Adventures in 6VPE and EVPN IPv6

David Freedman

UK IPv6 Council Meeting - 12-Dec-2019

claranet | helping our customers do amazing things

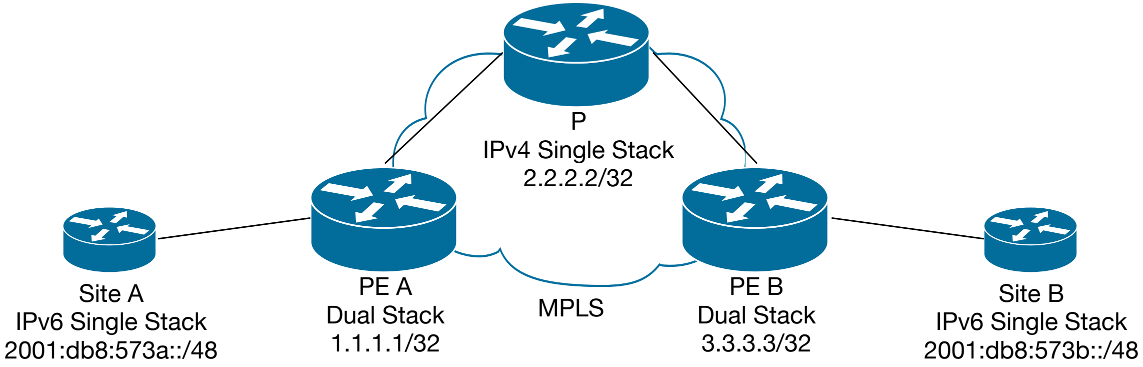
But first…

- **6VPE is a (Cisco) marketing term for IPv6 L3VPN (RFC4659) over 6PE (IPv6 over IPv4/MPLS – RFC4798)**
  - I'm sure you have all seen my 6PE talk, if not, go watch it (IPv6 Security Workshop 2017 - https://youtu.be/u-Igj5LMqCU)
  - My colleague Sandy Breeze also gave a recent talk about how bad Cisco 6PE is and how we plan to move to SR-MPLS – See **https://ptnog.pt/wp-content/uploads/2019/12/ptnog4_02_6PE_fails_and_other_short_stories.pdf**

    - **What is the point of IPv6 L3VPN given that IPv6 has unique addressing?**
      - Primary Reason : Private routing domains and routing behaviour – being able to move traffic in directions unintended and un-signalled by the GRT – all without any form of "SDN".
      - Secondary Reason: Parity with existing IPv4 L3VPN (RFC4364), customers are more likely to adopt if you can give them an easy, cheap (to them) and familiar - way of doing things.

Implementing 6VPE

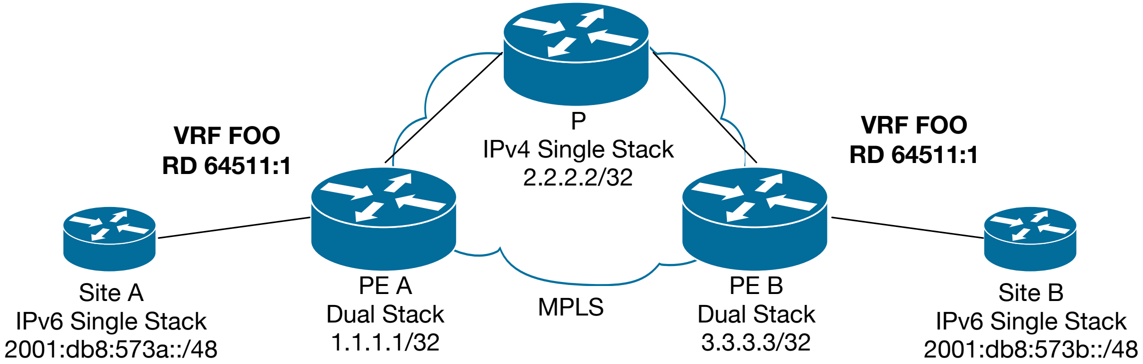
- **First you need 6PE**
  - 6PE connects islands of IPv6 with IPv4 MPLS, *that is to say, an LSP carrying IPv6 traffic is associated with an IPv4 FEC*.
- **6PE mandates two labels**
  - RFC4798 admits that single label operation is possible (s.3) but mandates that two labels are employed.
  - One label for the IPv6 prefix, signalled through MP-BGP.
  - One label for the IPv4 NH, resulting from an IPv4 FEC.
  - Having two labels avoids inter-op problems with PHP (and allows IPv6 un-aware routers to do link hashing)



```
PEA# show bgp ipv6 unicast 2001:db8:573b::/48
BGP routing table entry for 2001:db8:573b::/48, version 7
Paths: (1 available, best #1, table Global-IPv6-Table)
  Advertised to update-groups:
        1
  65002
    ::FFFF:3.3.3.3 (metric 4) from 2.2.2.2 (2.2.2.2)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 3.3.3.3
      mpls labels in/out nolabel/57

PEA# show ip cef 3.3.3.3
3.3.3.3/32, version 60, epoch 0, cached adjacency 10.1.1.2
0 packets, 0 bytes
  tag information set
    local tag: 18
    fast tag rewrite with Te0/0, 10.1.1.2, tags imposed: {33}
  via 10.1.1.2, TenGigabitEthernet0/0, 0 dependencies
    next hop 10.1.1.2, TenGigabitEthernet0/0
    valid cached adjacency
    tag rewrite with Te0/0, 10.1.1.2, tags imposed: {33}

PEA#show ipv6 cef 2001:db8:573b::/48
2001:db8:573b::/48
     nexthop ::FFFF:3.3.3.3
    fast tag rewrite with Te0/0, 10.1.1.2, tags imposed: {33 57}
```

