方形拓扑中采用CloudSec的多站点VXLAN故障排除

目录

简介

本文档介绍以方形拓扑连接的边界网关之间的VXLAN多站点配置和CloudSec故障排除。

先决条件

要求

思科建议您熟悉以下主题：

- Nexus NXOS ©软件。
- VXLAN EVPN技术。
- BGP和OSPF路由协议。

使用的组件

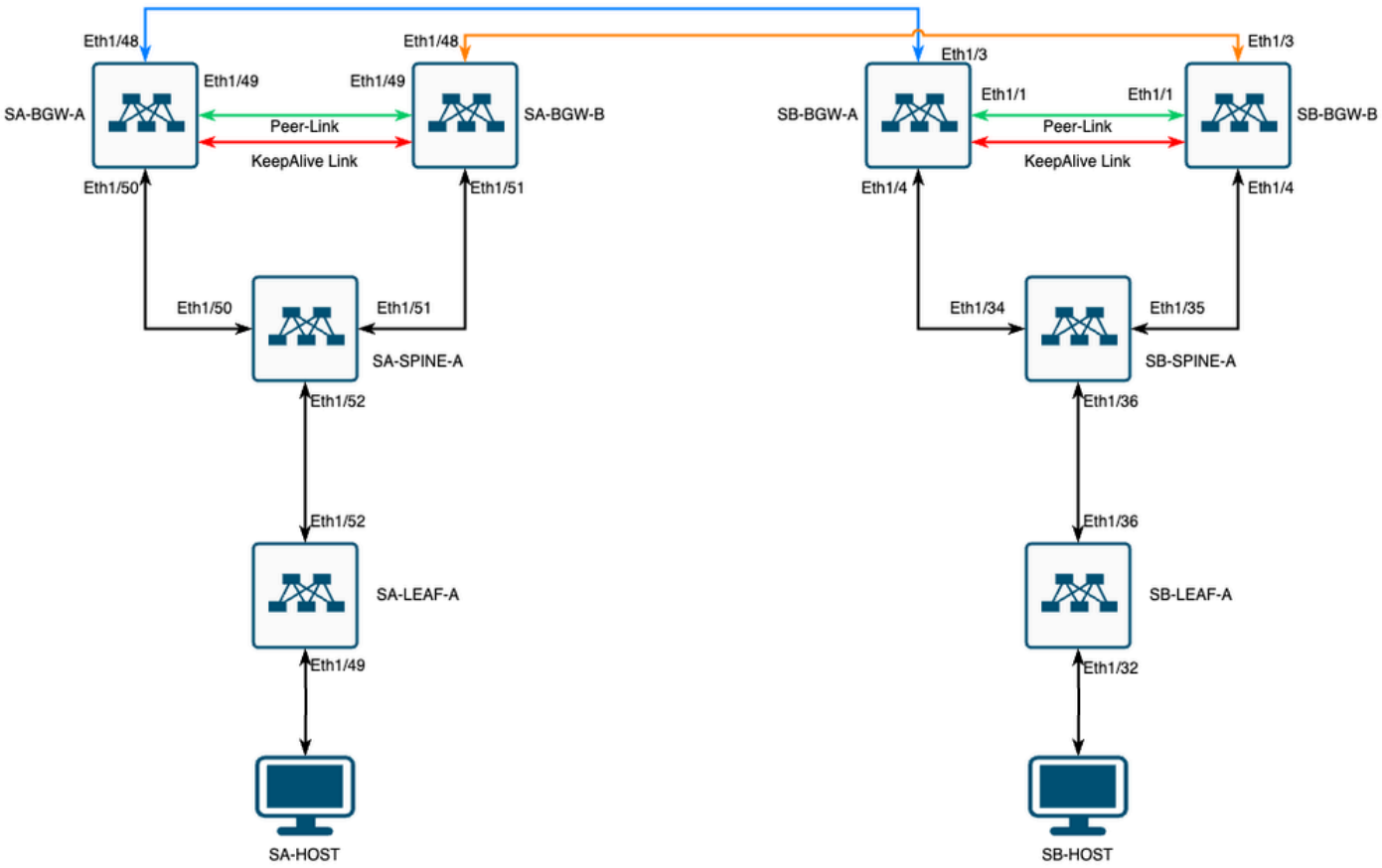
本文档中的信息基于以下软件和硬件版本：

- Cisco Nexus 9000。
- NXOS 10.3(4a)版。

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。如果您的网络处于活动状态，请确保您了解所有命令的潜在影响。

