Troubleshoot Multisite VXLAN with CloudSec in Square Topology

Contents

Introduction

This document describes VXLAN Multisite configuration and troubleshooting with CloudSec between border gateways connected in square topology.

Prerequisites

Requirements

Cisco recommends that you are familiar with these topics:

- Nexus NXOS © Software.
- VXLAN EVPN technology.
- BGP and OSPF routing protocols.

Components Used

The information in this document is based on the these software and hardware versions:

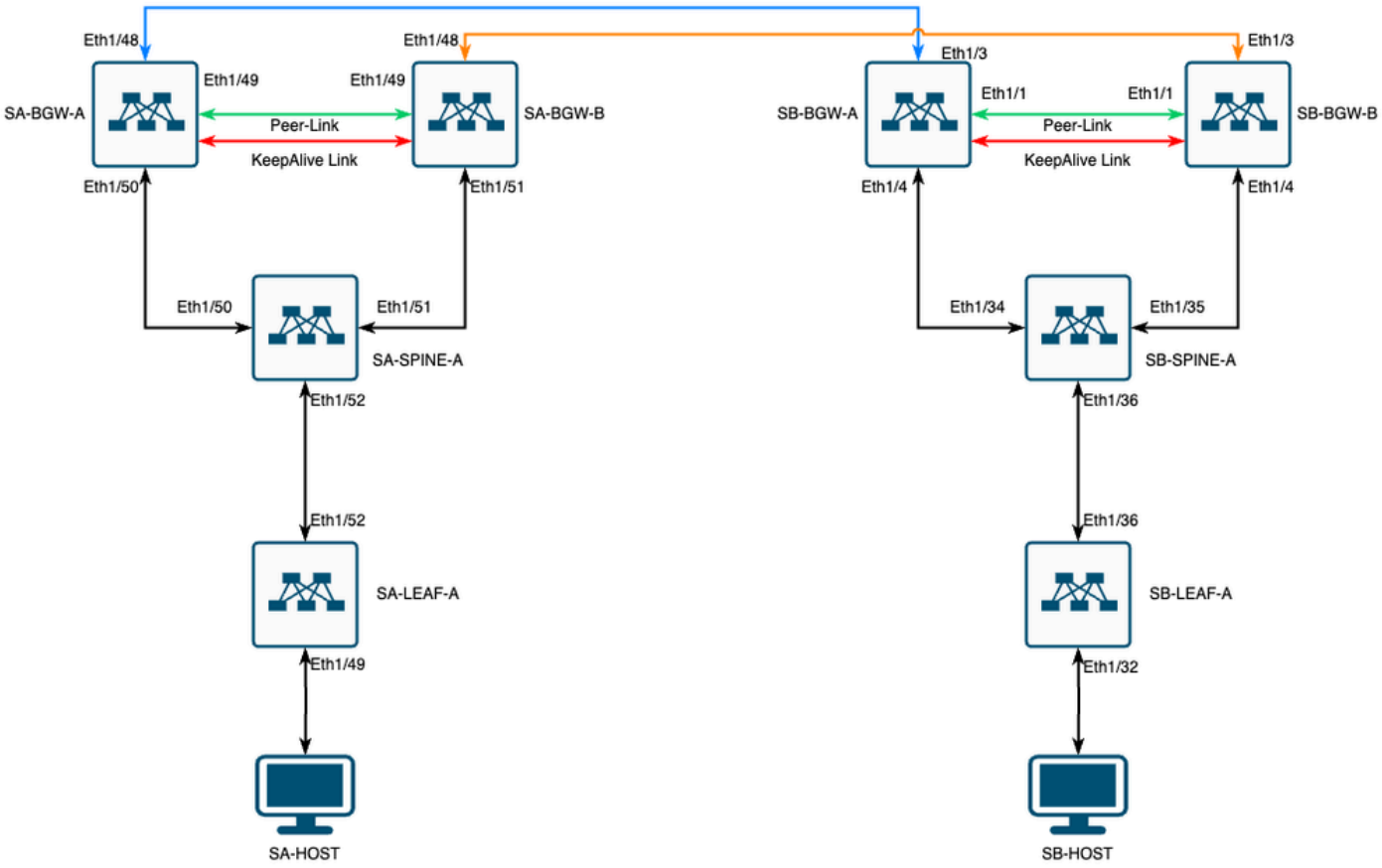
- Cisco Nexus 9000.
- NXOS version 10.3(4a).