Implementing 6VPE

- **Now add an L3VPN (VRF)**
  - Interface is constrained to Virtual Routing and Forwarding Context..
- **Again, two labels**
  - And again, outer label used for the IPv4 NH, resulting from an IPv4 FEC.
  - Now inner label for the IPv6 prefix, signalled through MP-BGP – this time with RT extended communities.
  - Inner label can either be per-prefix or per-VRF
    - Per VRF requires additional lookup at remote PE (bad for data plane scale) but can result in fewer labels associated with the VRF (good for control plane scale).
    - Per prefix results in more labels associated with the VRF (bad for control plane scale), but does not require extra lookup at remote PE as forwarding operation happens directly in LFIB (good for data plane scale).

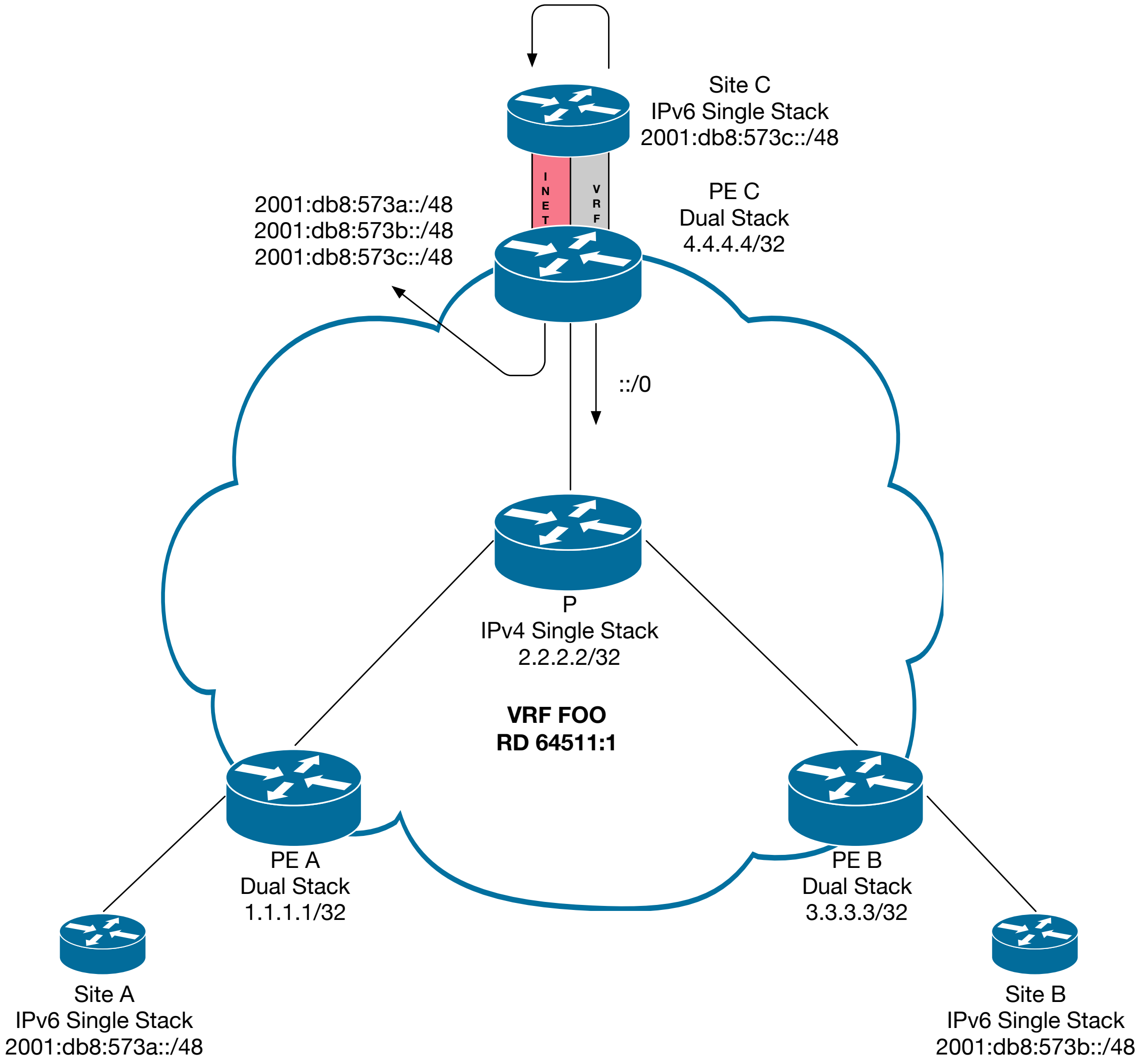


```
PEA# show bgp vpnv6 unicast vrf FOO 2001:db8:573b::/48
BGP routing table entry for 2001:db8:573b::/48, version 7
Paths: (1 available, best #1, table Global-IPv6-Table)
  Advertised to update-groups:
        1
  65002
    ::FFFF:3.3.3.3 (metric 4) from 2.2.2.2 (2.2.2.2)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 3.3.3.3
      mpls labels in/out nolabel/67

PEA# show ip cef 3.3.3.3
3.3.3.3/32, version 60, epoch 0, cached adjacency 10.1.1.2