配置

网络图



VXLAN MultiSite，采用CloudSec方形拓扑

拓扑的详细信息

- 双站点多站点VXLAN EVPN交换矩阵。
- 两个站点都配置了vPC边界网关。
- 终端托管在VLAN 1100中。
- 每个站点的边界网关在SVI接口Vlan3600上彼此之间具有IPv4 iBGP邻居关系。
- 一个站点上的边界网关仅与另一个站点上的直接连接的边界网关具有eBGP IPv4邻居关系。
- 站点A的边界网关与站点B的边界网关具有eBGP L2VPN EVPN邻居关系。

寻址计划

表中的IP地址在配置过程中使用：

	站点 A	站点 B				
设备角色	接口 ID	物理接口IP	RID环路IP	NVE环路IP	MSITE-VIP	备份SVI IP

枝叶	Eth1/52	192.168.1.1/30	192.168.2.1/32	192.168.3.1/32	不适用	不适用
主干	Eth1/52	192.168.1.2/30			不适用	
Eth1/50	192.168.1.5/30	192.168.2.2/32	不适用	不适用	不适用	Eth1/34
Eth1/51	192.168.1.9/30			不适用		Eth1/35
BGW-A	Eth1/51	192.168.1.6/30	192.168.2.3/32	192.168.3.2/32	192.168.100.1/32	192.168.4.1/3
Eth1/48	10.12.10.1/30		192.168.3.254/32			Eth1/3
BGW-B	Eth1/51	192.168.1.10/30	192.168.2.4/32	192.168.3.3/32	192.168.100.1/32	192.168.4.2/3
Eth1/48	10.12.10.5/30		192.168.3.254/32			Eth1/3

配置

- 请注意，本指南中仅显示与多站点相关的配置。对于完整配置，您可以使用VXLAN的思科官方文档指南Cisco Nexus 9000系列NX-OS VXLAN配置指南，版本10.3(x)

要启用CloudSec，必须在evpn multisite border-gateway下配置dci-advertise-pip 命令：

SA-BGW-A和SA-BGW-B	SB-BGW-A和SB-BGW-B
evpn multisite border-gateway 65001 dci-advertise-pip	evpn multisite border-gateway 65002 dci-advertise-pip

BGP配置

此配置特定于站点。

SA-BGW-A和SA-BGW-B	SB-BGW-A和SB-BGW-B
router bgp 65001 address-family ipv4 unicast maximum-paths 64 address-family l2vpn evpn maximum-paths 64 additional-paths send additional-paths receive	router bgp 65002 address-family ipv4 unicast maximum-paths 64 address-family l2vpn evpn maximum-paths 64 additional-paths send additional-paths receive

- **maximum-path** 命令允许从邻居接收多个eBGP L2VPN EVPN路径。
- **additional-path** 命令指示BGP进程通告设备有能力发送/接收其他路径

对于边界网关上的所有L3VNI VRF，还必须配置多路径：

SA-BGW-A和SA-BGW-B	SB-BGW-A和SB-BGW-B

router bgp 65001 vrf tenant-1 address-family ipv4 unicast maximum-paths 64 address-family ipv6 unicast maximum-paths 64	router bgp 65002 vrf tenant-1 address-family ipv4 unicast maximum-paths 64 address-family ipv6 unicast maximum-paths 64

隧道加密配置

所有边界网关上的此配置必须相同：

key chain CloudSec_Key_Chain1 tunnel-encryption key 1000 key-octet-string Cl0udSec! cryptographic-algorithm AES_128_CMAC feature tunnel-encryp

