

Virtual PortChannels: Building Networks without Spanning Tree Protocol

What You Will Learn

This document provides an in-depth look at Cisco's virtual PortChannel (vPC) technology, as developed for the Cisco® NX-OS Software on the Cisco Nexus™ 7000 Series Switches platform. It provides a short overview of the goals of the vPC technology, followed by a detailed discussion of the technology and its features. This document also briefly discusses network deployment models and failure response and recovery.

Layer 2 Scaleout in the Data Center

Virtualization technologies such as VMware ESX Server and clustering solutions such as Microsoft Cluster Service currently require Layer 2 Ethernet connectivity to function properly. With the increased use of these types of technologies in data centers and now even across data center locations, organizations are shifting from a highly scalable Layer 3 network model to a highly scalable Layer 2 model. This shift is causing changes in the technologies used to manage large Layer 2 network environments, including migration away from Spanning Tree Protocol as a primary loop management technology toward new technologies such as vPC and IETF TRILL (transparent interconnection of lots of links).

In early Layer 2 Ethernet network environments, it was necessary to develop protocol and control mechanisms that limited the disastrous effects of a topology loop in the network. Spanning Tree Protocol was the primary solution to this problem, providing a loop detection and loop management capability for Layer 2 Ethernet networks. This protocol has gone through a number of enhancements and extensions, and while it scales to very large network environments, it still has one suboptimal principle: to break loops in a network, only one active path is allowed from one device to another, regardless of how many actual connections might exist in the network. Although Spanning Tree Protocol is a robust and scalable solution to redundancy in a Layer 2 network, the single logical link does create two problems. One problem is that half (or more) of the available system bandwidth is off-limits to data traffic, and the other problem is that a failure of the active link tends to cause multiple seconds of system wide data loss while the network reevaluates the new "best" solution for network forwarding in the Layer 2 network. Although enhancements to Spanning Tree Protocol reduce the overhead of the rediscovery process and allow a Layer 2 network to reconverge far faster, the delay can still be too great for some networks. In addition, no efficient dynamic mechanism exists for using all the available bandwidth in a robust network with Spanning Tree Protocol loop management.

An early enhancement to Layer 2 Ethernet networks was PortChannel technology (now standardized as IEEE 802.3ad PortChannel technology), in which multiple links between two participating devices can use all the links between the devices to forward traffic by using a load-balancing algorithm that equally balances traffic across the available Inter-Switch Links (ISLs) while also managing the loop problem by bundling the links as one logical link. This logical construct keeps the remote device from forwarding broadcast and unicast frames back to the logical link, thereby breaking the loop that actually exists in the network. PortChannel technology has one other primary benefit: it can potentially deal with a link loss in the bundle in less than a second, with very little loss of traffic and no effect on the active Spanning Tree Protocol topology.

Introducing vPC

The biggest limitation in classic PortChannel communication is that the PortChannel operates only between two devices. In large networks, the support of multiple devices together is often a design requirement to provide some form of hardware failure alternate path. This alternate path is often connected in a way that would cause a loop, limiting the benefits gained with PortChannel technology to a single path. To address this limitation, the Cisco NX-OS Software platform provides a technology called virtual PortChannel, or vPC. Although a pair of switches acting as a vPC peer endpoint looks like a single logical entity to PortChannel-attached devices, the two devices that act as the logical PortChannel endpoint are still two separate devices. This environment combines the benefits of hardware redundancy with the benefits of PortChannel loop management. The other main benefit of migration to an all-PortChannel-based loop management mechanism is that link recovery is potentially much faster. Spanning Tree Protocol can recover from a link failure in approximately 6 seconds, while an all-PortChannel-based solution has the potential for failure recovery in less than a second.

Although vPC is not the only technology that provides this solution, other solutions tend to have a number of deficiencies that limit their practical implementation, especially when deployed at the core or distribution layer of a dense high-speed network. All multichassis PortChannel technologies still need a direct link between the two devices acting as the PortChannel endpoints. This link is often much smaller than the aggregate bandwidth of the vPCs connected to the endpoint pair. Cisco technologies such as vPC are specifically designed to limit the use of this ISL specifically to switch management traffic and the occasional traffic flow from a failed network port. Technologies from other vendors are not designed with this goal in mind, and in fact are dramatically limited in scale specifically because they require the use of the ISL for control traffic and approximately half the data throughput of the peer devices. For a small environment, this approach might be adequate, but it will not suffice for an environment in which many terabits of data traffic may be present.

vPC in Detail

To properly understand vPC, consider Figure 1, which shows a sample network. The figure illustrates vPC functions and features.