The information in this document was created from the devices in a specific lab environment. All of the

devices used in this document started with a cleared (default) configuration. If your network is live, ensure that you understand the potential impact of any command.

Configure

Network Diagram



VXLAN MultiSite with CloudSec in square topology

Details of the topology

- Two Site Multisite VXLAN EVPN Fabric.
- Both sites are configured with vPC Border Gateways.
- Endpoints are hosted in VLAN 1100.
- Border gateways on each site have IPv4 iBGP neighborship between each other over the SVI interface Vlan3600.
- Border gateways on one site have eBGP IPv4 neighborship only with directly connected border gateway on the other site.
- Border gateways on site A have eBGP L2VPN EVPN neighborship with border gateways on site B.

Addressing plan

The IP addresses in the table are used during the configuration:

	SITE A	SITE B					
Device Role	Interface ID	Physical Int IP	RID Loop IP	NVE Loop IP	MSITE-VIP	BackUp SVI IP	Interfa
LEAF	Eth1/52	192.168.1.1/30	192.168.2.1/32	192.168.3.1/32	N/A	N/A	Eth1

SPINE	Eth1/52	192.168.1.2/30			N/A		Eth1
Eth1/50	192.168.1.5/30	192.168.2.2/32	N/A	N/A	N/A	Eth1/34	192.168
Eth1/51	192.168.1.9/30			N/A		Eth1/35	192.168
BGW-A	Eth1/51	192.168.1.6/30	192.168.2.3/32	192.168.3.2/32	192.168.100.1/32	192.168.4.1/30	Eth
Eth1/48	10.12.10.1/30		192.168.3.254/32			Eth1/3	10.12.1
BGW-B	Eth1/51	192.168.1.10/30	192.168.2.4/32	192.168.3.3/32	192.168.100.1/32	192.168.4.2/30	Eth
Eth1/48	10.12.10.5/30		192.168.3.254/32			Eth1/3	10.12.1

Configurations

- Note that in this guide only multisite-related configuration is shown. For full configuration, you can use Cisco official documentation guide for VXLAN Cisco Nexus 9000 Series NX-OS VXLAN Configuration Guide, Release 10.3(x)

In order to enable CloudSec the `dci-advertise-pip` command must be configured under the evpn multisite border-gateway:

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

BGP configuration

This configuration is site-specific.

SA-BGW-A and SA-BGW-B	SB-BGW-A and 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

- The **maximum-path** command allows to receive multiple eBGP L2VPN EVPN paths from the neighbour.
- The **additional-path** command instructs the BGP process to advertise that the device is capable send/receive additional paths

For all L3VNI VRFs on border gateways, multipath must be configured also:

SA-BGW-A and SA-BGW-B	SB-BGW-A and SB-BGW-B
router bgp 65001	router bgp 65002

vrf tenant-1 address-family ipv4 unicast maximum-paths 64 address-family ipv6 unicast maximum-paths 64	vrf tenant-1 address-family ipv4 unicast maximum-paths 64 address-family ipv6 unicast maximum-paths 64

Tunnel-encryption configuration

This configuration must be the same on all border gateways:

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

feature tunnel-encryption

tunnel-encryption must-secure-policy
tunnel-encryption source-interface loopback0
tunnel-encryption policy CloudSec_Policy1
```

This configuration is site-specific. The tunnel-encryption command must be applied only to the interface which has evpn multisite dci-trackingcommand.

SA-BGW-A and SA-BGW-B	SB-BGW-A and 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

After enabling the tunnel-encryption additional attributes are added to the local loopback while advertising routes to the neighbour and all the eBGP IPv4 unicast neighbours must see this attribute:

<#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 192.168.2.3/32, version 1320
Paths: (2 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in urib
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    0.0.0.0 (metric 0) from 0.0.0.0 (192.168.2.3)
```

Origin IGP, MED not set, localpref 100, weight 32768
Tunnel Encapsulation attribute: Length 152

!---

This is a new attribute

```
  Path type: redist, path is valid, not best reason: Locally originated, no labeled nexthop
  AS-Path: NONE, path locally originated
    0.0.0.0 (metric 0) from 0.0.0.0 (192.168.2.3)
      Origin incomplete, MED 0, localpref 100, weight 32768

  Path-id 1 advertised to peers:
    10.12.10.2         192.168.4.2

SA-BGW-A#
```