0 packets, 0 bytes
  tag information set
    local tag: 18
    fast tag rewrite with Te0/0, 10.1.1.2, tags imposed: {33}
  via 10.1.1.2, TenGigabitEthernet0/0, 0 dependencies
    next hop 10.1.1.2, TenGigabitEthernet0/0
    valid cached adjacency
    tag rewrite with Te0/0, 10.1.1.2, tags imposed: {33}

PEA#show ipv6 cef verf FOO 2001:db8:573b::/48
2001:db8:573b::/48, epoch 0, flags [rib defined all labels]
    recursive via 3.3.3.3 label 67
       nexthop 10.1.1.2 TenGigabitEthernet0/0 label 33
```

claranet

So what do people use this for?

- **Simple Traffic Engineering**
  - Announce default in the VRF via a central or branch site.
- **Steer traffic through a mid-box**
  - Companies and Schools like to steer Internet traffic through mid-boxes.
  - Some of these boxes are ‘optimisers’, but most of them are filters or transparent proxies of some kind.
  - The No-NAT66 paradigm of IPv6 means the entire VRF should be returned via this site.
    - If routing is asymmetric then you rule out simple strict uRPF (and probably your optimiser or filter won’t work or be effective)
    - Also bear in mind, returning all traffic through this site and not the branches can lead to bottlenecking. Central site therefore needs more capacity.

