

The Lenovo logo is displayed in white text on a black rectangular background.

Flex System Interconnect Fabric Technical Overview and Planning Considerations

Last Update: July 2014

Explains the benefits of using Flex System Interconnect Fabric (FSIF)

Provides technical details of FSIF including topology, traffic flows, high availability features and performance features

Describes upstream networking and storage integration

A guide for networking professionals looking for expert advice

Ilya Krutov
Scott Lorditch



Abstract

To respond to the pressures of change in today's business environment, organizations need more flexible and cost-effective IT infrastructures. Many are taking a new approach by consolidating and virtualizing IT resources, such as servers, storage, networks, and even desktops. As the data center evolves, however, data center managers face new challenges. They must manage and control virtualization sprawl and ensure the security of the increasing number of virtual machines (VMs). In addition, they must address technology constraints while migrating workloads and determine how to manage a network infrastructure that cannot scale with the explosion of network traffic brought about by the consolidation of workloads in virtualized environments.

This paper describes the key factors that are driving changes in a data center design approach, and how Flex System™ Interconnect Fabric, the new data center networking (DCN) fabric solution, helps transform a data center into a scalable, manageable, and easy to deploy environment.

This paper is intended for networking IT professionals who want to learn more about Flex System Interconnect Fabric, the most recent addition to the Lenovo® portfolio of DCN technologies and building blocks.

At Lenovo Press, we bring together experts to produce technical publications around topics of importance to you, providing information and best practices for using Lenovo products and solutions to solve IT challenges. See a list of our most recent publications at the Lenovo Press web site:

<http://lenovopress.com>

Contents

Data center networking overview	3
Introduction to Flex System Interconnect Fabric	5
Flex System Interconnect Fabric components	7
Flex System Interconnect Fabric technical details	18
Point of delivery network integration	30
Point of delivery storage integration	38
Point of delivery management	44
Interconnect Fabric initial configuration	47
Appendix: Example Interconnect Fabric POD bill of materials	48
Related publications	49
Authors	49
Notices	51
Trademarks	52

Data center networking overview

Consolidation began as a trend toward centralizing the scattered IT assets of an enterprise for better cost control, operational optimization, and efficiency. Virtualization introduced an abstraction layer between hardware and software, which allowed enterprises to consolidate even further and getting the most out of each physical server platform in the data center by running multiple virtual servers on it. A shift toward new workload types, such as cloud computing and big data, requires higher performance and lower latency. Converged network solutions are becoming more popular because they dramatically reduce network complexity, simplify network management, and increase overall data center operational efficiency. All these trends have direct impact on the data center networking infrastructure by changing network traffic flow patterns and increasing bandwidth and quality of service (QoS) requirements.

For example, there is a significant shift from primarily client/server traffic flow (also known as *north-south* traffic) to a mixed server-to-server (or *east-west*) and client/server traffic flow, which has negative impact on application response time because of client/server network optimization and segmentation in most data centers. Also, there is a need for scalable, high performance, flattened (Layer 2), and highly available network infrastructure to support greater VM mobility for better allocation and usage of pooled resources. At the same time, satisfying these requirements should not come at the expense of significant capital investments and operational costs.

Often, the point of delivery (POD) term is used to describe a physical repeatable pattern that is used as a data center building block with the following characteristics:

- ▶ Server and network hardware are co-resident in the same POD
- ▶ Network functionality is typically Layer 2 only in the form of network access/aggregation
- ▶ All server NIC to network port connections stay within the POD
- ▶ All PODs are connected to a data center network core

Traditional data center networks are built using the tree-layer (access, distribution, and core) hierarchical model, as shown in Figure 1.

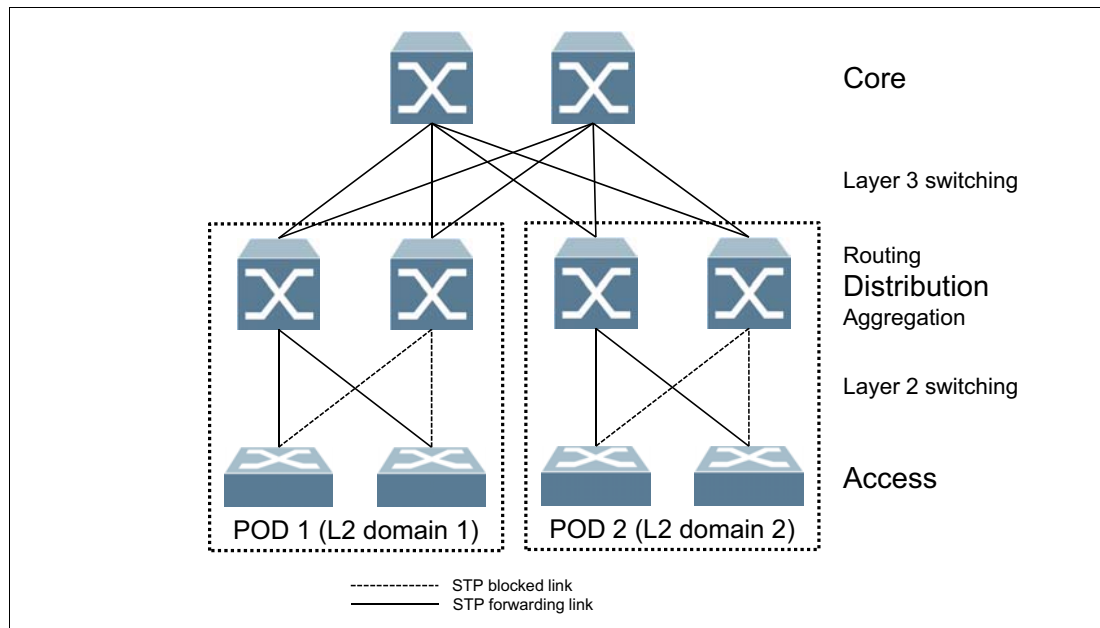


Figure 1 Traditional network design (three-tier)

Hierarchical data centers have some limitations for achieving performance and scalability requirements for virtualized workloads and distributed applications. This is because the three-tier design implies data center scaling using Layer 3 routing and switching, while VM mobility and distributed applications require Layer 2 scaling.

In addition, the network switches at the distribution and access layers are connected to each other using redundant links, and the spanning tree protocol (STP) is used to make sure that there are no loops in the Layer 2 network by disabling certain links, which effectively limits the available uplink bandwidth. Another STP drawback is slow convergence if there is an uplink failure, which might cause a network outage because of a traffic interruption for some period. There are protocols that can help STP converge faster and load balance traffic across the uplinks; however, this requires additional administration and support efforts.

Finally, each access layer switch (and distribution layer switches) is treated as a separate management entity that requires configuration and monitoring, which adds complexity to the overall network management operations.

To overcome these issues and to provide a highly scalable and resilient Layer 2 network, network *fabric* connectivity was introduced. Typically, fabric design is based on a leaf-spine architecture, as shown in Figure 2.

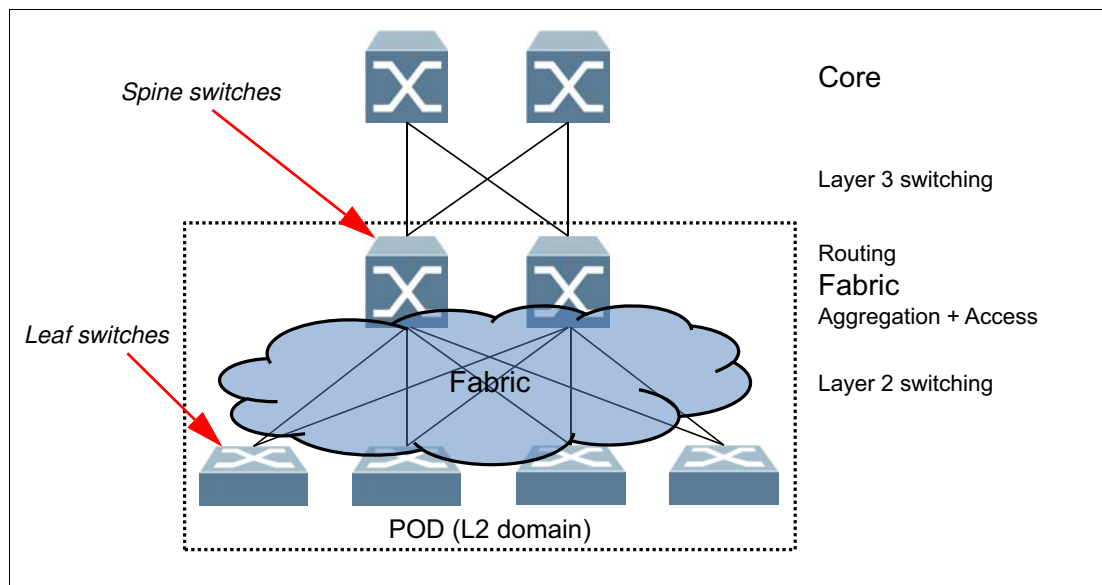


Figure 2 Data center fabric design

Leaf switches connect servers, and spine switches connect upstream networks. Each leaf switch is connected to all spine switches, which forms a scalable loop-free fabric. Typically, a proprietary protocol runs within the fabric that ensures that all links are active and are used for traffic forwarding. If there is a fabric link failure, the protocol quickly and automatically reconfigures the fabric so that the topology changes (and the actual fabric topology) are hidden from the rest of the network.

Another major advantage of the fabric topology is that the upstream network sees the entire fabric as one large logical switch that has many ports and is managed as a single entity, which simplifies the overall network management. Also, if the upstream network uses some sort of distributed link aggregation, the fabric network can be connected to the upstream network through a single logical aggregated link while still providing the required levels of performance and availability.

Typically, the entire fabric represents a Layer 2 domain, so VM mobility and distribution application scaling are simplified.

Lenovo offers Flex System Interconnect Fabric, an innovative Flex System based networking technology that helps build flexible, scalable, and easily managed data center networks that deliver high performance.

Introduction to Flex System Interconnect Fabric

Flex System Interconnect Fabric offers a solid foundation of compute, network, storage, and software resources in a Flex System POD.

The Interconnect Fabric solution supports two Lenovo RackSwitch™ G8264CS top of rack (TOR) switches, up to nine Flex System chassis that are populated with two Flex System Fabric SI4093 System Interconnect Modules in each chassis, up to 126 half-wide Flex System compute nodes with 2-port LAN on system board (LOM) or 4-port or 8-port 10 Gb converged network adapters (CNAs) or up to 282 high-density Flex System x222 Compute Nodes with 2-port LOM, and one or more internal Flex System V7000 Storage Nodes or external Storwize V7000 storage systems as a highly integrated POD solution for data centers.

The entire POD integrates as a seamless network fabric for compute node and storage under single IP management, and it attaches to the upstream data center network as a loop-free Layer 2 network fabric with a single Ethernet uplink connection or aggregation group to each Layer 2 network. The POD requires only network provisioning for uplink connections to a data center network, downlink connections to compute nodes, and storage connections to storage nodes.

Flex System Interconnect Fabric overview

The Flex System Interconnect Fabric solution has the following key elements:

- ▶ Hardware elements
 - RackSwitch G8264CS (10/40 GbE, 4/8 Gb FC uplink) as Aggregation
 - Flex System Fabric SI4093 System Interconnect Module (10 GbE to server) as Access
 - Embedded VFA, CN4054, or CN4058 adapters
 - Flex System V7000 Storage Node or Storwize V7000
- ▶ Software elements
 - Single IP managed multi-rack cluster (hDFP)
 - Automated rolling (staggered) upgrades of individual switches
 - Per-server link redundancy (LAG or active/passive teaming)
 - Dynamic bandwidth within and out of the POD
 - Multi-rack Flex System Interconnect mode
 - Integration of UFP and VMready®
- ▶ Management elements
 - Switch Center Management application (fabric management)
 - Flex System Manager configuration patterns (compute node NIC configuration)

Figure 3 on page 6 provides an overview of the Interconnect Fabric solution elements.

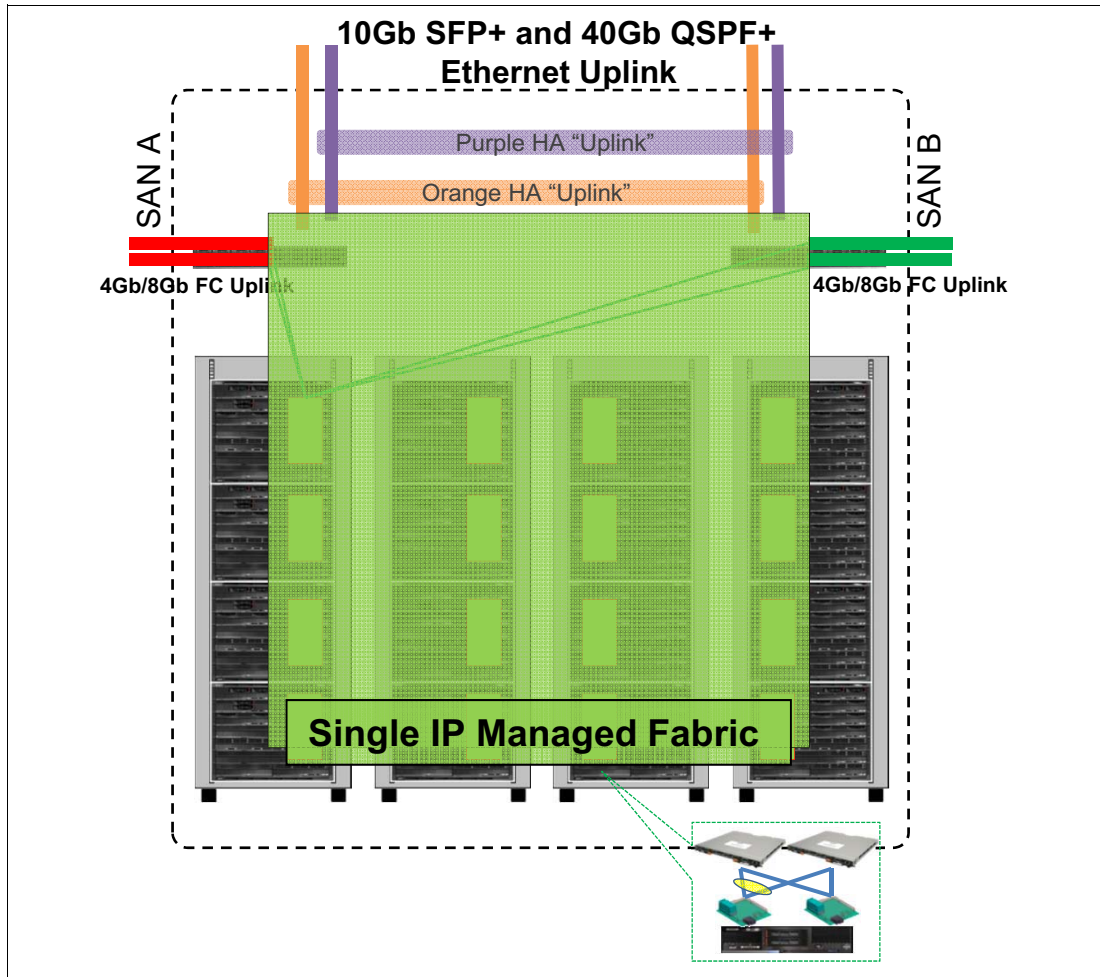


Figure 3 Flex System Interconnect Fabric overview

Flex System Interconnect Fabric benefits

The Flex System Interconnect Fabric solution offers the following benefits:

- ▶ Network simplification:
 - Provisions a seamless network fabric for compute node and storage connectivity in the data center.
 - Offers a loop-free network fabric without STP complexity for fast network convergence.
 - Minimizes network latency by local Layer 2 switching at every interconnect component and minimizes loss of data during network failover within the fabric.
 - Converges Ethernet for lossless storage traffic.
 - Integrates FCF to provide end-to-end FCoE storage functionality within the POD without needing an expensive Fibre Channel switch.
 - Supports single fabric mode topology and dual fabric mode topology.

- ▶ Management simplification:
 - Offers high availability with master and backup TOR switches in the fabric and hitless upgrade with no downtime for services.
 - Minimizes managed network elements with single point of management of the entire fabric at the master TOR switch.
 - Establishes a clear administrative boundary in a data center by pushing traditional networking configuration outside of the POD.
 - Integrates physical and virtual infrastructure management for compute, network, storage, and software elements.
- ▶ Storage integration:
 - Simplifies integration of storage and storage virtualization with Flex System V7000 Storage Node.
 - Provides access to an external SAN storage infrastructure
- ▶ Scalable POD design:
 - Enables the size of the POD to grow without adding management complexity.
 - Adds chassis resources up to the maximum configuration under the single IP management of the POD.

Flex System Interconnect Fabric components

This section describes features and ordering information for the Interconnect Fabric components. It also quickly reviews the Flex System I/O architecture for better understanding of inter-chassis and intra-chassis traffic flow using the SI4093. The following topics are covered:

- ▶ “Flex System I/O architecture overview”
- ▶ “Flex System Fabric SI4093 System Interconnect Module” on page 10
- ▶ “Lenovo RackSwitch G8264CS” on page 13
- ▶ “SFP+ transceivers and cables” on page 14
- ▶ “Converged fabric adapters” on page 15

Flex System I/O architecture overview

The networking I/O architecture for the Flex System includes an array of connectivity options for server nodes that are installed in the enclosure. Users can decide to use a local switching model that provides superior performance, cable reduction, and a rich feature set, or use pass-through technology and allow all Ethernet networking decisions to be made external to the Enterprise Chassis.

From a physical I/O module bay perspective, the Flex System Enterprise Chassis has four I/O bays in the rear of the chassis. The physical layout of these I/O module bays is shown in Figure 4 on page 8.

