Lenovo Flex System Interconnect Fabric
Product Guide

Lenovo® Flex System™ Interconnect Fabric offers a solid foundation of compute, network, storage, and software resources in a Flex System point of delivery (POD). The entire POD integrates a seamless network fabric for compute node and storage under single IP management. 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, as shown in the following figure. The POD requires only network provisioning for uplink connections to a data center network, downlink connections to compute nodes, and storage connections to external storage.

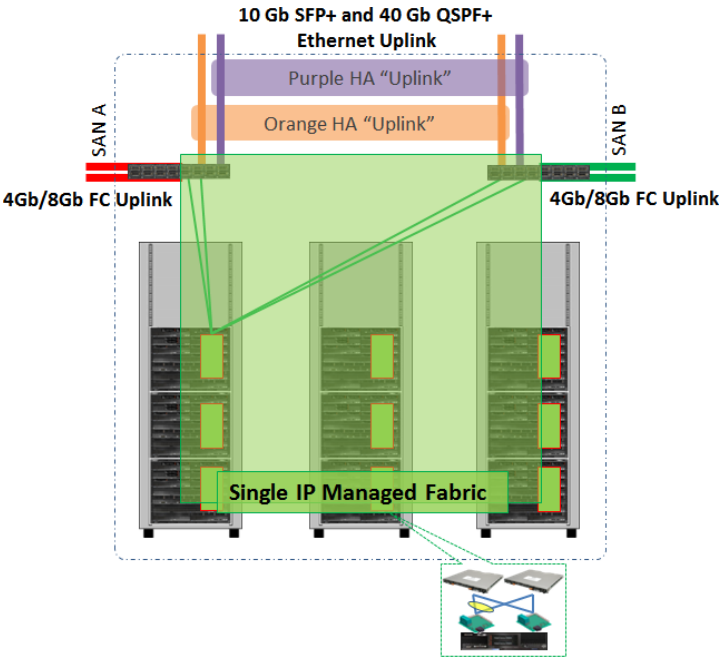

Figure 1. Lenovo Flex System Interconnect Fabric

Did you know?

Flex System Interconnect Fabric reduces communication latency and improves application response time with support for local switching within the chassis. It also reduces networking management complexity without compromising performance by reducing the number of devices that must be managed by 95% (managing one device instead of 20).

Flex System Interconnect Fabric simplifies POD integration into an upstream network by transparently interconnecting hosts to a data center network and representing the POD as a large compute element that isolates the POD's internal connectivity topology and protocols from the rest of the network.

Business value

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 Top-of-Rack (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 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.
  - 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.

Solution overview

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

- Hardware:
  - Aggregation: Lenovo RackSwitch™ G8264CS (10/40 GbE, 4/8 Gb FC uplinks)
  - Access (10 GbE to compute nodes):
    - Lenovo Flex System Fabric SI4093 System Interconnect Module
    - Lenovo Flex System Fabric EN4093R 10Gb Scalable Switch
  - Embedded VFA or CN4054/CN4054R adapters
  - IBM Storwize V7000 for Lenovo (optional)
- Software:
  - 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: Lenovo Switch Center management application (optional)

Flex System Interconnect Fabric supports the following networking software features:

- Single IP managed cluster
- Up to 1,024 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 maximum of two priorities
- UFP virtual port support (four per 10 Gb physical port)
- 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
- IGMPv2 Snooping for multicast optimization (up to 3,000 IGMP groups)
- A total of 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, or Manual Monitor (MMON)
- Hot Links
- Full private VLAN support
- VLAN-based load distribution in hotlinks (for active/active connectivity with non-vPC uplink)
- 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
- Network Time Protocol (NTP)
- Dynamic Host Configuration Protocol (DHCP)/Domain Name System (DNS) client

Solution architecture

The SI4093 and EN4093R embedded modules include 42 10GBASE-KR ports that connect to the compute nodes in the Flex System chassis through the midplane. There are three 10 Gb ports that connect to each of the 14 slots in the chassis.

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

The SI4093, EN4093R, and G8264CS have PHY interfaces for SFP+ transceivers and QSFP+ transceivers that can run as a single 40 Gb port or as a set of four 10 Gb ports by using a breakout cable. The interconnection between the SI4093 or EN4093R embedded modules and the G8264CS aggregation switches is configured to run over standard 10 Gb connections. Similar connections are used between the pair of G8264CS aggregation switches.