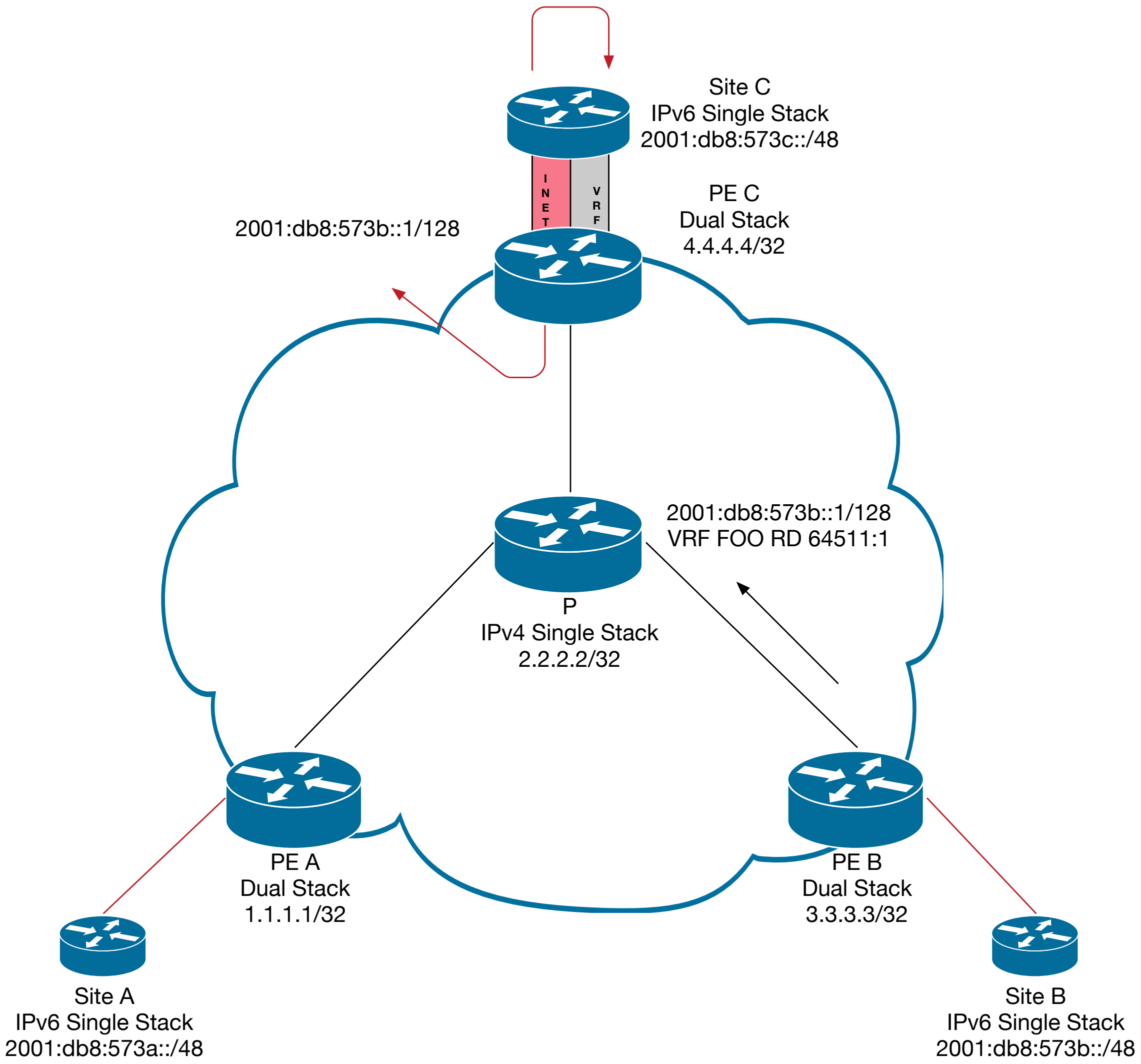


claranet

So what do people use this for? (#2)