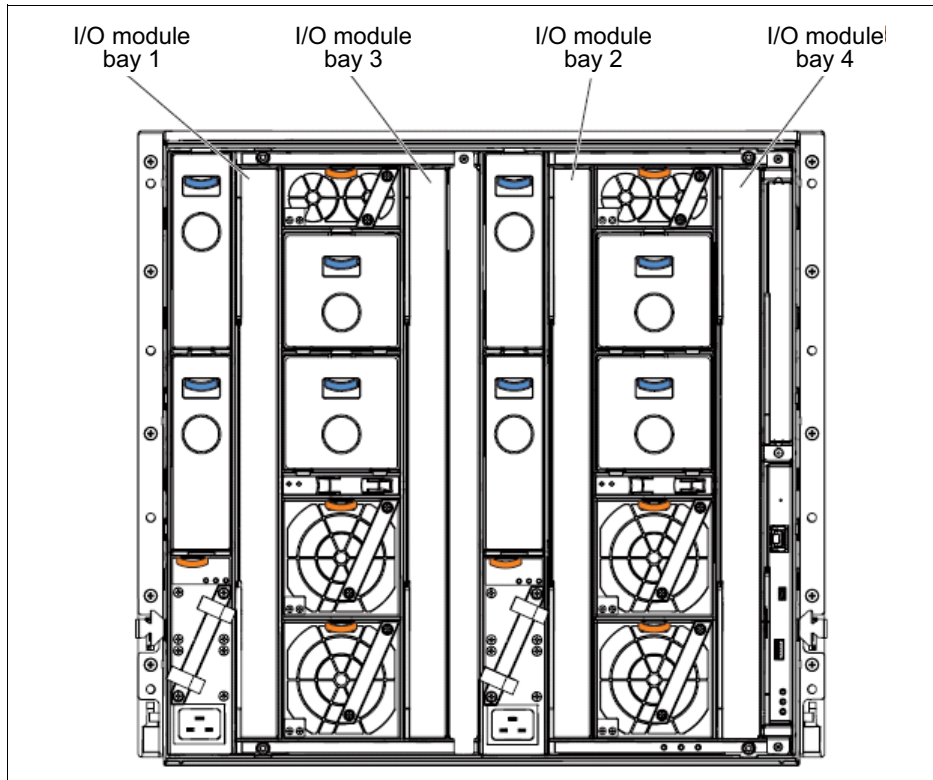


Figure 4 Rear view of the Enterprise Chassis showing I/O module bays

From a midplane wiring point of view, the Enterprise Chassis provides 16 lanes out of each half-wide node bay (toward the rear I/O bays) with each lane capable of 16 Gbps or higher speeds. How these lanes are used is a function of which adapters are installed in a node, which I/O module is installed in the rear, and which port licenses are enabled on the I/O module.

How the midplane lanes connect between the node bays upfront and the I/O bays in the rear is shown in Figure 5. The concept of an I/O module Features on Demand (FoD) Upgrade is also shown in Figure 5. From a physical perspective, an FoD upgrade in this context is a bank of 14 ports and some number of uplinks that can be enabled and used on a switch module. By default, all I/O modules include the base set of ports, and thus have 14 internal ports, one each connected to the 14 compute node bays in the front. By adding an upgrade license to the I/O module, it is possible to add more banks of 14 ports (plus some number of uplinks) to an I/O module. The node needs an adapter that has the necessary physical ports to connect to the new lanes enabled by the upgrades. Those lanes connect to the ports in the I/O module that is enabled by the upgrade.

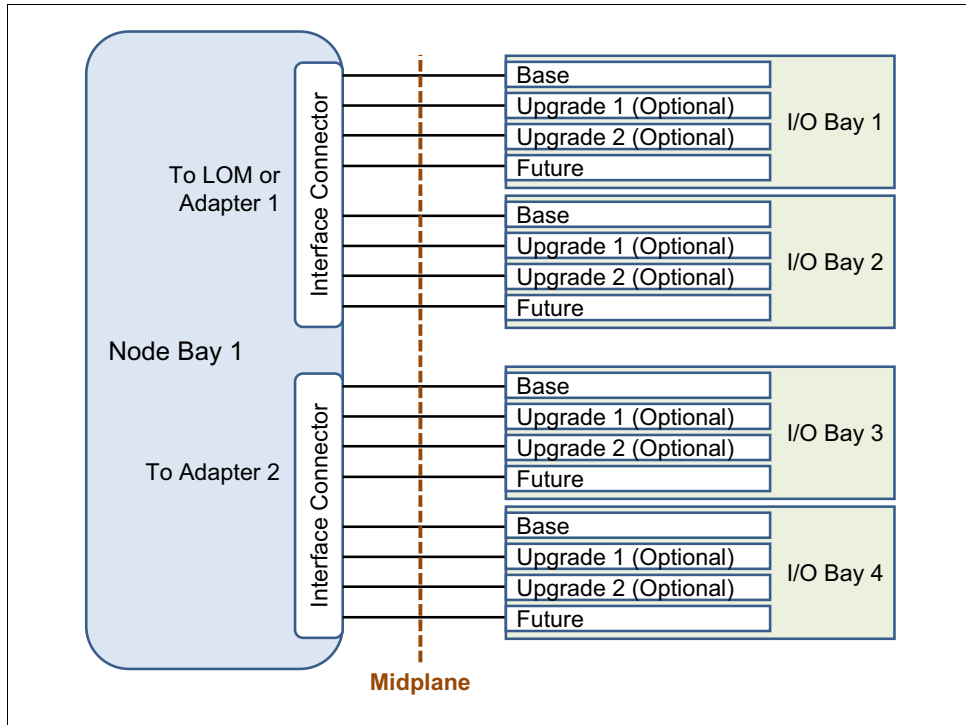


Figure 5 Sixteen lanes total of a single half-wide node bay toward the I/O bays

For example, if a node is installed with only the dual port LOM adapter, only two of the 16 lanes are used (one to I/O bay 1 and one to I/O bay 2), as shown in Figure 6 on page 10.

If a node is installed without LOM and two quad port adapters are installed, eight of the 16 lanes are used (two to each of the four I/O bays).

This installation can potentially provide up to 320 Gb of full duplex Ethernet bandwidth (16 lanes x 10 Gb x 2) to a single half-wide node and over half a terabit (Tb) per second of bandwidth to a full-wide node.

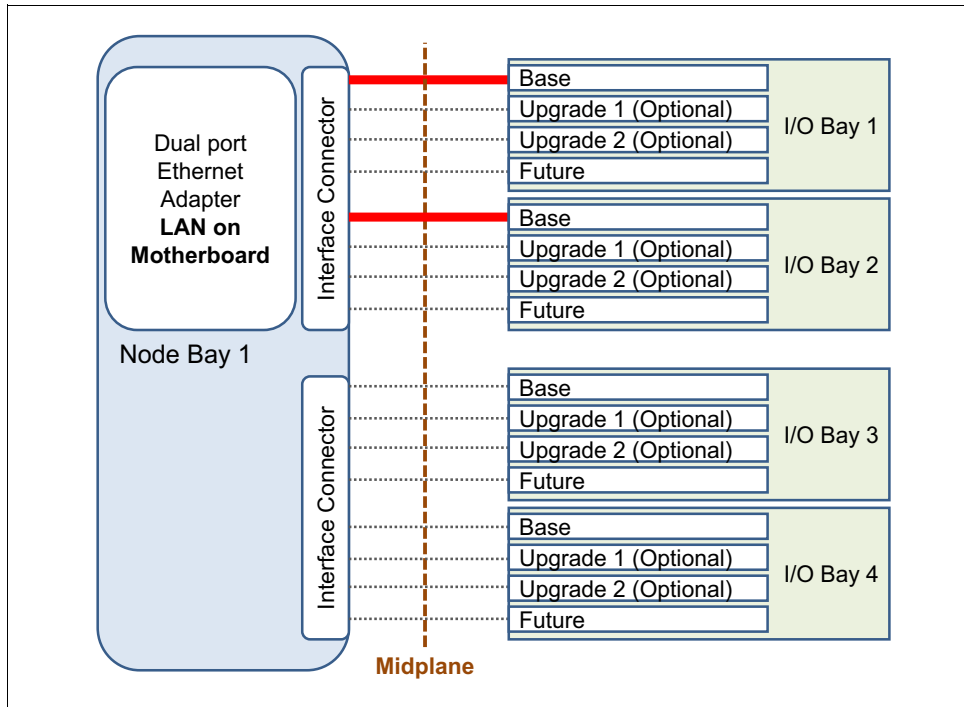


Figure 6 Dual port LOM connecting to ports on I/O bays 1 and 2 (all other lanes are unused)

Today, there are limits on the port density of the current I/O modules, in that only the first three lanes are potentially available from the I/O module.

By default, each I/O module provides a single connection (lane) to each of the 14 half-wide node bays upfront. By adding port licenses, an SI4093 System Interconnect Module that is used in the Interconnect Fabric solution can provide up to three 10 Gb ports to each of the 14 half-wide node bays.

As an example, if an 8-port adapter is installed and two I/O modules are installed in I/O bays 1 and 2 with all upgrades, the end node has access to six 10 Gb lanes (three to each switch). On the 8-port adapter, two lanes cannot be used at this time.

Concerning port licensing, the default available upstream connections also are associated with port licenses. For more information about these connections and the node that faces links, see “Flex System Fabric SI4093 System Interconnect Module”.

Flex System Fabric SI4093 System Interconnect Module

The IBM Flex System Fabric SI4093 System Interconnect Module enables simplified integration of Flex System into your existing networking infrastructure.

The SI4093 System Interconnect Module requires no management for most data center environments. This eliminates the need to configure each networking device or individual ports, which reduces the number of management points. It provides a low latency, loop-free interface that does not rely upon spanning tree protocols, which removes one of the greatest deployment and management complexities of a traditional switch.

The SI4093 System Interconnect Module offers administrators a simplified deployment experience while maintaining the performance of intra-chassis connectivity.

The SI4093 System Interconnect Module is shown in Figure 7.



Figure 7 IBM Flex System Fabric SI4093 System Interconnect Module

The SI4093 System Interconnect Module is initially licensed for fourteen 10 Gb internal ports enabled and ten 10 Gb external uplink ports enabled. More ports can be enabled, including 14 internal ports and two 40 Gb external uplink ports with Upgrade 1, and 14 internal ports and four SFP+ 10 Gb external ports with Upgrade 2 license options. Upgrade 1 must be applied before Upgrade 2 can be applied.

Table 1 shows the part numbers for ordering the switches and the upgrades.

Table 1 SI4093 ordering information

Description	Part number	Feature code (x-config / e-config)
Interconnect module		
Flex System Fabric SI4093 System Interconnect Module	95Y3313	A45T / ESWA
Features on Demand upgrades		
SI4093 System Interconnect Module (Upgrade 1)	95Y3318	A45U / ESW8
SI4093 System Interconnect Module (Upgrade 2)	95Y3320	A45V / ESW9

Important: SFP and SFP+ (small form-factor pluggable plus) transceivers or cables are not included with the switch. They must be ordered separately. For part number information, see Table 4 on page 14.

The following base switch and upgrades are available:

- ▶ 95Y3313 is the part number for the physical device, and it comes with 14 internal 10 Gb ports enabled (one to each node bay) and 10 external 10 Gb ports enabled for connectivity to an upstream network. All external 10 Gb ports are SFP+ based connections.
- ▶ 95Y3318 (Upgrade 1) can be applied on the base interconnect module to make full use of 4-port adapters that are installed in each compute node. This upgrade enables 14 more internal ports, for a total of 28 internal ports. The upgrade also enables two 40 Gb external ports with QSFP+ connectors. These QSFP+ ports can also be converted to four 10 Gb SFP+ DAC cable connections by using the appropriate fan-out cable. This upgrade requires the base interconnect module.

- ▶ 95Y3320 (Upgrade 2) can be applied on top of Upgrade 1 when you want more external bandwidth on the interconnect module or if you want more internal bandwidth to the compute nodes with the adapters capable of supporting six ports (like CN4058). The upgrade enables the remaining four external 10 Gb uplinks with SFP+ connectors, plus 14 internal 10 Gb ports, for a total of 42 ports (three to each compute node).

Table 2 lists the supported port combinations on the interconnect module and the required upgrades.

Table 2 Supported port combinations

Supported port combinations	Quantity required		
	Base switch, 95Y3313	Upgrade 1, 95Y3318	Upgrade 2, 95Y3320
14x internal 10 GbE 10x external 10 GbE	1	0	0
28x internal 10 GbE 10x external 10 GbE 2x external 40 GbE	1	1	0
42x internal 10 GbE ^a 14x external 10 GbE 2x external 40 GbE	1	1	1

a. This configuration uses six of the eight ports on the CN4058 adapter that are available for IBM Power Systems compute nodes.

Note: Flexible port mapping (introduced in Networking OS V7.8) is not supported on the SI4093 in the Flex System Interconnect Fabric mode.

The SI4093 has the following interfaces:

- ▶ Internal ports:
 - A total of 42 internal full-duplex 10 Gigabit ports; 14 ports are enabled by default. Optional FoD licenses are required to activate the remaining 28 ports.
 - Two internal full-duplex 1 GbE ports that are connected to the chassis management module.
- ▶ External ports:
 - A total of 14 ports for 1 Gb or 10 Gb Ethernet SFP+ transceivers (support for 1000BASE-SX, 1000BASE-LX, 1000BASE-T, 10GBASE-SR, or 10GBASE-LR) or SFP+ copper direct-attach cables (DAC). A total of 10 ports are enabled by default. An optional FoD license is required to activate the remaining four ports. SFP+ modules and DACs are not included and must be purchased separately.
 - Two ports for 40 Gb Ethernet QSFP+ transceivers or QSFP+ DACs. (Ports are disabled by default. An optional FoD license is required to activate them.) QSFP+ modules and DACs are not included and must be purchased separately.
 - One RS-232 serial port (mini-USB connector) that provides another means to configure the switch module.

For more information, see *Flex System Fabric SI4093 System Interconnect Module*, TIPS1045, which is available at this website:

<http://1enovopress.com/tips1045>

Lenovo RackSwitch G8264CS

The Lenovo RackSwitch G8264CS is an enterprise-class switch that offers high-bandwidth performance with thirty-six 1/10 Gb SFP+ connections, 12 Omni Ports™ that can be used for 10 Gb SFP+ connections, 4/8 Gb Fibre Channel connections or both, plus four 40 Gb QSFP+ connections.

It simplifies the deployment with its innovative Omni Port technology and offers the flexibility to choose 10 Gb Ethernet, 4/8 Gb Fibre Channel, or both for upstream connections. In FC mode, Omni Ports provide convenient access to FC storage.

The G8264CS provides 100% line rate performance with low latency and 1.28 Tbps non-blocking switching throughput (full duplex) on Ethernet ports, making it an optimal choice for managing dynamic workloads across the network. It provides a rich Layer 2 and Layer 3 feature set that is ideal for many of today's data centers.

Also, its Omni Port technology helps in consolidating the enterprise storage, networking, data, and management into a simple to manage single fabric, and reduces costs that are associated with energy and cooling, management and maintenance, and capital costs.

Figure 8 shows the Lenovo RackSwitch G8264CS.



Figure 8 Lenovo RackSwitch G8264CS

These key features are included:

- ▶ Omni Ports provide the flexibility of using 10 Gb Ethernet or 4/8 Gb Fibre Channel connections
- ▶ 36× 10 GbE SFP+ ports, 12 Omni Ports, and 4× 40 GbE QSFP+ ports in a 1U form factor
- ▶ G8264CS is optimized for High Performance Computing and other applications requiring high bandwidth and low latency
- ▶ Hot-swappable redundant power supplies and fans
- ▶ Software is based on Internet standards for optimal interoperability with Cisco or other vendors' networks
- ▶ Two airflow options to manage hot- and cold-aisle applications
- ▶ VMready and Virtual Fabric for virtualized networks

Table 3 shows the part numbers for ordering the switches and the upgrades.

Table 3 G8264CS ordering information

Description	Part number
Lenovo RackSwitch G8264CS (Rear-to-Front)	7309DRX
Lenovo RackSwitch G8264CS (Front-to-Rear)	7309DFX

Important: SFP and SFP+ (small form-factor pluggable plus) transceivers or cables are not included with the switch. They must be ordered separately. For part number information, see Table 4 on page 14.

The G8264CS switch has the following interfaces:

- ▶ 36 SFP+ ports (1 Gb or 10 Gb Ethernet)
- ▶ 12 Omni Ports (10 Gb Ethernet or 4/8 Gb Fibre Channel)
- ▶ Four QSFP+ ports (40 Gb Ethernet)
- ▶ One 10/100/1000 Ethernet RJ45 port for out-of-band management
- ▶ One USB port for mass storage device connection
- ▶ One mini-USB Console port for serial access

For more information, see *RackSwitch G8264CS*, TIPS0970, which is available at this website:

<http://lenovopress.com/tips0970>

SFP+ transceivers and cables

The G8264CS switches and the SI4093 I/O modules support for 10 Gb and 40 Gb interface modules and cables that are used for HiGig links and uplinks in Interconnect Fabric are shown in Table 4.

