Nexus 9000：配置并检验VXLAN Xconnect

目录

简介

本文档介绍如何在Nexus 9000交换机上配置和验证VXLAN Xconnect的快速参考。

先决条件

要求

思科建议您了解VXLAN EVPN。

使用的组件

本文档中的信息基于以下软件和硬件版本：

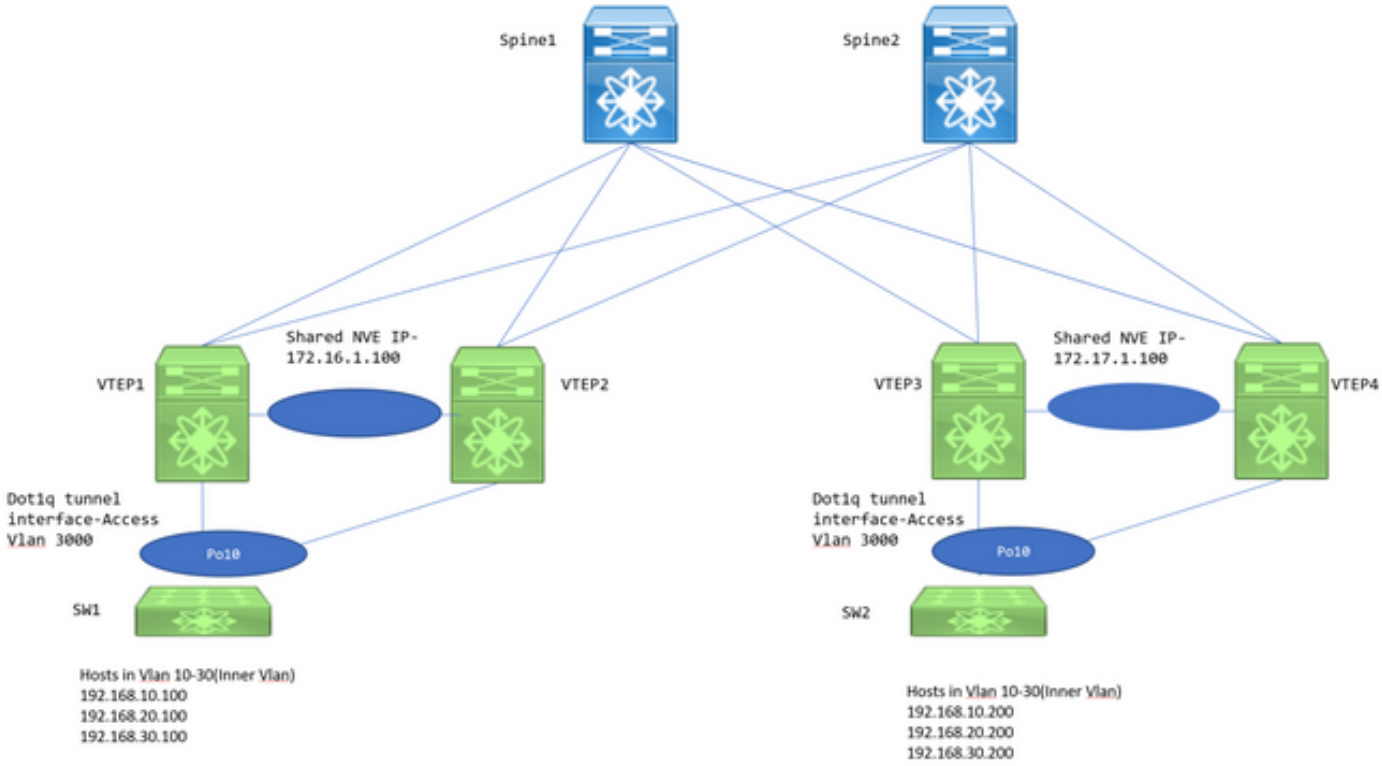
- N9K-C93180YC-EX
- NXOS 9.2(1)

本文档中的信息都是基于特定实验室环境中的设备编写的。本文档中使用的所有设备最初均采用原始（默认）配置。如果您使用的是真实网络，请确保您已经了解所有命令的潜在影响。