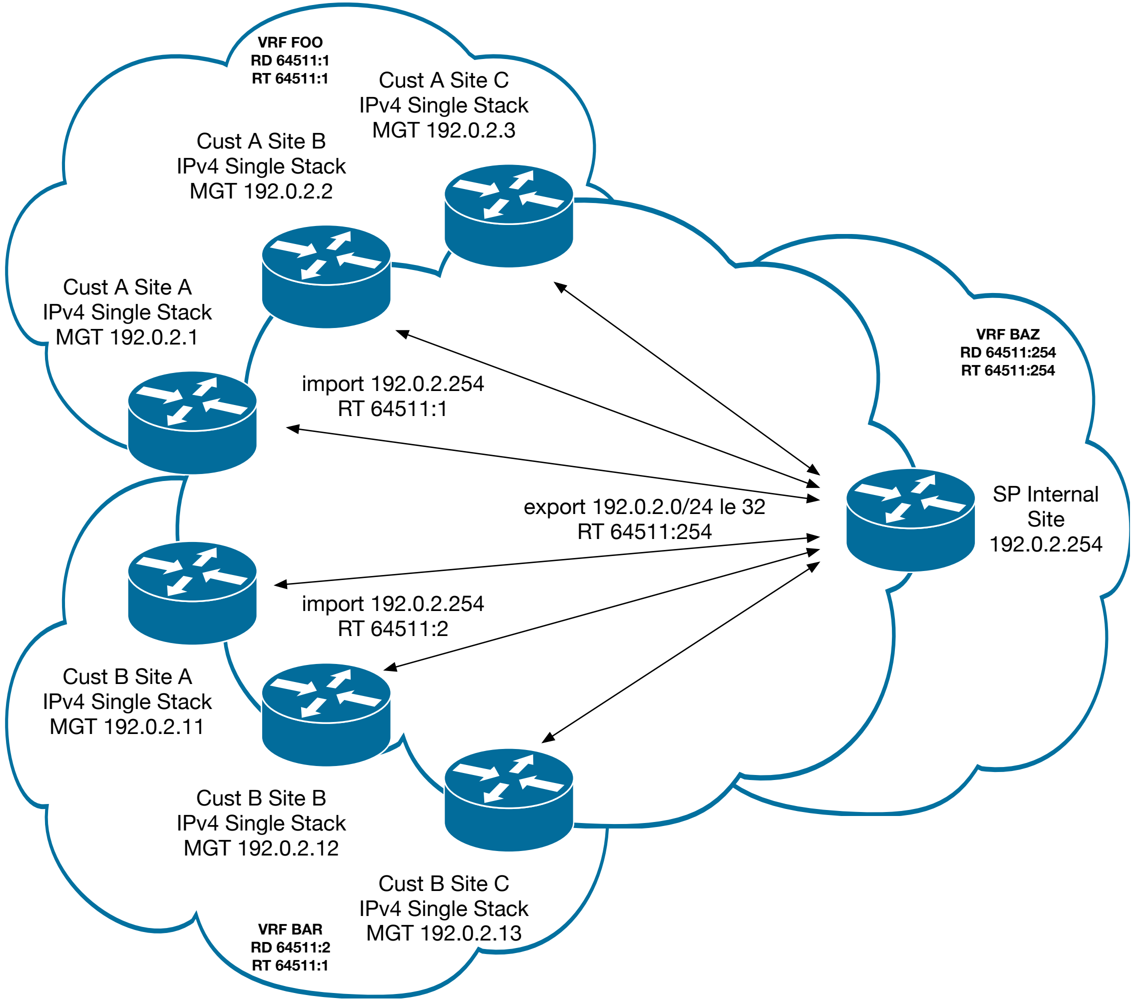
- **Simple Traffic Engineering**
  - Announce a more specific /128 to GRT
- **Steer traffic for DoS mitigation**
  - ”Dirty” traffic (mix of attack and legitimate) attracted to more specific /128 in GRT.
  - Traffic sent through some kind of cleaning platform.
  - “Clean” traffic (only legitimate traffic) returned back in VRF.
    - This is needed because if it was returned via GRT then it would encounter the “Dirty” version of itself, and thus a loop would ensue.
    - Remote end attracts the /128 in VRF and this is what actually produces the GRT announcement.
    - At the remote end, traffic should be broken out to GRT, only for forwarding on the last hop (where the “Dirty” prefix is not seen)

claranet



So is this just an IPv4 analogue?

