VXLAN ARP/ND Suppression

Speakers:

- Daryl Wan
- Ankit Kumar Sinha

a Hewlett Packard Enterprise company

Agenda: VXLAN ARP/ND Suppression

1 Overview

2 Use Cases

3 Details/Caveats

4 Configuration and Best Practices

5 Troubleshooting

6 Resources

7 Demo

Overview

aruba
a Hewlett Packard
Enterprise company

VXLAN ARP/ND Suppression Overview

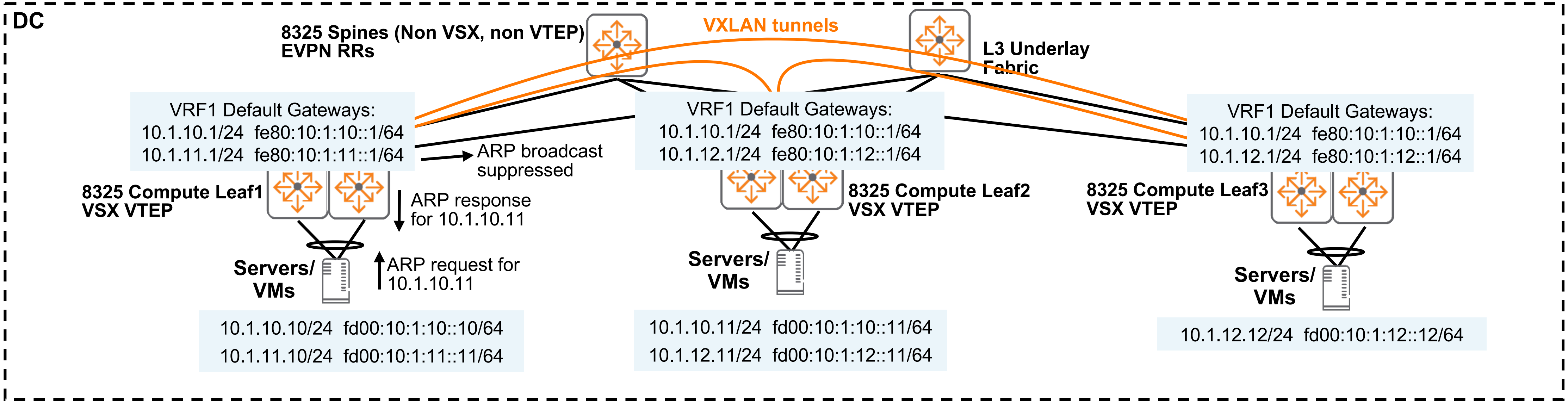
- 10.5 did not support ARP/ND Suppression in the VXLAN EVPN overlay network
- 10.6 adds support for ARP/ND Suppression in the VXLAN EVPN overlay network
  - On 6300/6400/8325/8360/8400

- This feature reduces the flooding of host IPv4 ARP and IPv6 Neighbor Discovery (ND) messages in the VXLAN overlay network, which results in a more efficient use of the underlay network

Use Cases

aruba
a Hewlett Packard
Enterprise company

VXLAN ARP/ND Suppression Use Case


DC
8325 Spines (Non VSX, non VTEP)
EVPN RRs
VXLAN tunnels
L3 Underlay
Fabric
VRF1 Default Gateways:
10.1.10.1/24 fe80:10:1:10::1/64
10.1.11.1/24 fe80:10:1:11::1/64
ARP broadcast
suppressed
VRF1 Default Gateways:
10.1.10.1/24 fe80:10:1:10::1/64
10.1.12.1/24 fe80:10:1:12::1/64
VRF1 Default Gateways:
10.1.10.1/24 fe80:10:1:10::1/64
10.1.12.1/24 fe80:10:1:12::1/64
8325 Compute Leaf1
VSX VTEP
ARP response
for 10.1.10.11
8325 Compute Leaf2
VSX VTEP
8325 Compute Leaf3
VSX VTEP
Servers/
VMs
ARP request for
10.1.10.11
Servers/
VMs
Servers/
VMs
10.1.10.10/24 fd00:10:1:10::10/64
10.1.11.10/24 fd00:10:1:11::11/64
10.1.10.11/24 fd00:10:1:10::11/64
10.1.12.11/24 fd00:10:1:12::11/64
10.1.12.12/24 fd00:10:1:12::12/64

- In a VXLAN EVPN distributed L3 gateway deployment, a host MAC/IP is normally discovered by the connected source VTEP when the host comes online or sends traffic, that MAC/IP info is shared with other VTEPs via MP-BGP, e.g. 10.10.10.11 on Leaf2
- That host MAC/IP entry is only removed when the source VTEP no longer sees the MAC (it does not timeout)