Table 4 10 Gb and 40 Gb modules and cables that are supported in Interconnect Fabric

Part number	Description	G8264CS	SI4093
10 GbE SFP+ transceivers and optical cables			
46C3447	SFP+ SR Transceiver	Yes	Yes
88Y6851	1m LC-LC Fiber Cable (networking) - Optical	Yes	Yes
88Y6854	5m LC-LC Fiber Cable (networking) - Optical	Yes	Yes
88Y6857	25m LC-LC Fiber Cable (networking) - Optical	Yes	Yes
10 GbE SFP+ DAC cables			
90Y9427	1m Passive DAC SFP+ Cable	Yes	Yes
00AY764	1.5m Passive DAC SFP+ Cable	Yes	Yes
00AY765	2m Passive DAC SFP+ Cable	Yes	Yes
90Y9430	3m Passive DAC SFP+ Cable	Yes	Yes
90Y9433	5m Passive DAC SFP+ Cable	Yes	Yes
00D6151	7m Passive DAC SFP+ Cable	Yes ^a	Yes

Part number	Description	G8264CS	SI4093
40 GbE QSFP+ transceivers and optical cables			
49Y7884	QSFP+ SR Transceiver	Yes	Yes
90Y3519	10m QSFP+ MTP Optical cable	Yes	Yes
90Y3521	30m QSFP+ MTP Optical cable	Yes	Yes
40 GbE QSFP+ to 4x 10 GbE SFP+ DAC breakout cables (uplink ports only)			
49Y7886	1m QSFP+ DAC Break Out Cable	Yes	Yes
49Y7887	3m QSFP+ DAC Break Out Cable	Yes	Yes
49Y7888	5m QSFP+ DAC Break Out Cable	Yes	Yes
40 GbE QSFP+ DAC cables			
49Y7890	1m QSFP+-to-QSFP+ Cable	Yes	Yes
49Y7891	3m QSFP+-to-QSFP+ Cable	Yes	Yes
00D5810	5m QSFP+ to QSFP+ Cable	Yes	Yes
00D5813	7m QSFP+ to QSFP+ Cable	Yes	Yes
4/8 Gb FC SFP+ transceivers (Omni Ports only)			
44X1964	8Gb SFP+ SW Optical Transceiver	Yes	No

a. Can be used on 10 GbE SFP+ ports only.

Converged fabric adapters

This section describes converged network adapters that can be used in the Flex System Interconnect Fabric solutions. The following adapters are described:

- ▶ “Flex System Embedded 10Gb Virtual Fabric Adapter”
- ▶ “Flex System CN4054/CN4054R 10Gb Virtual Fabric Adapters” on page 17

Flex System Embedded 10Gb Virtual Fabric Adapter

Some models of the Flex System x240 and x440 Compute Nodes include an Embedded 10Gb Virtual Fabric Adapter (VFA, also known as LAN on Motherboard (LOM)) built into the system board.

Each x240 model that includes the embedded 10 Gb VFA also has the Compute Node Fabric Connector installed in I/O connector 1 (and physically screwed onto the system board) to provide connectivity to the Enterprise Chassis midplane. Each x440 model that includes two embedded 10 Gb VFAs also has the Compute Node Fabric Connectors installed in each of I/O connectors 1 and 3 (and physically screwed onto the system board) to provide connectivity to the Enterprise Chassis midplane. The Fabric Connector enables port 1 on the embedded 10Gb VFA to be routed to I/O module bay 1 and port 2 to be routed to I/O module bay 2.

Each server in the x222 compute node includes an Embedded two-port 10Gb Virtual Fabric Adapter that is built in to the system board. The x222 has one Fabric Connector (which is physically on the lower server) and the Ethernet connections from both Embedded 10 Gb VFAs are routed through it to I/O module bays 1 and 2.

Table 5 lists the ordering information for the Virtual Fabric Advanced Software Upgrade (LOM), which enables the iSCSI and FCoE support on the Embedded 10 Gb Virtual Fabric Adapter.

Table 5 Feature on Demand upgrade for FCoE and iSCSI support

Part number	x-config feature code	Description
90Y9310	A2TD	Virtual Fabric Advanced Software Upgrade (LOM)

The Flex System Embedded 10 Gb VFA has the following features and specifications:

- ▶ Models with Intel Xeon E5-2400, E5-2600, and 4600 processors: Emulex BladeEngine 3 (BE3) ASIC
- ▶ Models with Intel Xeon E5-2600 v2 processors: Emulex BladeEngine 3R (BE3R) ASIC
- ▶ Operates as a 2-port 1/10 Gb Ethernet adapter or supports up to eight Virtual Network Interface Controllers (vNICs)
- ▶ Supports NIC virtualization:
 - Modes of operation:
 - Unified Fabric Port (UFP)
 - Virtual Fabric Mode
 - Switch Independent Mode
 - Virtual port bandwidth allocation in 100 Mbps increments
 - Up to eight virtual ports per adapter (four per port)
 - With the optional Advanced Upgrade, two of the eight vNICs (one per port) are transformed into iSCSI or FCoE HBAs
- ▶ Wake On LAN support
- ▶ FCoE and iSCSI HBA function support with the optional Advanced Upgrade
- ▶ PCI Express 2.0 x8 host interface
- ▶ Full-duplex capability
- ▶ DMA support
- ▶ PXE support
- ▶ IPv4/IPv6 TCP, UDP checksum offload:
 - Large send offload
 - Large receive offload
 - RSS
 - IPv4 TCP Chimney offload
 - TCP Segmentation offload
- ▶ VLAN insertion and extraction
- ▶ Jumbo frames up to 9000 bytes
- ▶ Load balancing and failover support, including AFT, SFT, ALB, and LACP
- ▶ Converged Enhanced Ethernet (draft):
 - Enhanced Transmission Selection (ETS) (P802.1Qaz)
 - Priority-based Flow Control (PFC) (P802.1Qbb)
 - Data Center Bridging eXchange protocol (DCBX) (P802.1Qaz)
- ▶ Support Serial over LAN (SoL)

Flex System CN4054/CN4054R 10Gb Virtual Fabric Adapters

The Flex System CN4054 and CN4054R 10Gb Virtual Fabric Adapters from Emulex are 4-port 10 Gb converged network adapters. They can scale to up to 16 virtual ports and support multiple protocols, such as Ethernet, iSCSI, and FCoE. The CN4054R adds support for compute nodes with the Intel Xeon E5-2600 v2 processors.

Table 6 lists the ordering part numbers and feature codes.

Table 6 Flex System CN4054 and CN4054R 4-port 10 Gb Ethernet Adapter ordering information

Part number	Feature code	Description
90Y3554	A1R1	CN4054 10Gb Virtual Fabric Adapter
00Y3306	A4K2	CN4054R 10Gb Virtual Fabric Adapter
90Y3558	A1R0	CN4054 Virtual Fabric Adapter Upgrade

The Flex System CN4054 and CN4054R 10Gb Virtual Fabric Adapters have the following features and specifications:

- ▶ CN4054: Dual-ASIC Emulex BladeEngine 3 (BE3) controller
- ▶ CN4054R: Dual-ASIC Emulex BladeEngine 3R (BE3R) controller
- ▶ Operates as a 4-port 1/10 Gb Ethernet adapter or supports up to 16 Virtual Network Interface Controllers (vNICs)
- ▶ Supports NIC virtualization:
 - Modes of operation:
 - Unified Fabric Port (UFP)
 - Virtual Fabric Mode
 - Switch Independent Mode
 - Virtual port bandwidth allocation in 100 Mbps increments
 - Up to eight virtual ports per adapter (four per port)
 - With the optional Advanced Upgrade, two of the eight vNICs (one per port) are transformed into iSCSI or FCoE HBAs
- ▶ Wake On LAN support
- ▶ FCoE and iSCSI HBA function support with the optional Advanced Upgrade
- ▶ PCI Express 3.0 x8 host interface
- ▶ Full-duplex capability
- ▶ DMA support
- ▶ PXE support
- ▶ IPv4/IPv6 TCP, UDP checksum offload:
 - Large send offload
 - Large receive offload
 - RSS
 - IPv4 TCP Chimney offload
 - TCP Segmentation offload
- ▶ VLAN insertion and extraction
- ▶ Jumbo frames up to 9000 bytes
- ▶ Load balancing and failover support, including AFT, SFT, ALB, and LACP

- ▶ Converged Enhanced Ethernet (draft):
 - Enhanced Transmission Selection (ETS) (P802.1Qaz)
 - Priority-based Flow Control (PFC) (P802.1Qbb)
 - Data Center Bridging eXchange protocol (DCBX) (P802.1Qaz)
- ▶ Support Serial over LAN (SoL)

Figure 9 shows the IBM Flex System CN4054 10Gb Virtual Fabric Adapter.



Figure 9 CN4054 10Gb Virtual Fabric Adapter

For more information, see *Flex System CN4054 and CN4054R 10Gb Virtual Fabric Adapters*, TIPS0868, which is available at this website:

<http://1enovopress.com/tips0868>

Flex System Interconnect Fabric technical details

This section describes the technical details of Flex System Interconnect Fabric. The following topics are covered:

- ▶ “Flex System Interconnect Fabric topology”
- ▶ “Hardware connectivity” on page 21
- ▶ “Traffic flow patterns” on page 23
- ▶ “High availability features” on page 25
- ▶ “Performance and scalability” on page 27
- ▶ “Advanced server-facing features” on page 28
- ▶ “Rolling firmware upgrade” on page 29

Flex System Interconnect Fabric topology

A typical Flex System Interconnect Fabric configuration includes the following components:

- ▶ Two G8264CS switches that serve as aggregation switches, and they connect the Interconnect Fabric POD to the upstream network. These two switches provide no single point of failure configuration, and they are interconnected with each other.

- ▶ Up to nine Flex System chassis, with two SI4093 System Interconnect modules in each chassis. Each SI4093 is connected to both of the G8264CS switches.
- ▶ Compute nodes in the Flex System chassis, x86 or IBM POWER, with the standard LOM or with EN4054, CN4054/CN4054R, or CN4058 adapters.
- ▶ Appropriate cabling to connect the components. It is likely that copper DAC cables can be used for most of the connections within the configuration.

It is possible to add additional chassis to the configuration and to add additional compute nodes to the chassis without disrupting the environment.

The G8264CS switches and SI4093 modules should be upgraded with the special Interconnect Fabric software image to enable Interconnect Fabric operations.

Flex System Interconnect Fabric uses double spine-leaf topology where each embedded leaf (member) switch is connected to two spine switches (master and backup), as shown in Figure 10. Master and backup switches are also connected to each other. All inter-switch links are dual 10 Gb connections.

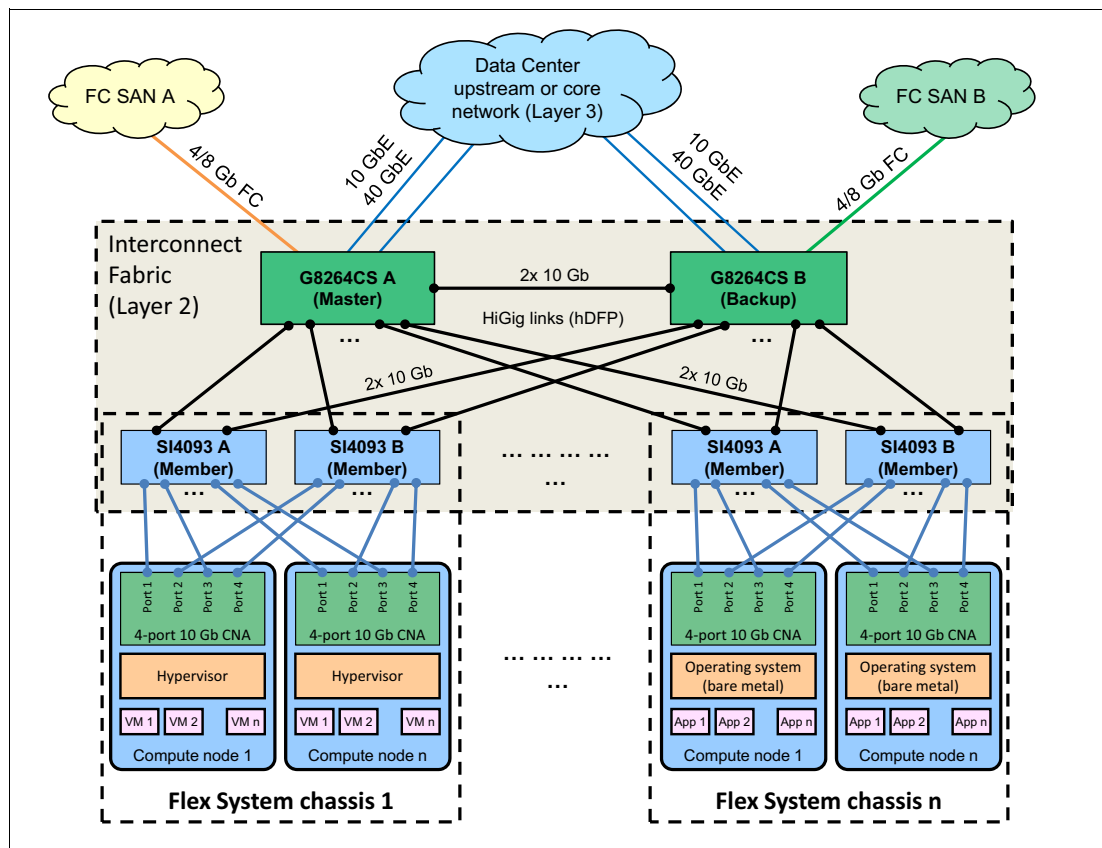


Figure 10 Interconnect Fabric topology

Configuration of the networking component is controlled through a master switch that propagates configuration information to all member switches. From a user configuration perspective, Interconnect Fabric appears as a large switching fabric where network administrators are responsible for provisioning server-facing ports on the SI4093s and upstream network ports on the G8264CS.

Flex System Interconnect Fabric software features

Flex System Interconnect Fabric supports the following software features:

- ▶ Single IP managed cluster
- ▶ Up to 1024 VLANs
- ▶ Up to 128,000 MAC addresses
- ▶ Layer 2 loop free solution with upstream data center core
- ▶ FCoE and native Fibre Channel support:
 - Up to 2,000 FCoE sessions
 - Up to 24 FC Forwarders (FCFs)
 - FIP Snooping
- ▶ Eight unicast traffic classes and four multicast traffic classes with configurable bandwidth
- ▶ Priority flow control for a maximum of two priorities
- ▶ UFP virtual port support (four per 10 Gb physical port)
- ▶ Jumbo frames up to 12 KB
- ▶ VMready:
 - Up to 4,096 Virtual Elements (VEs)
 - Up to 1,024 VM groups
 - Up to 2,048 VMs per local VM group
 - Up to 4,096 VMs per distributed VM group
- ▶ VMready and FCoE interoperability with UFP
- ▶ Tunneled VLAN domain (Q-in-Q) for multi-tenant customer VLAN isolation
- ▶ IGMP v2 snooping for multicast optimization (up to 3,000 IGMP groups)
- ▶ Up to 256 access lists and 128 VLAN maps for security and rate limiting policing
- ▶ Static port channel and static LACP (user assigned trunk ID):
 - Up to 32 uplink port channels (up to 16 links per port channel) to an upstream network
 - Up to 252 downlink port channels (up to six links per port channel) to compute nodes
- ▶ L2 Failover (Manual Monitor -(MMON))
- ▶ Hot Links
- ▶ VLAN-based load distribution for Hot Links (for active/active connectivity with non-vPC/non-vLAG upstream network)
- ▶ Industry-Standard Command-Line Interface (ISCLI)
- ▶ SNMP
- ▶ IPv6 support for management
- ▶ Staggered upgrade
- ▶ HiGig fabric:
 - Up to 64 HiGig links
 - Up to 32 HiGig trunks
 - Up to eight 10 Gb or two 40 Gb HiGig links per trunk
- ▶ Local preference for unicast traffic
- ▶ Port mirroring
- ▶ sFlow

- ▶ NTP
- ▶ DHCP/DNS client

Hardware connectivity

The SI4093, an embedded module, has 42 10GBASE-KR ports that connect to the compute nodes in the Flex System chassis through the midplane; there are three 10 Gb ports, which connect to each of the 14 slots in the chassis.

The G8264CS has 12 Omni Ports, which can be configured to operate either as 4/8 Gb Fibre Channel ports or as 10 Gb Ethernet ports. It also has an internal hardware module with a dedicated ASIC, which provides the FC gateway functionality (FCF, NPV).

Both the SI4093 and G8264CS have PHY interfaces for SFP+ transceivers and also QSFP+ transceivers that can run either as a single 40 GbE port or as a set of four 10 GbE ports.

Both the SI4093 and the G8364CS share an architecture that relies on two key chips:

- ▶ The Switch Processor (SP) chip in both of these devices is a Broadcom switching ASIC. This chip handles forwarded traffic through the switch, under the control of directives that are pushed in to it from the MP.
- ▶ The Management Processor (MP) is where the switch firmware runs. The MP processes the configuration text and pushes appropriate directives into the SP. Management of the switch or the entire fabric is run on the MP, including the CLI, browser interface, and SNMP. Protocols such as LLDP, ICMP, and others also run on the MP.

In the Interconnect Fabric, one of the G8264CS switches is the master and the other is a backup for purposes of managing the environment. This means that the MP on the master switch sends directives to all of the SP chips throughout the configuration. If the master switch fails, the MP on the backup (the other G8264CS) takes on this task.

A representation of the internal architecture of the switching elements is shown in Figure 11.