For Route Type-2 there is also new attribute:

<#root>

SA-BGW-A# show bgp l2vpn evpn 00ea.bd27.86ef
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 65002:31100
BGP routing table entry for [2]:[0]:[0]:[48]:[00ea.bd27.86ef]:[0]:[0.0.0.0]/216, version 7092
Paths: (2 available, best #2)
Flags: (0x000202) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW
Multipath: eBGP iBGP

Path type: external, path is valid, not best reason: Router Id, multipath, no labeled nexthop
Imported to 1 destination(s)
Imported paths list: L2-31100
AS-Path: 65002 , path sourced external to AS
192.168.13.3 (metric 0) from 192.168.12.4 (192.168.12.4)
Origin IGP, MED 2000, localpref 100, weight 0
Received label 31100
Received path-id 1
Extcommunity: RT:65001:31100 ENCAP:8
ESI: 0300.0000.00fd.ea00.0309

Advertised path-id 1
Path type: external, path is valid, is best path, no labeled nexthop
Imported to 1 destination(s)
Imported paths list: L2-31100
AS-Path: 65002 , path sourced external to AS
192.168.13.2 (metric 0) from 192.168.12.3 (192.168.12.3)
Origin IGP, MED 2000, localpref 100, weight 0
Received label 31100
Received path-id 1
Extcommunity: RT:65001:31100 ENCAP:8
ESI: 0300.0000.00fd.ea00.0309

Path-id 1 not advertised to any peer

Route Distinguisher: 192.168.2.3:33867 (L2VNI 31100)
BGP routing table entry for [2]:[0]:[0]:[48]:[00ea.bd27.86ef]:[0]:[0.0.0.0]/216, version 7112
Paths: (2 available, best #1)

```
Flags: (0x000212) (high32 0x000400) on xmit-list, is in l2rib/evpn, is not in HW
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, in rib
          Imported from 65002:31100:[2]:[0]:[0]:[48]:[00ea.bd27.86ef]:[0]:[0.0.0.0]/216
  AS-Path: 65002 , path sourced external to AS
    192.168.13.2 (metric 0) from 192.168.12.3 (192.168.12.3)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 31100
      Received path-id 1
      Extcommunity: RT:65001:31100 ENCAP:8
      ESI: 0300.0000.00fd.ea00.0309

  Path type: external, path is valid, not best reason: Router Id, multipath, no labeled nexthop, in rib
          Imported from 65002:31100:[2]:[0]:[0]:[48]:[00ea.bd27.86ef]:[0]:[0.0.0.0]/216
  AS-Path: 65002 , path sourced external to AS
    192.168.13.3 (metric 0) from 192.168.12.4 (192.168.12.4)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 31100
      Received path-id 1
      Extcommunity: RT:65001:31100 ENCAP:8
      ESI: 0300.0000.00fd.ea00.0309
```

!---

Ethernet Segment Identifier (ESI) is also new attribute

```
  Path-id 1 (dual) advertised to peers:
    192.168.2.2

SA-BGW-A#
```

Verify

Before enable cloudsec, it is good to check if the setup is working fine without it:

```
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 NTP
SA-BGW-A(config)# show tunnel-encryption session
Tunnel-Encryption Peer   Policy                          Keychain                        RxStatus        TxStatus
------------------------ ------------------------------- ------------------------------- ----------------- -----------------
==============================================
SA-HOST-A# show clock
Warning: No NTP peer/server configured. Time may be out of sync.
10:02:21.592 UTC Fri Jul 19 2024
Time source is NTP
SA-HOST-A# ping 10.100.20.10 count unlimited interval 1
PING 10.100.20.10 (10.100.20.10): 56 data bytes
64 bytes from 10.100.20.10: icmp_seq=0 ttl=254 time=1.583 ms
64 bytes from 10.100.20.10: icmp_seq=1 ttl=254 time=10.407 ms
64 bytes from 10.100.20.10: icmp_seq=2 ttl=254 time=1.37 ms
```

```
64 bytes from 10.100.20.10: icmp_seq=3 ttl=254 time=1.489 ms
64 bytes from 10.100.20.10: icmp_seq=4 ttl=254 time=6.685 ms
64 bytes from 10.100.20.10: icmp_seq=5 ttl=254 time=1.547 ms
64 bytes from 10.100.20.10: icmp_seq=6 ttl=254 time=1.859 ms
64 bytes from 10.100.20.10: icmp_seq=7 ttl=254 time=5.219 ms
64 bytes from 10.100.20.10: icmp_seq=8 ttl=254 time=1.337 ms
64 bytes from 10.100.20.10: icmp_seq=9 ttl=254 time=3.528 ms
64 bytes from 10.100.20.10: icmp_seq=10 ttl=254 time=4.057 ms
```

