

虚拟端口捆绑组： 展望无生成树协议的网络构建



您将了解到的内容

本文深入介绍专为在 Cisco Nexus™ 7000 系列交换机平台上的 Cisco NX-OS 软件开发的思科虚拟端口捆绑组(vPC)技术。首先简要概述 vPC 技术的目标，然后详细讨论该项技术及其特性。本文还大致介绍网络部署模式，以及故障响应和恢复措施。

数据中心的第二层扩展

虚拟化技术，如 VMware ESX 服务器和 Microsoft Cluster Service 等集群解决方案，目前需要第二层以太网连接才能正确运行。随着这些类型的技术在数据中心日益普及，现在甚至跨多个数据中心地点实施，机构正从高度可扩展的第三层网络模式向高度可扩展的第二层模式转变。这一转变也引发用于管理大型第二层网络环境的技术的改变，包括从将生成树协议作为主要的环路管理技术，升级到 vPC 和数据中心以太网等新技术。

在早期的第二层以太网环境中，需要开发协议和控制机制，来限制网络中的拓扑环路带来的灾难性影响。生成树协议是针对此问题的主要解决方案，为第二层以太网提供环路检测和管理功能。该协议虽然经过多次增强和扩展，但当扩展到极大型网络环境时，它仍有一个不太理想的做法：即打破网络中的环路，无论网络中实际存在多少条连接，在一个设备到另一设备之间仅允许有一条在用路径。尽管生成树协议是一个强大、可扩展、针对第二层网络冗余性的解决方案，但单一逻辑链路确实带来两个问题。第一个问题是，一半（或更多）的可用系统带宽禁止数据流量进入，另一问题是在用链路故障很可能造成数秒系统数据丢失，此时网络为第二层网络中的网络转发而重新评估新的“最佳”解决方案。生成树协议的改进降低重发现流程耗时，使第二层网络能在大约 6 秒重收敛，但这一延迟对于部分网络来说仍然过大。而且，采用生成树协议管理环路时，不存在高效、灵活地使用强大网络中所有可用带宽的机制。

第二层以太网的一项早期改进即是端口捆绑组技术（现在已通过标准化，成为 IEEE 802.3ad 端口捆绑组技术），其中两个网络设备间的多条链路能使用设备间的所有链路来转发流量，利用负载均衡算法，在可用的交换机间链路（ISL）上公平地均衡流量，并将多条链路捆绑为单一逻辑链路，来解决环路问题。此逻辑结构使远程设备不必再将广播和单播帧转发回逻辑链路，因此打破实际存在于网络中的环路。端口信道技术还有一个重要优势：它能在不到 1 秒内处理捆绑中的某条链路中断问题，几乎不会丢失流量，对当前在用的生成树协议拓扑无影响。

vPC 简介

传统端口信道通信的最大限制在于端口信道只能在两个设备之间运行。在大型网络中，设计中常常需要同时支持多个设备，来提供某些形式的硬件故障备用路径。这一备用路径的连接方式常常会导致环路，从而限制对单一路径实施端口信道技术的优势。为突破此限制，Cisco NX-OS 软件平台提供一种名为虚拟端口捆绑组，或即 vPC 的技术。尽管对于与端口信道相连的设备来说，一对作为 vPC 对等终端的交换机就像是单一逻辑实体，但这两个作为逻辑端口信道终端的设备仍是两个独立设备。该环境结合硬件冗余性和端口信道环路管理的优势。升级到一个完全基于端口信道的环路管理机制，所能获得的另一主要优势是，链路恢复速度大大加快。生成树协议从链路故障中恢复的时间大约为 6 秒，而完全基于虚拟端口捆绑组的解决方案则有可能在不到 1 秒能完成故障恢复。

尽管 vPC 不是实施此解决方案的唯一技术，但其他解决方案都有很多缺陷，限制它们的实际使用，特别是当部署在密集高速网络的核心或分布层时更是如此。所有的多机箱端口信道技术都仍需要在作为端口信道终端的两个设备间具有一条直接链路。该链路所占的带宽通常要远远小于与此终端对相连的 vPC 的总带宽。vPC 等思科技术进行专门设计，仅限将此 ISL 用于交换机管理流量和偶尔来自于故障网络端口的流量。其他厂商的技术设计时并未以此为目标，所以实际上在扩展规模方面有很大限制，因为它们需要使用 ISL 处理控制流量以及对等设备间接近一半的数据吞吐率。对于小型环境来说，这种方法可能足够，但对于可能有数 Tb 数据流量的环境来说，就无法满足需要。