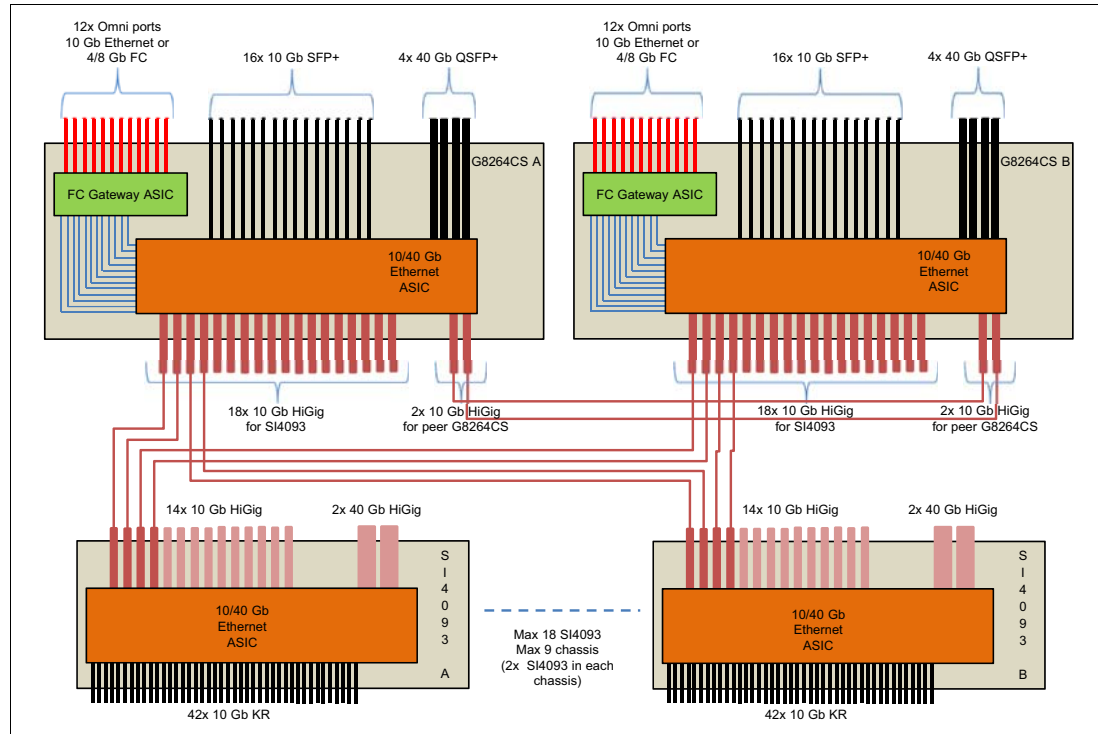


Figure 11 Interconnect Fabric hardware architecture

Flex System Interconnect Fabric uses standard 10 Gb connections between the SI4093 embedded modules and the G8264CS aggregation switches. Similar connections are used between the pair of G8264CS switches. Alternatively, 40 Gb connections can be used.

A Broadcom proprietary protocol, hDFP, is used over these links, which are referred to in the diagrams in this paper as HiGig links. This protocol carries proprietary control information and the content of the network traffic, and it enables the multiple SP chips in the different switches to operate as though they are part of a single switch.

Important: All HiGig links in Interconnect Fabric must operate at the same speed, that is, either 10 Gbps or 40 Gbps speeds but not both. At 10 Gbps fabric speeds, 40 Gb ports are always used as 4x 10 Gb HiGig links. At 40 Gb fabric speeds, 10 Gb ports cannot be used as HiGig links.

The links between switching elements in an Interconnect Fabric configuration are known as *Fabric Ports*. Fabric Ports must be explicitly configured on the G8264CS switches, and they are assigned to the VLAN 4090 by default. If two or more Fabric Ports are connected between the same two devices, then all of them are used because the aggregated link is formed automatically. After the Fabric Ports are selected, a reboot is required to make them active as Fabric Ports.

Tip: If you must add a chassis to Interconnect Fabric, complete the following steps:

1. Configure Fabric Ports on the G8264CS switches.
2. Reboot the master switch and wait until it comes back and rejoins the fabric.
3. Reboot the backup switch and wait until it comes back and rejoins the fabric.
4. Power on the new chassis.

As shown in Figure 11 on page 22, a typical Interconnect Fabric configuration uses four 10 Gb ports as Fabric Ports from each SI4093, two to each of the aggregation switches, providing 1:3.5 or 1:7 oversubscription ratios for the compute nodes with 2-port or 4-port CNAs, respectively.

The maximum of nine chassis with a total of 18 SI4093 modules is supported in a single Interconnect Fabric domain. For smaller Interconnect Fabric deployments, consider using more than two HiGig links from the SI4093 module to each of the G8264CS switches, for a total of up to eight HiGig links. Also, if you expect that the north-south traffic will prevail over the east-west traffic, consider adding more than two HiGig links between the G8264CS switches, for a total of up to eight HiGig links, because this HiGig trunk might be heavily used because of distributed uplink aggregation.

Important: In the Interconnect Fabric mode, all external ports on the SI4093 modules are configured as Fabric Ports by default, and this cannot be changed.

Fabric Ports do not need additional configuration other than what identifies them as Fabric Ports. They carry the hDFP proprietary protocol, which allows them to forward control information and substantive data traffic from one switching element to another within the Interconnect Fabric environment.

G8264CS ports that are not configured as Fabric Ports can be used as uplink ports.

Important: Omni Ports cannot be configured as Fabric Ports.

Uplink ports are used for connecting the Interconnect Fabric POD to the upstream data network (standard ports and Omni Ports) and to the storage networks (Omni Ports only). By default, all uplink ports are added to the black hole VLAN 4091.

Traffic flow patterns

Interconnect Fabric is formed by establishing HiGig links across the leaf and spine switches using the configured Fabric Ports. All HiGig links are active, and they carry network traffic. The traffic flow path between the different members in the fabric is established when a member joins the fabric or other topology change occurs. The data path should be balanced as well as possible.

Note: No changes in the traffic flow patterns are made until the topology change occurs. If a topology change is detected, the fabric reconfigures automatically.

Figure 12 shows the example traffic flows from different clients across the fabric.

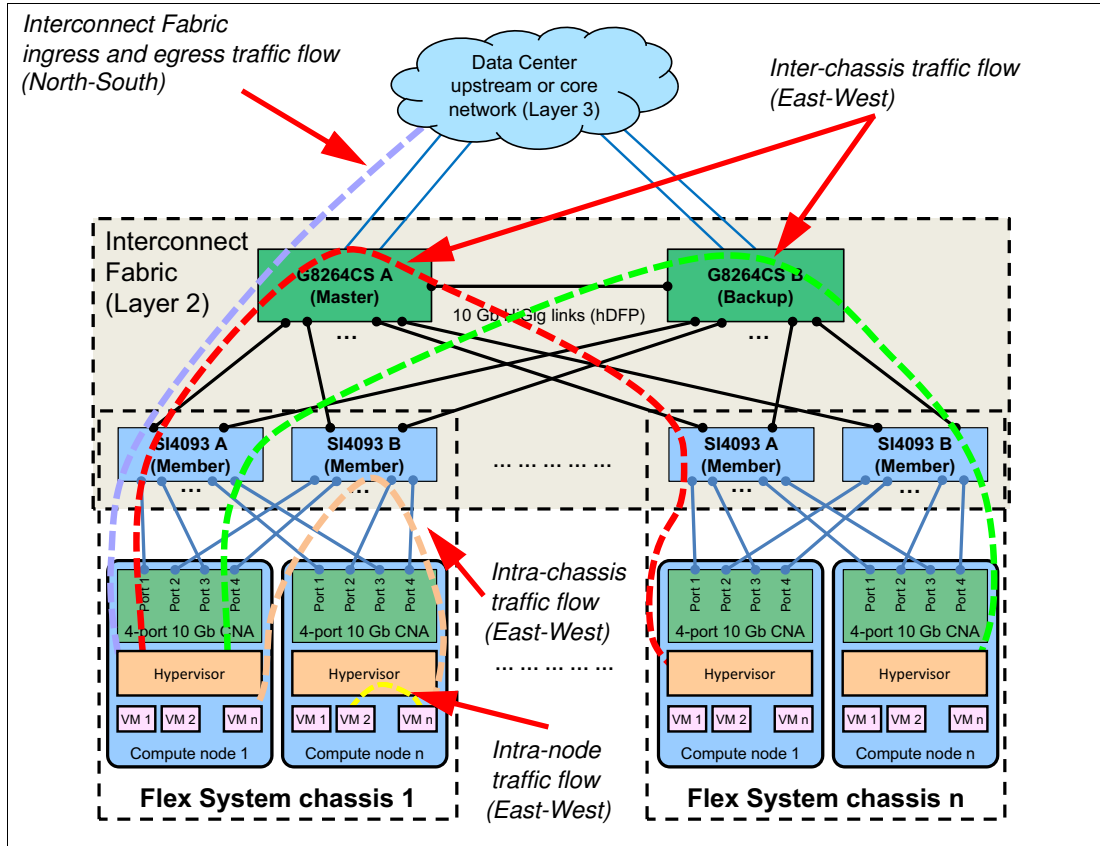


Figure 12 Interconnect Fabric internal traffic flow

For east-west type of traffic, in the topology where all embedded switches are connected to both aggregation switches, the underlying fabric tries to statically load balance the known unicast traffic between a pair of embedded switches in different chassis. For example, the clients that are connected to the SI4093 A module in the Flex System chassis 1 communicate with the clients that are connected to the SI4093 A modules in other chassis through the G8264CS A master switch using the following path:

SI4093 A (chassis 1) → G8264CS A → SI4093 A (chassis n)

At the same time, the same clients that are connected to the SI4093 A module in the Flex System chassis 1 communicate with the clients that are connected to the SI4093 B modules in other chassis through the G8264CS B backup switch using the following path:

SI4093 A (chassis 1) → G8264CS B → SI4093 B (chassis n)

Traffic across chassis (labeled *inter-chassis* in Figure 12) always flow through one of the G8264CS switches. Traffic across the compute nodes in the same chassis (labeled *intra-chassis* in Figure 12) is switched locally without leaving the chassis, which helps decrease server-to-server communication latency and improve overall response time. Traffic across VMs that are on the same compute node (labeled *intra-node* in Figure 12) does not leave the compute node, and switching is performed by the vSwitch in the hypervisor.

For north-south type of traffic (labeled *ingress* and *egress* in Figure 12 on page 24), when the packet from the compute node enters the fabric, the ingress SI4093 applies trunk hashing rules to determine the egress uplink trunk port that is used by the packet to leave the fabric. If the trunk port is on the G8264CS A, then the packet takes the nearest fabric path from the ingress SI4093 toward G8264CS A.

High availability features

Clients might require continuous access to their network-based resources and applications. Providing high availability (HA) for client network-attached resources can be a complex task that involves fitting multiple pieces together on a hardware and software level.

Network infrastructure availability can be achieved by using certain techniques and technologies. Most techniques and technologies are widely used standards, but some are specific to the Interconnect Fabric. This section reviews the most common technologies that can be implemented in an Interconnect Fabric environment to provide high availability to the network infrastructure.

The first step in achieving HA is to provide physical redundancy of components that are connected to the infrastructure. Interconnect Fabric has redundant components throughout. Here are the necessary tasks:

- ▶ Deploy dual-, quad-, or eight-port NICs.
- ▶ Deploy leaf switches in pairs.
- ▶ Deploy spine switches in a pair.
- ▶ Connect the pair of node NIC ports to separate leaf switches in the Flex System chassis.
- ▶ Provide redundant connections from each leaf switch to the spine switches.
- ▶ Provide redundant connections from each spine switch to an upstream infrastructure.

The environment survives failures of the following types if it is configured properly:

- ▶ Failure or loss of uplink to the upstream network in one of the two G8264CS spine switches
- ▶ Failure or loss of a HiGig link to the G8264CSs in one of the two SI4093 modules in a given chassis (built-in fabric resiliency feature)
- ▶ Failure of a single NIC port on a server, which is an unlikely event
- ▶ Failure of either the SAN A or SAN B networks or the connection from the Interconnect Fabric environment to them

Interconnect Fabric achieves survivability from these failures through the following features:

- ▶ Distributed LAGs for upstream connections and for connectivity to the compute nodes
- ▶ Hot Links
- ▶ Layer 2 failover
- ▶ Fabric path monitoring

Consideration: With distributed LAGs (see “Distributed link aggregation” on page 27) or Hot Links with VLAN load balancing (see “VLAN-based load distribution on Hot Links” on page 27), it is possible to configure traffic distribution across all available links, thus providing higher overall network bandwidth and dynamic redundancy at the same time.

Hot Links

Hot Links provide basic link redundancy with fast recovery. In Interconnect Fabric, Hot Links consist of up to 200 triggers. A *trigger* is a pair of Layer 2 interfaces, each containing an individual port or aggregated link. One interface is the *master*, and the other is the *backup*.

The master interface is set to the active state and forwards traffic, and the backup interface is set to the standby state and blocks traffic until the master interface fails. If the master interface fails, the backup interface is set to active and forwards traffic. By default (preemption enabled), after the master interface is restored, it goes back into the active state and begins forwarding traffic, and the backup interface becomes standby and blocks the traffic.

Interconnect Fabric supports VLAN load balancing over Hot Links. For each Hot Link pair, one uplink can be the master for one set of VLANs and the backup for another set of VLANs and vice versa (for more information about Hot Links and VLAN load balancing over Hot Links, see “Standard STP domain integration” on page 34).

Interconnect Fabric supports G8264CS uplink interfaces as Hot Links. They can be physical interfaces and static or LACP link aggregations.

Layer 2 Failover

Layer 2 Failover (also known as *trunk failover* or *link state tracking*) is primarily used with active/standby NIC teaming to ensure the compute nodes can detect an uplink failure on the upstream network-facing switches and reroute the traffic accordingly.

You can configure a set of interfaces (ports or aggregated links) to monitor for link failures (a monitor list), and another set of interfaces to disable when the link fails (a control list). When the switch detects a link failure on the monitor list, it automatically disables the interfaces in the control list. When the compute node-facing ports are disabled, the corresponding compute node’s network adapter can detect the disabled link and trigger a network adapter failover to another switch in the chassis. The switch automatically enables the control list items when the monitor list items return to service.

You can use Interconnect Fabric to monitor uplink ports on the master and backup switches for failures; the controlled interfaces are compute node-facing interfaces on the SI4093 modules.

Note: Layer 2 Failover monitoring is not supported on Hot Link interfaces.

With Interconnect Fabric, it is not common to use Layer 2 Failover for the upstream network connectivity because distributed LAGs or Hot Links with VLAN load balancing provide higher network bandwidth and dynamic high availability for the uplink connections. However, Layer 2 Failover can be used to monitor external storage connections on the G8264CS switches and to provide faster failover of FCoE traffic if there is an external storage link failure. In such a case, when an external link failure is detected, the compute node-facing links in the control list are disabled, and then the MPIO driver in the operating system detects this HBA port link failure and switches storage traffic to other available HBA ports.

Fabric path monitoring

When Layer 2 Failover is configured for the external FC interfaces (Omni Ports running FC), the system can additionally monitor HiGig connections between the master or backup switch and the member switches where the controlled interfaces are configured. This feature is called *fabric path monitoring*, and it allows faster failover of FCoE traffic if there is a failure of the master or backup switch.

Fabric path monitoring determines which HiGig links of the member switch (where controlled interfaces are) are connected to the master or backup switch. When a failure is detected on all monitored HiGig links in the failover profile, the member switch immediately brings down associated controlled ports.

Performance and scalability

Another major topic that is part of Interconnect Fabric network planning is network performance and scalability. The commonly used features include link aggregation, distributed link aggregation, VLAN load distribution across Hot Links, and active/active NIC teaming.

Link aggregation

Sometimes referred to as *trunking*, *port channel*, or *Etherchannel*, link aggregation involves taking multiple physical links and binding them into a single common link for use between two devices. The primary purposes of aggregation are to improve HA and increase performance.

The two most common types of aggregation that are supported by Interconnect Fabric are static port channel and static Link Aggregation Control Protocol (LACP) port channel.

Static aggregation does not use any protocol to create the aggregation. Instead, static aggregation combines the ports based on the aggregation configuration that is applied on the ports and assumes that the other side of the connection does the same. LACP is a dynamic way of determining whether both sides of the link agree they should be aggregating. Static LACP means that although the trunk negotiation is dynamic, the trunk ID is manually assigned by the network administrator.

The decision to use static aggregation or static LACP is usually a question of what a client uses in their network. Static aggregation is the quickest and easiest way to build an aggregated link. The use of static aggregation can also be advantageous in mixed vendor environments because it can help prevent possible interoperability issues.

One of the downsides to static aggregation is that it lacks a mechanism to detect if the other side is correctly configured for aggregation. So, if one side is static and the other side is not configured, configured incorrectly, or is not connected to the correct ports, it is possible to cause a network outage by bringing up the links. Static LACP has the inherent safety that a protocol brings to the link aggregation process, which helps protect the network from an outage.

For upstream network connections, Interconnect Fabric supports up to 32 link aggregation groups (LAGs), and each LAG supports up to 16 physical uplink interfaces. For server-side connectivity, Interconnect Fabric supports up to 252 LAGs with up to six ports per LAG.

Distributed link aggregation

Aside from the standard point-to-point aggregations, there is a technology that provides multi-switch aggregation, called *distributed link aggregation*. The primary goals of distributed link aggregation are to provide better performance and availability compared to standard point-to-point aggregation by combining ports on different switches in a pair into a single LAG.

In Interconnect Fabric, LAGs can be distributed across a pair of the G8264CS switches for the upstream network connectivity. Also, distributed LAGs are supported across a pair of the SI4093 modules in the same chassis for the server-side LAGs.

VLAN-based load distribution on Hot Links

VLAN load-balancing is a feature that allows both Hot Link interfaces to simultaneously forward the traffic for mutually exclusive VLANs. When VLAN load-balancing is enabled, the configured VLANs are automatically distributed over the two interfaces, unless preferred VLANs are configured.

For example, if a Hot Link pair is configured for VLANs 1-10, the traffic for five VLANs (VLANs 1 - 5) can be forwarded through one interface and for the remaining five VLANs (VLANs 6 - 10) through the other interface. If one of the active interfaces goes down, the other active interface forwards all the traffic. When the failed interface recovers, it resumes forwarding traffic in its assigned VLANs.

NIC teaming

The terms *bonding* and *teaming* are different words for the same thing. In general, in Linux, it is referred to as bonding, and in Windows and VMware, it is referred to as teaming. Regardless of the term, these technologies provide a way to allow two or more NICs to appear and operate as a single logical interface, for the purpose of high availability, increased performance, or both.

Each OS has its own way of providing these services, with most having native built-in support, but some older operating systems still require a third-party application to provide this function.

Most OSes support both an LACP and a static form of teaming/bonding, and these are all forms of active-active teaming/bonding, usually load balancing traffic on a per-session basis (what constitutes a *session* is controlled by settings on each side of the device supporting this mode of teaming/bonding).

