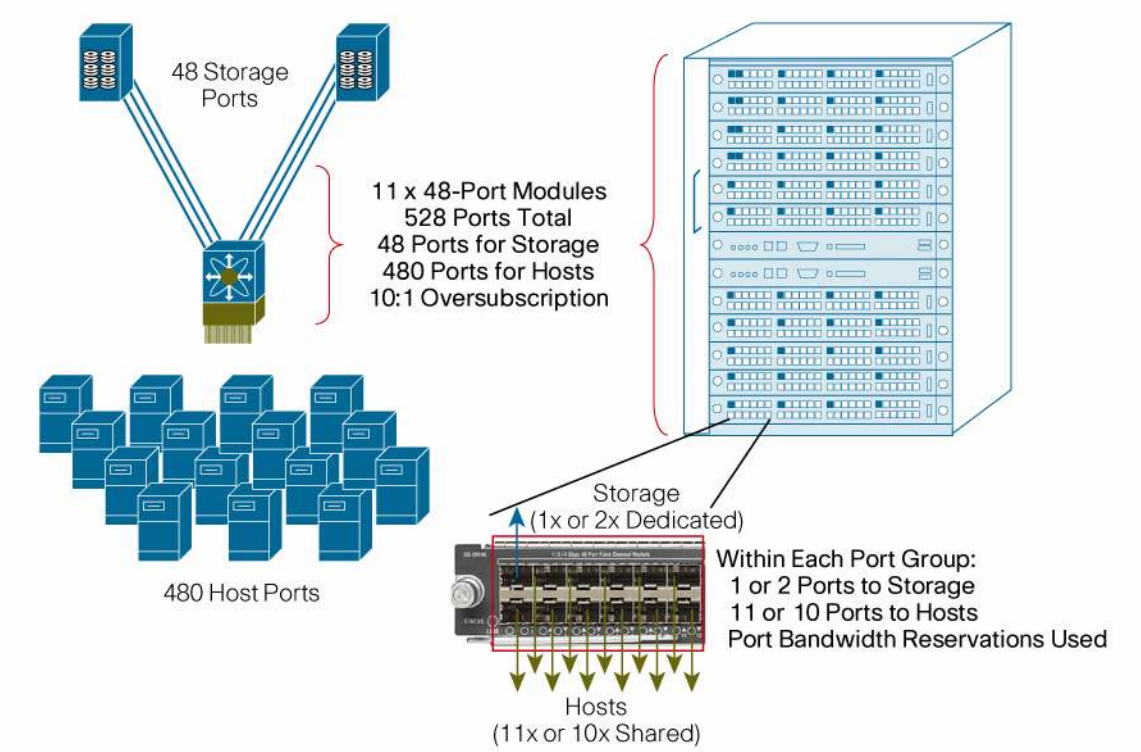
Although all SAN designs focus on the end goal of what the SAN would look like at its maximum configuration, this does not mean that all equipment needs to be provisioned initially. Rather, SAN designs grow to their maximum configuration and typically start out with a minimal number of switches and linecards. As the SAN design grows, the equipment list incrementally grows with it.

SAN designs should always use two isolated fabrics for high availability, with both hosts and storage connecting to both fabrics. Multipathing software should be deployed on the hosts to manage connectivity between the host and storage so that I/O uses both paths, and there is non-disruptive failover between fabrics in the event of a problem in one fabric. Fabric isolation can be achieved using either VSANs, or dual physical switches. Both provide separation of fabric services, although it could be argued that multiple physical fabrics provide increased physical protection (e.g. protection against a sprinkler head failing above a switch) and protection against equipment failure. Cisco MDS 9500 Series Director switches are fully redundant internally with no single point of failure. They are built to exceed 99.999 percent uptime, thus equipment failure is unlikely. Experience shows us that the majority of outages are as a result of human error rather then equipment failure or environmental factors (power, heat, water).