vPC 详细信息

为正确了解 vPC，请参见图 1，其中给出一个网络示例。图中显示 vPC 功能和特性。

图 1 典型的多层以太网

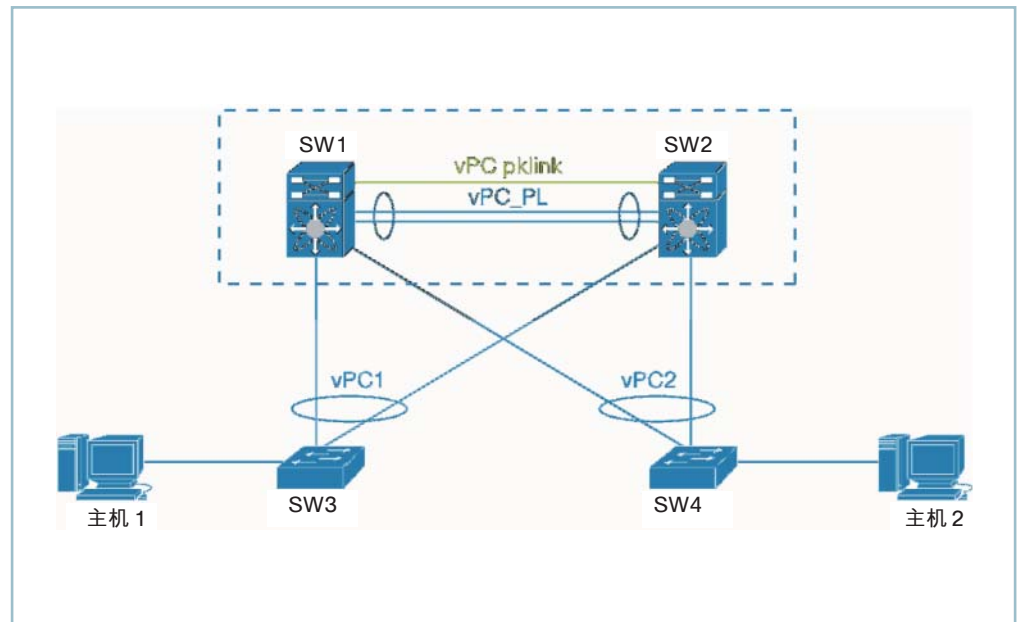


图 1 显示大量组件：

- **vPC 对等交换机:** 交换机 S1 和 S2 作为对等交换机，通过对等链路连接，它们构成一个 vPC 的单一逻辑终端。这些设备需要运行 Cisco NX-OS，才能运行 vPC 协议。
- **vPC 对等链路:** 这是两个 vPC 对等交换机间的多端口万兆以太网端口信道链路。该链路是一个标准 IEEE 802.3ad 端口信道，修改生成树协议权值，能够使用对等链路，将数据包标记为从本地对等设备发出。
- **vPC 对等设备保持激活链路:** 对等设备保持激活链路是一条通常运行在带外管理网络上的逻辑链路。它提供一条第三层通信协议，能用作辅助测试，来确定远程对等设备是否正确运行。不通过 vPC 对等设备保持激活链路发送任何数据或同步流量，只会发送一个表示原始交换机正在运行 vPC 的帧。

- **vPC 成员端口:** 一个vPC成员端口是一个物理端口, 位于属于vPC成员的某一vPC对等交换机之上。为运行vPC实例, 则至少与每个对等交换机上的一个成员端口间有一条端口信道连接。
- **思科交换阵列服务:** 思科交换阵列服务协议是一个可靠的消息传递协议, 专为支持迅速状态化配置消息传递和同步化而设计。vPC服务使用思科交换阵列服务来传输一份系统配置拷贝, 用于比较流程, 和实现两个vPC对等交换机间的MAC和IGMP状态信息同步化。

在运行中, 首先必须启用vPC特性, 然后必须建立vPC对等链路。在此之后, 两个vPC对等交换机间交换思科交换阵列服务消息, 向远程交换机提供一份本地交换机配置拷贝, 以确定在启动vPC前是否有需要处理的配置不一致问题。这种不一致包括生成树协议、(HSRP, PIM)和vPC配置不一致等。在确定这些协议同步, 所有不一致问题都得到解决后, vPC系统进入就绪状态, 能够向系统添加vPC成员端口。

