
TURN
IT
UP
cisco Live!
#CiscoLive

The bridge to possible

Advanced Case Studies on Troubleshooting VXLAN BGP EVPN Multi-Site

Kallol Bosu, Technical Leader, CX
Manoj Kumar Shukla, Data Center BU Escalation Engineer

BRKDCN-3003

This is an advanced session. Goal of this session is to discuss a few advanced case studies on troubleshooting VXLAN BGP EVPN Multi-Site based scenarios (standalone NXOS), using real world examples.

Intended audience is network engineers and admins who are interested in deep dive troubleshooting.

Kallol Bosu

Technical Leader, CX

CISCO CERTIFIED CCIE SERVICE PROVIDER

CCIE# 62833

kbosu@cisco.com

Kallol is a Technical Leader in CX, with 7+ years of experience in Enterprise and Data Center Networking. Within Cisco, he specializes in Enterprise and Data Center Switching/Routing technologies across various platforms.

Since joining Cisco, he has been handling customer service requests and Escalations. He is also driving technology & platform specific trainings within CX and quite a few initiatives related to troubleshooting documentation with Engineering teams.

Kallol holds a Masters degree in Software and Telecommunication. He is also a CCIE# 62833 in Service Provider track.

Manoj Kumar Shukla

Data Center BU Escalation Engineer

CISCO CERTIFIED CCIE DATACENTER

#40911

manoshuk@cisco.com

Manoj Kumar Shukla, CCIE No. 40911, is a Senior Escalation Engineer in Data Center BU, with over 9 years of experience in Data Center Networking.

Within Cisco, Manoj specializes in Routing and Switching portfolio on Enterprise and Data Centre products. Since joining Cisco, he has been handling customer service requests, Escalations, design discussions with customers. He is also been instrumental in driving technology, platform specific trainings within CX , BU and to external customers.

Agenda

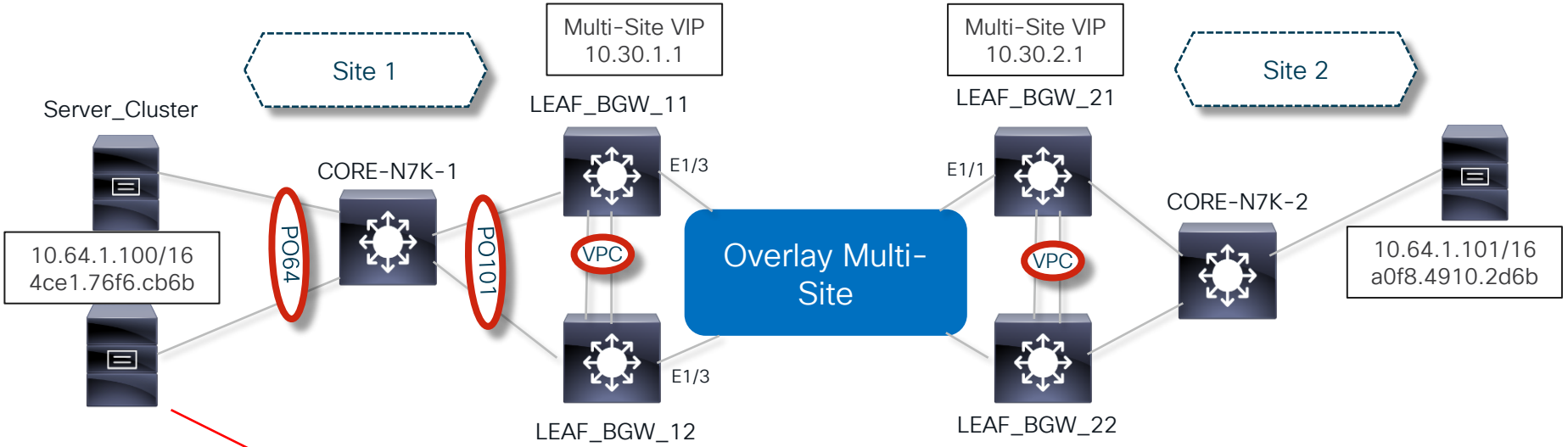
- Case Study #1
  - -Connectivity issue is seen, right after bringing up VXLAN Multi-site
  - -Solution and Take Away
- Case Study #2
  - -Traffic is black holed after DCI link fails on BGW in VXLAN Multi-site
  - -Solution and Take Away

Case Study #1

Problem Statement

Right after bringing up VXLAN BGP EVPN Multi-Site, few critical servers inside existing LAN network, started exhibiting 50-60% packet loss for all sort of communications to/from those.

Topology



Any communication to and from the server 10.64.1.100/16 is exhibiting 50-60% packet loss

```
Server-Cluster#ping 10.64.1.101
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.64.1.101, timeout is 2
seconds:
!.!..
Success rate is 40 percent (2/5), round-trip min/avg/max =
1/1/1 ms
```

Problem Isolation

It is not 100% packet loss here, which means both unicast and BUM traffic (used for ARP) are working for some time but getting disrupted/broken intermittently.

Check the ARP cache on both end hosts/servers (**arp -a**) for each other's IP address, to see if they are stable OR getting missed/refreshed in between.

In our example here, the ARP cache on both servers looked stable. Which means the issue is likely with unicast packet loss.

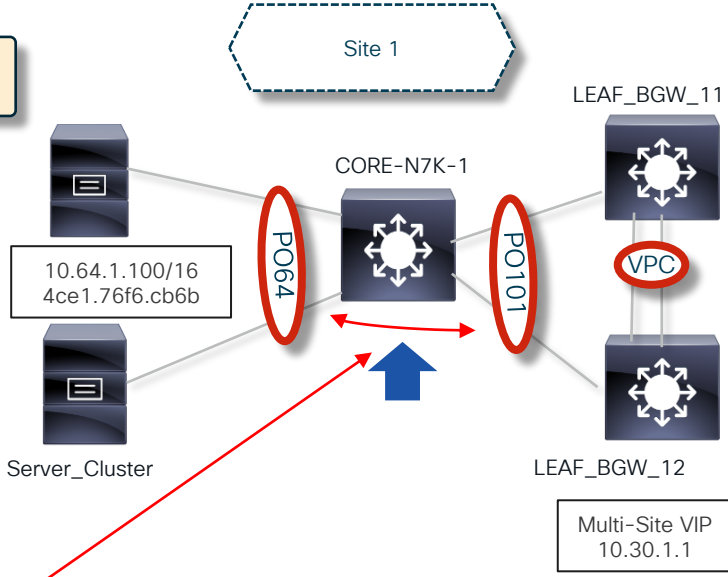
Please remember that SPAN is your best friend ☺. If feasible, try to setup SPAN capture on both end hosts/servers to understand the direction of packet loss.

In our example, SPAN capture from both ends indicated that, packet destined to Server_Cluster (10.64.1.100/4ce1.76f6.cb6b) is being lost/blackholed in DCI OR Fabric

Troubleshooting

Always try to isolate which switch in fabric/DCI is blackholing the traffic

```
CORE-N7K-1# sh mac address-table address 4ce1.76f6.cb6b
 VLAN/BD   MAC Address      Type      age     Secure NTFY Ports/SWID.SSID.LID
---------+----------------+--------+---------+------+----+------------------
•  64       4ce1.76f6.cb6b    dynamic     ~~~      F    F  Po64
!
CORE-N7K-1# sh mac address-table address 4ce1.76f6.cb6b
 VLAN/BD   MAC Address      Type      age     Secure NTFY Ports/SWID.SSID.LID
---------+----------------+--------+---------+------+----+------------------
* 64       4ce1.76f6.cb6b    dynamic     ~~~      F    F  Po101
!
CORE-N7K-1# sh mac address-table address 4ce1.76f6.cb6b
 VLAN/BD   MAC Address      Type      age     Secure NTFY Ports/SWID.SSID.LID
---------+----------------+--------+---------+------+----+------------------
* 64       4ce1.76f6.cb6b    dynamic     ~~~      F    F  Po64
```




```
CORE-N7K-1#
sh system internal l2fm event-history debugs | in cb6b
*snip*
[104] l2fm_macdb_entry_insert_snake_list(811): Moving MAC 4ce1.76f6.cb6b FROM (t:IF_MACDB_LIST_ENTRY
if:0x1600003f) to (t:IF_MACDB_LIST_ENTRY if:0x16000064)
```

Troubleshooting (Continued)

Site-1

MAC address is stable, verify the same from l2fm event-history as well
"sh system internal l2fm event-history debugs | in cb6b"

```
LEAF_BGW_11# sh mac address-table address 4ce1.76f6.cb6b
 V VLAN     MAC Address       Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------
*  64     4ce1.76f6.cb6b   dynamic  0         F      F    Po101
!
LEAF_BGW_11# sh mac address-table address 4ce1.76f6.cb6b
 V VLAN     MAC Address       Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------
*  64     4ce1.76f6.cb6b   dynamic  0         F      F    Po101
!
```

```
LEAF_BGW_11# sh bgp l2vpn evpn 4ce1.76f6.cb6b
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.30.10.1:32831    (L2VNI 30007)
BGP routing table entry for
[2]:[0]:[0]:[48]:[4ce1.76f6.cb6b]:[0]:[0.0.0.0]/216,
 version 127
Paths: (1 available, best #1)
Flags: (0x000102) (high32 00000000) on xmit-list, is not in l2rib/evpn

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    10.30.1.1 (metric 0) from 0.0.0.0 (10.30.10.1)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 30007
      Extcommunity: RT:65010:30007 SOO:10.30.1.1:0 ENCAP:8

  Path-id 1 advertised to peers:
    10.30.20.1          10.30.20.2
```

Multi-Site VIP
10.30.1.1

Verify the same on other member of VPC pair as well.
LEAF_BGW_12 in our example.