After cloudsec configuration as well, endpoint on SA must successfully ping the endpoint on site B. But, in some cases the ping can be unsuccessful. It is depending on which cloudsec peer selected by the local device to send cloudsec encrypted traffic.

```
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 timed out
Request 4 timed out

--- 10.100.20.10 ping statistics ---
5 packets transmitted, 0 packets received, 100.00% packet loss
SA-HOST-A#
```

Troubleshoot

Check the local ARP table on the source endpoint:

```
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 packets received, 100.00% packet loss
SA-HOST-A# clear ip arp delete-force
SA-HOST-A# show ip arp

Flags: * - Adjacencies learnt on non-active FHRP router
       + - Adjacencies synced via CFSoE
       # - Adjacencies Throttled for Glean
       CP - Added via L2RIB, Control plane Adjacencies
       PS - Added via L2RIB, Peer Sync
       RO - Re-Originated Peer Sync Entry
       D - Static Adjacencies attached to down interface

IP ARP Table for context default
Total number of entries: 1
Address         Age       MAC Address     Interface       Flags
10.100.20.10    00:00:02  00ea.bd27.86ef  Vlan1100
SA-HOST-A#
```

This output proves that, the BUM traffic is passing and Control-Plane is working. The next step is checking the tunnel-encryption status:

```
SA-BGW-A# show tunnel-encryption session
Tunnel-Encryption Peer   Policy                    Keychain                  RxStatus       TxStatus
------------------------ ------------------------------------- ------------------------------------- ---------------- ----------------
192.168.13.2             CloudSec_Policy1          CloudSec_Key_Chain1             Secure (AN: 0)   Secure (AN: 0)
192.168.13.3             CloudSec_Policy1          CloudSec_Key_Chain1             Secure (AN: 0)   Secure (AN: 0)
SA-BGW-A#
```

This output shows that the CloudSec session is established. As a next step you can run unlimited ping on SA-HOST-A:

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

From this point you must check on devices on site A and see if traffic is reaching this devices. You can accomplish this task with ELAM on all devices along the path on site A. Changing in-select from default value of 6 to 9 allows to match based on inner headers. You can read more about ELAM on this link: [Nexus 9000 Cloud Scale ASIC (Tahoe) NX-OS ELAM.]

ELAM on SA-LEAF-A

In production network more than one SPINE devices are exist. To understand to which spine the traffic was sent, you must take an ELAM on LEAF first. Despite that in-select 9 used, at the LEAF connected to the source, the outer ipv4 header must be used, as the traffic reached this LEAF is not VXLAN encrypted. In real network, it can be hard to catch the exact packet you generated. In such cases, you can run ping with specific length and use the Pkt len header to identify your packet. By default, icmp packet is 64 byte length. Plus 20 byte of IP header, which in summary gave you 84 byte 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-select 9, out-select 0
SA-LEAF-A(TAH-elam-insel9)# set outer ipv4 src_ip 10.100.10.10 dst_ip 10.100.20.10
SA-LEAF-A(TAH-elam-insel9)# start
SA-LEAF-A(TAH-elam-insel9)# report
ELAM not triggered yet on slot - 1, asic - 0, slice - 0
SUGARBOWL ELAM REPORT SUMMARY
slot - 1, asic - 0, slice - 1
==========================

Incoming Interface: Eth1/49
Src Idx : 0xc1, Src BD : 1100
Outgoing Interface Info: dmod 1, dpid 64

!---Note dpid value

Dst Idx : 0xcd, Dst BD : 1100
```

```
Packet Type: IPv4

Outer Dst IPv4 address: 10.100.20.10
Outer Src IPv4 address: 10.100.10.10
Ver     =  4, DSCP    =    0, Don't Fragment = 0
Proto   =  1, TTL     =  255, More Fragments = 0
Hdr len = 20,
```

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:
Final Drops:

SA-LEAF-A(TAH-elam-insel9)# show system internal ethpm info all | i i "dpid=64"
```