A Broadcom proprietary protocol, hDFP, is used over these links, which are referred to in the figures in this solution guide as HiGig links. This protocol carries proprietary control information and the content of the network traffic. It also enables the multiple switching processors 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, 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 40 GbE QSFP+ to 4x 10 GbE SFP+ DAC breakout cables can be used on uplink ports only.

In the Flex System Interconnect Fabric, one of the G8264CS aggregation switches is the master and the other is a backup for purposes of managing the environment. If the master switch fails, the backup G8264CS aggregation switch takes on this task.

The links between switching elements in a Flex System 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. All external ports on the SI4093 or EN4093R embedded modules are configured as Fabric Ports by default, which cannot be changed. If two or more Fabric Ports are connected between the same two devices, all of the Fabric Ports are used as the aggregated link that forms automatically.

The Flex System Interconnect Fabric solution architecture is shown in the following figure.

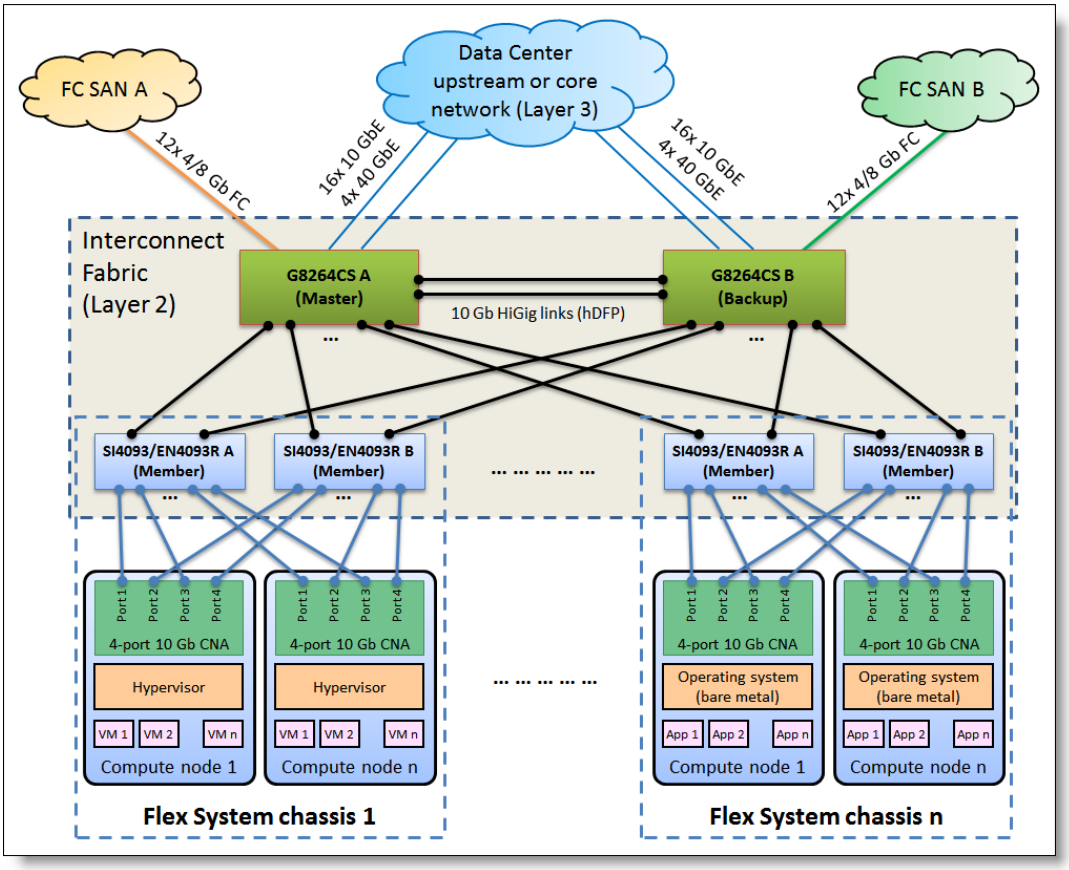

Figure 2. Flex System Interconnect Fabric solution architecture

A typical Flex System Interconnect Fabric configuration uses four 10 Gb ports as Fabric Ports from each SI4093 or EN4093R embedded module, two ports to each of the aggregation switches (up to 1:3.5 oversubscription with 2-port Embedded VFAs or up to 1:7 oversubscription with 4-port CNAs). The maximum of nine chassis (with a total of 18 SI4093 or EN4093R embedded modules) use 36 ports on each of the G8264CS aggregation switches. More ports can be used from the SI4093 or EN4093R embedded modules to the aggregation switches if there are ports available, up to a total of eight.