Potential Reasons?

Remote site is likely not advertising the same MAC through L2VNI/EVPN, since BGP EVPN table looks quite stable on site-1's LEAF_BGW_11 and LEAF_BGW_12

What if some packets sourced from server cluster at site 1, are being reflected by one of the local LEAF_BGW OR Remote LEAF_BGW in data plane?

Troubleshooting (Continued)

Site-1

```
CORE-N7K-1# ethanalyzer local interface inband capture-filter "host 10.64.1.100" limit-captured-frames 1 detail
Capturing on inband
**snip**
Ethernet II, Src: 4c:e1:76:f6:cb:6b (4c:e1:76:f6:cb:6b), Dst: IPv4mcast_00:00:12
 (01:00:5e:00:00:12)
    Destination: IPv4mcast_00:00:12 (01:00:5e:00:00:12)
        Address: IPv4mcast_00:00:12 (01:00:5e:00:00:12)
        .... ..0. .... .... .... .... = LG bit: Globally unique address (factory default)
        .... ...1 .... .... .... .... = IG bit: Group address (multicast/broadcast)
    Source: 4c:e1:76:f6:cb:6b (4c:e1:76:f6:cb:6b)
        Address: 4c:e1:76:f6:cb:6b (4c:e1:76:f6:cb:6b)
        .... ..0. .... .... .... .... = LG bit: Globally unique address (factory default)
        .... ...0 .... .... .... .... = IG bit: Individual address (unicast)
    Type: IP (0x0800)
Internet Protocol Version 4, Src: 10.64.1.100 (10.64.1.100), Dst: 224.0.0.18 (224.0.0.18)
    Version: 4
**snip**
```

Server_Cluster seems to be exchanging keep-alive internally using 224.0.0.18 every few msec, which is eventually getting reflected from remote site

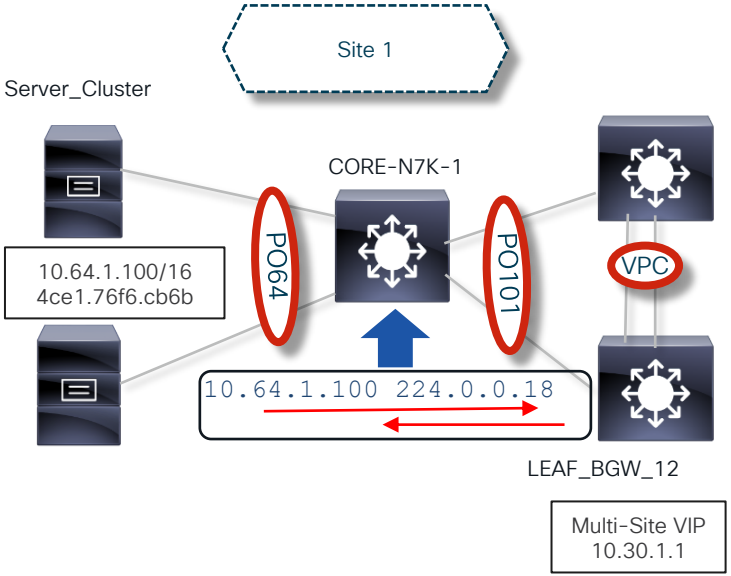
Idea here is to understand what packet is coming back to CORE-N7K-1, which is causing it move the MAC from PO64 to Po101/ LEAF_BGW_11 (N9Ks). There are several other options as well to get that information, for example ingress ELAM on N7K OR SPAN etc.

Troubleshooting (Continued)

Site-1

```
N7K-A# show port-channel summary interface po101
-----------------------------------------------
Group Port-        Type      Protocol  Member Ports
      Channel
-----------------------------------------------
101   Po101(SU)    Eth       LACP      Eth3/6(P)
```

```
CORE-N7K-1# show ip access-list FIND_PACKET
  statistics per-entry
  10 permit ip 10.64.1.100/32 224.0.0.18/32
  20 permit ip any any
!
interface port-channel101
ip port access-group FIND_PACKET in <<<
!

CORE-N7K-1#sh access-lists FIND_PACKET
IP access list FIND_PACKET
        statistics per-entry
        10 permit ip 10.64.1.100/32 224.0.0.18/32 [match=254]
         20 permit ip any any [match=255]
```



This confirms BUM traffic is reflected by LEAF_BGW_11 towards N7K, which is causing N7K to move the MAC back and forth between PO64 and PO101

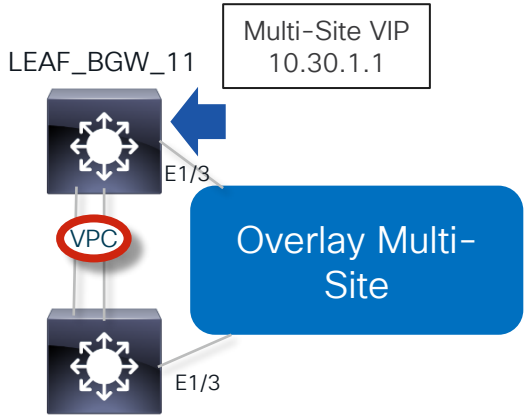
Troubleshooting (Continued)

Site-1

Let's take an ingress ELAM on LEAF_BGW_11 (N9K Tahoe based), to verify if it is indeed receiving the same packet back over DCI link, from remote LEAF_BGW_21.

```
N9K-1A# show hardware internal tah interface ethernet 1/3
#########################################
IfIndex: 0x1a000400
DstIndex: 6136
IfType: 26
Asic: 0
Asic: 0
AsicPort: 22
SrcId: 44
Slice: 0
PortOnSlice: 22
Table entries for interface Ethernet1/3
```

Alternatively, you may take SPAN on respective interface




```
N9K-1A# attach module 1
module-1# debug platform internal tah elam asic 0
module-1(TAH-elam)# trigger init in-select 9 use-src-id 44
Slot 1: param values: start asic 0, start slice 0, lu-a2d 1, in-select 9, out-select 0
module-1(TAH-elam-insel9)# reset
module-1(TAH-elam-insel9)# set inner ipv4 src_ip 10.64.1.100 dst_ip 224.0.0.18
module-1(TAH-elam-insel9)# start
module-1(TAH-elam-insel9)# report
```

Depending on packet flow, we might have needed to take the ELAM on other member of VPC pair as well

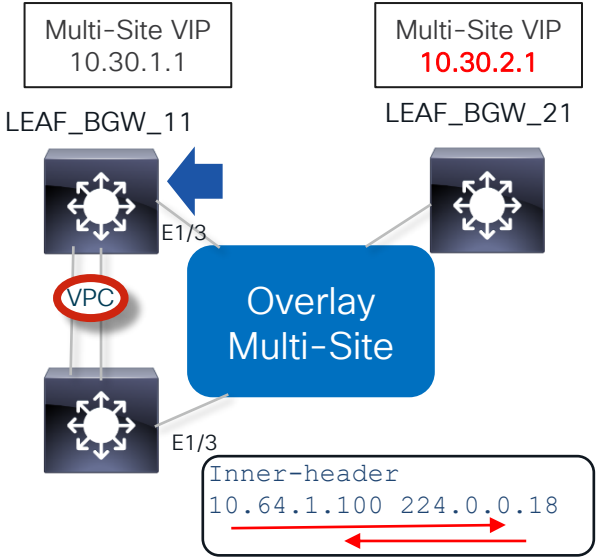
https://www.cisco.com/c/en/us/support/docs/switches/nexus-9000-series-switches/213848-nexus-9000-cloud-scale-asic-tahoe-nx-o.html

Troubleshooting (Continued)

```
Site-1, ingress ELAM done on LEAF_BGW_11
```

```
module-1(TAH-elam-insel9)# report
HOMEWOOD ELAM REPORT SUMMARY
slot - 1, asic - 0, slice - 0
============================

Incoming Interface: Eth1/3
Src Idx : 0x9, Src BD : 64
Outgoing Interface Info: met_ptr 0

Packet Type: IPv4

Outer Dst IPv4 address: 239.1.1.0
Outer Src IPv4 address: 10.30.2.1
Ver     =  4, DSCP    =    0, Don't Fragment = 0
Proto   = 17, TTL     =  253, More Fragments = 0
Hdr len = 20, Pkt len =  150, Checksum       = 0x6b8c

Inner Payload
Type: IPv4

Inner Dst IPv4 address: 224.0.0.18
Inner Src IPv4 address: 10.64.1.100
**snip**
```

Multicast is used for the replication of BUM traffic between sites. Is that correct?