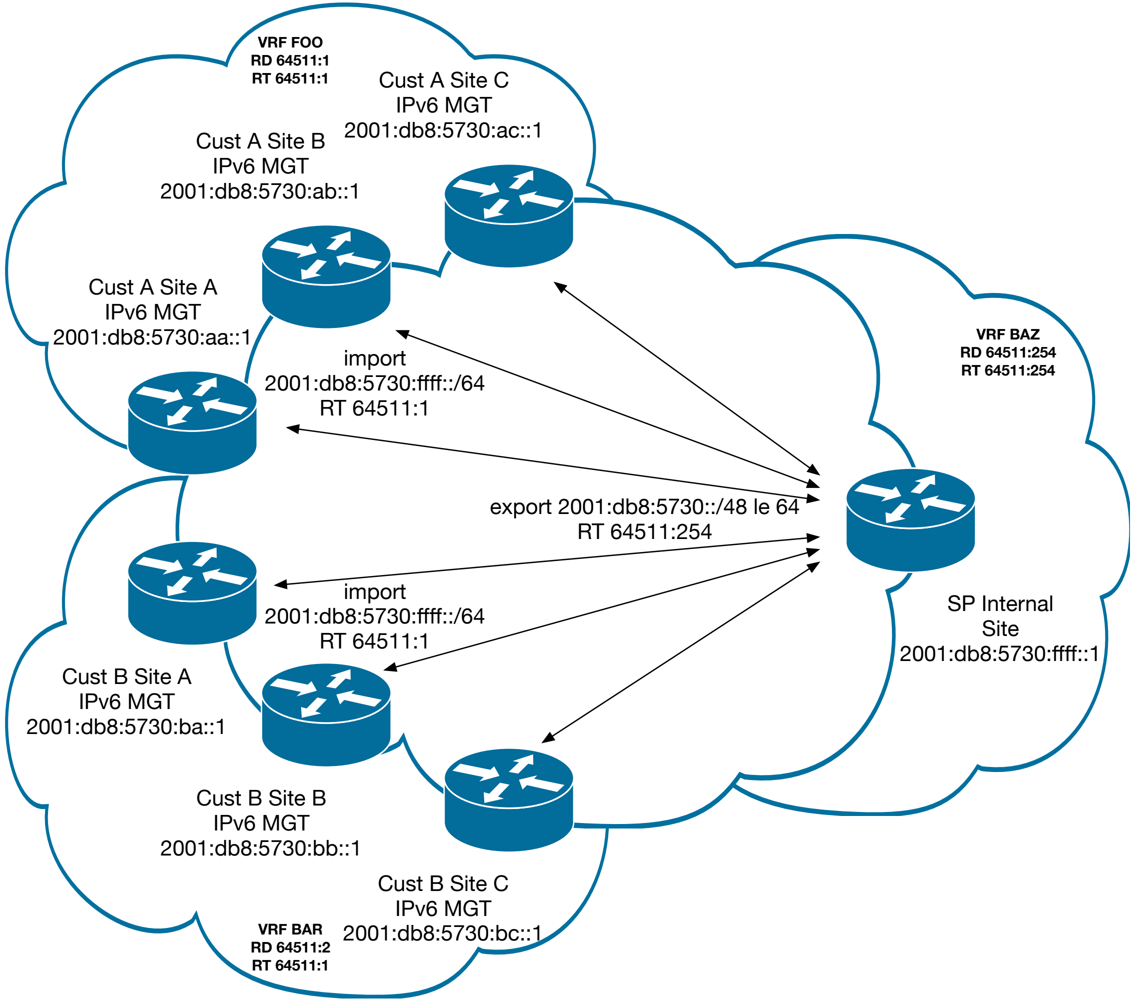
- **Take Managed CE as an example.**
  - SP wants to securely manage CE in all VPNs.
- **Usually this is done in-band.**
  - SP defines unique management IPv4 space.
    - Public IPv4 at scale not commercially viable, so SP usually uses some form of RFC1918 which does not conflict with customer.
  - SP Defines a management VRF.
    - Prefix of management platform imported by all customers as part of their VRF build, using management VRF RT.
    - CE management addresses matching approved prefix all exported using management VRF RT, again as part of VRF build.
    - This separation is needed such as not to contaminate customers with reachability to each other's VRFs.



claranet

So is this just an IPv4 analogue? (#2)

- **Look at what changes with IPv6..**
  - Suddenly there are plenty of addresses.
- **You can use public, global space.**
  - Sites can have an entire prefix (/64) to themselves.
  - The management platform can have it’s own prefix.
  - The same rules of engagement apply:
    - Prefix of management platform imported by all customers as part of their VRF build, using management VRF RT.
    - CE management addresses matching approved prefix all exported using management VRF RT, again as part of VRF build.
    - This separation is needed such as not to contaminate customers with reachability to each other's VRFs.