概述

VXLAN Xconnect是点对点隧道的机制，用于将数据包从一个枝叶传输到另一个枝叶。内部Dot1q标记保留，VXLAN封装在外部VNID中，该VNID被指定为Xconnect VNID。第2层控制帧(如链路层发现协议(LLDP)、思科发现协议(CDP)、生成树协议(STP))是VXLAN封装的，并发送到隧道的其他端。

拓扑



VTEP1、VTEP2、VTEP3和VTEP4是两个vPC VTEP对，这样，来自下游交换机的内部dot1q标记将保留，并且当VXLAN封装时，使用外部VLAN ID的VXLAN VNID将其发送到远程VTEP。 所有VTEP均为N9K-C93180YC-EX。

下游交换机是Nexus 3k，在各自的VLAN中配置了交换机虚拟接口(SVI)来模拟主机。

配置

1.此Xconnect拓扑中使用的外部VLAN是3000。这将是VNID和Xconnect配置。

```
VTEP1# sh run vlan 3000

vlan 3000
  vn-segment 1003000
  xconnect
```

2.必须启用功能NGOAM，并需要此配置。

```
VTEP1# sh run ngoam

feature ngoam

ngoam install acl
ngoam xconnect hb-interval 5000
```

3.指向下游交换机的Dot1q隧道配置。

```
VTEP1# sh run int po10

interface port-channel10
```

```
  switchport
  switchport mode dot1q-tunnel
  switchport access vlan 3000
  speed 40000
  no negotiate auto
  vpc 10
```

仅当VTEP部署为vPC时，才需要vPC配置。否则，请跳过本文档中提到的vPC配置。VXLAN Xconnect也可在独立VTEP上配置。

4.必须在NVE接口下定义组播组以处理转发。请注意，在相**关上行链路上启**用ip pim sparse-mode，并定义PIM RP，以便正确交换组播路由和PIM消息。通常，PIM RP在主干层定义。

```
VTEP1# sh run int nve1

  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  member vni 1003000 mcast-group 239.30.30.30
```

5.需要指定基础设施VLAN并允许其作为对等链路中的本征VLAN。vPC VTEP需要此步骤。

```
VTEP1# sh run span|infra
no spanning-tree vlan 3000
system nve infra-vlans 999

VTEP1# sh run int po1

interface port-channel1
  switchport
  switchport mode trunk
  switchport trunk native vlan 999
  spanning-tree port type network
  vpc peer-link
```

6. BGP/EVPN配置：枝叶/主干之间需要L2VPN EVPN邻居关系来交换建立VXLAN Xconnect所需的第3类路由。

- 此处，IP地址192.168.100.1和192.168.100.2是拓扑中的主干。通常，L2VPN EVPN邻居关系会形成到主干。主干在iBGP方案中将所有枝叶交换机配置为路由反射器客户端。
- 建议使用单独的环回用于BGP/OSPF和NVE。

```
feature bgp

router bgp 65000
  router-id 192.168.100.3
  neighbor 192.168.100.1
    remote-as 65000
    update-source loopback0
    address-family l2vpn evpn
      send-community
      send-community extended
  neighbor 192.168.100.2
    remote-as 65000
    update-source loopback0
    address-family l2vpn evpn
send-community
```

```
send-community extended evpn vni 1003000 l2 rd auto route-target import auto route-target export
auto
```

注意：必须在Xconnect VLAN中禁用STP。MAC学习不会在Xconnect VLAN内发生，这实质上意味着MAC地址没有第2类bgp l2vpn evpn更新。因此，来自一个vtep的流量将使用为Xconnect VLAN定义的Mcast组(239.30.30.30)的外部目标IP地址进行封装。

验证

使用本部分可确认配置能否正常运行。

1. BGP邻居关系。

```
 VTEP1# sh bgp l2vpn evpn sum
BGP summary information for VRF default, address family L2VPN EVPN
BGP router identifier 192.168.100.3, local AS number 65000
BGP table version is 14, L2VPN EVPN config peers 2, capable peers 1
4 network entries and 5 paths using 756 bytes of memory
BGP attribute entries [3/492], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [2/8]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
192.168.100.1   4 65000      92      90       14    0    0 01:21:41 2
```