Confirms that packet was reflected by remote BGW with it's own encapsulation, 10.30.2.1 is Multi-Site VIP of remote BGW

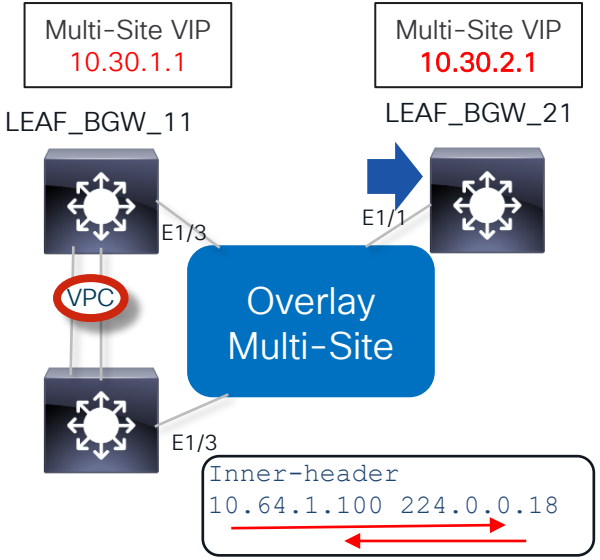
Troubleshooting (Continued)

```
Site-2, ingress ELAM done on LEAF_BGW_21
```

Alternatively, you may take SPAN. Goal here is to verify what is outer header of the overlay BUM traffic to 224.0.0.18

```
(TAH-elam-insel9)# set inner ipv4 src_ip 10.64.1.100 dst_ip 224.0.0.18
(TAH-elam-insel9)# start
(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: Eth1/1
Src Idx : 0x1, Src BD : 64
Outgoing Interface Info: met_ptr 0

Packet Type: IPv4

Outer Dst IPv4 address: 239.1.1.0
Outer Src IPv4 address: 10.30.1.1
**snip**

Inner Dst IPv4 address: 224.0.0.18
Inner Src IPv4 address: 10.64.1.100

L4 Protocol  : 17
L4 info not available
```

Confirms that LEAF_BGW_11 sent the BUM traffic with multicast encapsulation at first place.
10.30.1.1 is Multi-Site VIP of LEAF_BGW_11



Troubleshooting (Continued)

Snip from VXLAN multi-site white paper and guideline/restrictions

```
Note: Site-external BUM replication always uses ingress replication. Site-internal
BUM replication can use multicast (PIM ASM) or ingress replication.
```

```
Multicast Underlay between sites is not supported
```

Reference-

[Cisco Nexus 9000 Series NX-OS VXLAN Configuration Guide]

[VXLAN EVPN Multi-Site Design and Deployment White Paper]

Troubleshooting (Continued)

Why are the BGWs using multicast for replicating overlay-BUM traffic, if that should have used ingress-replication at first place?

```
configure profile VLAN64
  vlan 64
    vn-segment 30007
  interface nve1
    member vni 30007
      mcast-group 239.1.1.0
   evpn
    vni 30007 l2
      rd auto
      route-target import auto
      route-target export auto
!
interface Ethernet1/3
  description TRANSPORT-C9300-Gi1/0/1
  ip access-group TEST in
  ip address 10.30.254.129/30
  ip router isis UNDERLAY
  ip pim sparse-mode
  no shutdown
```

BGWs are missing the configuration of "multisite ingress-replication" under L2VNIs and DCI-tracking on DCI interface.

```
LEAF_BGW_11# show nve multisite dci-links
Interface       State
---------       -----
Ethernet1/3     Down <<<

LEAF_BGW_11# show nve interface nve 1 detail
Interface: nve1, State: Up, encapsulation: VXLAN
***snip**
10.30.100.2, admin: Up, oper: Down)
 Multisite bgw-if oper down reason:  DCI isolated
```

If all DCI-tracking interfaces are down, it converts the BGW to a traditional VTEP (the PIP address stays up).

Solution

Let's fix this

```
configure profile VLAN64
  vlan 64
    vn-segment 30007
  interface nve1
        multisite ingress-replication <<<
   **snip**
!
interface Ethernet1/3
evpn multisite dci-tracking <<<
```

```
LEAF_BGW_11# show nve multisite dci-links
Interface       State
---------       -----
Ethernet1/3     Up <<<

LEAF_BGW_11# show nve int nve 1 detail
Interface: nve1, State: Up, encapsulation: VXLAN
***snip**
10.30.100.2, admin: Up, oper: Up)
 Multisite bgw-if oper down reason:
```

Ingress ELAM on LEAF_BGW_21 confirms that ingress-replication is being used now, outer header is Unicast as opposed to Multicast as we have seen before

```
Incoming Interface: Eth1/1
Src Idx : 0x1, Src BD : 64
Outgoing Interface Info: dmod 0, dpid 12
Dst Idx : 0x0, Dst BD : 64
Packet Type: IPv4
Outer Dst IPv4 address: 10.30.2.1
Outer Src IPv4 address: 10.30.1.1
Ver = 4, DSCP = 0, Don't Fragment = 0
Proto = 17, TTL = 254, More Fragments = 0
Hdr len = 20, Pkt len = 90, Checksum = 0x4fab
Inner Payload
Type: IPv4
Inner Dst IPv4 address: 224.0.0.18
Inner Src IPv4 address: 10.64.6.4
**snip**
```

```
Server-Cluster#ping 10.64.1.101
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.64.1.101, timeout is 2
seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max =
1/1/1 ms
```

Case Study#1 What Did We Learn?

BGWs' L2VNIs must have the configuration "multisite ingress-replication", to enable ingress-replication for overlay BUM traffic, between the sites.

BGW's DCI interface must be configured with "evpn multisite dci-tracking" and fabric facing L3 interface should be configured with "evpn multisite fabric-tracking"

Additional Note (FYI)-
Make sure "ip igmp snooping vxlan" is enabled for Tennant Routed Multicast (TRM) L3 solution