Fabric Ports do not need more 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 Flex System Interconnect Fabric environment. G8264CS ports that are not configured as Fabric Ports can be used as uplink ports. (Omni Ports cannot be configured as Fabric Ports.)

Uplink ports are used for connecting the Flex System Interconnect Fabric POD to the upstream data network (standard ports and Omni Ports) and to the storage networks (Omni Ports only).

Flex System Interconnect Fabric is formed on its own by establishing HiGig links by using the configured Fabric Ports. All HiGig links are active, and they carry network traffic. The actual 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.

Usage scenarios

The Flex System Interconnect Fabric can be VLAN-aware or VLAN-agnostic, depending on specific client requirements.

In VLAN-aware mode (as shown in the following figure), client VLAN isolation is extended to the fabric by filtering and forwarding VLAN tagged frames that are based on the client VLAN tag. Client VLANs from the upstream network are configured within the Flex System Interconnect Fabric and on virtual switches (vSwitches) in hypervisors.

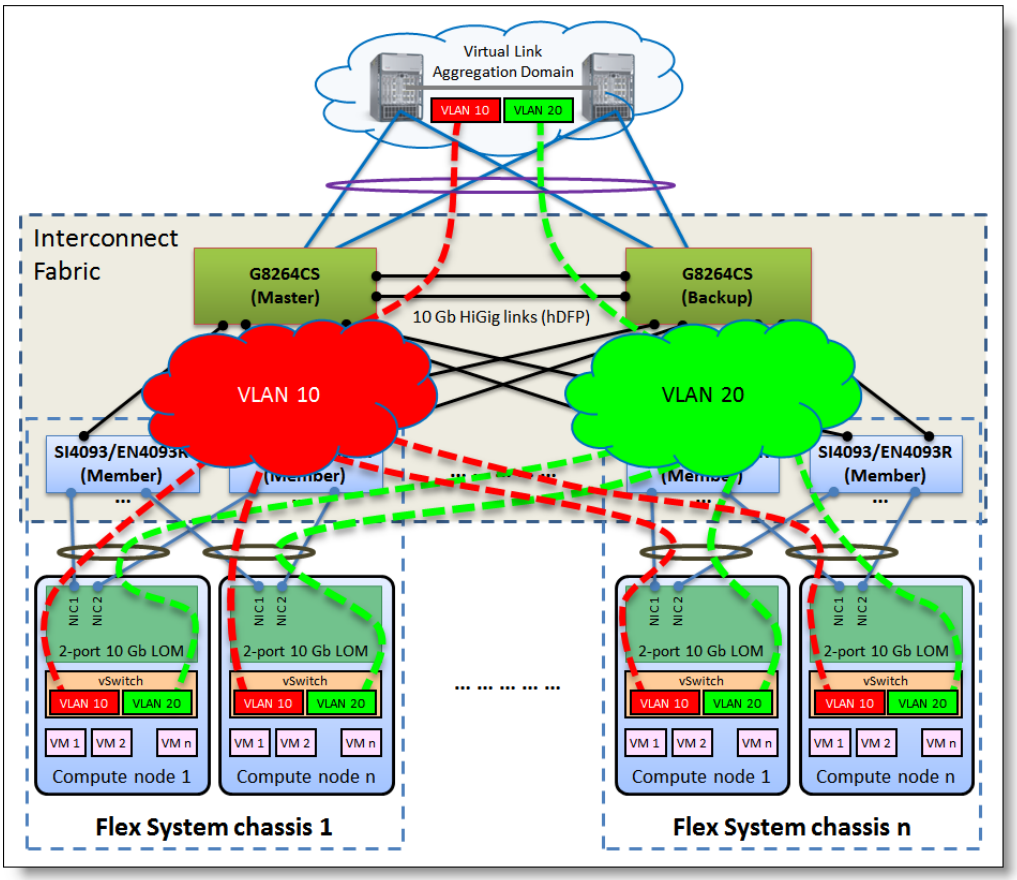

Figure 3. VLAN-aware Flex System Interconnect Fabric

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

The VLAN-agnostic mode of the Flex System 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, the tagpvid-ingress mode is the simplest way to configure the fabric. However, if you want to use FCoE storage, you cannot use the tagpvid-ingress feature and must switch to UFP tunnel mode.