添加成员端口操作很简单, 包括通知实际交换机本地端口信道, 告知它们有vPC成员, 以及通知vPC流程, 告知有一个新端口信道可用。注意即使某一特定交换机上只有一个端口希望成为特定vPC的成员, 也必须配置一个端口信道。如果某个vPC包括多个本地成员端口, 端口转发决策根据端口信道负载均衡机制作出, 该机制综合考虑源或目的地以太网MAC地址、IP地址、IP端口和VLAN ID。如果只有一个本地成员端口可用, 则无需执行更多负载均衡, 来确保输出流量不会因穿越对等链路而花费更多时间。

第二层组播转发也很类似, 一个交换机上所学习到的信息源或接收到的客户端加入请求被转发到另一个交换机, 以便能确定统一(S,G)的组播状态。该状态包括输出接口(OIF)信息, 并对此进行调整以便将vPC对等链路仅限于组播流量。

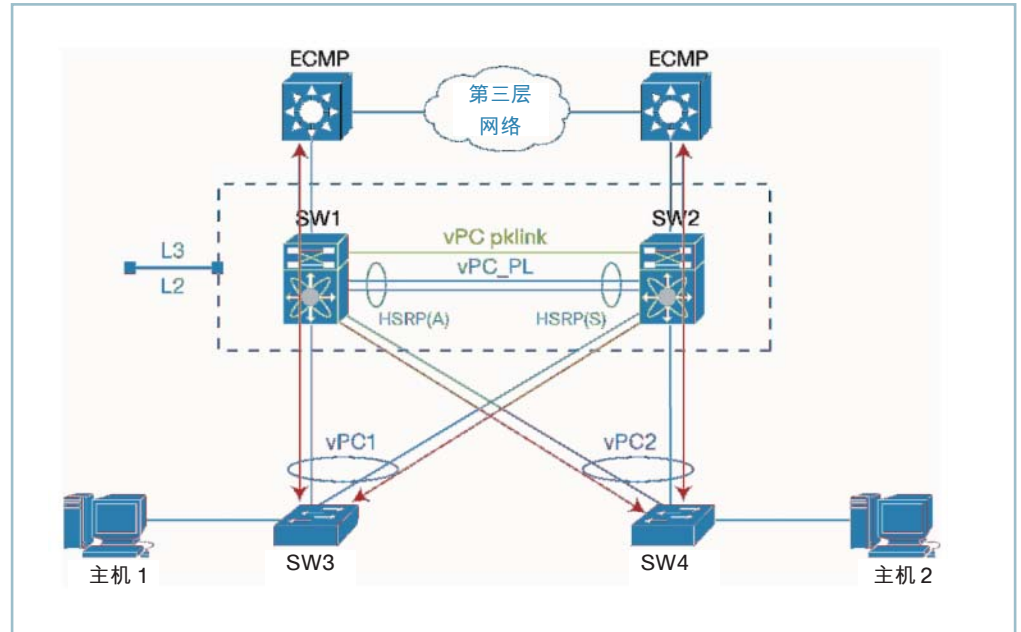
vPC 第三层交互

尽管vPC主要是一种第二层技术, Cisco Nexus 7000系列交换机也是功能全面的第三层网络设备。因此, vPC解决方案已进行很多改进, 以便能与Cisco Nexus 7000系列的第三层特性相集成。两个关键领域通过增强, 能提供可扩展性最高的vPC环境; 所修改的HSRP和PIM交互提高可扩展性和系统永续性。

对HSRP来说, 主要针对转发引擎进行改进, 能在主用和备用HSRP对等设备上进行本地第三层转发。这一改进实际上提供主用-主用HSRP配置, 且无需更改当前的HSRP配置建议或最佳实践, 也不必修改HSRP。HSRP控制协议仍作为主用-备用对运行, 因此只有主用设备响应ARP请求, 但一个目的地为共享HSRP MAC地址的数据包将被接受为主用或备用HSRP设备的本地数据包。

图2显示一个上述流程的示例, 其中主机2的请求被导向作为HSRP备用设备的交换机, 但该数据包仍转发到第三层云。主机1的数据包(根据端口信道负载均衡)被发送到作为HSRP主用设备的交换机, 同时也转发到第三层云。