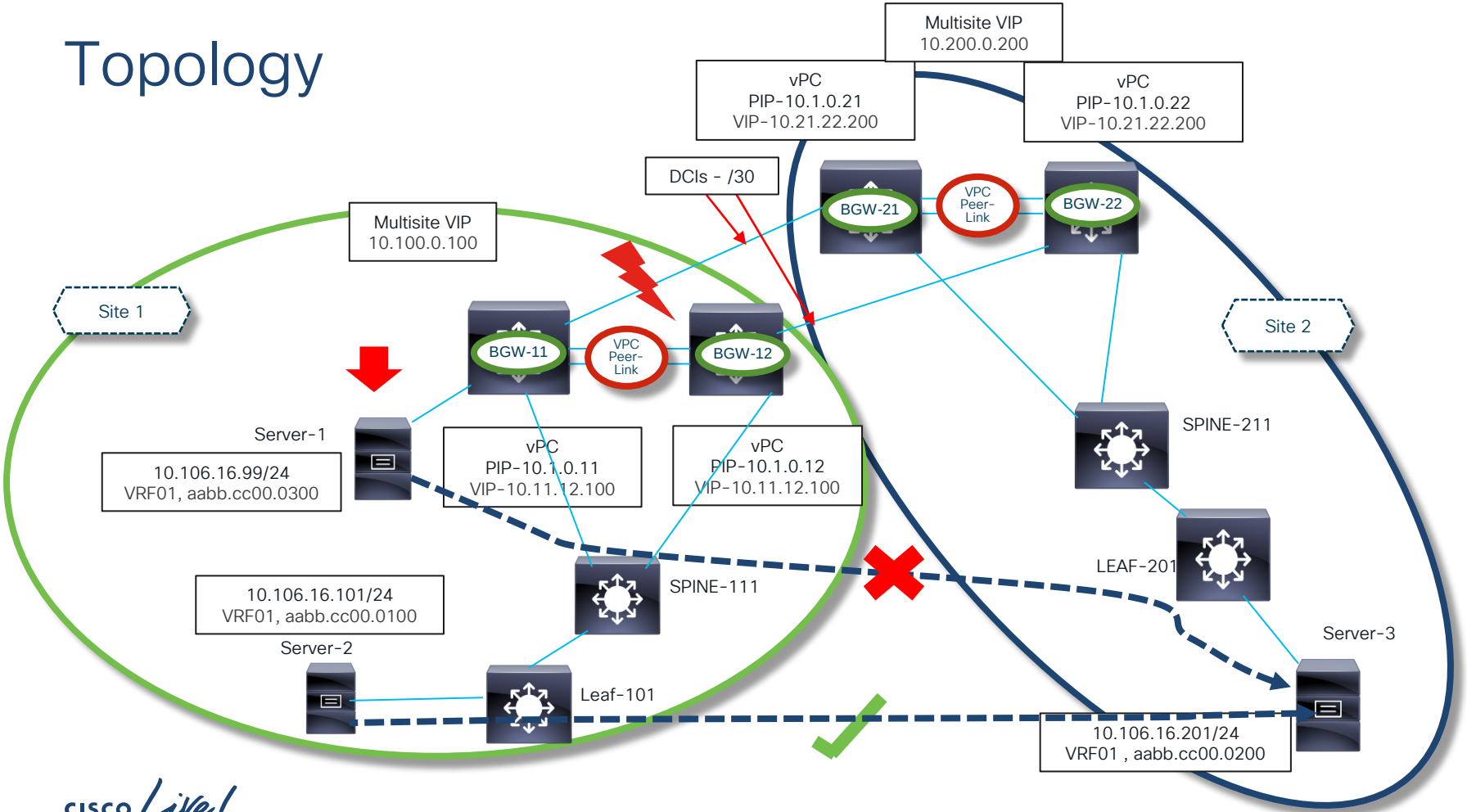
Case Study #2

Problem Statement

A VXLAN Multi-Site deployment with vPC border gateways (BGW) and back-to-back DCI connection between them. The vPC BGW is also a leaf.

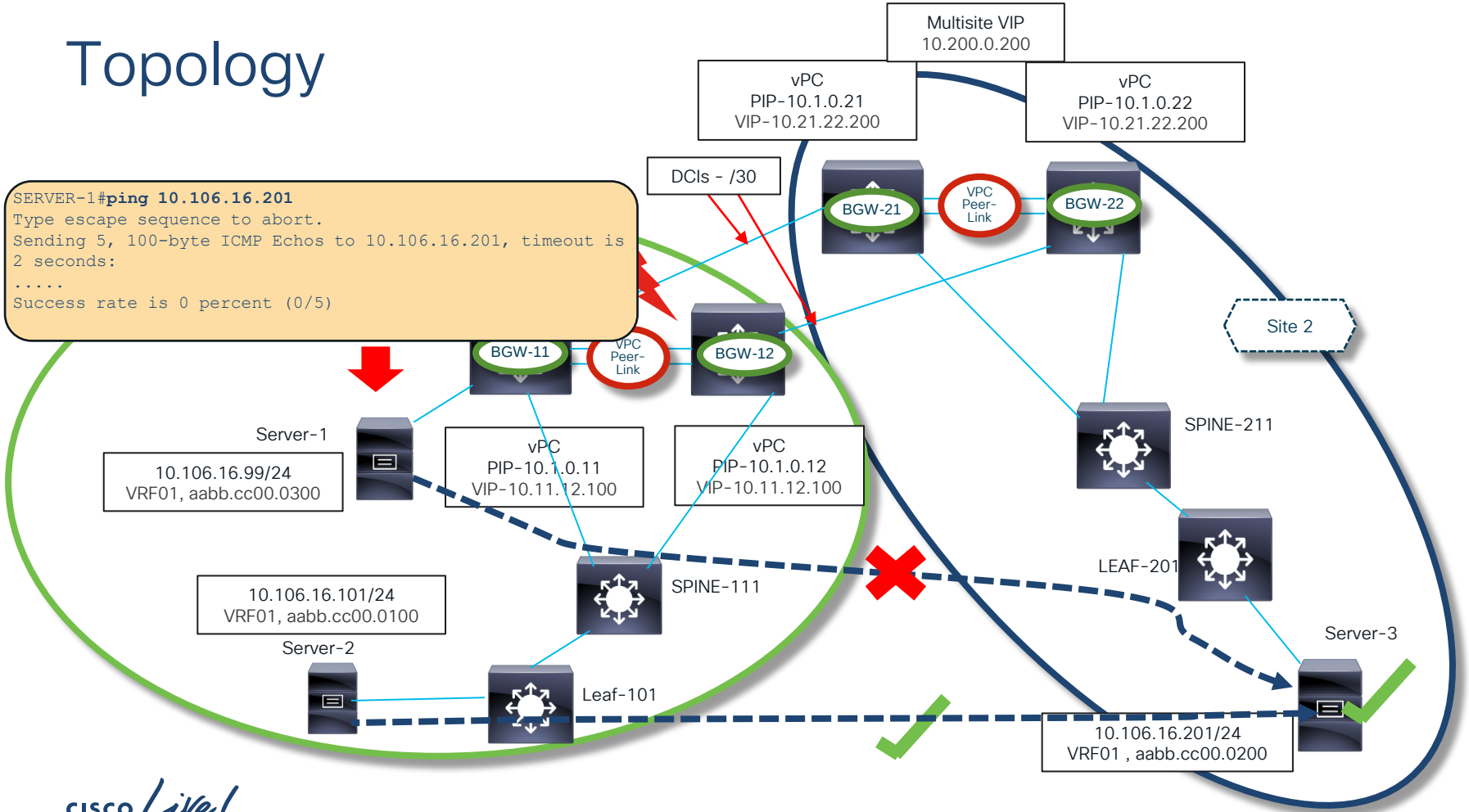
When one DCI (only one available on each BGWs) is shut or failed, then orphan systems locally attached to that failed-BGW leaf, can not communicate to hosts in another site.

Topology



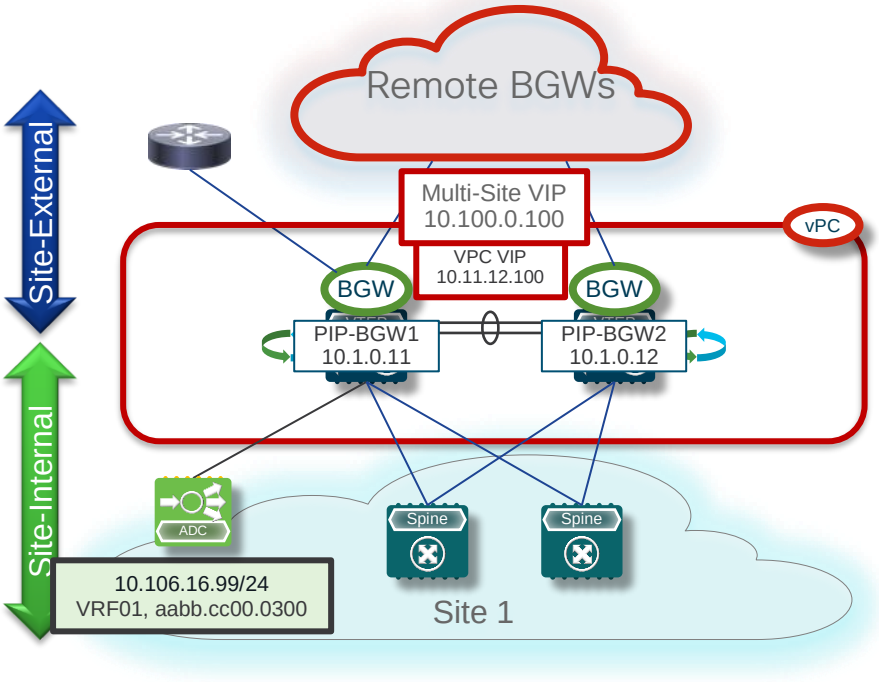
Topology

```
SERVER-1#ping 10.106.16.201
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.106.16.201, timeout is
2 seconds:
.....
Success rate is 0 percent (0/5)
```