此配置特定于站点。tunnel-encryption命令必须只应用于带有evpn multisite dci-tracking命令的接口。

SA-BGW-A和SA-BGW-B	SB-BGW-A和SB-BGW-B
tunnel-encryption peer-ip 192.168.13.2 keychain CloudSec_Key_Chain1 policy CloudSec_Policy1 tunnel-encryption peer-ip 192.168.13.3 keychain CloudSec_Key_Chain1 policy CloudSec_Policy1 interface Ethernet1/48 tunnel-encryption	tunnel-encryption peer-ip 192.168.3.2 keychain CloudSec_Key_Chain1 policy CloudSec_Policy1 tunnel-encryption peer-ip 192.168.3.3 keychain CloudSec_Key_Chain1 policy CloudSec_Policy1 interface Ethernet1/3 tunnel-encryption

启用隧道加密后，向邻居通告路由时，其他属性将添加到本地环回，并且所有eBGP IPv4单播邻居必须看到此属性：

<#root>

SA-BGW-A# show ip bgp 192.168.2.3 BGP routing table information for VRF default, address family IPv4 Unicast BGP routing table entry for 192.168.2

!---

This is a new attribute

Path type: redist, path is valid, not best reason: Locally originated, no labeled nexthop AS-Path: NON

对于路由类型2，还有新属性：

<#root>

SA-BGW-A# show bgp l2vpn evpn 00ea.bd27.86ef BGP routing table information for VRF default, address family L2VPN EVPN Route Distinguisher: 65

!---

Ethernet Segment Identifier (ESI) is also new attribute

```
 Path-id 1 (dual) advertised to peers: 192.168.2.2 SA-BGW-A#
```

验证

在启用cloudsec之前，最好检查一下没有它的设置是否运行正常：

SA-BGW-A(config)# show clock Warning: No NTP peer/server configured. Time may be out of sync. 10:02:01.016 UTC Fri Jul 19 2024 Time source is N

配置完cloudsec后，SA上的终端必须成功ping通站点B上的终端。但是，在某些情况下，ping操作可能不成功。这取决于本地设备选择用于发送cloudsec加密流量的cloudsec对等体。

SA-HOST-A# ping 10.100.20.10 PING 10.100.20.10 (10.100.20.10): 56 data bytes Request 0 timed out Request 1 timed out Request 2 timed out Request 3

故障排除

检查源终端上的本地ARP表：

SA-HOST-A# ping 10.100.20.10 count unlimited interval 1 Request 352 timed out Request 353 timed out Request 354 timed out 356 packets transmitted, 0

此输出证明BUM流量正在传递且控制平面工作正常。下一步是检查隧道加密状态：

SA-BGW-A# show tunnel-encryption session Tunnel-Encryption Peer Policy Keychain RxStatus TxStatus ---------- ---------- ----------

此输出显示CloudSec会话已建立。下一步可以在SA-HOST-A上运行无限制ping：

SA-HOST-A# ping 10.100.20.10 count unlimited interval 1

从此时起，您必须检查站点A上的设备，查看流量是否到达此设备。您可以通过ELAM在站点A路径上的所有设备上完成此任务。 in-select 将默认值6更改为9可根据内部报头进行匹配。您可以通过此链接了解有关ELAM的更多信息：Nexus 9000云扩展ASIC (Tahoe) NX-OS ELAM。

SA-LEAF-A上的ELAM

在生产网络中，存在多个SPINE设备。要了解流量发送到哪个主干，您必须首先在LEAF上获取ELAM。尽管使用了 in-select 9 这种方法，但在连接到源的枝叶上，必须使用外部ipv4报头，因为到达此枝叶的流量不是VXLAN加密的。在真实网络中，可能很难捕获到您生成的确切数据包。在这种情况下，您可以运行特定长度的ping，并使用Pkt len报头来识别数据包。默认情况下，icmp数据包的长度为64字节。加上20字节的IP报头，总而言之，为您提供了84字节的PKT Len：

```
<#root>

SA-LEAF-A# debug platform internal tah elam SA-LEAF-A(TAH-elam)# trigger init in-select 9 Slot 1: param values: start asic 0, start slice 0, lu-a2d 1, in

!---Note dpid value

 Dst Idx : 0xcd, Dst BD : 1100 Packet Type: IPv4 Outer Dst IPv4 address: 10.100.20.10 Outer Src IPv4 ad

Pkt len = 84

, Checksum = 0xb4ae

!---64 byte + 20 byte IP header Pkt len = 84

 Inner Payload Type: CE L4 Protocol : 1 L4 info not available Drop Info: ---------- LUA: LUB: LUC: LUD:

!---

Put dpid value here

 IF_STATIC_INFO: port_name=Ethernet1/52,if_index:0x1a006600,ltl=5940,slot=0, nxos_port=204,dmod=1,dpid=
```