- If 10.1.10.10 connected to Leaf1 tries to ARP for 10.1.10.11 on Leaf2
- Leaf1 will suppress ARP from being flooded out to other VTEPs and provide ARP response on behalf of 10.1.10.11, as it has 10.1.10.11 MAC/IP info from Leaf2

- Similar benefits will be seen for ND Solicitation requests for devices on the same VLAN/subnet

Details

aruba

a Hewlett Packard
Enterprise company

VXLAN ARP/ND Suppression Details

- Before 10.6, only MAC(s) were advertised by EVPN Type 2 (MAC/IP Advertisement route)
- Starting from 10.6, locally learned neighbor IP(v4/v6) addresses are also advertised along with MAC by EVPN Type 2 to remote VTEP(s)
- Advertised neighbors are updated in Neighbor table of received VTEP(s) as PERMANENT entry if corresponding SVI exists. It builds the ARP/ND cache at Remote VTEP(s) for proxy
- Above is done irrespective of ARP/ND suppression configuration

- Once ARP/ND suppression is enabled
  - All ARP/NS request received from local/front plane ports are stolen to VTEPs' CPU and suppress/proxy if possible
  - ARP/NS request received from VXLAN fabric as *NOT* (never) stolen and *NOT* (never) attempted to suppress/proxy
- On reception of ARP/NS request,
  - Lookup is performed on Neighbor cache and proxy reply is generated by local VTEP if target address exists in cache.
  - Proxy reply contains source MAC of actual target MAC and not the VTEP MAC
  - If target address doesn't exists in cache ARP/NS is allowed to go over VXLAN fabric
  - Proxy function is performed using complete Neighbor cache which contain DYNAMIC (local) and PERMANENT (remote) neighbor entries.

- AOS-CX implementation is done to keep flood in VXLAN fabric minimal by suppressing them local VTEP and sending proxy reply from that, but at the same time it allows certain some flood in network for faster convergence
- IEFT document for suppression is still in draft version

VXLAN ARP/ND Suppression Details

- Configure Distributed Gateway along with ARP/ND suppression under EVPN configuration context to enable suppression
- **Once local neighbors are learnt**, then both advertisement via EVPN and suppression are possible

- Local neighbor can be learnt via different means,
  - IPv4: once ARP suppression is enabled
    - On receiving Gratuitous ARP request, source address is learned proactively
    - On receiving ARP request if target is L3 gateway address
    - On receiving inter VLAN traffic, when destination is not present on VTEPs cache, ARP request is generated to learn the destination
  - IPv6: once ND suppression is enabled
    - On receiving Router Solicitation (RS), source IP is learnt (if unspecified), we have observed most hosts use Link Local address as source in RS packet
    - On receiving Neighbor Solicitation (NS), source IP is learnt if target is L3 gateway address
    - On receiving inter VLAN traffic, when destination is not present on VTEPs cache; NS is generated to learn the destination

VXLAN ARP/ND Suppression Caveats

- Only recommended for distributed L3 gateway deployments
- Only ARP or NS received from local front plane ports are suppressed
- ARP request or NS received from VXLAN fabric is not suppressed considering the fact that packet has been already flooded and cross VXLAN fabric
- On VSX setup, ARP request or NS received over ISL are not stolen and not suppressed/proxy

- Not recommended for centralized L3 gateway deployments as it degrades neighbor resolution
- In centralized L3 gateway deployments only a VTEP (or VSX pair) will have L3 configuration
- Hence all ARP/NS are already flooded in the VXLAN fabric and no gain is achieved with ARP/ND suppression

Configuration

aruba
a Hewlett Packard
Enterprise company

VXLAN ARP/ND Suppression Configuration

- Configured under evpn context
- Applies to all vlans under evpn context

```
evpn
    arp-suppression
    nd-suppression
    vlan 10
        rd auto
        route-target export auto
        route-target import auto
        redistribute host-route
```