Let's review some key concepts related to DCI failure scenarios on vPC Border Gateways

VXLAN Multi-Site
Failure Detection on vPC BGWs – DCI Isolation

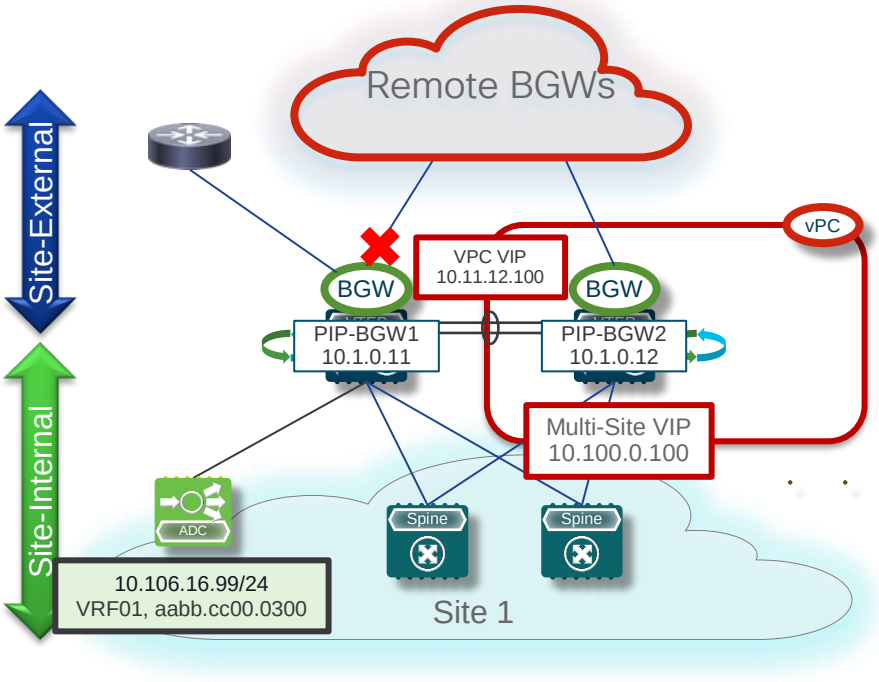


- The Site-External interfaces on BGW nodes are tracked to determine their status ('**evpn multisite dci-tracking**' command)

VXLAN Multi-Site

Failure Detection on vPC BGWs – DCI Isolation

NXOS Release 9.2(1)



- The Site-External interfaces on BGW nodes are also tracked to determine their status ('**evpn multisite dci-tracking**' command)
- If all the Site-External interfaces are detected as down:
  - The isolated BGW keeps advertising PIP/vPC VIP addresses toward the Site-External network (via vPC Peer-Link) and toward the Site-Internal network (for External Connectivity and Local Hosts)
  - The Multi-Site VIP is shut down on the isolated BGW, but it continues to advertise it toward the Site-Internal network (as it learns it via vPC Peer-Link from the peer BGW)

Before DCI was Shutdown

```
BGW-11#show bgp l2vpn evpn aabb.cc00.0200
**snip*
BGP routing table entry for
[2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272, version 605
Paths: (1 available, best #1)
Flags: (0x000212) (high32 0x000400) on xmit-list, is in l2rib/evpn, is not
in HW

  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, in
rib
              Imported from
2:102801:[2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272
  AS-Path: 65002 , path sourced external to AS
    10.200.0.200 (metric 0) from 10.0.0.21 (10.0.0.21)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 102801 901111
      Extcommunity: RT:102801:1 RT:901111:1 ENCAP:8 Router
MAC:0200.0ac8.00c8

  Path-id 1 (dual) advertised to peers:
    10.0.0.111
```

Next-hop is Multisite VIP of Remote BGWs

```
BGW-21# show nve interface nve 1 detail
Interface: nve1, State: Up, encapsulation: VXLAN
 VPC Capability: VPC-VIP-Only [notified]
 Local Router MAC: 0c1f.f800.1b08
 Host Learning Mode: Control-Plane
 Source-Interface: loopback1 (primary: 10.1.0.21,
secondary: 10.21.22.200)
 Source Interface State: Up
 Virtual RMAC Advertisement: Yes
 NVE Flags:
 Interface Handle: 0x49000001
 Source Interface hold-down-time: 180
 Source Interface hold-up-time: 30
 Remaining hold-down time: 0 seconds
 Virtual Router MAC: 0200.0a15.16c8
 Virtual Router MAC Re-origination: 0200.0ac8.00c8
 Interface state: nve-intf-add-complete
 Multisite delay-restore time: 30 seconds
 Multisite delay-restore time left: 0 seconds
 Multisite dci-advertise-pip configured: False
 Multisite bgw-if: loopback100 (ip: 10.200.0.200,
admin: Up, oper: Up)
 Multisite bgw-if oper down reason:
```

Before DCI was Shutdown

```
BGW-11# show nve peers
Interface Peer-IP                                       State LearnType Uptime   Router-Mac
--------- ---------------------------------------------- ----- --------- -------- -----------------
nve1      10.1.0.21                                      Up    CP        00:13:37 0c1f.f800.1b08
nve1      10.1.0.101                                     Up    CP        04:42:40 0c88.5400.1b08
nve1      10.200.0.200                                   Up    CP        00:13:37 0200.0ac8.00c8
nve1      10.21.22.200                                   Up    CP        00:13:37 0200.0a15.16c8
```

<< PC PIP of Remote BGW21
<< Leaf-101's Loopback1
<< Multi-site VIP of Remote site
<< VPC VIP of Remote BGWs

Type-2 EVPN (MAC and MAC-IP) route imported into respective L2/L3VNI

```
BGW-11# show bgp l2vpn evpn vni-id 102801 route-type 2
**snip**
BGP routing table entry for [2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272, version 605
Paths: (1 available, best #1)
Flags: (0x000212) (high32 0x000400) on xmit-list, is in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, in rib
             Imported from 2:102801:[2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272
  AS-Path: 65002 , path sourced external to AS
    10.200.0.200 (metric 0) from 10.0.0.21 (10.0.0.21)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 102801 901111
      Extcommunity: RT:102801:1 RT:901111:1 ENCAP:8 Router MAC:0200.0ac8.00c8

  Path-id 1 (dual) advertised to peers:
    10.0.0.111
**snip**
```

Before DCI was Shutdown

MAC and MAC-IP route present into the L2RIB as a remote/BGP entry

Vlan-ID 2801 is mapped to L2VNI 102801 and L3VNI 901111

```
BGW-11# show l2route evpn mac evi 2801
**snip*
Topology    Mac Address    Prod   Flags          Seq No     Next-Hops
----------- -------------- ------ -------------- ---------- --------------------------------
2801        0cc0.3900.1b08 VXLAN  Stt,Nho,       0          10.11.12.100
2801        aabb.cc00.0100 BGP    SplRcv         0          10.1.0.101 (Label: 102801)
2801        aabb.cc00.0200 BGP    SplRcv         0          10.200.0.200 (Label: 102801)
2801        aabb.cc00.0300 Local  L,             0          Eth1/2
```

MAC address present in the L2FM

```
BGW-11# show mac address-table address aabb.cc00.0200
* - 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
   VLAN     MAC Address       Type       age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
C 2801     aabb.cc00.0200   dynamic  0         F      F    nve1(10.200.0.200)
```

MAC-IP route imported into L3RIB of the tenant VRF

```
BGW-11# show ip route 10.106.16.201/32 vrf VRF01
IP Route Table for VRF "VRF01"
**snip**
10.106.16.201/32, ubest/mbest: 1/0
    *via 10.200.0.200%default, [20/2000], 00:36:07, bgp-65001, external, tag
65002, segid: 901111 tunnelid: 0xac800c8 encap: VXLAN
```

After DCI on BGW11 is Shutdown

```
BGW-11#show bgp l2vpn evpn aabb.cc00.0200
**snip*

  Path-id 1 (dual) not advertised to any peer
BGP routing table entry for
[2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272, version 693
Paths: (1 available, best #1)
Flags: (0x000212) (high32 0x000400) on xmit-list, is in l2rib/evpn, is not
in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop, in
rib
             Imported from
1:102801:[2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272
  AS-Path: 65002 , path sourced external to AS
    10.100.0.100 (metric 41) from 10.0.0.111 (10.0.0.111)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 102801 901111
      Extcommunity: RT:102801:1 RT:901111:1 ENCAP:8 Router
MAC:0200.0a64.0064
      Originator: 10.0.0.12 Cluster list: 10.0.0.111
**snip*
```