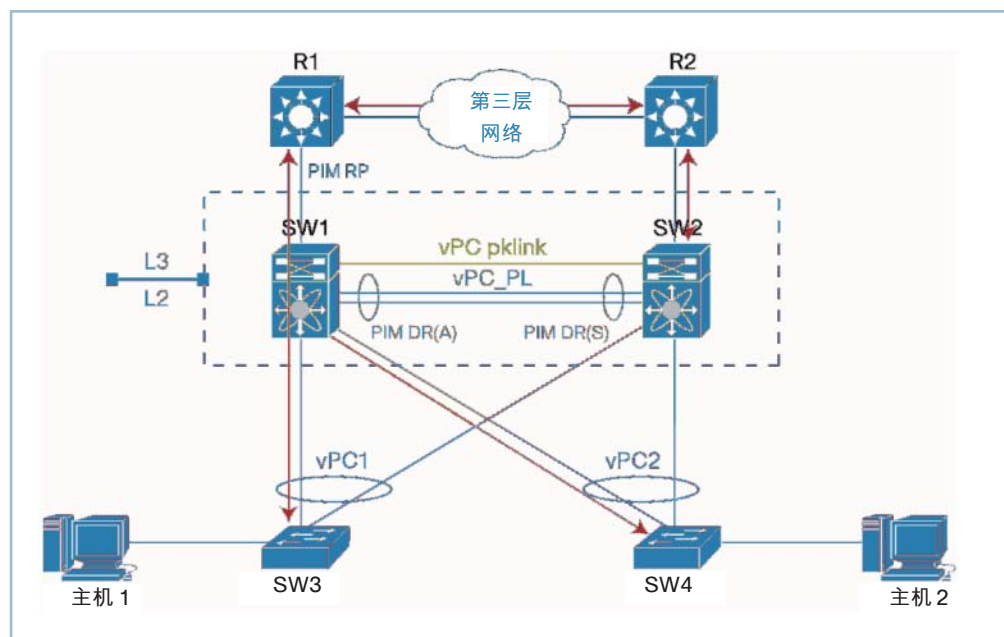
图 2 与等成本多路径(ECMP)和 HSRP 相连的第二层网络



另一个重要的第三层交互与PIM路由相关。vPC支持对于PIM特定源组播(SSM)和PIM任意源组播(ASM)路由的改进,在发生对等设备级故障时,能进一步提高永续性,并确保vPC相连环境中输入或输出的组播转发最为高效。

特别是,因为需要两个第三层实例(每个对等交换机上一个),必须配置两个PIM路由器:每个设备上一个。一般来说,这两个设备都会断言自己是本地第二层组播环境中的指定路由器。在此情况下,最后胜出的路由器随后加入第三层组播环境,开始向任意所加入的客户端转发组播流量,或如果第三层网络中有客户端请求,如图3所示,则开始转发本地网络上的任意组播流量源。

图 3 典型第二层和第三层网络中的 PIM 树



而在上述流程中失败的路由器将不会加入第三层网络，而是等待下一个Assert Interval（“断言间隔”）过去，再次尝试从测试中胜出。而使用 vPC 时，PIM 流程得到改进，使这两个路由器都能加入第三层网络，但只有主 vPC 对等设备上的路由器能胜出，另一设备作为备用，如果主用设备在默认 Reassertion Interval（“重新断言间隔”）过去后失败，则该备用设备向组播网络转发数据包。该模式能够在发生转发故障时更快恢复，而且因为这两个设备已进行第二层状态同步，所以指定的备用路由器也无需再重建本第二层转发状态。该解决方案能提高组播故障恢复性能，充分利用 Cisco Nexus 7000 系列平台的全部三层功能。

vPC 故障恢复

vPC 进行环路管理的方法的优势之一是，万一出现端口信道成员端口链路故障的情况，恢复将依靠端口信道恢复机制而非整个网络的生成树协议以 Relearning 的形式进行。生成树协议最快须 6 秒恢复响应，但端口信道恢复通常只需不到一秒。仅速度这一项就能很好地说明 vPC 是比生成树协议所管理的第二层拓扑更为高效的扩展机制。

相对来说，成员端口故障是最有可能发生的，即某个接入交换机的成员端口发生故障。当 vPC 对等设备确定某个成员端口发生故障（且此 vPC 没有其他可用的本地成员端口）时，故障 vPC 成员端口所在的对等设备会通知远程对等设备，告知它不再拥有所配置 vPC 的可用成员端口。远程对等设备将在此 vPC 上对穿越此对等链路的数据包执行转发。该机制有助于确保数据包可达性，同时提供环路管理。