Best Practices

aruba

a Hewlett Packard Enterprise company

VXLAN ARP/ND Suppression Best Practices

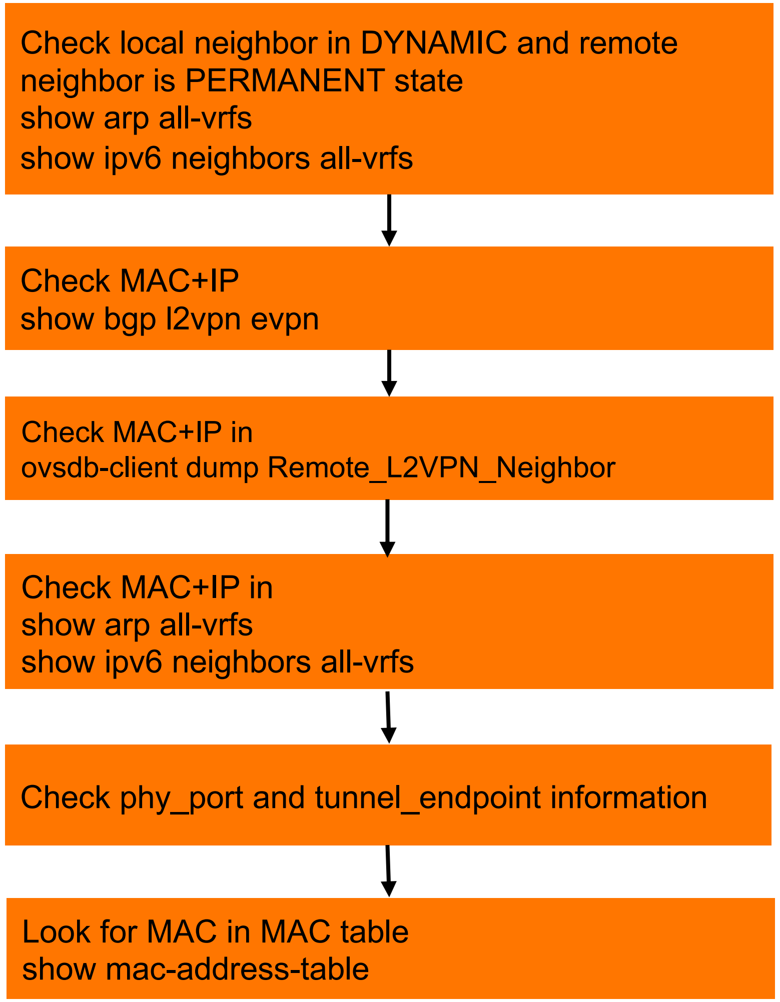
- Recommended for distributed L3 gateway deployments
- Not recommended for centralized L3 gateway deployments as it degrades neighbor resolution

Troubleshooting

aruba
a Hewlett Packard
Enterprise company

VXLAN ARP/ND Suppression Troubleshooting

- Flowchart with general recommendations



Check local neighbor in DYNAMIC and remote neighbor is PERMANENT state
show arp all-vrfs
show ipv6 neighbors all-vrfs

Check MAC+IP
show bgp l2vpn evpn

Check MAC+IP in
ovsdb-client dump Remote_L2VPN_Neighbor

Check MAC+IP in
show arp all-vrfs
show ipv6 neighbors all-vrfs

Check phy_port and tunnel_endpoint information

Look for MAC in MAC table
show mac-address-table

VXLAN ARP/ND Suppression Troubleshooting Details

Note: Assuming host H1 and H2 are remotely connected over VXLAN fabric

ARP/ND Suppression debugging

1. If ARP/NS is send from H1 -> H2 and is crossing VXLAN fabric
   a) Check at VTEP local to H2, that H2 is present in its neighbor table via “show arp all-vrf” or “show ipv6 neighbor all-vrfs” in DYNAMIC state
   b) Then check at VTEP local to H1 for neighbor entry of H2 via “show arp all-vrf” or “show ipv6 neighbor all-vrfs” in PERMANENT state
   c) Then follow steps give in “EVPN Type 2 troubleshooting steps”

EVPN Type 2 troubleshooting steps

1. Check local neighbors are advertised in RT-2 via evpn using "show bgp l2vpn evpn".
   a) If local neighbor not present, EVPN issue
2. Check Remote_L2VPN_Neighbor table is updated with IP and MAC details and from=evpn_l3.
   a) start-shell; ovsdb-client dump Remote_L2VPN_Neighbor
   b) If not then EVPN issue
3. Check remote neighbor are present in Neighbor table using "show arp all-vrfs" and "show ipv6 neighbor all-vrfs" and these are updated with phy_port with tunnel_endpoint info (e.g. vxlan1(1.1.1.1)) and state PERMANENT
   a) If "show arp all-vrfs" and "show ipv6 neighbor all-vrfs" doesn't contain the entry it is NDMD/ARP bug.
   b) If phy_port/tunnel_endpoint is missing, dump the MAC table using "show mac-address-table" and check of the same MAC
      i. If MAC table doesn't have the entry, either it is EVPN or MAC table issue. Please work with respective team.
      ii. if MAC table has entry then it is NDMD bug