Next-hop is Multisite VIP of Local BGWs

Learned from SPINE-111 (Route-Reflector) of local site, iBGP

```
BGW-11# show nve interface nve 1 detail
Interface: nve1, State: Up, encapsulation: VXLAN
 VPC Capability: VPC-VIP-Only [notified]
 Local Router MAC: 0cc0.3900.1b08
 **snip**
 Multisite dci-advertise-pip configured: False
 Multisite bgw-if: loopback100 (ip: 10.100.0.100,
admin: Up, oper: Down)
 Multisite bgw-if oper down reason:  DCI isolated.
```

```
BGW-11# show nve multisite dci-links
Interface       State
---------       -----
Ethernet1/10    Down   <<<
```

```
BGW-11# show interface lo100
loopback100 is down (Administratively down)
admin state is up,
```

After DCI on BGW11 is Shutdown

Multisite NVE peering's are removed from BGW-11

```
BGW-11# show nve peers
Interface Peer-IP                                State LearnType Uptime   Router-Mac
--------- --------------------------------------  ----- --------- -------- -----------------
nve1      10.1.0.101                             Up    CP        05:23:42 0c88.5400.1b08
```

```
BGW-11# show bgp l2vpn evpn vni-id 102801 route-type 2
**snip**
BGP routing table entry for [2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272, version 693
Paths: (1 available, best #1)
Flags: (0x000212) (high32 0x000400) on xmit-list, is in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop, in rib
             Imported from 1:102801:[2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272
  AS-Path: 65002 , path sourced external to AS
    10.100.0.100 (metric 41) from 10.0.0.111 (10.0.0.111)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 102801 901111
      Extcommunity: RT:102801:1 RT:901111:1 ENCAP:8 Router MAC:0200.0a64.0064
      Originator: 10.0.0.12 Cluster list: 10.0.0.111

  Path-id 1 (dual) not advertised to any peer **snip**
```

After DCI on BGW11 is Shutdown

Vlan-ID 2801 is mapped to L2VNI 102801 and L3VNI 901111

```
BGW-11# show l2route evpn mac evi 2801
**snip*
Topology     Mac Address    Prod   Flags          Seq No     Next-Hops
----------- -------------- ------ ------------- ---------- --------------------------------
-------
2801         0cc0.3900.1b08 VXLAN  Stt,Nho,       0          10.11.12.100
2801         aabb.cc00.0100 BGP    SplRcv         0          10.1.0.101 (Label: 102801)
2801         aabb.cc00.0200 BGP    SplRcv         0          10.100.0.100 (Label: 102801)
2801         aabb.cc00.0300 Local  L,             0          Eth1/2
BGW-11#
```

```
BGW-11# show mac address-table address aabb.cc00.0200
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
   VLAN     MAC Address      Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
BGW-11#
```

Traffic being blackholed. No entries in L2FM and L3RIB

```
BGW-11# show ip route 10.106.16.201/32 vrf VRF01
IP Route Table for VRF "VRF01"

Route not found
BGW-11#
```

Potential Solution

Establish Backdoor iBGP connection between local BGWs

Consider advertising DCI link prefix into underlay BGP IPv4 AFI (Or Alternative)

Verify reachability of loopback IPs of remote BGWs from failed BGW-11, through VPC peer-link/ backdoor L3 connection

Establish L2VPN EVPN peering (fabric-external) with other remote BGW (BGW-22) , leveraging the VPC peer-link /iBGP connection between local BGWs

Potential Solution- Configuration Example

Snip from major configuration added on failed BGW-11. Relevant configuration needed to added on other BGWs too

```
interface Vlan1152
  description BACKDOOR_SVI_PEER_LINK
  no shutdown
  mtu 9216
  no ip redirects
  ip address 100.11.12.11/24
  no ipv6 redirects
  ip ospf network point-to-point
  no ip ospf passive-interface
  ip router ospf UNDERLAY area 0.0.0.1
!
```

```
router bgp 65001
neighbor 10.0.0.22
    remote-as 65002
    update-source loopback0
    ebgp-multihop 5
    peer-type fabric-external
    address-family l2vpn evpn
      send-community
      send-community extended
```

```
router bgp 65001
neighbor 100.11.12.12
    remote-as 65001
    description INFRA_PEERING
    address-family ipv4 unicast
      send-community
      send-community extended
```

This internal peering should help getting IP connectivity to loopback of other remote BGW and we will leverage that to establish l2vpn peering as shown at right side

```
BGW-11# show ip route 10.0.0.22
*snip**
10.0.0.22/32, ubest/mbest: 1/0
    *via 172.12.22.22, [200/0], 03:25:08, bgp-65001,
internal, tag 65002
BGW-11#
!
172.12.22.0/24, ubest/mbest: 1/0
    *via 100.11.12.12, [200/0], 03:25:53, bgp-65001,
internal, tag 65001
BGW-11#
```

How does it look after making the changes?

```
BGW-11#show bgp l2vpn evpn aabb.cc00.0200
**snip*

 BGP routing table entry for
[2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272, version 731
Paths: (1 available, best #1)
Flags: (0x000212) (high32 0x000400) on xmit-list, is in l2rib/evpn, is not
in HW

  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, in
rib
              Imported from
2:102801:[2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272
  AS-Path: 65002 , path sourced external to AS
    10.200.0.200 (metric 0) from 10.0.0.22 (10.0.0.22)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 102801 901111
      Extcommunity: RT:102801:1 RT:901111:1 ENCAP:8 Router
MAC:0200.0ac8.00c8

  Path-id 1 (dual) advertised to peers:
    10.0.0.111 **snip*
```

Next-hop is Multisite VIP of Remote BGWs

Learned from BGW-22 of remote site

```
BGW-11# show nve interface nve 1 detail
Interface: nve1, State: Up, encapsulation: VXLAN
 VPC Capability: VPC-VIP-Only [notified]
 Local Router MAC: 0cc0.3900.1b08
 **snip**
 Multisite dci-advertise-pip configured: False
 Multisite bgw-if: loopback100 (ip: 10.100.0.100,
admin: Up, oper: Down)
 Multisite bgw-if oper down reason:  DCI isolated.
```

```
BGW-11# show nve multisite dci-links
Interface       State
---------       -----
Ethernet1/10    Down   <<<
```

```
BGW-11# show interface lo100
loopback100 is down (Administratively down)
admin state is up,
```

Verification of Working State

Multisite NVE peering's are back on BGW-11

```
BGW-11# show nve peers
Interface Peer-IP                                State LearnType Uptime   Router-Mac
--------- ---------------------------------------  ----- --------- -------- -----------------
nve1      10.1.0.22                              Up    CP        00:16:05 0cd9.6700.1b08
nve1      10.1.0.101                             Up    CP        06:04:00 0c88.5400.1b08
nve1      10.200.0.200                           Up    CP        00:16:05 0200.0ac8.00c8
nve1      10.21.22.200                           Up    CP        00:16:05 0200.0a15.16c8
```

```
BGW-11# show bgp l2vpn evpn vni-id 102801 route-type 2
**snip**
BGP routing table entry for [2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272, version 731
Paths: (1 available, best #1)
Flags: (0x000212) (high32 0x000400) on xmit-list, is in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, in rib
             Imported from 2:102801:[2]:[0]:[0]:[48]:[aabb.cc00.0200]:[32]:[10.106.16.201]/272
  AS-Path: 65002 , path sourced external to AS
    10.200.0.200 (metric 0) from 10.0.0.22 (10.0.0.22)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 102801 901111
      Extcommunity: RT:102801:1 RT:901111:1 ENCAP:8 Router MAC:0200.0ac8.00c8

  Path-id 1 (dual) advertised to peers:
    10.0.0.111
**snip**
```

Verification of Working State

Vlan-ID 2801 is mapped to L2VNI 102801 and L3VNI 901111

```
BGW-11# show l2route evpn mac evi 2801
**snip*
Topology    Mac Address    Prod   Flags         Seq No     Next-Hops
----------- -------------- ------ ------------- ---------- -------------------------------
-------
2801        0cc0.3900.1b08 VXLAN  Stt,Nho,      0          10.11.12.100
2801        aabb.cc00.0100 BGP    SplRcv        0          10.1.0.101 (Label: 102801)
2801        aabb.cc00.0200 BGP    SplRcv        0          10.200.0.200 (Label: 102801)
2801        aabb.cc00.0300 Local  L,            0          Eth1/2
BGW-11#
```