!---

Put dpid value here

```
  IF_STATIC_INFO: port_name=Ethernet1/52,if_index:0x1a006600,ltl=5940,slot=0, nxos_port=204,dmod=1,dpid=
SA-LEAF-A(TAH-elam-insel9)# show cdp neighbors interface ethernet 1/52
Capability Codes: R - Router, T - Trans-Bridge, B - Source-Route-Bridge
                  S - Switch, H - Host, I - IGMP, r - Repeater,
                  V - VoIP-Phone, D - Remotely-Managed-Device,
                  s - Supports-STP-Dispute

Device-ID          Local Intrfce  Hldtme Capability  Platform      Port ID
SA-SPINE-A(FDO242210CS)
                    Eth1/52        130    R S s     N9K-C93240YC-FX2 Eth1/52

Total entries displayed: 1
SA-LEAF-A(TAH-elam-insel9)#
```

From this output you can see that traffic is reached SA-LEAF-A and forwarded out the interface Ethernet1/52, which is connected to SA-SPINE-A from the topology.

ELAM on SA-SPINE-A

On SPINE the Pkt Len value going to be more, since the 50 byte VXLAN header also added. By default, SPINE can not match on internal headers without vxlan-parse or feature nv overlay . So, you must use vxlan-parse enable command on SPINE:

<#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, lu-a2d 1, in-select 9, out-select 0
SA-SPINE-A(TAH-elam)# vxlan-parse enable
SA-SPINE-A(TAH-elam-insel9)# set inner ipv4 src_ip 10.100.10.10 dst_ip 10.100.20.10
SA-SPINE-A(TAH-elam-insel9)# start
SA-SPINE-A(TAH-elam-insel9)# report
ELAM not triggered yet on slot - 1, asic - 0, slice - 0
HEAVENLY ELAM REPORT SUMMARY
slot - 1, asic - 0, slice - 1
============================

Incoming Interface: Eth1/52
Src Idx : 0xcd, Src BD : 4153
Outgoing Interface Info: dmod 1, dpid 72
Dst Idx : 0xc5, Dst BD : 4151

Packet Type: IPv4

Outer Dst IPv4 address: 192.168.100.1
Outer Src IPv4 address: 192.168.3.1
Ver = 4, DSCP = 0, Don't Fragment = 0
Proto = 17, TTL = 255, More Fragments = 0
Hdr len = 20, Pkt len = 134, Checksum = 0x7d69

!---

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 Protocol : 17
L4 info not available

Drop Info:

LUA:
LUB:
LUC:
LUD:
Final Drops:

SA-SPINE-A(TAH-elam-insel9)# show system internal ethpm info all | i i "dpid=72"
  IF_STATIC_INFO: port_name=Ethernet1/50,if_index:0x1a006200,ltl=5948,slot=0, nxos_port=196,dmod=1,dpid=
SA-SPINE-A(TAH-elam-insel9)# show cdp neighbors interface ethernet 1/50
Capability Codes: R - Router, T - Trans-Bridge, B - Source-Route-Bridge
                  S - Switch, H - Host, I - IGMP, r - Repeater,
                  V - VoIP-Phone, D - Remotely-Managed-Device,
                  s - Supports-STP-Dispute

Device-ID   Local Intrfce Hldtme Capability Platform   Port ID
SA-BGW-A(FDO242210CX)
                    Eth1/50   169   R S s   N9K-C93240YC-FX2 Eth1/50

```
Total entries displayed: 1
SA-SPINE-A(TAH-elam-insel9)#
```

SA-SPINE-A sends traffic toward the SA-BGW-A according to the output.