VXLAN ARP/ND Suppression Troubleshooting Details

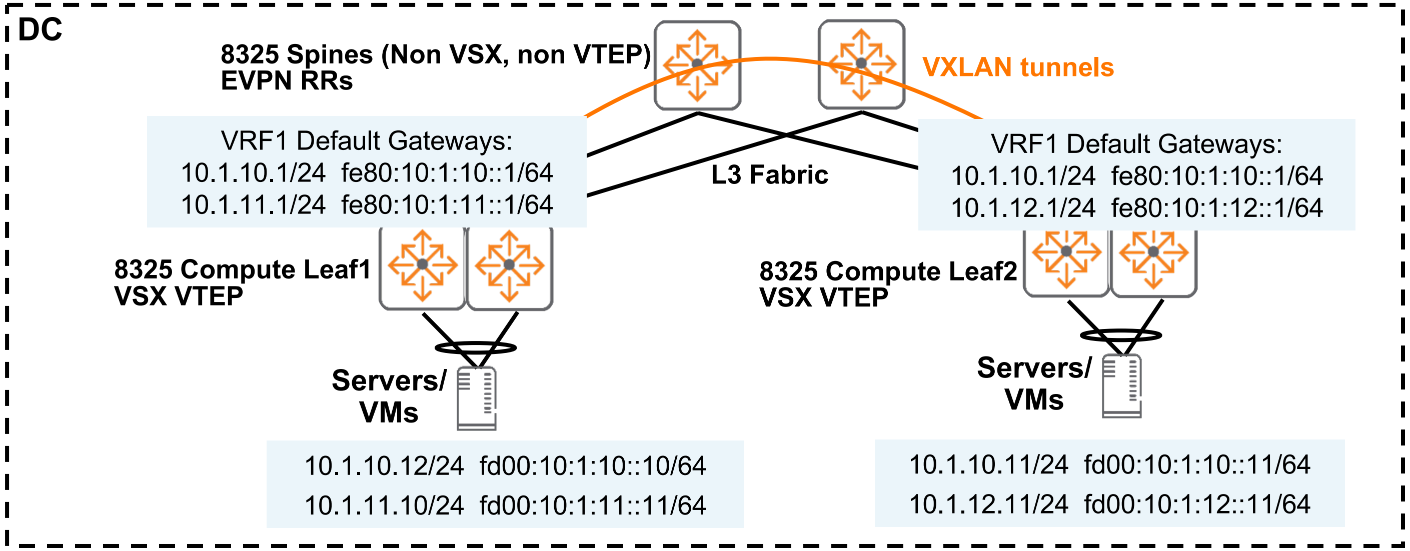
Missing ARP and traffic related debugging

1. If local ARP is missing at destination VTEP, say H2 is missing in local VTEP
    - a) If there is no traffic drop to H2, means all the flows going to H2 are L2
    - b) If there is traffic drop to H2
        - i. In case of L2 flow (intra vlan traffic) <-- **Neighbor learning may not happen but traffic will flow via MAC**
            1. Check the originator VTEP has H2 mac in “show mac”
            2. Check if packet is send from originator VTEP, which was encapsulated and send to VXLAN tunnel local to H2
            3. If packet is received at H2, check if packets coming are decapsulated
            4. Check if inner packet destination mac is present locally in “show mac”
        - ii. In case of L3 flow (inter vlan traffic) <-- **Neighbor learning is mandatory**
            1. Check the originator VTEP has H2 in neighbor table using “show arp all-vrfs” or “show ipv6 neighbor all-vrfs” OR Check if “host/subnet” route is present
            2. Check if packet is send from originator VTEP, which was encapsulated and send to VxLAN tunnel local to H2
            3. If packet is received at VTEP local to H2, check if packets coming are decapsulated
            4. Check if innet packet destination IP is present locally in using “show arp all-vrfs” or “show ipv6 neighbor all-vrfs
            5. Check if packet is coming to CPU to correct l3vni interface in associated VRF. To check that do following,

```
start-shell
sudo start_vrf_shell <vrf_name>
ip link show | grep vni <- find the destination vni interface
tcpdump -i <vni_interface_name> <- if packets are not seen, either VxLAN or CPU_Rx issue
ip monitor <- If “INCOMPLETE” notification is not seen, either VxLAN or CPU_Rx or kernel interface settings issue
              else, ndmd must be getting notification and it should have been sending ARP/NS on respective interface (SVI) for resolution.
                   Capture Tx and Rx and check if ARP request/responses OR NS/NA are seen.
```