从该输出中，您可以看到流量到达SA-LEAF-A并从接口Ethernet1/52转发出去，该接口从拓扑连接到SA-SPINE-A。

SA-SPINE-A上的ELAM

在SPINE上，Pkt Len值将更大，因为50字节的VXLAN报头也会添加。默认情况下，没有 vxlan-parse 或 feature nv overlay ，SPINE无法在内部报头上匹配。因此，您必须对SPINE使用 vxlan-parse enable 命令：

```
<#root>

SA-SPINE-A(config-if)# debug platform internal tah elam SA-SPINE-A(TAH-elam)# trigger init in-select 9 Slot 1: param values: start asic 0, start slice 0,

!---

84 bytes + 50 bytes VXLAN header Pkt len = 134

 Inner Payload Type: IPv4 Inner Dst IPv4 address: 10.100.20.10 Inner Src IPv4 address: 10.100.10.10 L4
```

SA-SPINE-A根据输出向SA-BGW-A发送流量。

SA-BGW-A上的ELAM

SA-BGW-A(TAH-elam-insel9)# set inner ipv4 src_ip 10.100.10.10 dst_ip 10.100.20.10 SA-BGW-A(TAH-elam-insel9)# start SA-BGW-A(TAH-elam-ins

根据SA-BGW-A的输出，流量从Ethernet1/48流向SB-BGW-A。下一步是检查SB-BGW-A：

<#root>

SB-BGW-A# debug platform internal tah elam SB-BGW-A(TAH-elam)# trigger init in-select 9 Slot 1: param values: start asic 0, start slice 0, lu-a2d 1, in-

!---Reset the previous filter and start again just in case if packet was not captured.

SB-BGW-A(TAH-elam-insel9)# reset SB-BGW-A(TAH-elam-insel9)# set inner ipv4 src_ip 10.100.10.10 dst_ip

根据SB-BGW-A的输出，ELAM甚至未触发。这意味着SB-BGW-B正在接收数据包，无法正确解密和解析这些数据包，或者根本无法接收这些数据包。要了解cloudsec流量发生的情况，可以再次在SB-BGW-A上运行ELAM，但是触发过滤器必须设置为用于cloudsec的外部IP地址，因为无法查看cloudsec加密传输数据包的内部报头。从先前的输出可以了解到，SA-BGW-A处理了流量，这意味着SA-BGW-A使用cloudsec加密流量。因此，您可以使用SA-BGW-A的NVE IP作为ELAM的触发过滤器。根据前面的输出，VXLAN加密ICMP数据包长度为134字节。加上摘要中的32字节的cloudsec报头，可得出166字节：

<#root>

SB-BGW-A(TAH-elam-insel9)# reset SB-BGW-A(TAH-elam-insel9)# set outer ipv4 src_ip 192.168.3.2 SB-BGW-A(TAH-elam-insel9)# start SB-BGW-

192.168.13.3 ***!---NVE IP address of SB-BGW-B***

Outer Src IPv4 address: 192.168.3.2 Ver = 4, DSCP = 0, Don't Fragment = 0 Proto = 17, TTL = 254, More

!---134 byte VXLAN packet + 32 byte cloudsec header Pkt len = 166

Inner Payload Type: CE L4 Protocol : 17 L4 info not available Drop Info: ---------- LUA: LUB: LUC: LUD