2.接收第3类前缀。

```
VTEP1# sh bgp l2vpn evpn
BGP routing table information for VRF default, address family L2VPN EVPN
BGP table version is 14, Local Router ID is 192.168.100.3
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 192.168.100.3:35767    (L2VNI 1003000)
*>l[3]:[0]:[32]:[172.16.1.100]/88
                      172.16.1.100                       100      32768 i
* i[3]:[0]:[32]:[172.17.1.100]/88<<< bgp type 3
                      172.17.1.100                       100          0 i
*>i                   172.17.1.100                       100          0 i

Route Distinguisher: 192.168.100.5:35767
*>i[3]:[0]:[32]:[172.17.1.100]/88
                      172.17.1.100                       100          0 i

Route Distinguisher: 192.168.100.6:35767
*>i[3]:[0]:[32]:[172.17.1.100]/88
                      172.17.1.100                       100          0 i
```

3. NVE对等。

```
 VTEP1# sh nve peer
Interface Peer-IP          State LearnType Uptime   Router-Mac
--------- ---------------  ----- --------- -------- -----------------
nve1      172.17.1.100     Up    CP        00:58:06 n/a
```

```
VTEP1# show nve vni
Codes: CP - Control Plane        DP - Data Plane
       UC - Unconfigured         SA - Suppress ARP
       SU - Suppress Unknown Unicast

Interface VNI      Multicast-group   State Mode Type [BD/VRF]      Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1      1003000  239.30.30.30      Up    CP   L2 [3000]                  Xconn  <<<
```

4. 恩戈姆检查。

```
VTEP1# show ngoam xconnect sess all

States: LD = Local interface down, RD = Remote interface Down
          HB = Heartbeat lost, DB = Database/Routes not present
          * - Showing Vpc-peer interface info
Vlan          Peer-ip/vni      XC-State       Local-if/State    Rmt-if/State
================================================================================
3000  172.17.1.100 / 1003000      Active          Po10 / UP          Po10 / UP


VTEP1# show ngoam xconnect sess 3000
Vlan ID: 3000
Peer IP: 172.17.1.100  VNI : 1003000
State: Active                                    <<< State should be active
Last state update: 12/10/2018 17:13:45.337
Local interface: Po10  State:  UP
Local vpc interface Po10  State:  UP
Remote interface: Po10  State:  UP
Remote vpc interface: Po10  State:  UP
```

NGOAM会话启动后，N3k将在CDP中看到彼此。STP BPDU也通过隧道传输，以便交换机也同意根网桥的放置。

5.验证下游交换机。

```
SW1(config)# sh span vl 10

VLAN0010
  Spanning tree enabled protocol rstp
  Root ID    Priority    32778
             Address     7079.b348.6cb7
             This bridge is the root
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID  Priority    32778  (priority 32768 sys-id-ext 10)
             Address     7079.b348.6cb7
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Po10             Desg FWD 1         128.4105 P2p

SW2(config)# sh span vl 10

VLAN0010
  Spanning tree enabled protocol rstp
  Root ID    Priority    32778
             Address     7079.b348.6cb7
             Cost        1
             Port        4105 (port-channel10)
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec
```

```
  Bridge ID  Priority    32778  (priority 32768 sys-id-ext 10)
             Address     707d.b964.9441
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Po10             Root FWD 1         128.4105 P2p

SW1(config)# show ip int b
IP Interface Status for VRF "default"(1)
Interface            IP Address      Interface Status
Vlan10               192.168.10.100  protocol-up/link-up/admin-up
Vlan20               192.168.20.100  protocol-up/link-up/admin-up
Vlan30               192.168.30.100  protocol-up/link-up/admin-up

SW2(config)# show ip int b
IP Interface Status for VRF "default"(1)
Interface            IP Address      Interface Status
Vlan10               192.168.10.200  protocol-up/link-up/admin-up
Vlan20               192.168.20.200  protocol-up/link-up/admin-up
Vlan30               192.168.30.200  protocol-up/link-up/admin-up

SW1(config)# ping 192.168.10.200
PING 192.168.10.200 (192.168.10.200): 56 data bytes
64 bytes from 192.168.10.200: icmp_seq=0 ttl=254 time=0.826 ms
64 bytes from 192.168.10.200: icmp_seq=1 ttl=254 time=0.531 ms
64 bytes from 192.168.10.200: icmp_seq=2 ttl=254 time=0.54 ms
64 bytes from 192.168.10.200: icmp_seq=3 ttl=254 time=0.522 ms
64 bytes from 192.168.10.200: icmp_seq=4 ttl=254 time=0.571 ms
```