Interconnect Fabric appears a single logical switch for all compute nodes in the POD, so all NIC ports on the same compute node are connected to the same logical switch, enabling static or dynamic link aggregation configurations between the compute node and the switch. All links in the bundle can be active and forwarding the traffic.

Advanced server-facing features

This section describes advanced compute node-facing features that can be implemented with Interconnect Fabric, including Unified Fabric Port (UFP), VMready with NMotion®, and Q-in-Q VLAN tunneling.

Unified Fabric Port

UFP allows a 10 Gb port on a compute node to be seen by the server's operating system as four virtualized NICs (called *vPorts*), or as three virtualized NICs and a virtual HBA for storage access.

All of the modes of UFP are supported in the Interconnect Fabric environment, and they are explicitly configured on the master aggregation switch (G8264CS). The configuration applies to the ports on the SI4093 modules.

Here are the available UFP modes:

- ▶ Access mode: Allows a single VLAN on the vPort.
- ▶ Trunk mode: Allows multiple VLANs on the vPort.
- ▶ Tunnel mode: Carries VLANs that are configured on the compute node to the upstream network through a Q-in-Q tunnel.
- ▶ FCoE mode: Carries FCoE storage traffic.
- ▶ Auto-VLAN mode: Discovers VMs through VMready and dynamically provisions VLAN for vPort depending on VMready discovery.

VMready

VMready is a unique solution that enables the network to be VM-aware. The network can be configured and managed for virtual ports (vPorts) rather than just for physical ports. With VMready, as VMs migrate across physical hosts, so do their network attributes. VMs can be added, moved, and removed while retaining the same ACLs, QoS, and VLAN attributes. VMready allows for a define-once-use-many configuration that evolves as the server and network topologies evolve. VMready works with all virtualization products, including VMware, Hyper-V, KVM, and PowerVM, without modification of VM hypervisors or guest operating systems.

VMready automatically can discover VMs on hypervisor-based hosts that are connected to internal switch ports and pre-provision network connections and group membership for Virtual Entities. In addition, VMready together with NMotion allows seamless migration/failover of VMs to different hypervisor hosts, preserving network connectivity configurations.

Q-in-Q VLAN tunneling

In a multi-tenant data center, clients might have specific requirements for VLAN IDs and the number of VLANs they want to use. Different clients might also use VLAN IDs that overlap with each other, which might cause mixing of traffic from different clients.

With 802.1Q tunneling, data center administrators can use a single VLAN within Interconnect Fabric to support a client who might need multiple VLANs. Client VLANs are encapsulated into a tunnel VLAN and transparently forwarded through the fabric. In such a case, different clients use different tunnel VLANs, and they are logically isolated from each other even if the client VLAN IDs are the same.

Each tunneled VLAN domain should be connected to a separate upstream network so that the traffic from overlapping VLANs is not mixed in the upstream network.

Another scenario for Q-in-Q tunneling might be when the network administrator relies on the upstream network for VLAN configuration, and the upstream VLANs transparently communicate to the servers.

Rolling firmware upgrade

A rolling firmware upgrade allows the firmware on the two G8264CS aggregation switches and the embedded SI4093 modules to be upgraded sequentially rather than having the entire cluster upgraded concurrently. By using this feature, you avoid the need for an outage, which takes the environment down for several minutes while all of the switching hardware reboots.

The upgrade is performed by first copying the new firmware to all members of the fabric. The master switch (G8264CS) then triggers the other members of the fabric to reboot sequentially. This is done in the following order:

1. Backup G8264CS
2. Master G8264CS
3. SI4093 switches in bay 1 of their chassis, all at the same time
4. SI4093 switches in bay 2 of their chassis, all at the same time

During the firmware upgrade process, there should be no time when connectivity to the upstream network is lost. From the point of view of the fabric, it is as though a series of switch and uplink failures are occurring; while the design is cabled and configured properly, the environment redirects traffic as described under “High availability features” on page 25.

Note: A rolling upgrade applies to the switch firmware and not to the boot loader. However, if a new boot loader and new firmware are installed together, both files are copied to each member of the fabric. The rolling reboot of the elements of the fabric then loads the new boot loader without a fabric-wide outage.

Point of delivery network integration

This section describes upstream network integration considerations for the Flex System Interconnect Fabric POD. The following integration scenarios are covered:

- ▶ “Compute node connectivity” on page 30
- ▶ “Virtual link aggregation domain integration” on page 32
- ▶ “Standard STP domain integration” on page 34
- ▶ “Disjoint VLANs” on page 36
- ▶ “Multi-tenancy” on page 37

Compute node connectivity

Figure 13 shows an example network connectivity topology for the compute nodes. This topology is common across the scenarios that are described in this paper. All deviations from the described topology are specifically mentioned, if any.

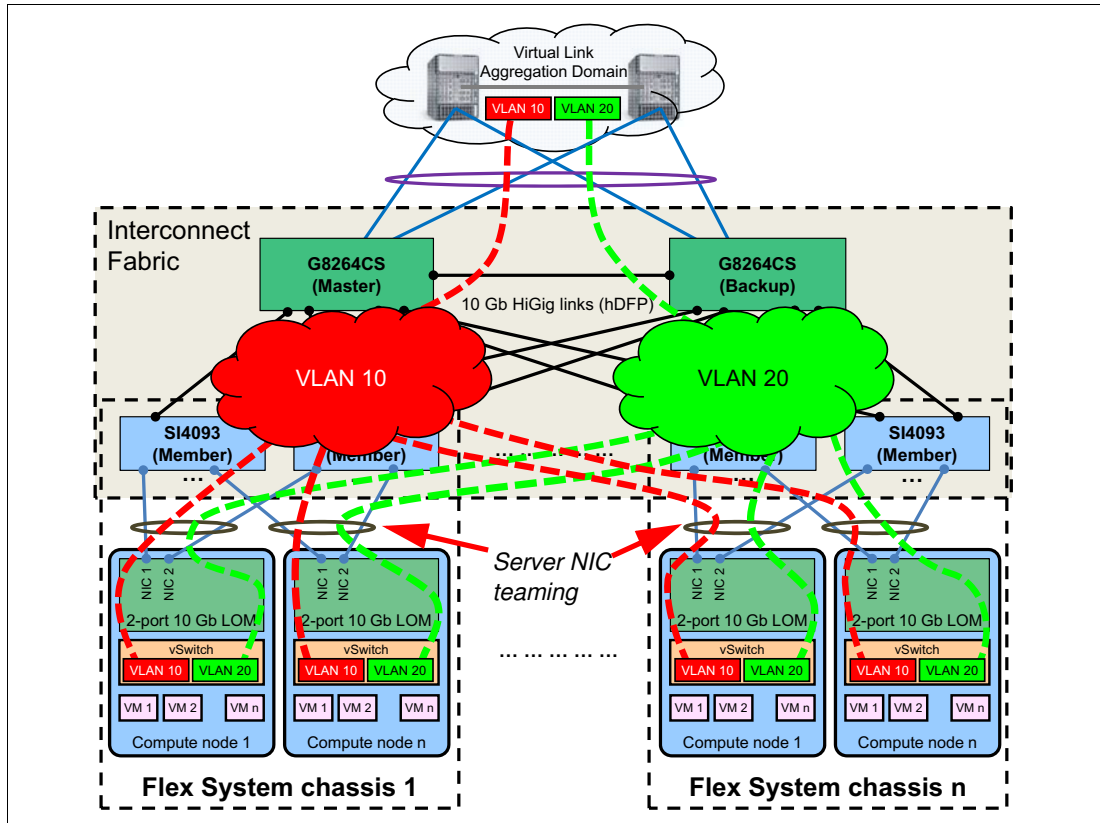


Figure 13 Compute node connectivity - VLAN-aware fabric

In this example topology, a virtualized infrastructure is deployed within the Interconnect Fabric POD. VMs are running on the compute nodes, and each VM is connected to one of the client VLANs (either VLAN 10 or VLAN 20).

The Interconnect Fabric can be VLAN-aware or VLAN-agnostic, depending on specific client requirements. In VLAN-aware mode (shown in Figure 13 on page 30), the client VLAN isolation is extended to the fabric by filtering and forwarding VLAN tagged frames based on the client VLAN tag, and client VLANs from the upstream network are configured within the Interconnect Fabric and on virtual switches (vSwitches) in hypervisors.

In VLAN-agnostic mode, the Interconnect Fabric transparently forwards VLAN tagged frames without filtering on the client VLAN tag, providing an end host view to the upstream network, where client VLANs are configured on vSwitches only. This is achieved by using a Q-in-Q type operation to hide user VLANs from the switch fabric in the POD so that the Interconnect Fabric acts as more of a port aggregator and is user VLAN-independent.

The VLAN-agnostic mode of the Interconnect Fabric can be implemented through the tagpvid-ingress feature or UFP vPort tunnel mode. If no storage access is required for the compute nodes in the POD, then the tagpvid-ingress mode is the simplest way to configure the fabric. However, if you want to use FCoE storage, you cannot use tagpvid-ingress and must switch to UFP tunnel mode.

As an example of how Q-in-Q tagging works, consider the tagpvid-ingress operation. When all internal ports and the wanted uplink ports are placed into a common PVID/Native VLAN and tagpvid-ingress is enabled on these ports (with any wanted aggregation protocol on the uplinks that are required to match the other end of the links), all ports with this Native or PVID setting are part of Q-in-Q tunnel. The Native/PVID VLAN acts as the outer tag (and switches traffic based on this VLAN). The inner customer tag rides through the fabric on the Native or PVID VLAN to the wanted port (or ports) in this tunnel. UFP tunnel mode operates in a similar way.

VLAN-agnostic Interconnect Fabric topology is shown in Figure 14. Tunnel VLAN 4001 is created within the POD, and it is used as the native VLAN for outer tagging. Multiple tunnels can be established in a multi-tenant environment, but each tunnel must have a dedicated uplink connection.

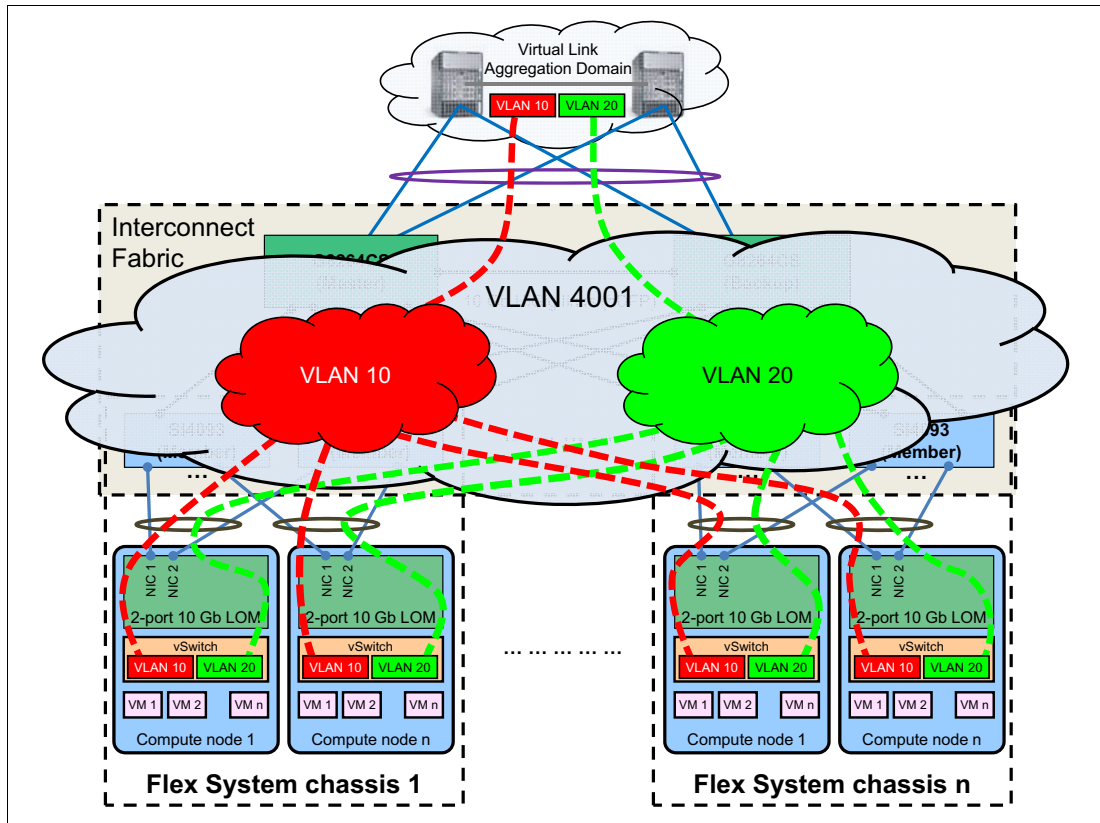


Figure 14 Compute node connectivity - VLAN-agnostic fabric

All VMs that are connected to the same client VLAN can communicate with each other within the POD and with other endpoints attached to this VLAN in the upstream network. VMs and endpoints that are connected to different VLANs cannot communicate with each other in the Layer 2 network.

vSwitches are connected to Interconnect Fabric through a teamed NIC interface that is configured on the compute node. The compute node's CNA NIC ports (either physical (pNIC) or virtual (UFP vPort) ports) are configured in a load balancing pair (static or static LACP aggregation) using the hypervisor's teaming/bonding feature. Respective compute node-facing ports on the SI4093 modules are also configured for static or dynamic aggregation, and VLAN tagging (802.1Q) is enabled on the aggregated link.

Virtual link aggregation domain integration

Distributed (or virtual) link aggregation (such as Cisco vPC or vLAG) is the most flexible way to connect the Interconnect Fabric POD to the upstream network. However, this requires support for some form of virtual link aggregation from the upstream switches. If upstream switches do not support virtual link aggregation, this integration scenario cannot be used.

Figure 15 shows example integration topology that uses virtual link aggregation.

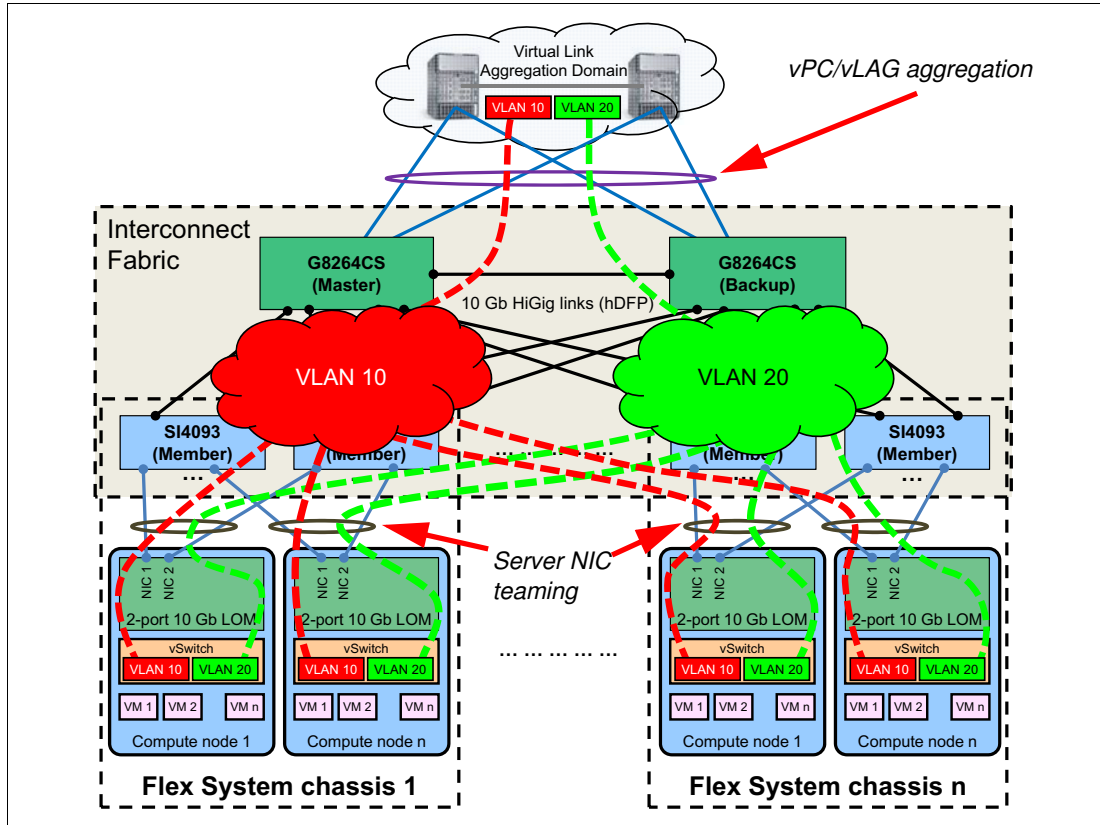


Figure 15 Virtual link aggregation integration (vPC/vLAG)

In this example topology, each G8264CS switch is connected to both upstream network switches in a single virtual link aggregation domain. The Interconnect Fabric POD sees the upstream network as one logical switch, and the upstream network sees the Interconnect Fabric POD as one logical switch as well.

A network administrator must configure a single aggregated port channel (static or dynamic) between these two logical switches using all the connected uplinks, and all these links in the aggregation carry traffic from all client VLANs. If any of physical links in the aggregation fails, the port channel continues to operate using the remaining active links.

If virtual link aggregation is not supported on the upstream network switches (that is, the upstream network operates in a standard STP domain), then Hot Link interfaces are used.

Standard STP domain integration

Figure 16 shows example integration topology that uses Hot Link interfaces in a standard STP domain.