!---To reach SB-BGW-B NVE IP traffic was sent out of Ethernet1/4 which is connected to SB-SPINE-A

SB-BGW-A(TAH-elam-insel9)# show system internal ethpm info all | i i "dpid=130" IF_STATIC_INFO: port_n
SB-BGW-A(TAH-elam-insel9)# show cdp neighbors interface ethernet 1/4 Capability Codes: R - Router, T - Trans-Bridge, B - Source-Route-Bridge S - S

192.168.13.3/32

, ubest/mbest: 1/0 *via 192.168.11.5,

Eth1/4

, [110/6], 00:56:13, ospf-UNDERLAY, intra via

192.168.14.2

, [200/0], 01:13:46, bgp-65002, internal, tag 65002

!---The device still have a route for SB-BGW-B NVE IP via SVI

SB-BGW-A(TAH-elam-insel9)# show ip route 192.168.14.2 IP Route Table for VRF "default" '*' denotes best

***via 192.168.14.2, Vlan3600**

, [250/0], 01:15:05, am SB-BGW-A(TAH-elam-insel9)# show ip arp 192.168.14.2 Flags: * - Adjacencies lear

ecce.1324.c803

Vlan3600

SB-BGW-A(TAH-elam-insel9)# show mac address-table address ecce.1324.c803 Legend: * - primary entry, G

3600

ecce.1324.c803

static - F F

vPC Peer-Link(R)

SB-BGW-A(TAH-elam-insel9)#

从该输出中，您可以看到，根据路由表，Cloudsec流量通过接口Ethernet1/4转发到SB-BGW-B。根据Cisco Nexus 9000系列NX-OS VXLAN配置指南，版本10.3(x)准则和限制：

- 流向交换机的CloudSec流量必须通过DCI上行链路进入交换机。

根据同一指南的vPC边界网关对Cloudsec的支持部分，如果vPC BGW了解对等vPC BGW的PIP地址并在DCI端进行通告，则来自两个vPC BGW的BGP路径属性将相同。因此，DCI中间节点最终可以从不拥有PIP地址的vPC BGW中选择路径。在此场景中，MCT链路用于来自远程站点的加密流量。但是，在这种情况下，使用的是指向SPINE的接口，尽管如此，BGW也通过BackUp SVI具有OSPF邻接关系。

SB-BGW-A(TAH-elam-insel9)# show ip ospf neighbors OSPF Process ID UNDERLAY VRF default Total number of neighbors: 2 Neighbor ID Pri State

问题的原因和修复

原因是SVI接口的OSPF开销。默认情况下，在NXOS上，自动成本参考带宽为40G。SVI接口的带宽为1Gbps，而物理接口的带宽为10Gbps：

<#root>

SB-BGW-A(TAH-elam-insel9)# show ip ospf interface brief OSPF Process ID UNDERLAY VRF default Total number of interface: 5 Interface ID Area C

<Output omitted>

Eth1/4 5 0.0.0.0 1 P2P 1 up

在这种情况下，对SVI成本进行管理更改可以解决此问题。必须在所有边界网关上进行调整。

<#root>

SB-BGW-A(config)# int vlan 3600 SB-BGW-A(config-if)# ip ospf cost 1 SB-BGW-A(config-if)# sh ip route 192.168.13.3 IP Route Table for VRF "defau

via 192.168.14.2

, Vlan3600, [110/2], 00:00:08, ospf-UNDERLAY, intra via 192.168.14.2, [200/0], 01:34:07, bgp-65002, int

!---The ping is started to work immediately

Request 1204 timed out Request 1205 timed out Request 1206 timed out 64 bytes from 10.100.20.10: icmp_seq=1207 ttl=254 time=1.476 ms 64 bytes from

关于此翻译
思科采用人工翻译与机器翻译相结合的方式将此文档翻译成不同语言，希望全球的用户都能通过各自的语言得到支持性的内容。

请注意：即使是最好的机器翻译，其准确度也不及专业翻译人员的水平。

Cisco Systems, Inc. 对于翻译的准确性不承担任何责任，并建议您总是参考英文原始文档（已提供链接）。