Medium Collapsed Core SAN Design

The collapsed core SAN design in Figure 3 shows how it is possible to scale a single-switch design up to 480 host ports and 48 storage ports, at an end-to-end oversubscription ratio of 10:1; eleven 48-port line-card modules are used on a single Cisco MDS 9513 chassis.

Figure 3. 528-Port Collapsed Core SAN Design



Port bandwidth reservation is used to dedicate forwarding resources to storage-facing ports, with host-facing ports remaining in shared mode. Using 48-port linecards, each port group contains 12 front-panel ports, and there are 44 port groups across the entire switch. Four port groups have 2 ports connected to storage and 10 ports for hosts; all 40 remaining port groups have 1 port connected to storage and 11 ports for hosts. Although the host-attached ports are operating in shared mode, the oversubscription in each port group does not exceed the end-to-end 10:1 oversubscription ratio between hosts and storage.

Table 2. 528-Port Collapsed Core SAN Design Metrics

Metric	Data Point
Ports Deployed	528
Usable Ports	528
Unused (Available) Ports	0
Design Efficiency	100%
End-to-End Oversubscription	10:1

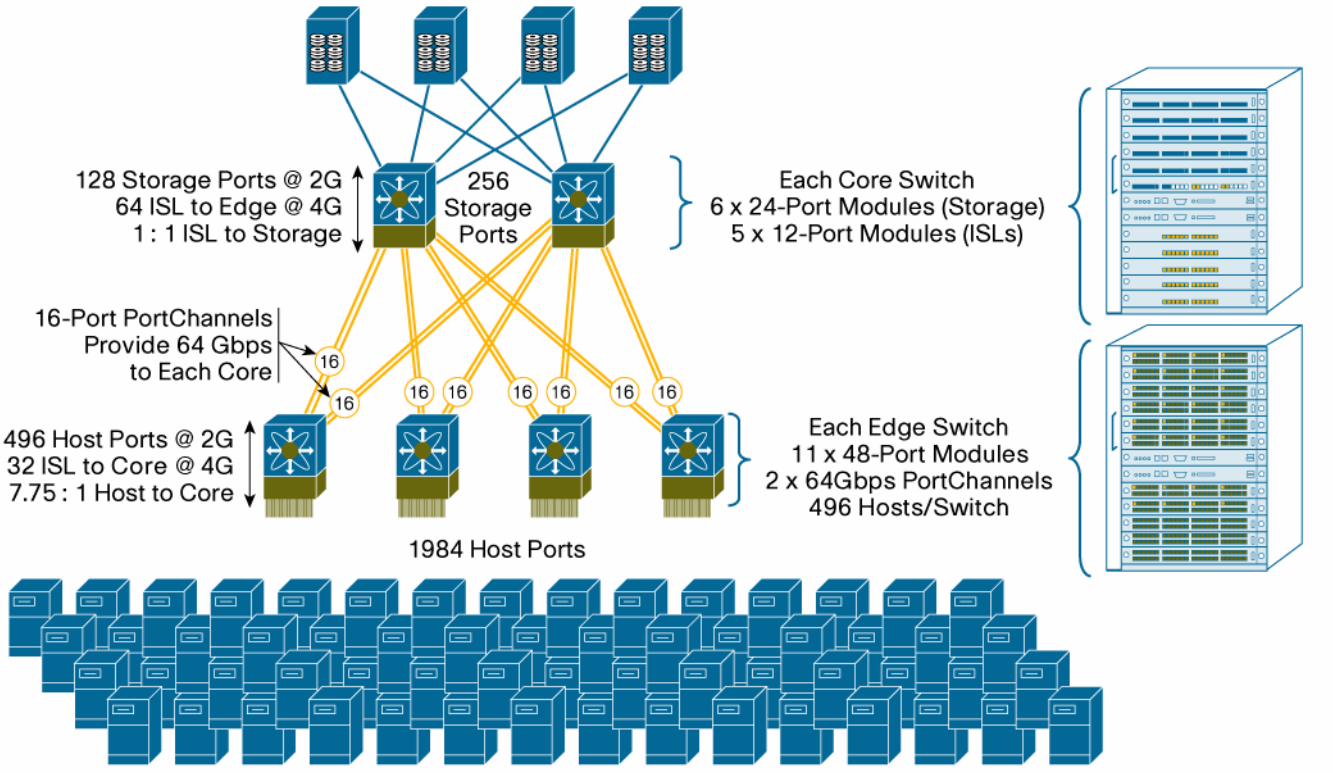
The same design using a Cisco MDS 9509 Multilayer Director can handle up to 305 host ports and 31 storage ports.