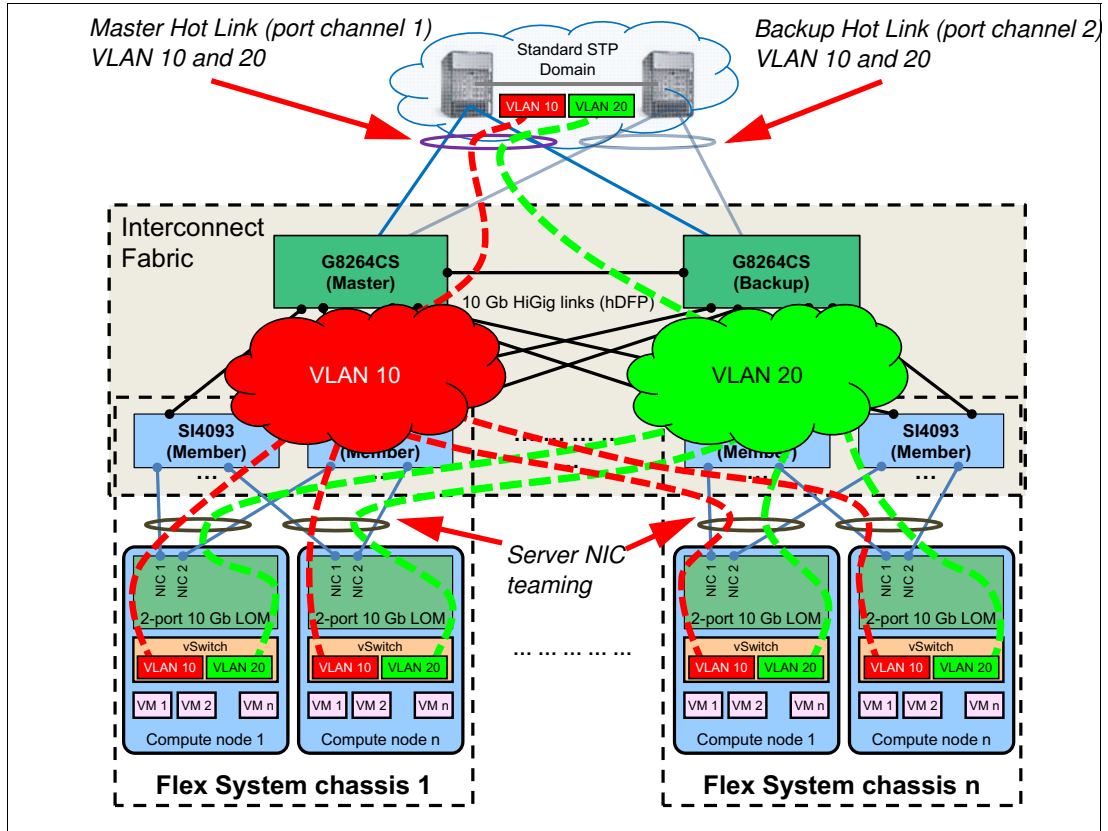


Figure 16 Standard STP domain - master and backup Hot Links

In this example topology, each G8264CS switch is connected to both upstream network switches in a single STP domain. The Interconnect Fabric POD sees the upstream network as two separate switches, and the upstream network sees the Interconnect Fabric POD as one logical switch.

The network administrator must configure two port channels (static or dynamic):

- ▶ Port channel 1 between the first upstream switch and the Interconnect Fabric logical switch
- ▶ Port channel 2 between the second upstream switch and the Interconnect Fabric logical switch

Port channel 1 is designated as the master Hot Link, and port channel 2 is configured as the backup Hot Link. The port channel 1 carries traffic from all client VLANs, and the port channel 2 is in blocking state. In case of a port channel 1 failure, port channel 2 becomes active and all traffic flows through port channel 2.

The downside of this approach is that only a half of available uplink bandwidth is used. Interconnect Fabric supports VLAN load distribution across Hot Links to maximize bandwidth usage.

Figure 17 shows example integration topology that uses Hot Link interfaces with VLAN load distribution.

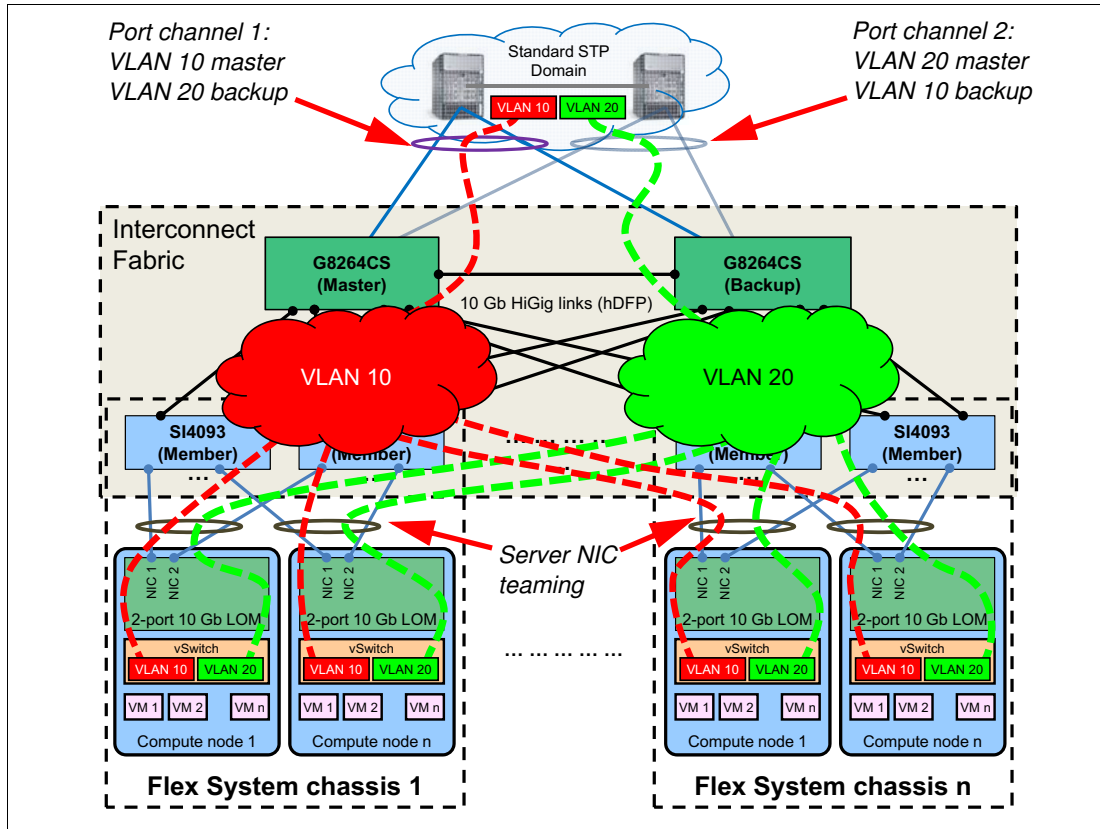


Figure 17 Standard STP domain - master and backup Hot Links with VLAN load distribution

The difference from the scenario that shown in Figure 16 on page 34 is that both Hot Links are master and backup for different VLANs at the same time.

Port channel 1 is the master Hot Link for VLAN 10 and the backup Hot Link for VLAN 20, and port channel 2 is the backup Hot Link for VLAN 10 and the master Hot Link for VLAN 20. In case of a Hot Link failure, the traffic from all VLANs is carried over the remaining master Hot Link.

Disjoint VLANs

It is not uncommon to have separate Layer 2 domains for different VLANs in the upstream network. For example, VLAN 10 is a public network and VLAN 20 is a production network, and they do not have adjacencies with each other. At the same time, VMs running on the Flex System compute nodes within the POD might need access to both networks. This scenario is shown in Figure 18.

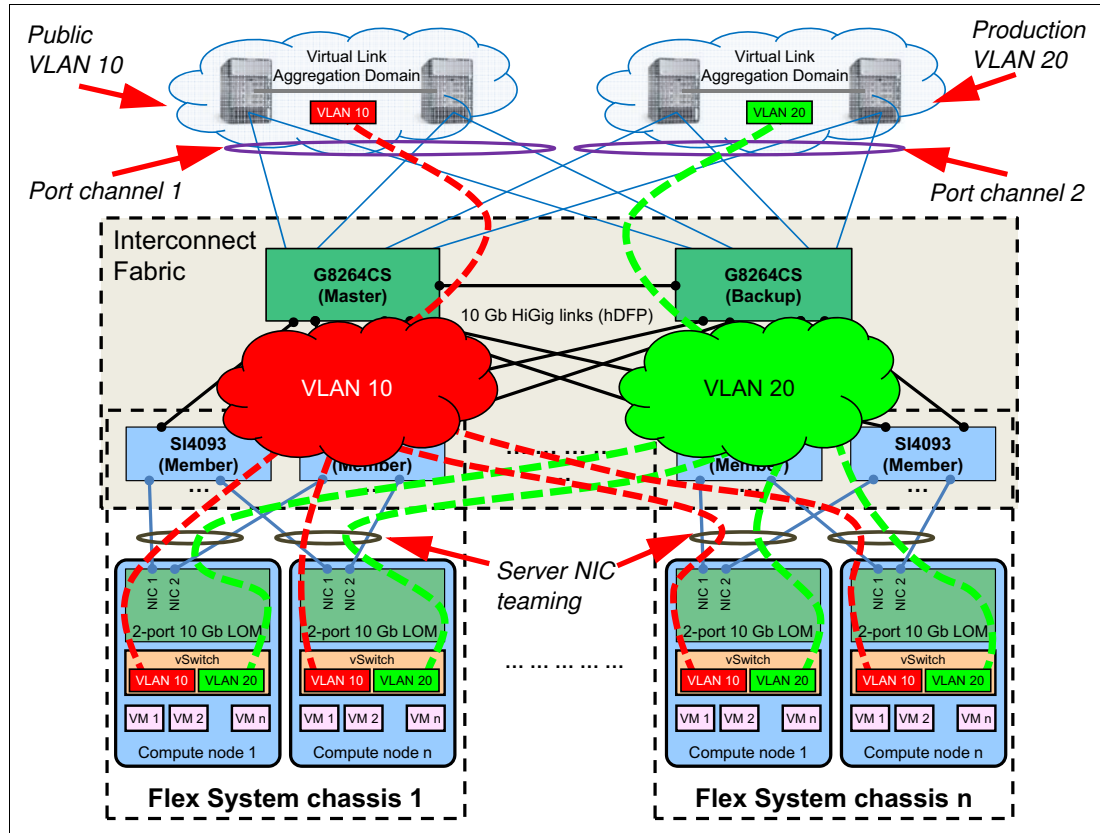


Figure 18 Disjoint VLANs

In this example topology, there are two isolated vPC domains, and each G8264CS switch is connected to all upstream network switches. The Interconnect Fabric POD sees two upstream logical switches (VLAN 10 domain and VLAN 20 domain), and each of the upstream network domains sees the Interconnect Fabric POD as one logical switch.

The network administrator must configure two port channels:

- ▶ Port channel 1 between the logical switch hosting VLAN 10 domain and the Interconnect Fabric logical switch
- ▶ Port channel 2 between the logical switch hosting VLAN 20 domain and the Interconnect Fabric logical switch

All uplinks in each port channel are active and they carry traffic from the respective VLANs (VLAN 10 for port channel 1 and VLAN 20 for port channel 2). If any of physical links in the aggregation fails, the port channel continues to operate using the remaining active links.

Note: VLANs in the upstream network cannot overlap.

Multi-tenancy

In a multi-tenant data center, clients might have specific requirements for VLAN IDs and the number of VLANs that they want to use. Different clients might also use VLAN IDs that overlap with each other. An example multi-tenant topology for the Interconnect Fabric POD is shown in Figure 19.

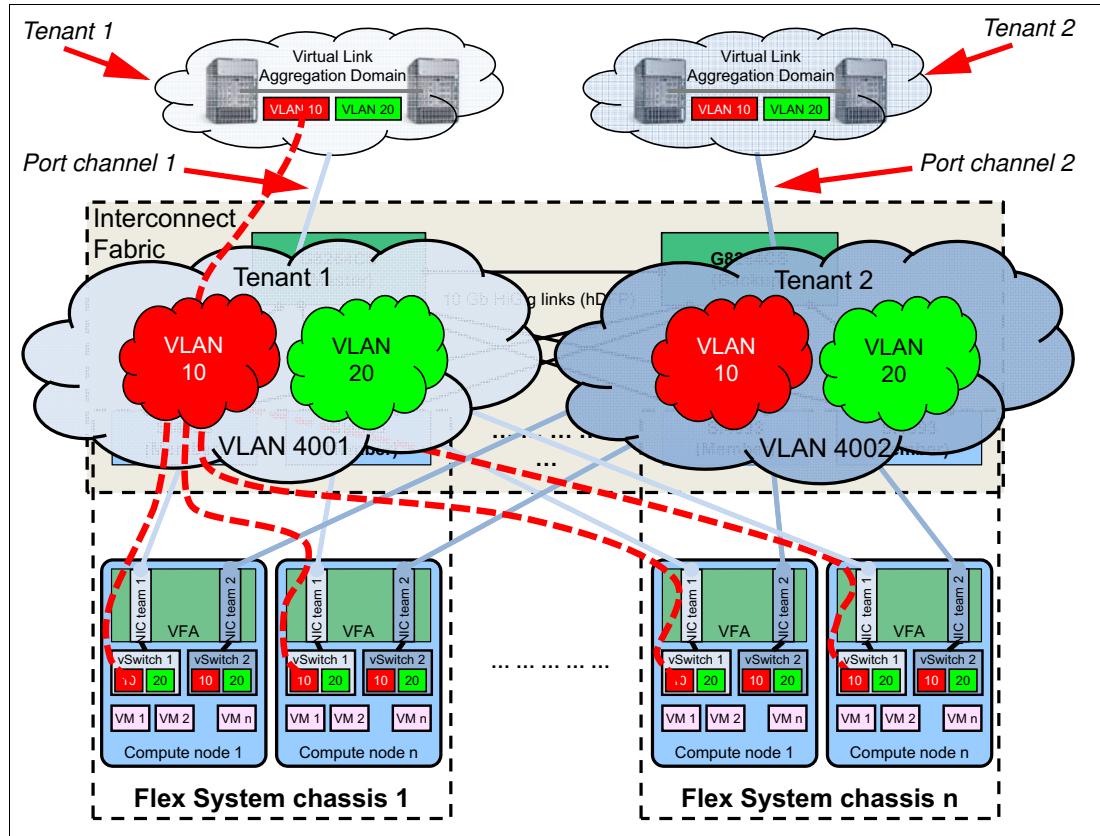


Figure 19 Multi-tenancy

In this example topology, there are two isolated vPC domains (for tenants 1 and 2), and each G8264CS switch is connected to all upstream network switches. The Interconnect Fabric POD sees two upstream logical switches, and each of the upstream network domains sees the Interconnect Fabric POD as one logical switch.

The network administrator must configure two port channels:

- ▶ Port channel 1 between the logical switch hosting the tenant 1 and the Interconnect Fabric logical switch
- ▶ Port channel 2 between the logical switch hosting the tenant 2 and the Interconnect Fabric logical switch

All uplinks in each port channel are active and they carry traffic from the respective tenant's VLANs (VLAN 10 and VLAN 20 for both tenants). If any of physical links in the aggregation fails, the port channel continues to operate using the remaining active links.

VLANs with the same ID but for different tenants are isolated from each other. This is achieved by configuring tunnel VLANs within Interconnect Fabric: VLANs 4001 and 4002 for the tenants 1 and 2, respectively.

A Q-in-Q tunnel is configured between the compute node-facing physical or virtual aggregated ports on the SI4093 that match the ports of the teamed NICs on the compute node and the aggregated uplink connection to the tenant's upstream network. An aggregated link that uses ports from NIC team 1 that face the SI4093 modules is assigned tunnel VLAN 4001. An aggregated link that uses ports from NIC team 2 that face the SI4093 modules is assigned tunnel VLAN 4002. Also, the port channel 1 interface on the Interconnect Fabric upstream side is assigned tunnel VLAN 4001, and the port channel 2 interface on the Interconnect Fabric upstream side is assigned tunnel VLAN 4002.

Hypervisor network configuration is also modified. Each hypervisor now has two isolated vSwitches, and each vSwitch has a dedicated uplink connection (two NIC teams are configured and each NIC team belongs to different vSwitch).

For example, the client VLAN 10 tagged traffic for the tenant 1 enters the Interconnect Fabric server-facing edge (the SI4093) and the tunnel VLAN 4001 tag is added to the frame. Then, the double-tagged frame is forwarded to all VLAN 4001 tunnel ports (port channel 1 uplinks and downlinks to the NIC team 1 on each compute node) where the VLAN 4001 tag is removed, and the original frame with only the client VLAN 10 tag is forwarded to the compute node and to the upstream network.

This ensures that traffic flow from different tenants do not interfere with each other. Tenants in the upstream network cannot overlap; each tenant must have its own domain.

Point of delivery storage integration

This section describes the POD storage architecture and different storage integration scenarios. The following topics are covered:

- ▶ "Storage connectivity architecture"
- ▶ "Internal SAN storage" on page 39
- ▶ "External FC SAN storage" on page 41

Storage connectivity architecture

The Flex System Interconnect Fabric converged network design enables shared access to the full capabilities of the FCoE-based V7000 Storage Node and Storwize V7000 storage systems while simultaneously providing connectivity to the client's enterprise FC SAN storage environment.

Interconnect Fabric introduces a new storage fabric capability that is called Hybrid Fabric in which there are two types of SANs, internal and external, that are on separate SAN fabrics. The internal SAN is used for the POD-wide V7000 Storage Node or Storwize V7000 connectivity, and the external SAN is used for the data center-wide storage connectivity. Both internal and external SANs are dual-fabric SANs. The total number of fabrics in the hybrid storage configuration is four.

The G8264CS switches support this capability by partitioning their internal FC switch fabric into two segments that support the following capabilities:

- ▶ FC SAN fabric services for FC targets that are attached to FC Omni Ports and FCoE targets attached to SI4093 or G8264CS FCoE ports. This mode is known as Full Fabric mode and is associated with an independent FCF VLAN.
- ▶ N_Port support for FC Omni Port attachment to FC SAN switch F_ports. This mode is known as NPV mode and is also associated with an independent FCF VLAN.