ELAM on SA-BGW-A

```
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-insel9)# report
ELAM not triggered yet on slot - 1, asic - 0, slice - 0
HEAVENLY ELAM REPORT SUMMARY
slot - 1, asic - 0, slice - 1
============================

Incoming Interface: Eth1/50
Src Idx : 0xc5, Src BD : 1100
Outgoing Interface Info: dmod 1, dpid 48
Dst Idx : 0xbd, Dst BD : 1100

Packet Type: IPv4

Outer Dst IPv4 address: 192.168.100.1
Outer Src IPv4 address: 192.168.3.1
Ver     =  4, DSCP    =    0, Don't Fragment = 0
Proto   = 17, TTL     =  254, More Fragments = 0
Hdr len = 20, Pkt len =  134, Checksum       = 0x7e69

Inner Payload
Type: IPv4

Inner Dst IPv4 address: 10.100.20.10
Inner Src IPv4 address: 10.100.10.10

L4 Protocol  : 17
L4 info not available

Drop Info:
----------

LUA:
LUB:
LUC:
LUD:
Final Drops:

SA-BGW-A(TAH-elam-insel9)# show system internal ethpm info all | i i "dpid=48"
  IF_STATIC_INFO: port_name=Ethernet1/48,if_index:0x1a005e00,ltl=5956,slot=0, nxos_port=188,dmod=1,dpid=48,unit=0,queue=65535,xbar_unitbmp
SA-BGW-A(TAH-elam-insel9)# show cdp neighbors interface ethernet 1/48
Capability Codes: R - Router, T - Trans-Bridge, B - Source-Route-Bridge
                  S - Switch, H - Host, I - IGMP, r - Repeater,
                  V - VoIP-Phone, D - Remotely-Managed-Device,
                  s - Supports-STP-Dispute

Device-ID          Local Intrfce  Hldtme Capability  Platform      Port ID
```

```
SB-BGW-A(FDO2452070B)
            Eth1/48       122    R S s     N9K-C93216TC-FX2 Eth1/3

Total entries displayed: 1
SA-BGW-A(TAH-elam-insel9)#
```

According to output from SA-BGW-A, traffic was went out Ethernet1/48 toward SB-BGW-A. The next step is to check on 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-select 9, out-select 0
SB-BGW-A(TAH-elam-insel9)# set inner ipv4 src_ip 10.100.10.10 dst_ip 10.100.20.10
SB-BGW-A(TAH-elam-insel9)# start
SB-BGW-A(TAH-elam-insel9)# report
ELAM not triggered yet on slot - 1, asic - 0, slice - 0
ELAM not triggered yet on slot - 1, asic - 0, slice - 1
```

!---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 10.100.20.10
SB-BGW-A(TAH-elam-insel9)# start
SB-BGW-A(TAH-elam-insel9)# report
ELAM not triggered yet on slot - 1, asic - 0, slice - 0
ELAM not triggered yet on slot - 1, asic - 0, slice - 1
SB-BGW-A(TAH-elam-insel9)#
```

According to the output from SB-BGW-A, ELAM was not even triggered. This means that either SB-BGW-B is receiving the packets and not being able to correctly decrypt and parse them, or it not receives them at all. To understand what happened with the cloudsec traffic, you can run an ELAM on SB-BGW-A again, but trigger filter must be set to outer IP address which is used for cloudsec, as there is no way to see the inner header of cloudsec encrypted transit packet. From the previous output you know, that the SA-BGW-A handled the traffic, which means that SA-BGW-A encrypts traffic with cloudsec. So, you can use NVE IP of SA-BGW-A as a trigger filter for ELAM. From the previous outputs the VXLAN encrypted ICMP packet length is 134 byte. Plus 32 byte cloudsec header in summary gives you 166 byte:

<#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-A(TAH-elam-insel9)# report
ELAM not triggered yet on slot - 1, asic - 0, slice - 0
HEAVENLY ELAM REPORT SUMMARY
slot - 1, asic - 0, slice - 1
==============================

Incoming Interface: Eth1/3
Src Idx : 0x9, Src BD : 4108
Outgoing Interface Info: dmod 1, dpid 130
```

Dst Idx : 0xd, Dst BD : 4109

Packet Type: IPv4

Outer Dst IPv4 address:

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 Fragments = 0
Hdr len = 20, Pkt len =  166, Checksum       = 0xd546
```

!---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:
Final Drops:
```

!---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_name=Ethernet1/4,if_index:0x1a000600,ltl=6132,slot=0, nxos_port=12,dmod=1,dpid=1
```

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 - Switch, H - Host, I - IGMP, r - Repeater,
V - VoIP-Phone, D - Remotely-Managed-Device,
s - Supports-STP-Dispute

Device-ID	Local Intrfce	Hldtme	Capability	Platform	Port ID
SB-SPINE-A(FDO22302CJ0)					
	Eth1/4	131	R S s	N9K-C9236C	Eth1/34

Total entries displayed: 1
SB-BGW-A(TAH-elam-insel9)# show ip route 192.168.13.3
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

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 ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