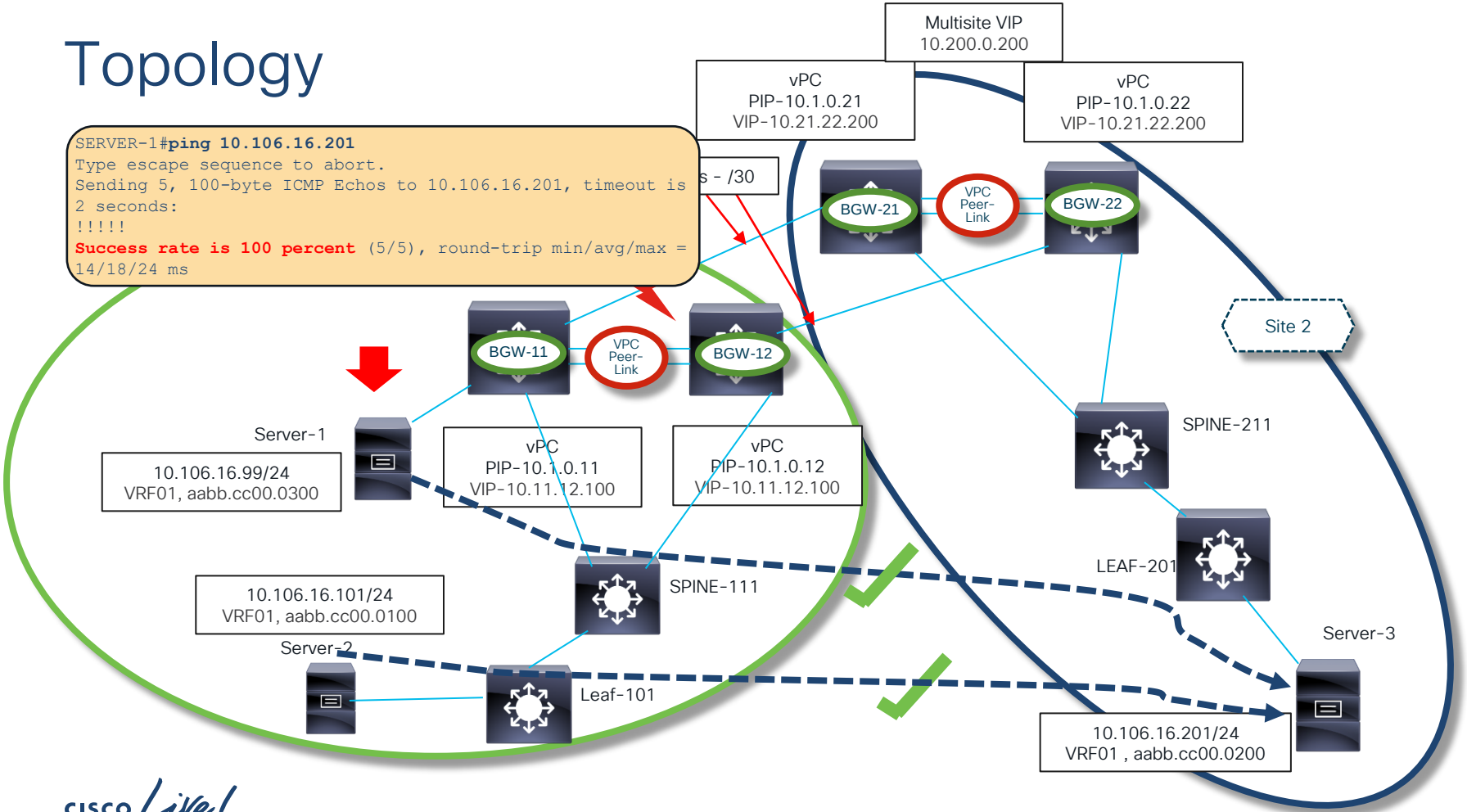
```
BGW-11# show mac address-table address aabb.cc00.0200
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link,
        (T) - True, (F) - False, C - ControlPlane MAC, ~ - vsan
   VLAN     MAC Address      Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
C 2801     aabb.cc00.0200   dynamic  0         F      F    nve1(10.200.0.200)
BGW-11#
```

Looks Better ☺

```
BGW-11# show ip route 10.106.16.201/32 vrf VRF01
IP Route Table for VRF "VRF01"

10.106.16.201/32, ubest/mbest: 1/0
    *via 10.200.0.200%default, [20/2000], 00:18:31, bgp-65001, external, tag
65002, segid: 901111 tunnelid: 0xac800c8 encap: VXLAN
```

Topology

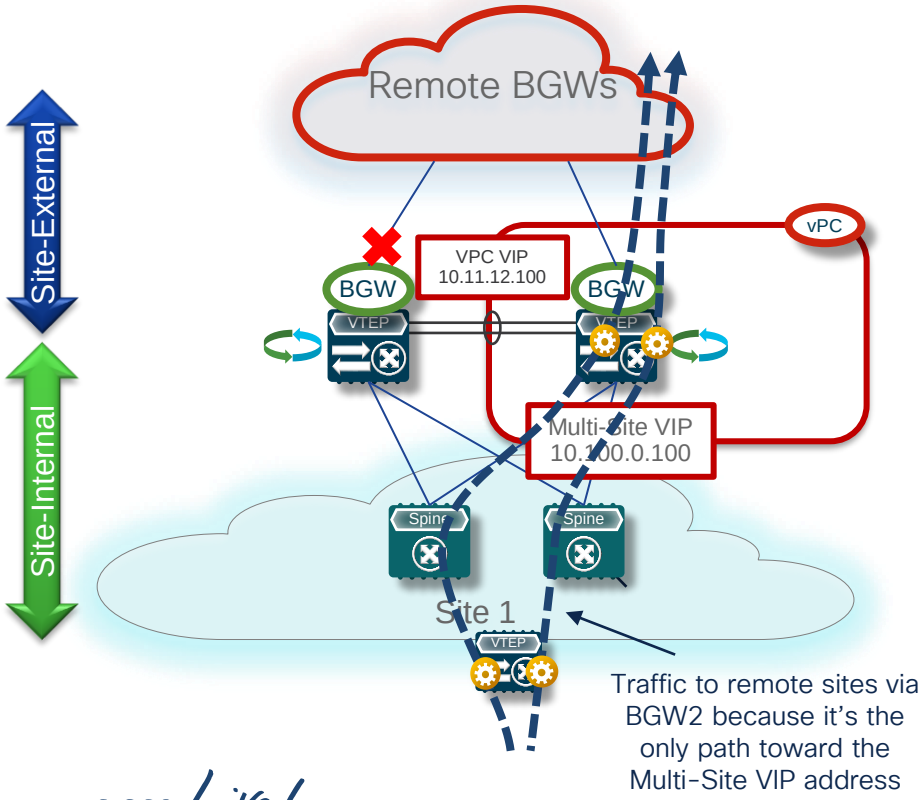
```
SERVER-1#ping 10.106.16.201
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.106.16.201, timeout is
2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max =
14/18/24 ms
```



Recap

How was the traffic working earlier from hosts behind other VTEPs in site-1 but not from locally connected orphan hosts on BGW-11 leaf ?