Figure 1. Classic Multi-tier Ethernet Network

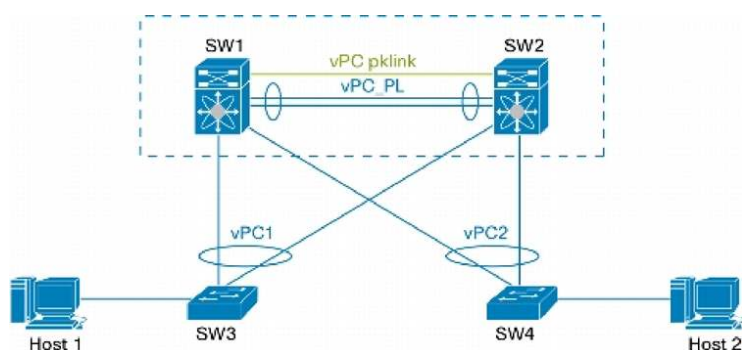


Figure 1 shows a number of components:

- **vPC peer switches:** Switches S1 and S2 are connected as peers through the peer link, and they form the single logical endpoint for a vPC. These devices need to run Cisco NX-OS to run the vPC protocol.
- **vPC peer link:** This is a multiport 10 Gigabit Ethernet PortChannel between the two vPC peer switches. This link is a standard IEEE 802.3ad PortChannel with a modified Spanning Tree Protocol weight and the capability to tag packets as having originated on the local peer using the peer link.

- **vPC peer keepalive link:** The peer keepalive link is a logical link that often runs over an out-of-band management network. It provides a Layer 3 communications path that is used as a secondary test to determine whether the remote peer is operating properly. No data or synchronization traffic is sent over the vPC peer keepalive link, just a frame that indicates that the originating switch is operating and running vPC.
- **vPC member port:** A vPC member port is a physical port on one of the vPC peer switches that is a member in a vPC. To have a running vPC instance, at least one PortChannel is needed with a member port on each peer switch.
- **Cisco Fabric Services:** The Cisco Fabric Services protocol is a reliable messaging protocol designed to support rapid stateful configuration message passing and synchronization. vPC services use Cisco Fabric Services to transfer a copy of the system configuration for a comparison process and to synchronize MAC and Internet Group Management Protocol (IGMP) state information between the two vPC peer switches.

In operation, the vPC feature first must be enabled and then the vPC peer link must be established. After this is done, Cisco Fabric Services messages are interchanged between the two vPC peer switches, providing a copy of the local switch configuration to the remote switch to determine whether any configuration inconsistencies exist that need to be addressed prior to starting vPC. Such inconsistencies include differences in the configuration of Spanning Tree Protocol, (Hot Standby Router Protocol (HSRP), Protocol Independent Multicast (PIM), and vPC. After it has been established that these protocols are in sync, and any consistency mismatches have been addressed, the vPC system will be in a ready state, available to add vPC member ports to the system.

Adding a member-port is a simple operation consisting of informing the actual switch local PortChannels that they are vPC members and informing the vPC process that a new PortChannel is available. Note that a PortChannel must be configured even if only a single port on a particular switch is expected to be a member of a specific vPC. If a vPC includes more than one local member port, port forwarding decisions depend on the PortChannel load-balancing mechanism, which is any combination of source or destination Ethernet MAC address, IP address, IP port, and VLAN ID. If only a single local member port is available, then no additional load balancing will be performed to help ensure that outbound traffic does not incur the additional forwarding cost of crossing the peer link.

Layer 2 multicast forwarding is similar in that learned sources or client join requests received on one switch are forwarded to the other so that a consistent (S, G) multicast state can be determined. This state includes outgoing interface (OIF) information and is tuned so as to limit the use of the vPC peer link for multicast traffic.

vPC Layer 3 Interaction

Although vPC is primarily a Layer 2 technology, the Cisco Nexus 7000 Series Switches are also full-featured Layer 3 network devices. Therefore a number of enhancements have been made to the vPC solution to integrate with the Layer 3 features of the Cisco Nexus 7000 Series. Two critical areas were enhanced to provide the most scalable vPC environment; specifically HSRP and PIM interaction were modified to improve scalability and system resiliency.

In the case of HSRP, the improvement was made to the forwarding engine specifically to allow local Layer 3 forwarding at both the active HSRP peer and the standby HSRP peer. This enhancement provides, in effect, an active-active HSRP configuration with no changes to current HSRP configuration recommendations or best practices and no changes to HSRP. The HSRP control protocol still acts like an active-standby pair, so that only the active device responds to Address Resolution Protocol (ARP) requests, but a packet destined for the shared HSRP MAC address is accepted as local on either the active or standby HSRP device.

Figure 2 shows an example of this process in which requests from Host 2 are directed to the switch that acts as the HSRP standby, but the packet is still forwarded to the Layer 3 cloud. Host 1's packets are sent (based on PortChannel load balancing) to the switch that is the HSRP active device and are also forwarded to the Layer 3 cloud.

solution improves multicast failure recovery performance and makes use of the full Layer 3 capability of the Cisco Nexus 7000 Series platform.

vPC Failure Recovery

One of the advantages of the vPC approach to loop management is that in the unlikely event of a PortChannel member port link failure, recovery relies on the PortChannel recovery mechanism rather than Spanning Tree Protocol relearning of the entire network. While Spanning Tree Protocol can potentially be tuned to respond in as little as 6 seconds, PortChannel recovery often takes less than one second. This speed factor alone is a primary reason why vPC provides a more efficient scaling mechanism than Layer 2 topologies managed by Spanning Tree Protocol.