192.168.14.2/32, ubest/mbest: 1/0, attached
```

***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 learnt on non-active FHRP router
       + - Adjacencies synced via CFSoE
       # - Adjacencies Throttled for Glean
       CP - Added via L2RIB, Control plane Adjacencies
       PS - Added via L2RIB, Peer Sync
       RO - Re-Originated Peer Sync Entry
       D - Static Adjacencies attached to down interface

IP ARP Table
Total number of entries: 1
Address         Age       MAC Address     Interface       Flags
192.168.14.2    00:00:13
```

ecce.1324.c803

Vlan3600

```
SB-BGW-A(TAH-elam-insel9)# show  mac address-table address ecce.1324.c803
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link,
        (T) - True, (F) - False, C - ControlPlane MAC, ~ - vsan,
        (NA)- Not Applicable
   VLAN     MAC Address      Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
G
```

3600

ecce.1324.c803

```
   static   -         F      F
```

vPC Peer-Link(R)

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

From this output, you can see, that the cloudsec traffic is forwarded toward the SB-BGW-B via the interface Ethernet1/4, based on the routing table. According to Cisco Nexus 9000 Series NX-OS VXLAN Configuration Guide, Release 10.3(x) guidlines and limitations:

- CloudSec traffic that is destined for the switch must enter the switch through the DCI uplinks.

According to vPC Border Gateway Support for Cloudsec section of the same guide, if vPC BGW learns peer vPC BGWs PIP address and advertises on DCI side, BGP path attributes from both vPC BGW going to be same. Hence the DCI intermediate nodes can end up choosing the path from vPC BGW which does not own the PIP address. In this scenario MCT link is used for encrypted traffic coming from the remote site. But in this case, interface toward the SPINE is used, despite that, the BGWs also have an OSPF adjacency via the BackUp SVI.

```
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          Up Time  Address         Interface
 192.168.12.4      1 FULL/ -        01:33:11 192.168.14.2    Vlan3600
 192.168.12.2      1 FULL/ -        01:33:12 192.168.11.5    Eth1/4
SB-BGW-A(TAH-elam-insel9)#
```

Reason of the issue and fix

The reason is the OSPF cost of the SVI interface. By default, on NXOS auto-cost reference bandwidth is 40G. SVI interfaces have bandwidth of 1Gbps, while the physical interface has a bandwidth of 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            Cost  State   Neighbors Status
 Vlan3600           3     0.0.0.0         40    P2P     1         up
```

<Output omitted>

```
 Eth1/4                   5       0.0.0.0          1       P2P       1          up
```

In such a case, the administrative change of the cost for SVI can resolve the issue. The tuning must be done on all border gateways.

<#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 "default"
```

'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

192.168.13.3/32, ubest/mbest: 1/0
    *

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, internal, tag 65002
SB-BGW-A(config-if)#