Using 256-port directors, the design could grow to a maximum of 232 host ports and 24 storage ports.

Large Core/Edge SAN Design

The core/edge SAN design in Figure 4 shows how it is possible to build a SAN for up to 1984 host ports and 256 storage ports with an end-to-end oversubscription ratio of 7.75:1 with a pair of core switches and four edge switches.

Figure 4. 2240-Port Core/Edge SAN Design



In this design, the port bandwidth reservation feature is used to dedicate forwarding resources on edge switches for ISL connectivity from the edge to the core. Host-facing ports remain in shared mode. With the 48-port linecards, each port group contains 12 front-panel ports. Because each Cisco MDS 9513 contains 11 linecards, there are 44 port groups across the entire switch: 32 of these 12-port port groups use a single ISL connected to the core, with the remaining 11 ports in the port group used for host connectivity. The 12 remaining 12-port port groups use all ports available for host connectivity.

Each edge switch connects to each of two core switches using 16 ISLs, each operating with 4 Gbps dedicated forwarding resource. These 32 ISLs are splayed across the entire chassis, but are configured to reside within two PortChannel bundles (one for each core switch). This results in two 64-Gbps logical ISLs, one to each of the two core switches. With each host-facing port operating at 2 Gbps, this results in an overall ratio of 992 Gbps (496 host ports at 2 Gbps) to 128 Gbps (two 16-port 4-Gbps PortChannels). Put another way, that is an oversubscription ratio of 7.75:1.

Each of the core switches is configured with five 12-port modules and six 24-port modules. This provides each core switch with:

- **60 non-over-subscribed ports at 4 Gbps (12-port modules)**—Used for ISLs
- **144 non-over-subscribed ports at 2 Gbps (24-port modules)**—128 ports used for storage, 4 ports used for ISLs, 12 unused

64 ports at 4-Gbps are used for ISL connectivity to each edge switch (4 edge switches each, with a single 16-port PortChannel). 60 ISLs are connected using 12-port module front-panel ports, and 4 are connected using 24-port module ports. Port bandwidth reservation is used to help ensure there is 4-Gbps dedicated forwarding resource for each of the 4 ports connected to 24-port linecards. 128 ports at 2-Gbps are used for storage connectivity. Each core switch has 12 remaining unused ports that can be used for any future expansion.

In summary, the design contains 1984 host-facing ports, 256 storage-facing ports, 7.75:1 oversubscription at 2 Gbps, built using 4 edge switches and 2 core switches.

Table 3. 2240-Port Core/Edge SAN Design Metrics

Metric	Data Point
Ports Deployed	2544
Usable Ports	2240
Unused (Available) Ports	24
Design Efficiency	89%
End-to-End Oversubscription	7.75:1

Although this design focuses on the end goal of what the SAN would look like at its maximum configuration, there is no requirement to deploy and provision all of the switches or ISLs upfront. Rather, the requirements grow to their maximum configuration and switches and ISLs are incrementally added as required. All Cisco MDS 9000 Family switches support non-disruptive incremental addition of ISL using Cisco PortChannels and port usage can be tracked using Cisco Fabric Manager or other RMON–based (threshold alarm) monitoring tools.

Attempting to build the same port count/oversubscription design using 256-port 2-Gbps director switches (or 256-port 4-Gbps 2:1 over-subscribed director switch) would require 11 fully configured switches (9 edge, 2 core) with an overall oversubscription ratio of 7.82:1.