故障排除

目前没有针对此配置的故障排除信息。

注意事项

1.如果vPC交换机内的配置不一致，则在Xconnect VXLAN设置**中，dot1q隧道接口将陷**入错误禁用状态。以下是某些接口将错误禁用的情况；

- 如果两台vPC交换机上未定义VLAN到VN网段。
- 如果两个vPC交换机上未定义到组播组的NVE。
- 如果未收到NGOAM心跳(使用ethanalyzer with filter=**cfm**捕获NGOAM心跳数据包)。
- 即使dot1q隧道接口在vPC设置中是孤立连接的，仍需要在NVE接口下为VN网段配置组播组，VN网段是两台交换机Xconnect的一部分。
- NGOAM心跳由vPC主交换机处理/发送。登录vPC辅助的心跳消息将同步到主

2.在VLAN中配置Xconnect时，从一个站点到另一个站点的流量将使用在特定vn网段的NVE接口下定义的外部目标地址=组播地址进行封装。建议为Xconnect VLAN使用唯一组播组。核心/主干中的组播必须正常运行。

3.组播流量可能会同时到达Xconnect远端的vPC框；但是，Decap获胜者（可解封BUM的盒子）将只是vPC对中的一台交换机。这可以使用命令 — **show forwarding multicast route group <Group address> source <SRC IP>来验证**。如果此处显示的标志是小写**v**，则表示该框是小写**V**，如果它是

大写V，则该框是小写赢，因此可以解封组播流量并进一步向下转发。

4.在基于93180YC的平台上，当主机孤立地连接到9k1，且9k1上没有S、G的OIL时，使用源IP->127.0.0.1和目标IP->共享NVE IP的特殊封装将组播数据包的副本发送到vPC对等体，如果9k2具有S、G条目的OIL，然后9k2将负责到远程站点的流量转发。

数据包捕获

以下是在主干交换机上捕获的数据包截图：

```
Frame 1: 152 bytes on wire (1216 bits), 152 bytes captured (1216 bits)
Ethernet II, Src: Cisco_2a:89:a7 (70:79:b3:2a:89:a7), Dst: IPv4mcast_1e:1e:1e (01:00:5e:1e:1e:1e)
Internet Protocol Version 4, Src: 172.17.1.100, Dst: 239.30.30.30
User Datagram Protocol, Src Port: 12860, Dst Port: 4789
Virtual eXtensible Local Area Network
  > Flags: 0x0800, VXLAN Network ID (VNI)
    Group Policy ID: 0
    VXLAN Network Identifier (VNI): 1003000
    Reserved: 0
Ethernet II, Src: Cisco_64:94:41 (70:7d:b9:64:94:41), Dst: Cisco_48:6c:b7 (70:79:b3:48:6c:b7)
802.1Q Virtual LAN, PRI: 0, DEI: 0, ID: 10
    000. .... .... .... = Priority: Best Effort (default) (0)
    ...0 .... .... .... = DEI: Ineligible
    .... 0000 0000 1010 = ID: 10
    Type: IPv4 (0x0800)
Internet Protocol Version 4, Src: 192.168.10.200, Dst: 192.168.10.100
```

- 保留内部dot1q报头=10
- 使用的VNI是1003000（即外部VLAN的VNID）
- 目的IP地址是在NVE接口下定义的组播组