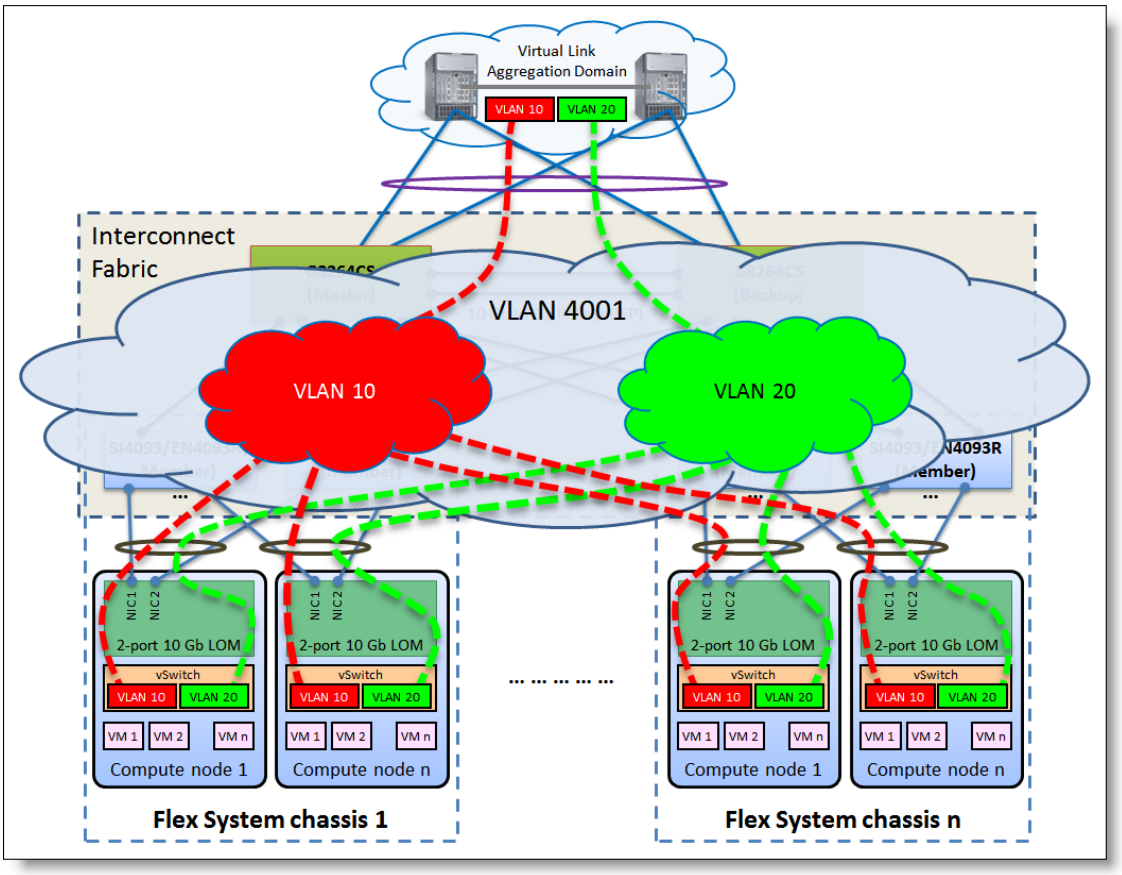

Figure 4. VLAN-agnostic Flex System Interconnect Fabric

All VMs that are connected to the same client VLAN can communicate with each other within the POD and with other endpoints that are 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 Flex System Interconnect Fabric through a teamed NIC interface that is configured on the compute node. The compute node's CNA NIC ports, which are physical (pNIC) or virtual (UFP vPort) ports, are configured in a load-balancing pair (static or LACP aggregation) by using the hypervisor's teaming and bonding feature. Respective compute node-facing ports on the SI4093 or EN4093R embedded modules are also configured for static or dynamic aggregation, and VLAN tagging (802.1Q) is enabled on the aggregated link.

If the upstream network supports distributed (or *virtual*) link aggregation, this connectivity type can be used to connect Flex System Interconnect Fabric to the core. Interconnect Fabric sees the upstream network as one logical switch, and the upstream network sees Interconnect Fabric as one logical switch. A single aggregated port channel (static or dynamic) is configured between these two logical switches by using all connected uplinks. All of these links in the aggregation carry traffic from all client VLANs.