If we wanted a maximum of 7.75:1 oversubscription this would convert the design into a 12-switch design (9 edge, 3 core) with a design efficiency of 80%:

Table 4. 2240-Port Core/Edge SAN Design Metrics (When Built Using 256-Port Director Switches)

Metric	Data Point
Ports Deployed	2848
Usable Ports	2240
Unused (Available) Ports	68 (18 in edge, 50 in core)
Design Efficiency	80%
End-to-End Oversubscription	7.75:1

SAN MIGRATION MADE SIMPLE

All Cisco MDS 9000 Family switches offer standard features that can be used to simplify migration and minimize disruptions and reconfigurations that are commonly required with SAN migrations.

Static FC_ID Allocation and Persistent FC_IDs

Cisco MDS 9000 Family switches do not lock FC_IDs to front-panel ports. Any port can have any FC_ID. By default, Cisco MDS switches will remember the FC_ID assigned to a given device (FCID is persistently mapped to device WWN) and will always attempt to allocate the same FC_ID to a device, regardless of which front-panel port it is connected to.

Static FC_ID allocation and persistent FC_ID configuration are important for some operating systems—for example, various versions of HP-UX and AIX—which otherwise require a reboot if the FC_IDs of storage arrays change.

Standards-Based Interop and Native Interop Modes

Cisco MDS 9000 Family switches have always supported industry-standard FC-SW-2 and FC-SW-3 switch-to-switch interoperability. Qualified and certified by all major storage vendors, standards-based interoperability enables multivendor fabrics to be deployed with full confidence that the configuration is valid and supported.

Beginning with Cisco MDS 9000 SAN-OS Software Release 1.3, Cisco added support for interoperability with Brocade's proprietary Core PID 0 and Core PID 1 modes. This enabled connectivity between Cisco MDS switches and Brocade switches, without requiring the Brocade switches to be converted into interop mode (a disruptive process). This allows organizations with existing Brocade switches to not lose any Brocade-proprietary functionality as a result of running a multivendor environment. Cisco MDS switches continue to support all functionality, including IVR between different VSANs operating in different interoperability modes.

With the introduction of Cisco MDS 9000 SAN-OS Software Release 3.0, Cisco also added support for interoperability with McData's proprietary mode, including supporting IVR between different VSANs regardless of interoperability mode.

These additional interoperability modes can be used to simplify SAN migration by removing the vendor lock-in associated with proprietary modes.

Zoning Migration

Cisco Fabric Manager can copy zoning configuration from existing Brocade and McData switches, without requiring switch-to-switch connectivity. The Zoning Migration wizard within Cisco Fabric Manager can import complete zone set information and aliases without any effect on existing SAN fabrics, simplifying SAN migration between different vendors.

CONCLUSION

The Cisco MDS 9000 Family of intelligent multilayer SAN switches continues to offer innovative cost-saving storage solutions and provide SAN connectivity using multiprotocol and multiservice intelligence. Cisco continues to deliver on its strategy of offering a comprehensive manageable, secure, scalable, and resilient SAN product line, with a consistent architecture and software feature set, embedded multiprotocol and multiservice intelligence, backed by an industry-leading support and services organization and a vast network of partners, each leaders in the storage industry.

From a SAN design standpoint, the Cisco MDS 9000 Family of switches offers significant SAN design flexibility, lowering the overall cost and leading to higher returns on investment.

Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters
Cisco Systems International BV
Haarlerbergpark
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands
www-europe.cisco.com
Tel: 31 0 20 357 1000
Fax: 31 0 20 357 1100

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

Asia Pacific Headquarters
Cisco Systems, Inc.
168 Robinson Road
#28-01 Capital Tower
Singapore 068912
www.cisco.com
Tel: +65 6317 7777
Fax: +65 6317 7799

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on **the Cisco Website at www.cisco.com/go/offices.**

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia • Cyprus Czech Republic • Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia • Ireland • Israel Italy • Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland • Portugal Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa • Spain • Sweden • Switzerland • Taiwan Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

Printed in the USA C11-335072-00 02/06