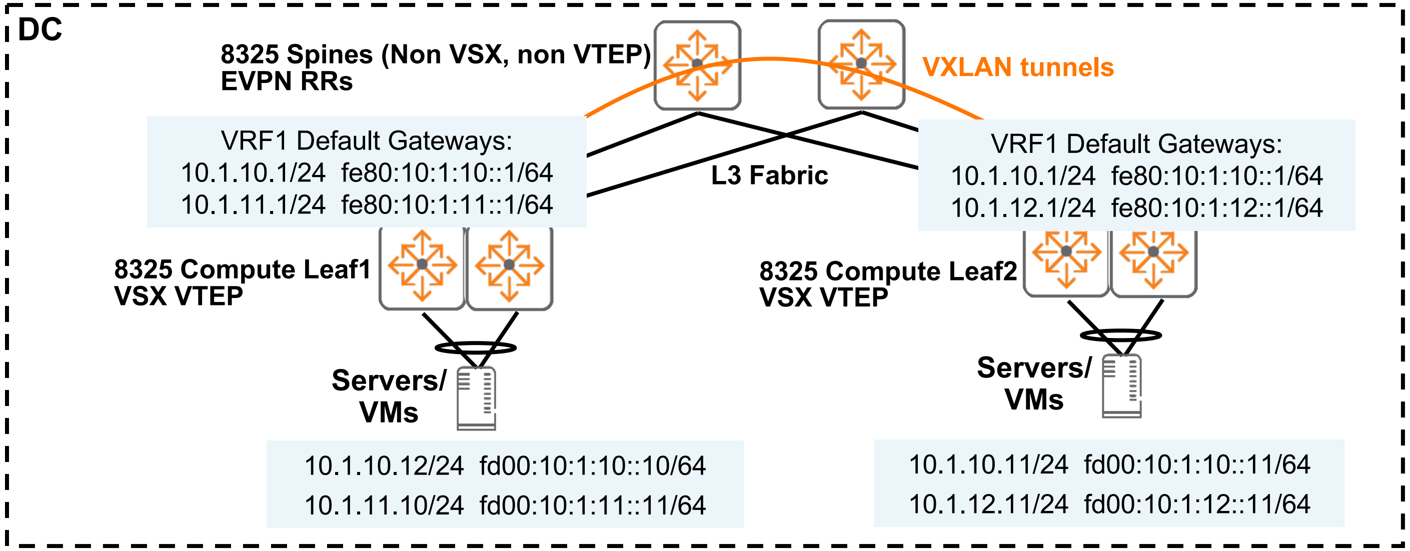
Demo

Demo#1 – VXLAN ARP Suppression with L3 Distributed L3 gateways


DC
8325 Spines (Non VSX, non VTEP)
EVPN RRs
VXLAN tunnels
VRF1 Default Gateways:
10.1.10.1/24 fe80:10:1:10::1/64
10.1.11.1/24 fe80:10:1:11::1/64
L3 Fabric
VRF1 Default Gateways:
10.1.10.1/24 fe80:10:1:10::1/64
10.1.12.1/24 fe80:10:1:12::1/64
8325 Compute Leaf1
VSX VTEP
8325 Compute Leaf2
VSX VTEP
Servers/
VMs
Servers/
VMs
10.1.10.12/24 fd00:10:1:10::10/64
10.1.11.10/24 fd00:10:1:11::11/64
10.1.10.11/24 fd00:10:1:10::11/64
10.1.12.11/24 fd00:10:1:12::11/64

- ARP suppression only beneficial for L2 connectivity on the same subnet
- ARP suppression not used during L3 connectivity between subnets
- Show wireshark on VTEP uplink with ARP requests/responses being sent/not being sent

Demo#2 – Troubleshooting VXLAN ARP Suppression



VRF1 Default Gateways:
10.1.10.1/24 fe80:10:1:10::1/64
10.1.11.1/24 fe80:10:1:11::1/64

VRF1 Default Gateways:
10.1.10.1/24 fe80:10:1:10::1/64
10.1.12.1/24 fe80:10:1:12::1/64

10.1.10.12/24 fd00:10:1:10::10/64
10.1.11.10/24 fd00:10:1:11::11/64

10.1.10.11/24 fd00:10:1:10::11/64
10.1.12.11/24 fd00:10:1:12::11/64

- Suppression scenario troubleshooting = Neighbor learning issue/traffic drop aruba

a Hewlett Packard
Enterprise company

Thank you