VRF FOO
RD 64511:1
RT 64511:1

Cust A Site C
IPv6 MGT
2001:db8:5730:ac::1

Cust A Site B
IPv6 MGT
2001:db8:5730:ab::1

Cust A Site A
IPv6 MGT
2001:db8:5730:aa::1

import
2001:db8:5730:ffff::/64
RT 64511:1

VRF BAZ
RD 64511:254
RT 64511:254

export 2001:db8:5730::/48 le 64
RT 64511:254

SP Internal
Site
2001:db8:5730:ffff::1

import
2001:db8:5730:ffff::/64
RT 64511:1

Cust B Site A
IPv6 MGT
2001:db8:5730:ba::1

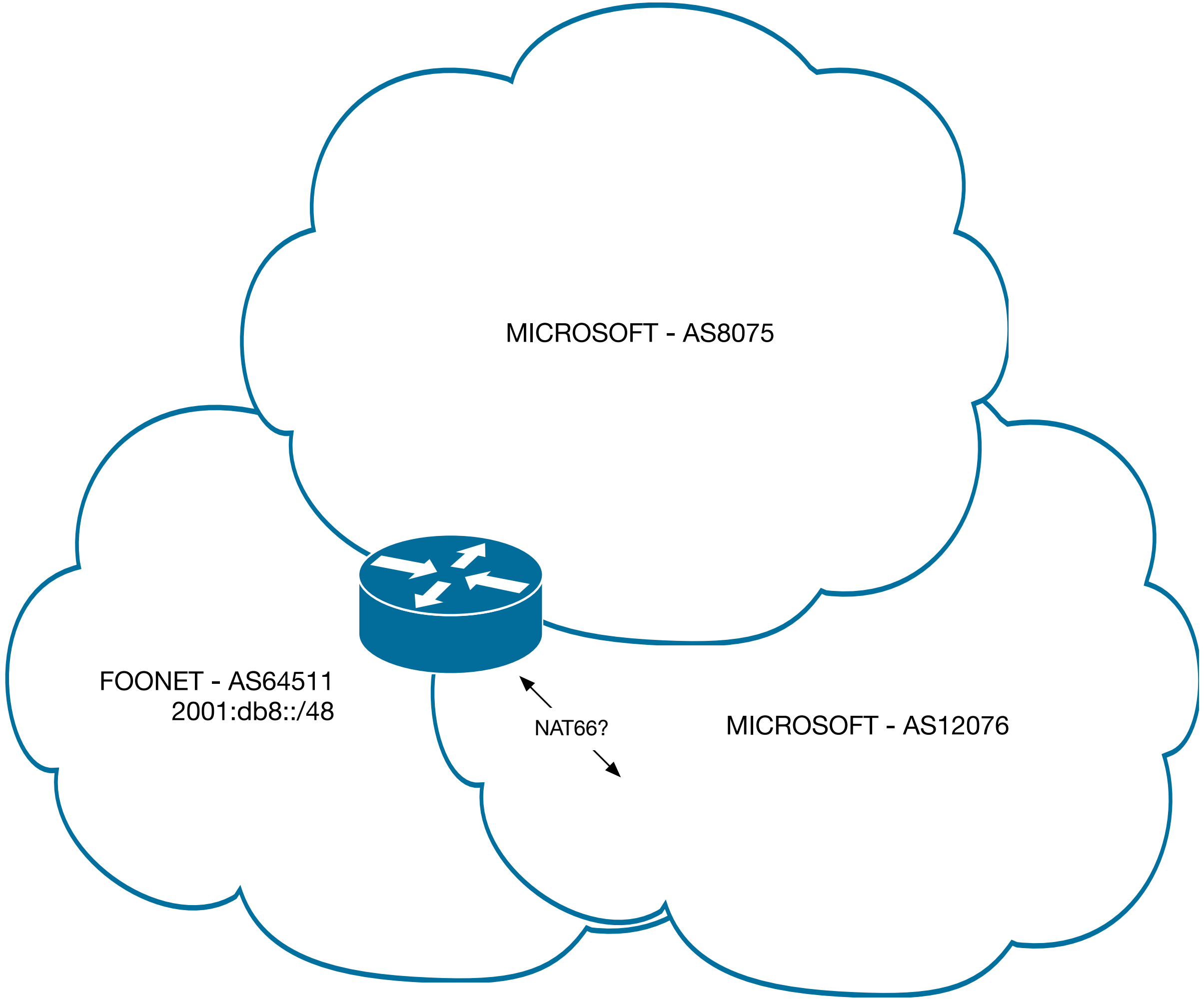
Cust B Site B
IPv6 MGT
2001:db8:5730:bb::1