Member port failure is perhaps the most likely scenario, in which a member port from an access switch fails. When the vPC peer determines that a member port has failed (and that no other local member ports for that vPC are available), the peer with the failed vPC member port notifies the remote peer that it no longer has an active member port for a configured vPC. The remote peer will then enable forwarding on that vPC for packets that traverse the peer link. This mechanism helps ensure reachability and at the same time provides loop management.

In the highly unlikely case that both ports and line cards in the peer link fail (two ports on two different line cards are the recommended minimum for the peer link, reducing the possibility of complete failure), or if the Cisco Fabric Services messaging infrastructure fails to communicate across the peer link, the vPC management system will look to the peer keepalive interface to determine whether the failure is a link-level failure or whether in fact the remote peer has failed entirely. If the remote peer is still alive (peer keepalive messages are still being received), the vPC secondary switch will disable its vPC member ports and any Layer 3 interfaces attached to a vPC-associated VLAN. If the peer keepalive messages are not being received, then the peer continues to forward traffic as it is then assuming that it is the last device available in the network. In either case, on recovery of the peer link or reestablishment of Cisco Fabric Services message forwarding, the systems will resynchronize any MAC addresses learned while communications was disrupted, and the system will continue forwarding normally.

vPC Network Benefits

vPC provides a number of important benefits to a Layer 2 network and a set of enhancements to the Layer 3 interconnect specifically resulting from the benefits derived from the Layer 2 capabilities. In a Layer 2 network, the following benefits are achieved:

- Enhanced system availability through redundant systems
- Loop management without use of Spanning Tree Protocol
- Full system bandwidth availability at all times
- Rapid link-failure recovery
- PortChannel connectivity to any IEEE 802.3ad-capable edge device

In addition, the following important Layer 3 features are enabled:

- Active-active Layer 3 forwarding through HSRP configuration
- Full Layer 3 bandwidth access through active-active HSRP
- Rapid Layer 3 multicast convergence through the active-active PIM designated router

While these capabilities enhance any data center environment, when they are deployed on the Cisco Nexus 7000 Series of 10 Gigabit Ethernet data center switches as is a requirement today, an additional set of capabilities provide the highest level of data center throughput and resiliency currently available. These features include:

- **In Service Software Upgrade (ISSU):** Nonstop forwarding (NSF) during full-system upgrades enables short or even nonexistent maintenance windows for software updating of mission-critical platforms.
- **Hardware forwarding:** All current Layer 2 and 3 forwarding functions are performed in hardware, providing a consistent forwarding environment regardless of features or system scale.
- **Virtual device contexts (VDCs):** The Cisco Nexus 7000 Series can currently be divided into four logically separate switch environments, providing efficient administrative domain deployments on Cisco's highest-density Gigabit Ethernet and 10 Gigabit Ethernet platforms.

Conclusion

This document provided an overview of the benefits of PortChannel technology and an in-depth discussion of the ways that vPC extends the PortChannel model to remove Spanning Tree Protocol as a loop management technology in large-scale Layer 2 Ethernet networks. It also described HSRP and PIM Layer 3 enhancements to the basic Layer 2 PortChannel model and the response of vPC to failures, and the use of vPC as a core technology in an advanced data center network. Through use of the vPC technology, Cisco will continue its commitment to support for networks of all types, and specifically its commitment to support for the ongoing shift toward large-scale Layer 2 network deployments both within a single data center and across multiple data centers.

For More Information

For more information, visit <http://www.cisco.com/go/nexus7000> or contact your local account representative.



Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

CCDE, CCSI, CCENT, Cisco Eos, Cisco HealthPresence, the Cisco logo, Cisco Lumin, Cisco Nexus, Cisco Nurse Connect, Cisco Stackpower, Cisco StadiumVision, Cisco TelePresence, Cisco WebEx, DCE, and Welcome to the Human Network are trademarks; Changing the Way We Work, Live, Play, and Learn and Cisco Store are service marks; and Access Registrar, Aironet, AsyncOS, Bringing the Meeting To You, Catalyst, CCDA, CCDP, CCIE, CCIP, CCNA, CCNP, CCSP, CCVP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unity, Collaboration Without Limitation, EtherFast, EtherSwitch, Event Center, Fast Step, Follow Me Browsing, FormShare, GigaDrive, HomeLink, Internet Quotient, IOS, iPhone, iQuick Study, IronPort, the IronPort logo, LightStream, Linksys, MediaTone, MeetingPlace, MeetingPlace Chime Sound, MGX, Networkers, Networking Academy, Network Registrar, PCNow, PIX, PowerPanels, ProConnect, ScriptShare, SenderBase, SMARtNet, Spectrum Expert, StackWise, The Fastest Way to Increase Your Internet Quotient, TransPath, WebEx, and the WebEx logo are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or website are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0903R)