If virtual link aggregation is not supported on the upstream network switches (that is, the upstream network operates in a standard STP domain), Hot Link interfaces are used. Flex System Interconnect Fabric sees the upstream network as two separate switches, and the upstream network sees Flex System Interconnect Fabric as one logical switch.

This logical switch is connected to the upstream switches by using the following aggregated port channels (static or dynamic):

- One port channel is configured between the first upstream switch and the Flex System Interconnect Fabric logical switch.
- Another port channel is configured between the second upstream switch and the Flex System Interconnect Fabric logical switch.

One port channel is designated as the master hot link, and the second port channel is configured as the backup hot link. The master port channel carries traffic from all client VLANs, and the backup port channel is in the blocking state. If there is a master port channel failure, the backup port channel becomes active and all traffic flows through it. The downside of this approach is that only half of the available uplink bandwidth is used. Flex System Interconnect Fabric supports VLAN load distribution over Hot Links to maximize bandwidth usage. Both hot links are masters and backups for different VLANs at the same time.

Storage integration

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

Flex System Interconnect Fabric introduces a new storage fabric capability that is called Hybrid Fabric, in which there are two types of SANs (internal and external) on separate SAN fabrics. The internal SAN is used for the POD-wide Storwize V7000 connectivity in Full Fabric mode, and the external SAN is used for the data center-wide storage connectivity in NPV Gateway mode. Internal and external SANs are dual-fabric SANs and the hybrid storage configuration features four fabrics.

Hybrid mode requires dual initiators per compute node connection to each SI4093 or EN4093R embedded module so that each initiator can discover one FC fabric. Each initiator can communicate only with one FCF VLAN and FC fabric. The 4-port CNAs offer the required number of ports to support dual switch path storage access. Dual-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 (internal or external, but not both). Each HBA port on the CNA that is installed in the compute node is connected to the dedicated fabric. Path redundancy is provided with the usage of MPIO software that is installed on the compute node (in the bare-metal operating system or in the hypervisor).

The hybrid storage configuration, which uses Storwize V7000 and the client's external storage, is shown in the following figure.