另外一种出现可能性极低的情况是，对等链路中的端口和线卡均发生故障（建议对等链路至少须在两个不同线卡上有两个端口，以降低完全故障的可能性），或是思科交换阵列服务消息传送基础设施无法通过对等链路通信，vPC 管理系统将查看对等设备保持激活接口，来确定是链路级故障还是远程对等设备全面故障。如果远程对等设备仍在运行（一直能够接收到对等设备保持激活消息），vPC 备用交换机将

禁用其 vPC 成员端口以及所有与 vPC 相关 VLAN 连接的第三层接口。如果未收到对等设备保持激活消息，则该对等设备继续转发流量，并随后认为它是网络中最后一个可用设备。无论出现上述哪种情况，为恢复对等链路或重建思科交换阵列服务消息转发，系统将在通信中断时重新同步所学习到的 MAC 地址，然后系统将恢复正常转发。

vPC 网络优势

vPC 为第二层网络提供大量重要优势，并借助第二层功能提供的优势，对第三层互联进行一系列改进。

在第二层网络中，能够实现以下优势：

- 通过冗余系统提高系统可用性
- 无需使用生成树协议，即能进行环路管理
- 始终提供完全系统带宽可用性
- 迅速恢复链路故障
- 为任意支持 IEEE 802.3ad 的边缘设备提供端口信道连接

此外，还支持以下重要的第三层特性：

- 通过 HSRP 配置进行主用 - 主用第三层转发
- 通过主用 - 主用 HSRP 进行完全第三层带宽访问
- 通过主用 - 主用 PIM 指定路由器进行第三层迅速组播融合

这些功能能够优化任意数据中心环境，当它们根据目前要求部署在 Cisco Nexus 7000 系列万兆以太网数据中心交换机之上时，另一套特性提供当前最高水平的数据中心吞吐率和永续性。这些特性包括：

- 运行中软件升级 (ISSU)：通过完全系统升级期间的不间断转发(NSF)，使关键任务平台的软件更新的维护窗口缩短甚或完全消除。
- 硬件转发：目前所有第二层和第三层转发功能都在硬件中执行，无论使用哪些特性或系统规模如何，都提供统一的转发环境。
- 虚拟设备环境 (VDC)：Cisco Nexus 7000 系列目前能划分为四个逻辑上独立的交换机环境，在思科最高密度的千兆以太网和万兆以太网平台上提供高效管理域部署。

总结

本文对端口信道技术的优势进行简要介绍，并深入讨论 vPC 扩展端口信道，以删除生成树协议，作为大型第二层以太网中的环路管理技术的方式。它也介绍基本第二层端口信道模式的 HSRP 和 PIM 第三层改进，vPC 对故障的响应，以及 vPC 作为核心技术在高级数据中心网络中的应用。通过使用 vPC 技术，思科将继续实现其支持所有网络类型的承诺，并特别致力于支持向单一数据中心或跨多个数据中心的大型第二层网络部署的长期迁移。

了解更多信息

如需了解更多信息，请访问 <http://www.cisco.com/go/nexus7000>，或联系您当地的客户代表。



北京

北京市朝阳区建国门外大街2号北京银泰中心银泰写字楼C座7-12层
邮编: 100022
电话: (8610)85155000
传真: (8610)85155960

上海

上海市淮海中路222号力宝广场32-33层
邮编: 200021
电话: (8621)23024000
传真: (8621)23024450

广州

广州市天河区林和西路161号中泰国际广场A塔34层
邮编: 510620
电话: (8620)85193000
传真: (8620)85193008

成都

成都滨江东路9号B座香格里拉中心办公楼12层
邮编: 610021
电话: (8628)86961000
传真: (8628)86528999

如需了解思科公司的更多信息, 请浏览<http://www.cisco.com/cn>

思科系统(中国)网络技术有限公司版权所有。

2009©思科系统公司版权所有。该版权和/或其它所有权利均由思科系统公司拥有并保留。Cisco, Cisco IOS, Cisco IOS标识, Cisco Systems, Cisco Systems标识, Cisco Systems Cisco Press标识等均为思科系统公司或其在美国和其他国家的附属机构的注册商标。这份文档中所提到的所有其它品牌, 名称或商标均为其各自所有人的财产。合作伙伴一词的使用并不意味着在思科和任何其他公司之间存在合伙经营的关系

2009年2月印刷