Troubleshoot BGP Flaps Between Ultra Packet Core and Nexus Switch Due to Incorrect Configuration

Contents

Introduction

This document describes the solution to the Border Gateway Protocol (BGP) flaps between Cisco Ultra Packet Core (UPC) and Nexus 9000 switch configured with redundant BGP connection.

Problem

BGP flaps are triggered when one of the redundant interfaces between the Cisco Ultra Packet Core and Nexus switch flaps.

Conditions

The Ultra Packet Core (UPC) node is connected to Nexus Leaf A and Leaf B on separate ports. The BGP IPv6 peers are established and the default routes are installed on the UPC node. Figure 1 shows the high-level network diagram with redundant path to Leaf switches.

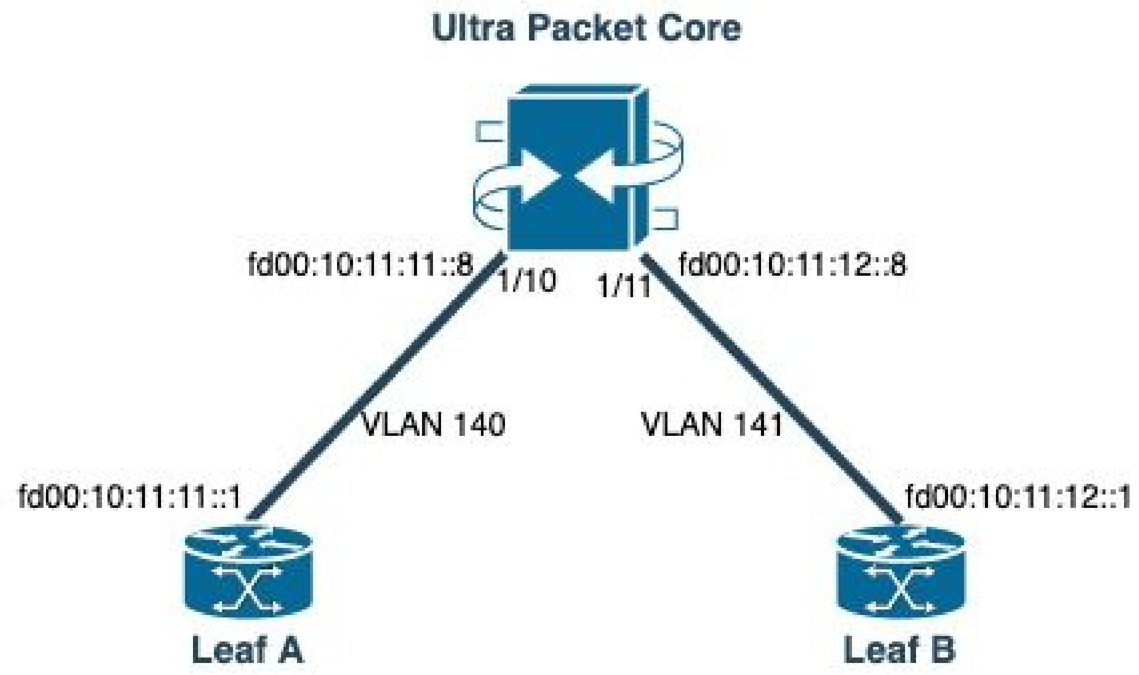

Figure 1: Network Diagram

Configuration

UPC port configuration with VLAN and interface binding:

```
port ethernet 1/10
    no shutdown
    vlan 140
        no shutdown
        bind interface saegw_vlan140_1/10 saegw
#exit

#exit
port ethernet 1/11
    no shutdown
    vlan 141
        no shutdown
        bind interface saegw_vlan141_1/11 saegw
#exit
#exit
end
```

UPC Interface configuration with IP addresses:

```
interface saegw_vlan140_1/10
  ip address 10.11.11..8 255.255.255.0
  ipv6 address fd00:10:11:11::8/64 secondary
  bfd interval 300 min_rx 300 multiplier 3
#exit
interface saegw_vlan141_1/11
  ip address 10.11.12.8 255.255.255.0
  ipv6 address fd00:10:11:12::8/64 secondary
  bfd interval 300 min_rx 300 multiplier 3
#exit
```

UPC BGP configuration:

```
router bgp 25949
  router-id 172.19.20.30
  maximum-paths ebgp 4
  neighbor 10.11.11..1 remote-as 25949
  neighbor 10.11.11..1 fall-over bfd
  neighbor 10.11.12.1 remote-as 25949
  neighbor 10.11.12.1 fall-over bfd
  neighbor fd00:10:11:11::1 remote-as 25949
  neighbor fd00:10:11:12::1 remote-as 25949
  address-family ipv4
    neighbor 10.11.11..1 route-map accept_default in
    neighbor 10.11.11..1 route-map gw-1-OUT out
    neighbor 10.11.12.1 route-map accept_default in
    neighbor 10.11.12.1 route-map gw-1-OUT out
    redistribute connected
#exit
address-family ipv6
  neighbor fd00:10:11:11::1 activate
  neighbor fd00:10:11:11::1 route-map accept_v6_default in
  neighbor fd00:10:11:11::1 route-map allow_service_ips_v6 out
  neighbor fd00:10:11:12::1 activate
  neighbor fd00:10:11:12::1 route-map accept_v6_default in
  neighbor fd00:10:11:12::1 route-map allow_service_ips_v6 out
  redistribute connected
#exit

ipv6 prefix-list name accept_v6_default_routes seq 10 permit ::/0
route-map accept_v6_default permit 10
  match ipv6 address prefix-list accept_v6_default_routes
#exit
```