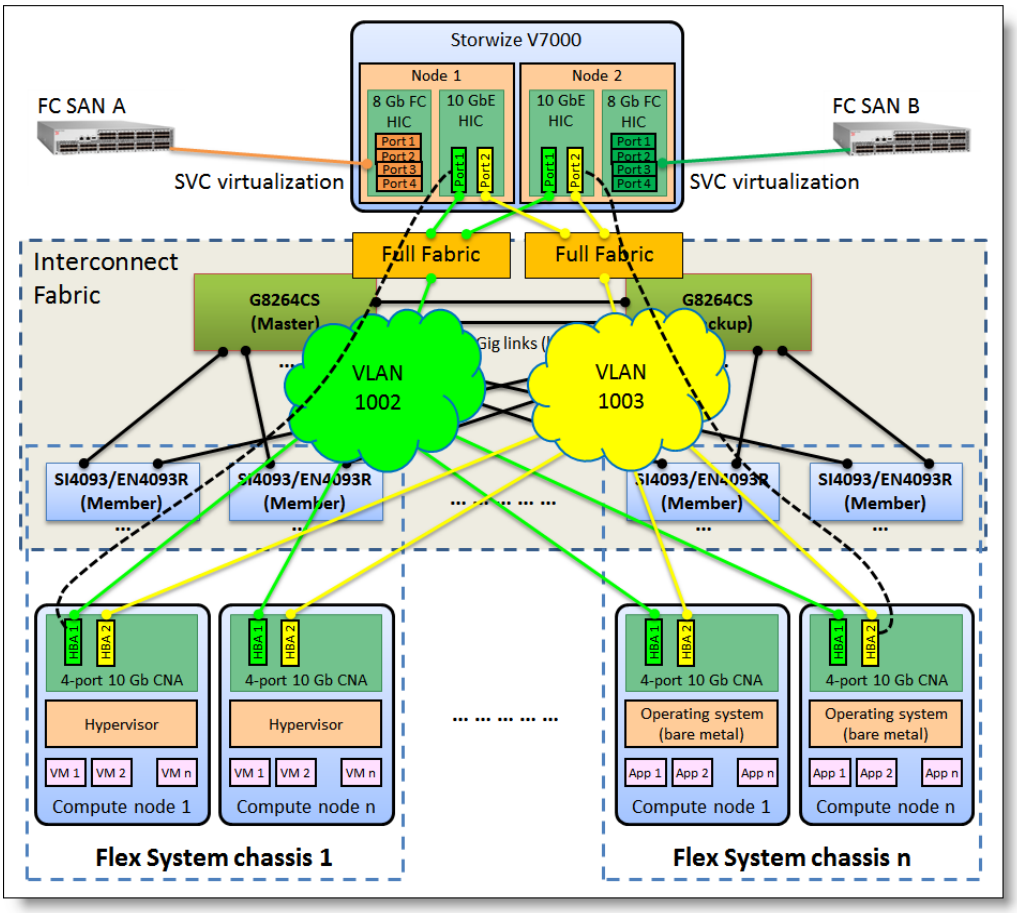

Figure 5. Flex System Interconnect Fabric storage integration

The compute node's HBA ports 1 and 2 are used to connect to the internal storage in the Full Fabric mode. HBA ports 3 and 4 provide connectivity to the external FC SAN storage in the NPV mode.

Supported platforms

Flex System Interconnect Fabric is supported by the following network devices:

- RackSwitch G8264CS
- Flex System Fabric SI4093 System Interconnect Module
- Flex System Fabric EN4093R 10Gb Scalable Switch

Note: Two G8264CS switches and 2 - 18 SI4093 or EN4093R modules are supported in a POD. G8264CS switches and SI4093 or EN4093R embedded modules in a POD run a special Flex System Interconnect Fabric software image of Networking OS.

The following adapters are supported:

- Flex System Embedded Virtual Fabric LOM
- Flex System CN4054/CN4054R 10Gb Virtual Fabric Adapters

Ordering information

The ordering information for the Flex System Interconnect Fabric solution components is listed in the following table.

Table 1. Ordering part numbers and feature codes

Description	Part number	Feature code	Quantity
Top-of-Rack (ToR) switches - G8264CS			
Lenovo RackSwitch G8264CS (Rear-to-Front)	7159DRX	ASY0	2
Optional components for ToR switches			
Lenovo RackSwitch Adjustable 19-inch 4 Post Rail Kit	00D6185	A3KP	1 per G8264CS
Air Inlet Duct for 483 mm RackSwitch	00D6060	A3KQ	1 per G8264CS
Flex System embedded modules			
Lenovo Flex System Fabric SI4093 System Interconnect Module	00FM518	ASUV	Up to 18#
Lenovo Flex System Fabric EN4093R 10Gb Scalable Switch	00FM514	ASUU	Up to 18#
Optional Features on Demand upgrades for Flex System modules			
Flex System Fabric SI4093 System Interconnect Module (Upgrade 1)	95Y3318	A45U	1 per SI4093*
Flex System Fabric EN4093 10Gb Scalable Switch (Upgrade 1)	49Y4798	A1EL	1 per EN4093R*
SFP+ DAC cables - Fabric connectivity			
Lenovo 1m Passive SFP+ DAC Cable	90Y9427	A1PH	Varies
Lenovo 1.5m Passive SFP+ DAC Cable	00AY764	A51N	Varies
Lenovo 2m Passive SFP+ DAC Cable	00AY765	A51P	Varies
Lenovo 3m Passive SFP+ DAC Cable	90Y9430	A1PJ	Varies
Lenovo 5m Passive SFP+ DAC Cable	90Y9433	A1PK	Varies
Lenovo 7m Passive SFP+ DAC Cable	00D6151	A3RH	Varies

Two embedded modules per Flex System chassis.
* Required for 4-port adapters.

Note: Cables or SFP+ modules for the upstream network connectivity are not listed.

Related publications and links

For more information, see the following resources:

- *Flex System Interconnect Fabric: Technical Overview and Planning Considerations* , REDP-5106: http://lenovopress.com/redp5106
- *NIC Virtualization on Flex System*, SG24-8223: http://lenovopress.com/sg248223
- *Lenovo Flex System Fabric SI4093 System Interconnect Module Product Guide* : http://lenovopress.com/tips1294
- *Lenovo Flex System Fabric EN4093R 10Gb Scalable Switch Product Guide* : http://lenovopress.com/tips1292
- *Lenovo RackSwitch G8264CS Product Guide* : http://lenovopress.com/tips1273

Related product families

Product families related to this document are the following:

- 10 Gb Embedded Connectivity
- Blade Networking Modules
- Network Management

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.
8001 Development Drive
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, TIPS1307, was created or updated on March 12, 2016.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/TIPS1307
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/TIPS1307.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
Flex System
Omni Ports
RackSwitch
VMready®

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