Hybrid mode requires dual initiators per compute node connection to each SI4093 module so that each initiator can discover one FC fabric. Each initiator can communicate with only one FCF VLAN and FC fabric. The 4-port and 8-port CNAs offer the required number of ports to support dual switch path storage access.

Note: 2-port CNAs (such as embedded VFA LOM) can also be used for storage connectivity, but only one type of dual-fabric SAN can be used (either internal or external but not both).

Figure 20 shows an Interconnect Fabric storage connectivity architecture.

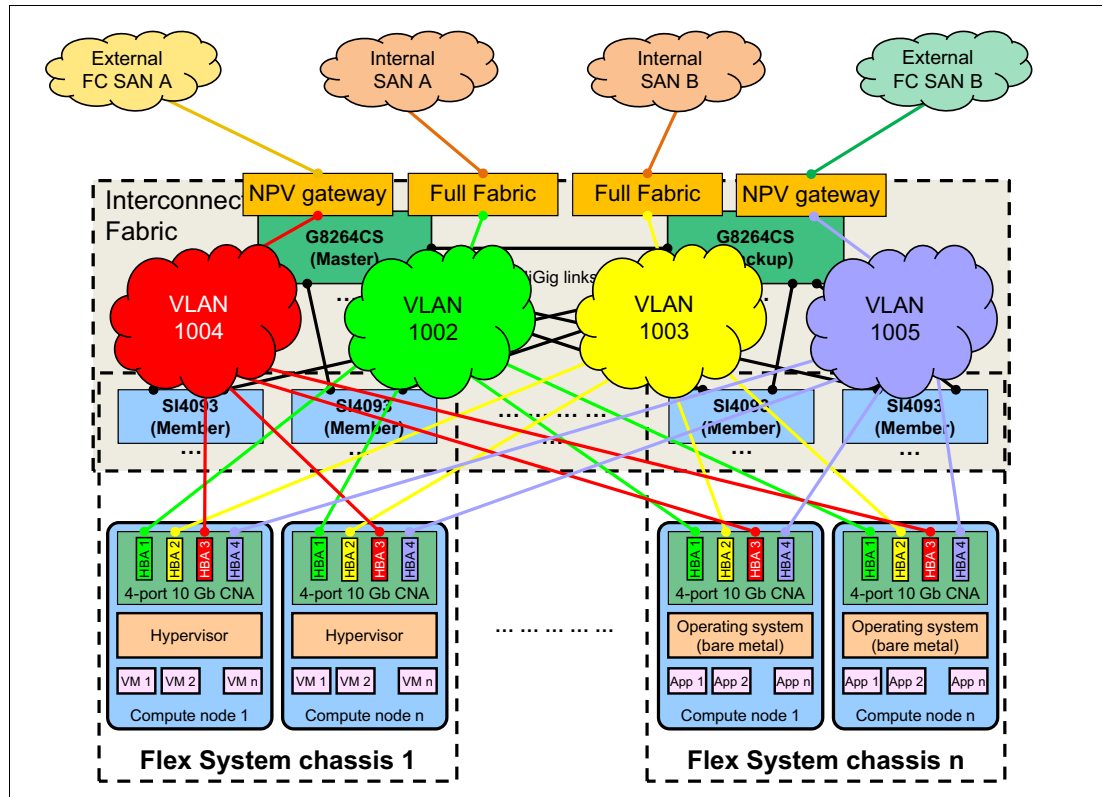


Figure 20 Interconnect Fabric storage connectivity architecture

Each HBA port on the CNA that is installed in the compute node is connected to a dedicated fabric. Path redundancy is provided with the usage of MPIO software that is installed on the compute node (on the bare-metal operating system or on the hypervisor).

Storage VLANs (one VLAN for each HBA port) are used in the fabric to logically isolate SANs from each other. Each G8264CS is connected to a separate isolated external FC SAN (external SAN A or external SAN B) in NPV mode, and up to four N_Port links per NPV VLAN (VLANs 1004 and 1005 in Figure 20) can be connected to an external SAN. If more than four N_Port links per G8264CS are required for the external FC SAN connectivity, consider using more NPV VLANs.

Internal SAN storage

Flex System V7000 Storage Node or Storwize V7000 are used for the internal SAN storage.

Flex System V7000 Storage Node

The V7000 Storage Node is an embedded storage that is installed in the Flex System chassis. Figure 21 shows an example connectivity topology for the internal SAN with the V7000 Storage Node.

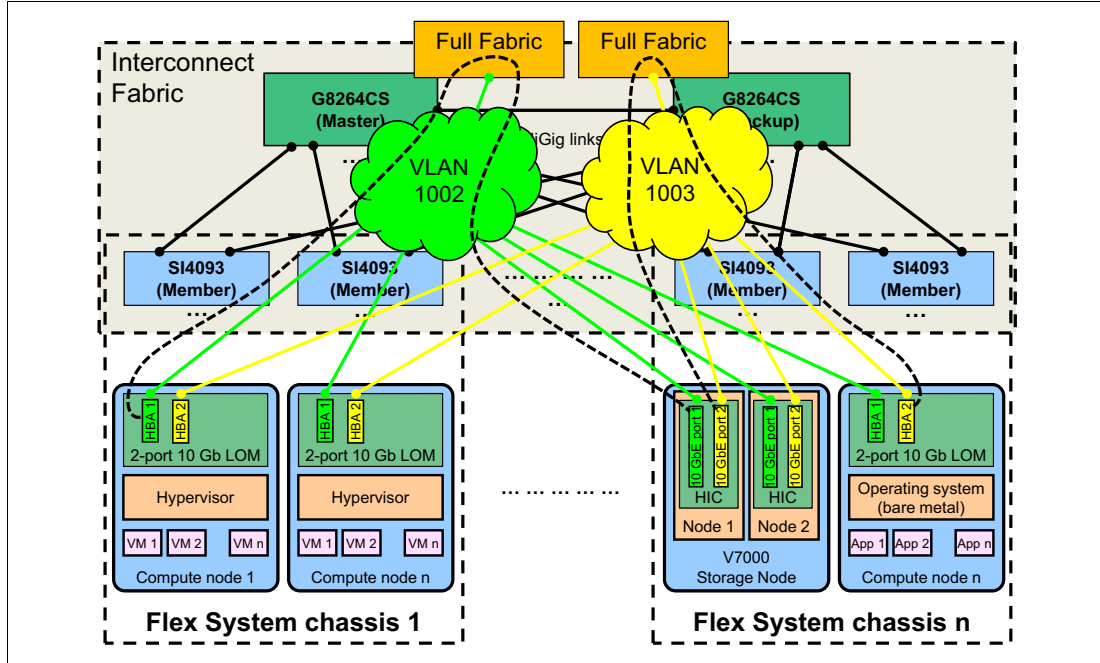


Figure 21 V7000 Storage Node connectivity topology

In this example, a 2-port embedded 10 Gb LOM on the compute node is used for the internal storage connectivity. Virtual Fabric Advanced Upgrade must be enabled on the LOM to support storage attachments. When the Advanced Upgrade is installed, FCoE HBAs (one HBA per physical port) are enabled on the compute node.

VLANs 1002 and 1003 are configured to support dual fabric connectivity, and they are associated with two FCFs that operate in the Full Fabric mode. In the Full Fabric mode, zoning configuration provides SAN isolation and enforces port security among zones. Compute node HBAs, and V7000 Storage Node host ports on the host interface card (HIC), are connected to the SI4093 leaf switches.

Storage traffic flow is as follows:

Compute node's HBA → SI4093 → G8264CS → SI4093 → V7000's HIC port

Alternatively, an external rack-mounted Storwize V7000 can be used instead of a V7000 Storage Node.

Storwize V7000

The Storwize V7000 connectivity topology is similar to the V7000 Storage Node. The only difference is that Storwize V7000 host ports are connected to the G8264CS switches, and the storage traffic flow is as follows:

Compute node's HBA → SI4093 → G8264CS → V7000's HIC port

This flow is illustrated in Figure 22 on page 41.

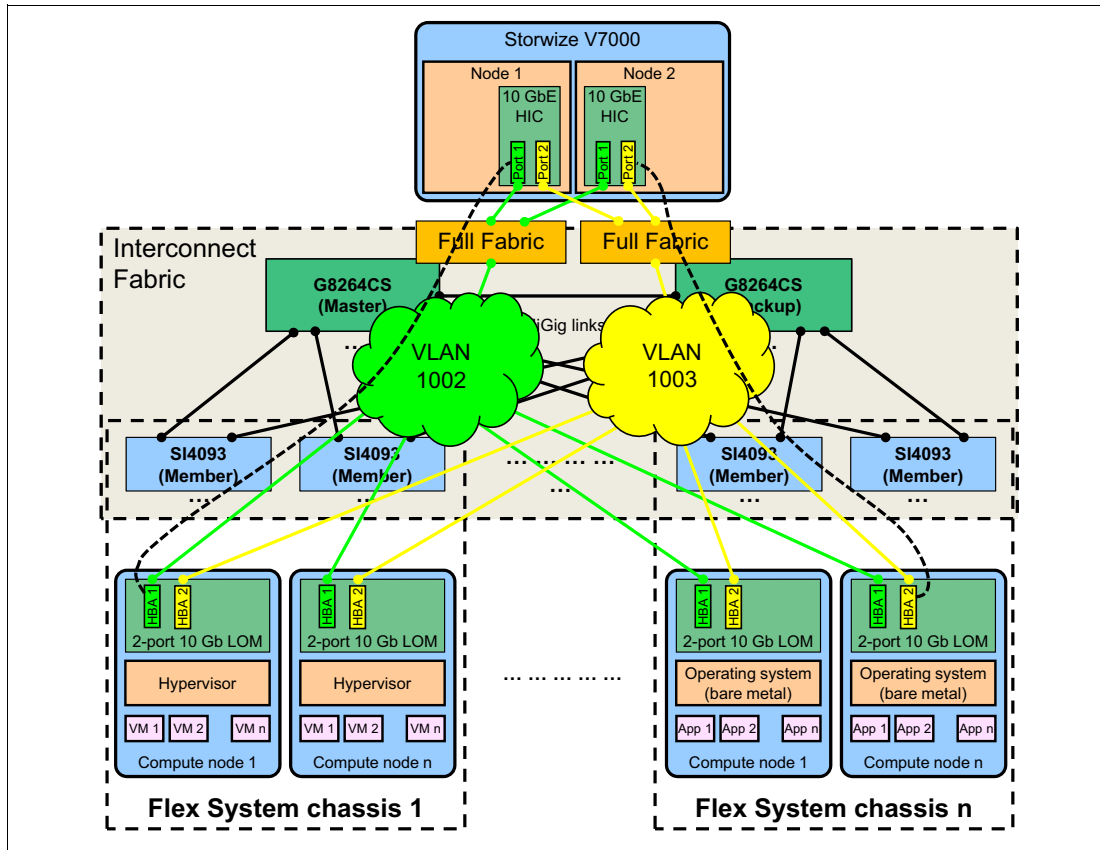


Figure 22 Storwize V7000 connectivity

External FC SAN storage

If external FC SAN storage connectivity is required, the following options are available:

- ▶ Storwize V7000 with SAN Volume Controller enabled
- ▶ Hybrid storage connectivity
- ▶ External dual isolated SAN fabric

Storize V7000 with SAN Volume Controller enabled

Figure 23 shows an example of Storize V7000 storage connectivity with SAN Volume Controller enabled.

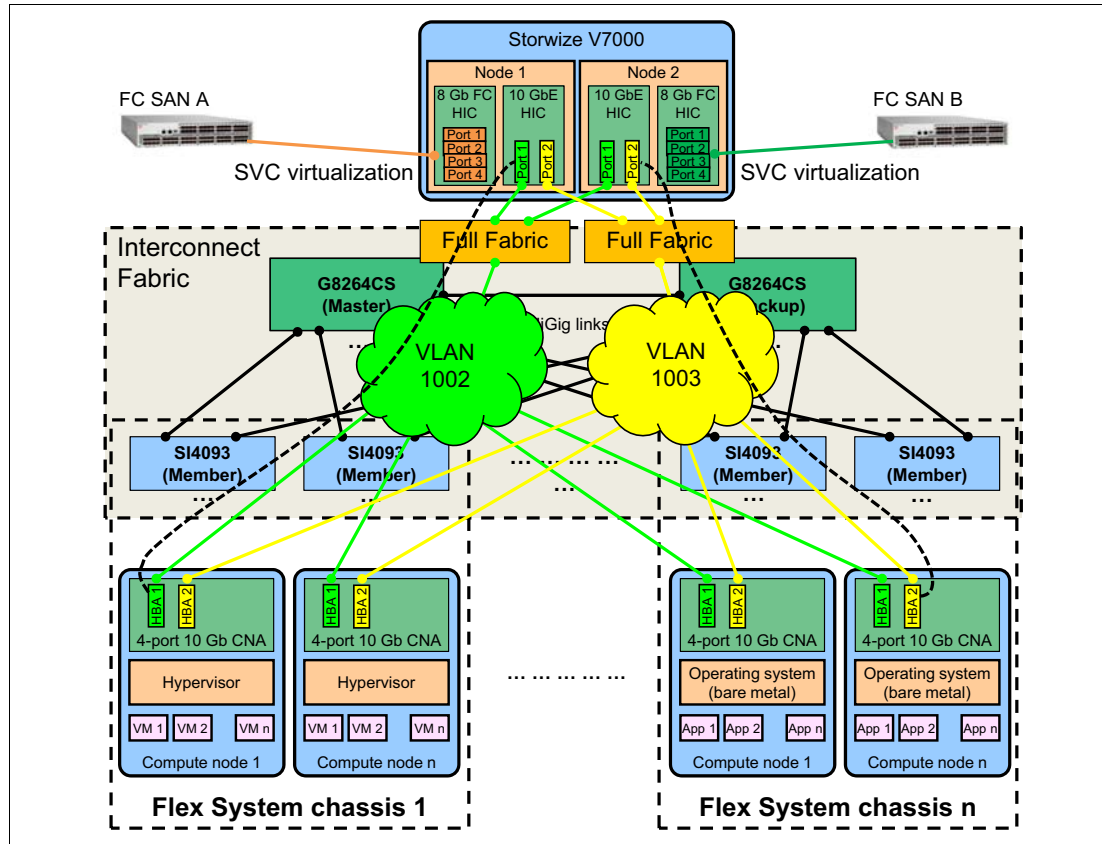


Figure 23 Storize V7000 connectivity with the SAN Volume Controller feature enabled

The Interconnect Fabric configuration is the same as in “Storize V7000” on page 40; the only difference is that Storize V7000 has native FC ports on the FC HIC that are connected to the client’s enterprise network, and these FC ports are used for SAN Volume Controller operations. The external FC storage is hidden from Interconnect Fabric, and the compute nodes interact with the external storage through Storize V7000.

Hybrid storage configuration

A hybrid storage configuration that uses both internal and external storage is shown in Figure 24.

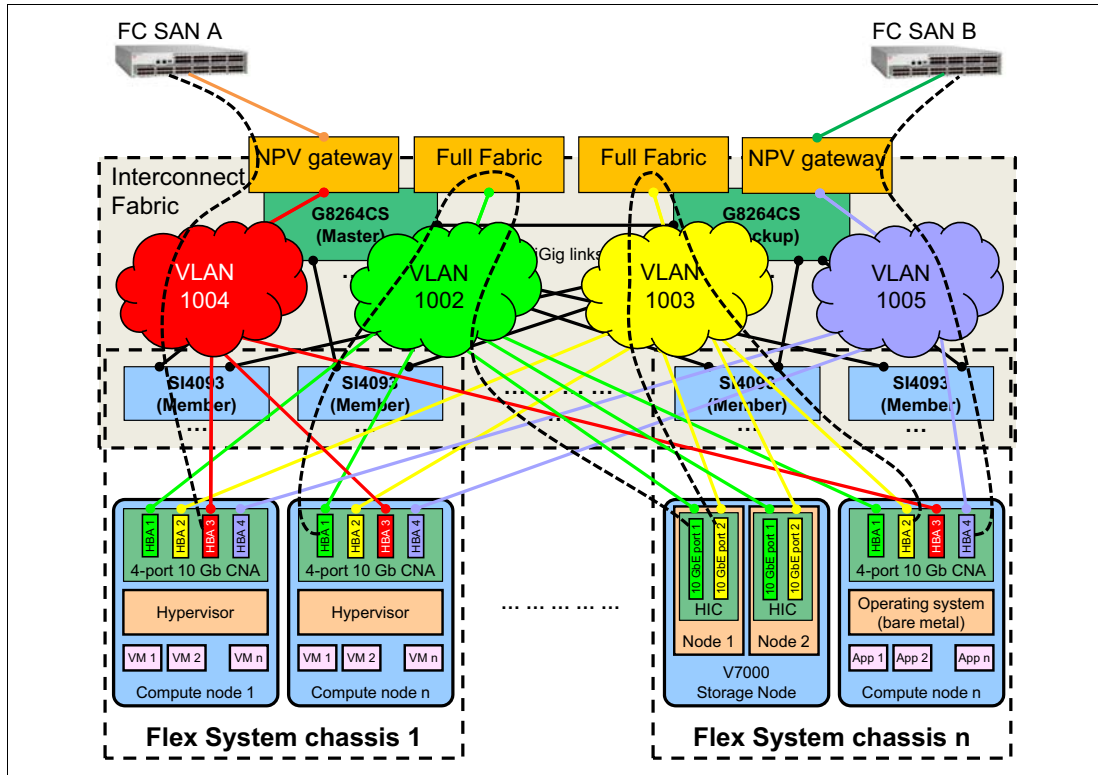


Figure 24 Hybrid storage connectivity - V7000 Storage Node and external FC SAN

In this configuration, 4-port or 8-port CNAs are required. Compute node's HBA ports 1 and 2 are used to connect to the internal storage in the Full Fabric mode, and HBA ports 3 and 4 provide connectivity to the external FC SAN storage in the NPV mode.