!---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 10.100.20.10: icmp_seq=1208 ttl=254 time=5.371 ms
64 bytes from 10.100.20.10: icmp_seq=1209 ttl=254 time=5.972 ms
64 bytes from 10.100.20.10: icmp_seq=1210 ttl=254 time=1.466 ms
64 bytes from 10.100.20.10: icmp_seq=1211 ttl=254 time=2.972 ms
64 bytes from 10.100.20.10: icmp_seq=1212 ttl=254 time=4.582 ms
64 bytes from 10.100.20.10: icmp_seq=1213 ttl=254 time=1.434 ms
64 bytes from 10.100.20.10: icmp_seq=1214 ttl=254 time=4.486 ms
64 bytes from 10.100.20.10: icmp_seq=1215 ttl=254 time=2.743 ms
64 bytes from 10.100.20.10: icmp_seq=1216 ttl=254 time=1.469 ms
64 bytes from 10.100.20.10: icmp_seq=1217 ttl=254 time=7.322 ms
64 bytes from 10.100.20.10: icmp_seq=1218 ttl=254 time=1.532 ms
64 bytes from 10.100.20.10: icmp_seq=1219 ttl=254 time=1.438 ms
64 bytes from 10.100.20.10: icmp_seq=1220 ttl=254 time=7.122 ms
64 bytes from 10.100.20.10: icmp_seq=1221 ttl=254 time=1.344 ms
64 bytes from 10.100.20.10: icmp_seq=1222 ttl=254 time=1.63 ms
64 bytes from 10.100.20.10: icmp_seq=1223 ttl=254 time=6.133 ms
64 bytes from 10.100.20.10: icmp_seq=1224 ttl=254 time=1.455 ms
64 bytes from 10.100.20.10: icmp_seq=1225 ttl=254 time=3.221 ms
64 bytes from 10.100.20.10: icmp_seq=1226 ttl=254 time=4.435 ms
64 bytes from 10.100.20.10: icmp_seq=1227 ttl=254 time=1.463 ms
64 bytes from 10.100.20.10: icmp_seq=1228 ttl=254 time=5.14 ms
64 bytes from 10.100.20.10: icmp_seq=1229 ttl=254 time=2.796 ms
64 bytes from 10.100.20.10: icmp_seq=1230 ttl=254 time=1.49 ms
64 bytes from 10.100.20.10: icmp_seq=1231 ttl=254 time=6.707 ms
64 bytes from 10.100.20.10: icmp_seq=1232 ttl=254 time=1.447 ms
64 bytes from 10.100.20.10: icmp_seq=1233 ttl=254 time=1.285 ms
64 bytes from 10.100.20.10: icmp_seq=1234 ttl=254 time=7.097 ms
64 bytes from 10.100.20.10: icmp_seq=1235 ttl=254 time=1.295 ms
64 bytes from 10.100.20.10: icmp_seq=1236 ttl=254 time=0.916 ms
64 bytes from 10.100.20.10: icmp_seq=1237 ttl=254 time=6.24 ms
64 bytes from 10.100.20.10: icmp_seq=1238 ttl=254 time=1.439 ms
64 bytes from 10.100.20.10: icmp_seq=1239 ttl=254 time=2.739 ms
64 bytes from 10.100.20.10: icmp_seq=1240 ttl=254 time=4.477 ms
64 bytes from 10.100.20.10: icmp_seq=1241 ttl=254 time=1.431 ms
64 bytes from 10.100.20.10: icmp_seq=1242 ttl=254 time=5.372 ms
64 bytes from 10.100.20.10: icmp_seq=1243 ttl=254 time=3.119 ms
64 bytes from 10.100.20.10: icmp_seq=1244 ttl=254 time=1.504 ms
64 bytes from 10.100.20.10: icmp_seq=1245 ttl=254 time=6.909 ms
64 bytes from 10.100.20.10: icmp_seq=1246 ttl=254 time=1.498 ms

64 bytes from 10.100.20.10: icmp_seq=1247 ttl=254 time=1.454 ms
64 bytes from 10.100.20.10: icmp_seq=1248 ttl=254 time=6.701 ms
64 bytes from 10.100.20.10: icmp_seq=1249 ttl=254 time=1.441 ms
64 bytes from 10.100.20.10: icmp_seq=1250 ttl=254 time=1.888 ms
64 bytes from 10.100.20.10: icmp_seq=1251 ttl=254 time=6.052 ms
64 bytes from 10.100.20.10: icmp_seq=1252 ttl=254 time=1.469 ms
64 bytes from 10.100.20.10: icmp_seq=1253 ttl=254 time=3.61 ms
64 bytes from 10.100.20.10: icmp_seq=1254 ttl=254 time=4.213 ms
64 bytes from 10.100.20.10: icmp_seq=1255 ttl=254 time=1.276 ms
64 bytes from 10.100.20.10: icmp_seq=1256 ttl=254 time=5.712 ms
64 bytes from 10.100.20.10: icmp_seq=1257 ttl=254 time=2.299 ms
64 bytes from 10.100.20.10: icmp_seq=1258 ttl=254 time=1.417 ms
64 bytes from 10.100.20.10: icmp_seq=1259 ttl=254 time=7.159 ms
64 bytes from 10.100.20.10: icmp_seq=1260 ttl=254 time=1.538 ms
64 bytes from 10.100.20.10: icmp_seq=1261 ttl=254 time=1.629 ms
64 bytes from 10.100.20.10: icmp_seq=1262 ttl=254 time=7.892 ms
64 bytes from 10.100.20.10: icmp_seq=1263 ttl=254 time=1.495 ms
64 bytes from 10.100.20.10: icmp_seq=1264 ttl=254 time=2.792 ms
^C
--- 10.100.20.10 ping statistics ---
1265 packets transmitted, 58 packets received, 95.42% packet loss
round-trip min/avg/max = 0.916/3.31/7.892 ms
SA-HOST-A#