Failure Detection on vPC BGWs
DCI Isolation – Outbound Traffic toward Remote Sites



- Outbound traffic toward remote sites originated by endpoints connected to local leaf nodes is always encapsulated to the local Multi-Site VIP address
- BGW2 receives all the outbound flows, decapsulates them and re-encapsulates to the Multi-Site VIP address of the destination site
  - BGW2 still advertises the Multi-Site VIP address prefix received by the peer BGW into the Site-Internal network, but it is seen as less preferable by the local leaf nodes

CLI Cheat Sheet - Examples

Local VTEP

- MAC Address present in L2FM the Local VTEP
  sh mac address-table address ac7a.565a.9999
  sh system internal l2fm l2dbg macdb address ac7a.565a.9999 vlan 2801
  sh sys inter l2fm event-hist deb | in ac7a.565a.9999

- ARP present and forwarding adjacency present
  sh ip arp vrf VRF01
  sh forwarding vrf VRF01 adjacency

- HMM route present
  sh ip route 10.106.16.201 hmm vrf VRF01

- MAC and MAC-IP route present in L2RIB
  sh l2route evpn mac evi 2801
  sh l2route evpn mac-ip evi 2801
  sh system internal l2rib event-history mac
  sh system internal l2rib event-history mac-ip

- MAC and MAC-IP route exported into BGP in correct L2VNI
  sh bgp l2vpn evpn vni-id 102801

- Type-2 EVPN route (MAC and MAC-IP) advertised towards the remote VTEP's
  sh bgp l2vpn evpn ac7a.565a.9999
  sh bgp internal event-history event | ac7a.565a.9999

Remote VTEP

- Type-2 EVPN (MAC and MAC-IP) route received from the nve peer
  sh bgp l2vpn evpn ac7a.565a.9999
  sh bgp internal event-history event | ac7a.565a.9999

- Type-2 EVPN (MAC and MAC-IP) route imported into correct L2VNI and L3VNI
  sh bgp l2vpn evpn vni-id 102801
  sh bgp l2vpn evpn vni-id 901111

- MAC and MAC-IP route present into the L2RIB as a remote/BGP entry
  sh l2route evpn mac evi 2801
  sh l2route evpn mac-ip evi 2801
  sh system internal l2rib event-history mac
  sh system internal l2rib event-history mac-ip

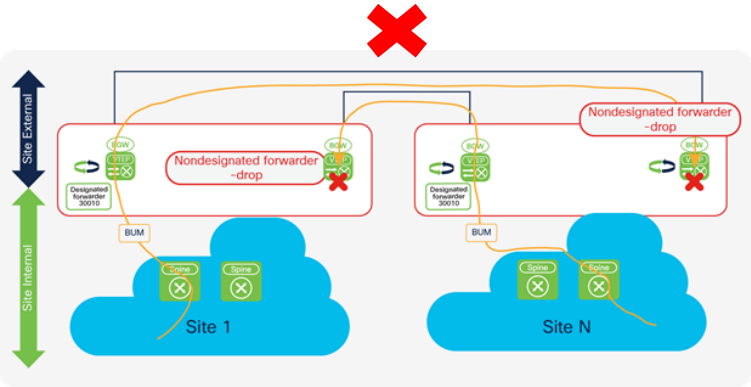
- MAC address present in the L2FM with N.H. nve peer
  sh mac address-table address ac7a.565a.9999
  sh system internal l2fm l2dbg macdb address ac7a.565a.
  9999 vlan 2801
  sh sys inter l2fm event-hist deb | in ac7a.565a.9999

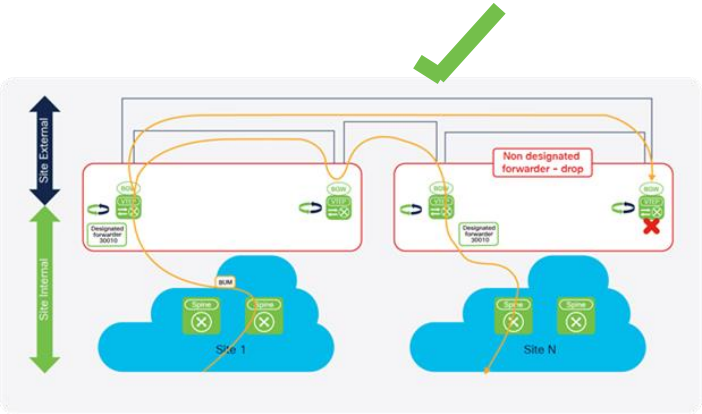
- MAC-IP route imported into L3RIB of the tenant VRF with N.H. nve peer
  sh ip route 10.106.16.201 vrf VRF01

Solution and Take Away

Consider resilient design for different failover scenarios

Some designs can impact BUM traffic badly, apart from regular unicast failure scenarios





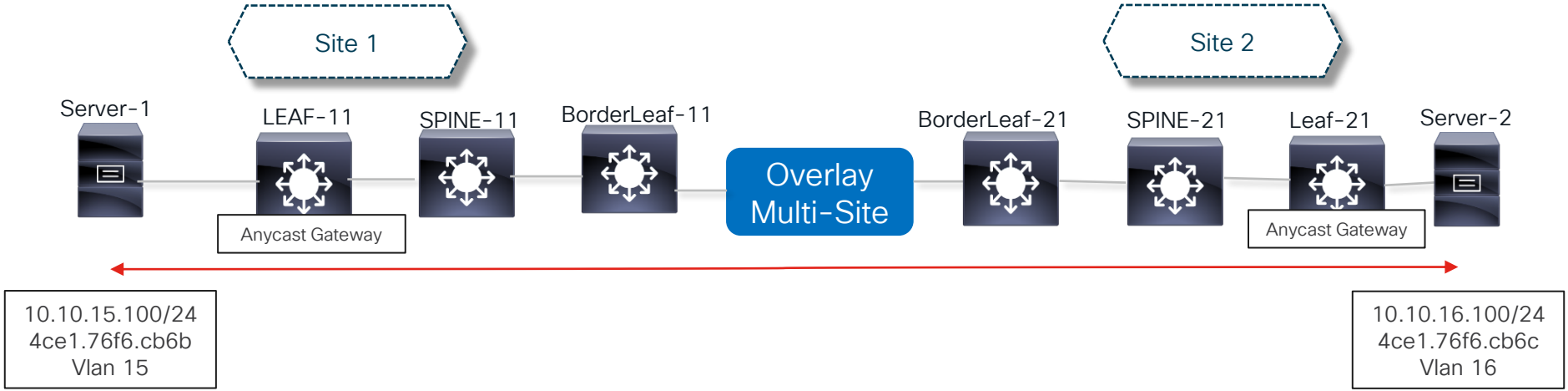
Case Study #3

Problem Statement

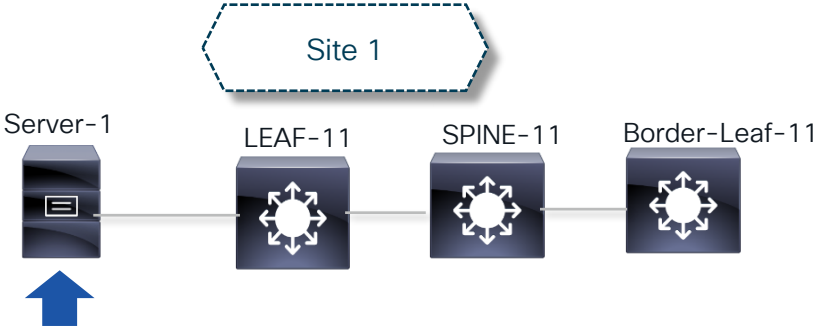
Delay sensitive applications ,those are communicating in inter-vlan between the two data centers (VXLAN BGP EVPN Multi-Site), are experiencing intermittent performance/slowness issue.

There is no problem seen for the intra-vlan communication between the sites.

Simplified Topology



Troubleshooting



A TCP dump on the server indicated that, destination MAC of outbound packets (**Inter-VLAN**) keep switching between two MAC addresses.

Further analysis revealed, that behavior is due to some random (but legit) ARP requests coming from the Gateway's IP but with two different source MAC

Troubleshooting (Continued)

Why are we receiving ARP packet from Gateway's IP with two different source MAC addresses ?


```
LEAF-11# sh run | in gateway
fabric forwarding anycast-gateway-mac 2020.0000.00aa
```

```
LEAF-21# sh run | in gateway
fabric forwarding anycast-gateway-mac 2020.0000.00bb
```

Solution and Take Away

Configure same anycast gateway-mac on all sites to avoid suboptimal traffic forwarding across sites.

We applied following change on site-2 in our use case and that fixed the issue

```
fabric forwarding anycast-gateway-mac 2020.0000.00aa
```

Alternate solution could have been, enabling ARP suppression for all VLANs that have been extended.

Reference-
Nexus 9000 NX-OS VXLAN Configuration Guide 9.3(x)

VXLAN Multi-Site Characteristics

- **Multiple** Overlay Domains – Interconnected and Controlled
- **Multiple** Overlay Control-Plane Domains – Interconnected and Controlled
- **Multiple** Underlay Domains - Isolated
- **Multiple** Replication Domains for BUM – Interconnected and Controlled
- **Multiple** VNI Administrative Domains

Underlay Isolation – Overlay Hierarchies

More details on VXLAN BGP EVPN Multi-Site

VXLAN BGP EVPN Based Multi-Site
DGTL-BRKDCN-2035 by Lukas Krattiger

Building DC Networks with VXLAN EVPN Overlays -VXLAN EVPN Multi-Site - BRKDCN-3378 by Lukas Krattiger

Troubleshooting VxLAN BGP EVPN
BRKDCN-3040 by Vinit Jain

Thank you


TURN
IT
UP
cisco Live!
#CiscoLive