Dual isolated SAN fabric (SAN air gap)

The Interconnect Fabric POD integration into an external dual isolated SAN fabric design (sometimes referred to as *SAN air gap*) is shown in Figure 25.

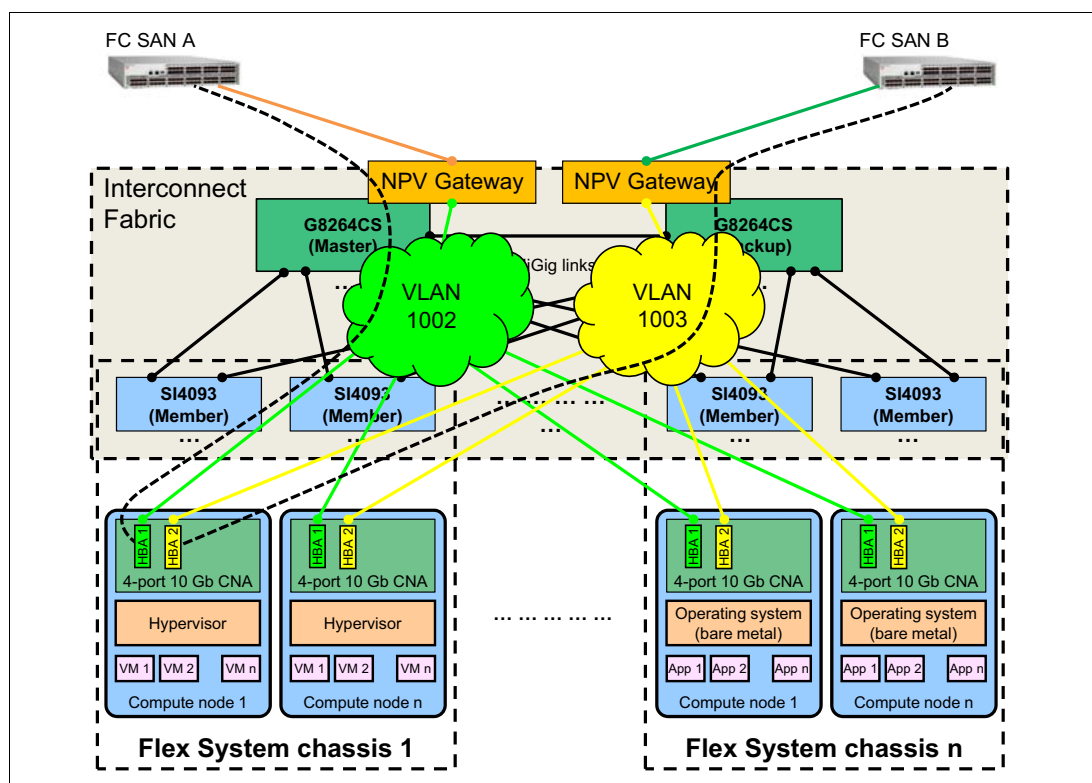


Figure 25 Dual isolated SAN fabric integration

In this configuration, a 2-port LOM with Advanced Upgrade can be used. The compute node's HBA port 1 is used to connect to the external isolated FC SAN A, and HBA port 2 provides connectivity to the external isolated FC SAN B. The POD uses NPV gateway mode for FC SAN connectivity.

Point of delivery management

This section provides a brief overview of the relevant tools and features to manage the Flex System Interconnect Fabric POD. The following topics are covered:

- ▶ "Flex System Interconnect Fabric management"
- ▶ "Compute node management" on page 46

Flex System Interconnect Fabric management

Flex System Interconnect Fabric management is performed through the G8264CS switches, and their 1 GbE management interfaces should be connected to the client's management network.

The various user interfaces that are available for the Interconnect Fabric management, whether through the CLI or the web-based GUI, offer the ability to fully configure and manage all the features that are available.

The best tool for a user often depends on that user's experience with different interfaces and their knowledge of networking features. Most commonly, the CLI is used by those who work with networks as part of their day-to-day jobs. The CLI offers the quickest way to accomplish tasks, such as scripting an entire configuration. The downside to the CLI is that it tends to be more cryptic to those users that do not use them every day. For those users that do not need the power of the CLI, the web-based GUI permits the configuration and management of all switch features.

Lenovo Switch Center

Lenovo offers System Networking Switch Center (SNSC), a tool that provides the following functions:

- ▶ Improves network visibility and drive availability, reliability, and performance
- ▶ Simplifies management of large groups of switches with automatic discovery of switches in the network
- ▶ Automates and integrates management, deployment, and monitoring
- ▶ Provides simple network management protocol (SNMP) based configuration and management
- ▶ Supports network policies for virtualization
- ▶ Provides authentication and authorization
- ▶ Provides fault and performance management
- ▶ Provides integration with IBM Systems Director and VMware Virtual Center and vSphere clients

Figure 26 shows an example Interconnect Fabric devices view in the Switch Center.

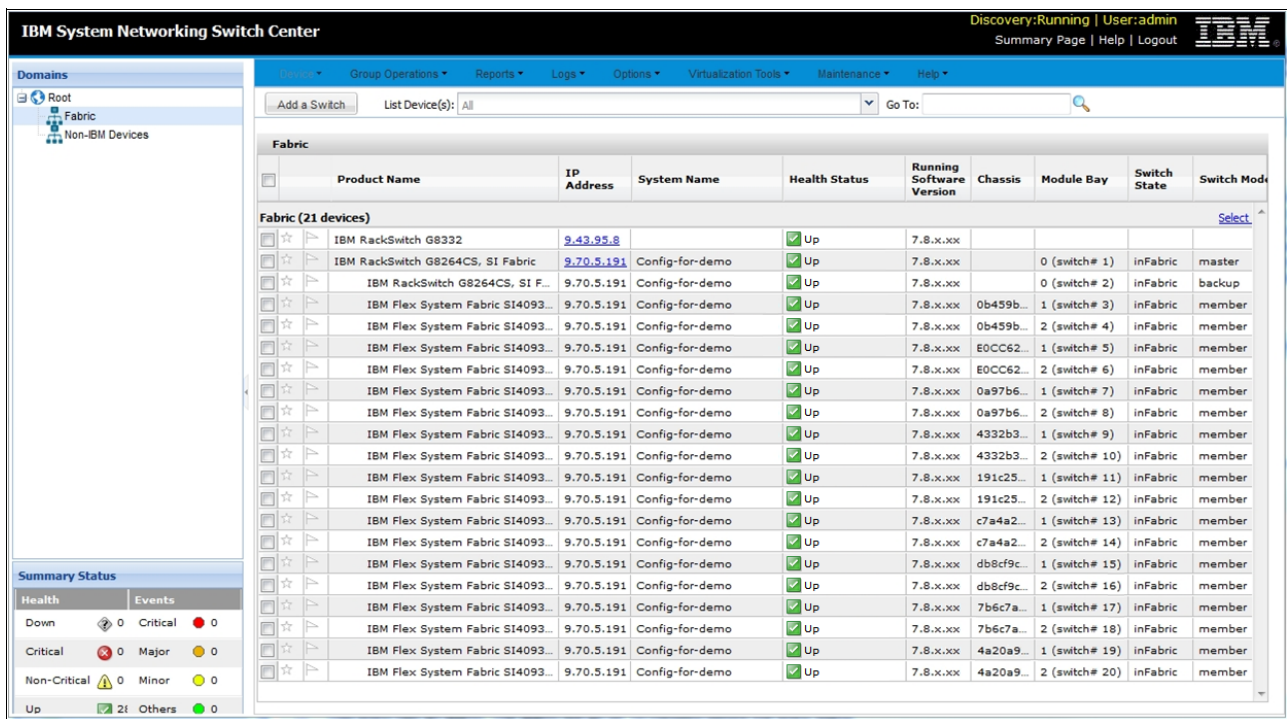


Figure 26 Flex System Interconnect Fabric devices view in the Switch Center

The Switch Center management of Flex System Interconnect Fabric eases the initial setup and configuration, and views and monitors the elements and links of the fabric. The following fabric configuration and monitoring tasks can be performed from the Switch Center:

- ▶ Configuration
 - Server port settings
 - Uplink port settings
 - Fibre Channel (Omni Ports and zoning)
 - Fabric Wide VLAN Domains
- ▶ Monitoring
 - Health status of stack members
 - hDFP port status, usage, and connectivity
 - Server port status, usage, and connectivity
 - Uplink port status, usage, and connectivity

Table 7 summarizes the steps that are needed to configure Interconnect Fabric using CLI and the Switch Center.

Table 7 Interconnect Fabric configuration steps

Step	CLI	Switch Center
1. Configure initially the IP addresses on switches.	Yes	
2. Discover switches with the Management tool.		Yes
3. Upgrade switches with the hDFP cluster image.	Yes	Yes
4. Configure the IP address on master and backup.	Yes	
5. Configure the fabric on master and backup.	Yes	Yes
6. Review members and stacking numbers.	Yes	Yes
7. Provision VLANs for compute nodes and storage nodes.	Yes	Yes
8. Create and associate uplinks with VLANs.	Yes	Yes
9. Associate downlinks and storage connections with VLANs.	Yes	Yes
10. Enable the FC/FCoE storage function and associate storage links with VLANs.	Yes	Yes
11. Enable the UFP function and associate the UFP ports with VLANs.	Yes	Yes
12. Enable the VMready feature in the Distributed VM Group model	Yes	Yes

For more information about the Switch Center, see the following website:

<http://www-03.ibm.com/systems/networking/software/snsc/index.html>

Any third-party management platforms that support SNMP also can be used to configure and manage the fabric.

Compute node management

Compute node management is performed by the Chassis Management Module (CMM) with the option to use a dedicated Flex System Manager (FSM) management node. CMM's 1 GbE management interfaces in all chassis in the POD should be connected to the client's management network.

One of the key features that is related to the compute node management in the Interconnect Fabric POD is the ability to automate the deployment of the NIC settings with certain Emulex NICs through the FSM Configuration Patterns. Here are some examples of items that can be automated through the Configuration Patterns:

- ▶ Change the personality between NIC, FCoE, or iSCSI (assuming FoDs are installed).
- ▶ Enable or disable a mode of Virtual NIC.
- ▶ For the vNIC™ modes of virtual NICs that offer other configuration options, such as LPVID or bandwidth, you can change those options,

Currently, the Embedded 10Gb Virtual Fabric Ethernet Controller (LOM) and Flex System CN4054/CN4054R 10Gb Virtual Fabric Adapter are supported by FSM configuration patterns.

The most important aspect of using configuration patterns is the ability to push out changes to many servers without having to perform the tedious process of manually going into the F1 setup on every server on which virtual NICs must be changed.

For more information about Flex System Manager, see *IBM Flex System Manager*, TIPS0862, which can be found at the following website:

<http://lenovopress.com/tips0862>

Interconnect Fabric initial configuration

This section outlines the required steps to initially configure an Interconnect Fabric environment. It is assumed that the necessary power and network cabling is done before beginning the configuration effort. This section describes the following steps:

1. Obtaining and installing the required firmware images.
2. Specifying the Fabric Ports on each element of the chassis (G8264CS, SI4093).
3. Choosing the Master switch in the fabric.
4. Binding all the members into the fabric.

Obtaining and installing firmware

Flex System Interconnect Fabric uses a specific firmware image which is different from the standard images used for the G8264CS and SI4093. The Interconnect Fabric image can be downloaded from Fix Central (<http://www.ibm.com/support/fixcentral>). In Fix Central, you should look for the software updates for the G8264CS and select the Interconnect Fabric image (**G8264CS_SI_Fabric_Image_7.8.5.0** or later version) for download.

In order to load the firmware onto the G8264CS and SI4093 hardware, the **copy si-fabric-mode** command must be issued before copying the firmware (OS image and boot image) into flash memory. The actual copy process can use the ISCLI **copy tftp boot** and **copy tftp image1|image2** commands. Once the firmware is copied into flash, the switches and interconnect modules must all be rebooted to activate the new firmware images.

Specifying fabric ports

The G8264CS switches will default to having ports 17 through 35 as Fabric ports. If these are the ports that will be used to connect to the other G8264CS and to SI4093 modules installed in Flex System chassis then no action is required.

If there is a need to use other ports as Fabric ports, or if additional ports are needed as uplink ports rather than fabric ports, then the **boot fabric ports <list|range>** command can be used to specify the ports desired. In order to make the change effective, the G8264CS switch must be rebooted, but this can be done after the master switch is chosen as shown below.

On the SI4093, all of the 10 GbE external ports are Fabric ports by default, and they can not be used for any other purpose.

Identifying the master switch

One of the G8264CS switches must be chosen as the master; by default, both will initially be members of the fabric. This is done with the command: **boot fabric mode master**.

In order to make this selection effective, the master G8264CS must be rebooted. The second G8264CS will be instructed to become the backup switch once the master switch completes rebooting; the second G8264CS does not need to be rebooted unless required for another purpose such as changing the selected fabric ports.

Binding the remaining members of the fabric

After the master G8264CS has been selected, and the second G8264CS has joined into the fabric as the backup, the following command will bind all of the SI4093s into the fabric:

fabric bind. SI4093s will be joined to the fabric as long as the following conditions are met:

- ▶ SI4093s must be running the Interconnect Fabric firmware.
- ▶ SI4093s must be physically connected to both G8264CS switches.
- ▶ The connection between an SI4093 and the G8264CSs must use ports that are designated as fabric ports on both devices.

Once the **fabric bind** command is issued, the status of the members of the fabric can be seen by using the **show fabric switch** command; all of the G8264CSs and SI4093s should be shown as IN_FABRIC.

Appendix: Example Interconnect Fabric POD bill of materials

Table 8 on page 48 shows an example bill of materials for a 3-chassis Flex System Interconnect Fabric POD consisting of the following components:

- ▶ Six SI4093 modules
- ▶ Two G8264CS switches with the required SFP+ DAC cables
- ▶ Required rack and PDU infrastructure
- ▶ Optional Lenovo Switch Center management application

Notes:

- ▶ Up to 42 half-wide or 84 high-density compute nodes can be used in this example, but they are not included in the table.
- ▶ Cables or SFP+ modules for the upstream network connectivity are not included.

Table 8 Example Flex System Interconnect Fabric POD bill of materials

Part Number	Description	Quantity
Rack and PDU infrastructure		
93604PX	42U 1200mm Deep Dynamic Rack	1
46M4143	0U 12 C19/12 C13 32A 3 Phase PDU	2
25R5559	1U Quick Install Filler Panel Kit	2
Top of Rack switches		

Part Number	Description	Quantity
7309DRX	Lenovo RackSwitch G8264CS (Rear-to-Front)	2
90Y9430	3m Passive DAC SFP+ Cable	16
90Y9427	1m Passive DAC SFP+ Cable	10
Flex System Enterprise Chassis with SI4093 modules		
8721A1G	Flex System Enterprise Chassis with 2x2500W PSU	3
43W9049	Flex System Enterprise Chassis 2500W Power Module	12
95Y3313	Flex System Fabric SI4093 System Interconnect Module	6
68Y7030	Flex System Chassis Management Module	3
43W9078	Flex System Enterprise Chassis 80mm Fan Module Pair	6
System Networking Switch Center management application (optional)		
00AE226	Lenovo Switch Center, per installation with 1-year software subscription and support for 20 switches	1

Related publications

For more information, see the following Lenovo Press publications:

- ▶ *Lenovo Flex System Interconnect Fabric Solution Guide*
<http://lenovopress.com/tips1307>
- ▶ *Flex System Networking in an Enterprise Data Center, REDP-4834*
<http://lenovopress.com/redp4834>
- ▶ *Flex System Network Implementation, SG24-8089*
<http://lenovopress.com/sg248089>
- ▶ *NIC Virtualization on Lenovo Flex System, SG24-8223*
<http://lenovopress.com/sg248223>

Authors

This paper was produced by a team of specialists from around the world:

Ilya Krutov is a Project Leader at Lenovo Press. He manages and produces pre-sale and post-sale technical publications on various IT topics, including x86 rack and blade servers, server operating systems, virtualization and cloud, networking, storage, and systems management. Ilya has more than 15 years of experience in the IT industry, backed by professional certifications from Cisco Systems, IBM, and Microsoft. During his career, Ilya has held a variety of technical and leadership positions in education, consulting, services, technical sales, marketing, channel business, and programming. He has written more than 200 books, papers, and other technical documents. Ilya has a Specialist's degree with honors in Computer Engineering from the Moscow State Engineering and Physics Institute (Technical University).

Scott Lorditch is a Consulting System Engineer for Lenovo. He performs network architecture assessments and develops designs and proposals for solutions that involve Lenovo Networking products. He also developed several training and lab sessions for technical and sales personnel. Scott joined IBM as part of the acquisition of Blade Network Technologies® and joined Lenovo as part of the System x® acquisition from IBM. Scott spent almost 20 years working on networking in various industries, as a senior network architect, a product manager for managed hosting services, and manager of electronic securities transfer projects. Scott holds a BS degree in Operations Research with a specialization in computer science from Cornell University.

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service.

Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
1009 Think Place - Building One
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk.

Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document was created or updated on September 4, 2016.

Send us your comments via the **Rate & Provide Feedback** form found at <http://lenovopress.com/redp5106>

Trademarks

Lenovo, the Lenovo logo, and For Those Who Do are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. These and other Lenovo trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by Lenovo at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of Lenovo trademarks is available on the Web at <http://www.lenovo.com/legal/copytrade.html>.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Blade Network Technologies®	Omni Ports™	VMready®
Flex System™	RackSwitch™	vNIC™
Lenovo®	Lenovo(logo)®	
NMotion®	System x®	

The following terms are trademarks of other companies:

Intel, Xeon, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Hyper-V, Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.