Nexus 9000 switch configuration:

```
Interface vlan140
ipv6 address fd00:10:11:11::1/64
no ipv6 redirects

interface vlan141
ipv6 address fd00:10:11:12::1/64
no ipv6 redirects

vrf upc
address-family ipv4 unicast
advertise l2vpn evpn
maximum-paths ibgp 2
address-family ipv6 unicast
advertise l2vpn evpn
maximum-paths ibgp 2
neighbor fd00:10:11:12::5
remote-as 25949
address-family ipv6 unicast
neighbor fd00:10:11:12::6
remote-as 25949
address-family ipv6 unicast
neighbor fd00:10:11:12::8
remote-as 25949
address-family ipv6 unicast
```

Analysis

Initially a normal BGP communication between one of the UPC interfaces (fd00:10:11:12::8) and the Nexus switch (fd00:10:11:12::1 belongs to vlan141) is observed that includes TCP ACK messages:

```
2023-01-01 01:01:59.000000 fd00:10:11:12::8 -> fd00:10:11:12::1 TCP 35813 > bgp [ACK] Seq=250
Ack=8664 Win=31744 Len=0 TSV=2412344062 TSER=531234647
2023-01-01 01:01:59.000087 fd00:10:11:12::8 -> fd00:10:11:12::1 TCP 35813 > bgp [ACK] Seq=250
Ack=11520 Win=37376 Len=0 TSV=2412344062 TSER=531234647
2023-01-01 01:01:59.000162 fd00:10:11:12::8 -> fd00:10:11:12::1 TCP 35813 > bgp [ACK] Seq=250
Ack=14376 Win=43008 Len=0 TSV=241234062 TSER=531234647
2023-01-01 01:01:59.000281 fd00:10:11:12::8 -> fd00:10:11:12::1 TCP 35813 > bgp [ACK] Seq=250
Ack=17232 Win=49152 Len=0 TSV=2412344062 TSER=531234647
2023-01-01 01:01:59.000936 fd00:10:11:12::8 -> fd00:10:11:12::1 TCP 35813 > bgp [ACK] Seq=250
Ack=20663 Win=48640 Len=0 TSV=2412344063 TSER=531234647
```

Upon failure of Leaf-B interface towards UPC, an incorrect behaviour is seen in the logs where a new BGP connection attempt is initiated by the UPC (source: fd00:10:11:12::8) towards the Leaf-A on interface fd00:10:11:11::1, which belongs to a different VLAN, vlan140.

```
2023-01-01 22:36:12.370117 fd00:10:11:12::8 -> fd00:10:11:11::1 TCP 41987 > bgp [SYN] Seq=0
Win=14400 Len=0 MSS=1440 TSV=2412347369 TSER=0 WS=9
```

Such invalid BGP SYN message sent on the wrong interface results in the BGP down. When the Nexus advertises its own connected route and UPC gets a route for the interface which was down over BGP, then UPC attempts connection via another interface with a different/wrong outgoing IP.

Solution

Due to the configuration referred in the Condition section of this article, since UPC receives the connected route information of both Leafs from both interfaces, when one of the interfaces is down, UPC attempts to communicate to that Leaf through the other interface.

To avoid UPC to send the BGP connection establishment messages from the wrong interface, here are the configuration changes for consideration:

1. In UPC configuration, add update-source for the neighbor. This configuration prevents the BGP connection from a different interface, if the main interface is down. For example, when saegw_vlan140_1/10 (fd00:10:11:11::1/64) is down then the node cannot use outgoing interface saegw_vlan141_1/11 for BGP peer fd00:10:11:11::8.
   Here is a sample configuration :

   ```
   neighbor fd00:10:11:11::1 update-source fd00:10:11:11::8
   neighbor fd00:10:11:12::1 update-source fd00:10:11:12::8
   ```
2. In the Nexus configuration, block the prefixes from the wrong interfaces.
   For example, we deny routes for the redundant leaf over neighbor fd00:10:11:11::1

   ```
   neighbor fd00:10:11:11::1
   update prefix list to deny fd00:10:11:12::8/64
   ```
3. In Nexus switch, the EBGP peering from the VTEP to a external node over VXLAN must be in a tenant VRF and must use the update-source of a loopback interface (peering over VXLAN) as recommended in the Cisco Nexus 9000 Configuration Guide