Cust B Site C
IPv6 MGT
2001:db8:5730:bc::1

VRF BAR
RD 64511:2
RT 64511:1

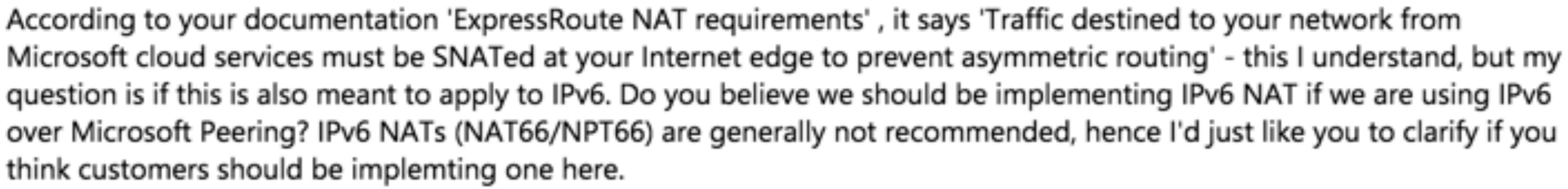
So is this just like IPv4 without NAT?

- **Not Quite.**
  - Let me tell you about Microsoft Expressroute…
- **Dedicated (and paid) circuit to MS**
  - Originally L3VPN into V-NET (Private Mode).
  - Microsoft global ASN is 8075.
  - But Expressroute terminates in AS12076.
  - Public services (Azure, O365 etc..) appear use routing table of AS12076 first.
  - If network has both a public peering and an Expressroute circuit , asymmetry can occur.
  - Asymmetry can cause anything from filtering, to overbilling (if the return traffic comes back via the (paid) Expressroute)
  - Microsoft solution is to SNAT the Expressroute in both directions!

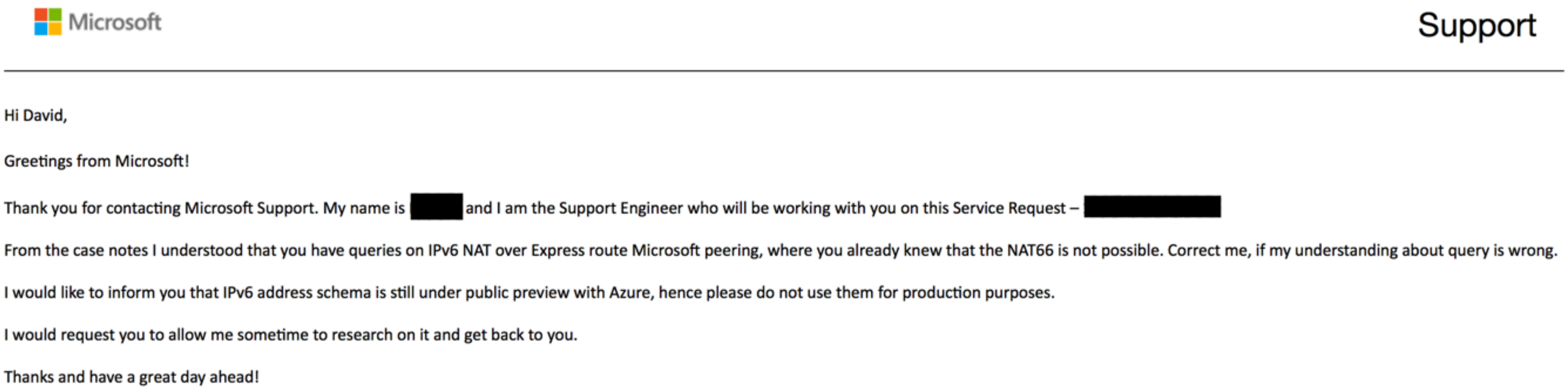


claranet

So is this just like IPv4 without NAT?



According to your documentation 'ExpressRoute NAT requirements' , it says 'Traffic destined to your network from Microsoft cloud services must be SNATed at your Internet edge to prevent asymmetric routing' - this I understand, but my question is if this is also meant to apply to IPv6. Do you believe we should be implementing IPv6 NAT if we are using IPv6 over Microsoft Peering? IPv6 NATs (NAT66/NPT66) are generally not recommended, hence I'd just like you to clarify if you think customers should be implemting one here.



Microsoft Support

Hi David,

Greetings from Microsoft!

Thank you for contacting Microsoft Support. My name is [redacted] and I am the Support Engineer who will be working with you on this Service Request – [redacted]

From the case notes I understood that you have queries on IPv6 NAT over Express route Microsoft peering, where you already knew that the NAT66 is not possible. Correct me, if my understanding about query is wrong.

I would like to inform you that IPv6 address schema is still under public preview with Azure, hence please do not use them for production purposes.

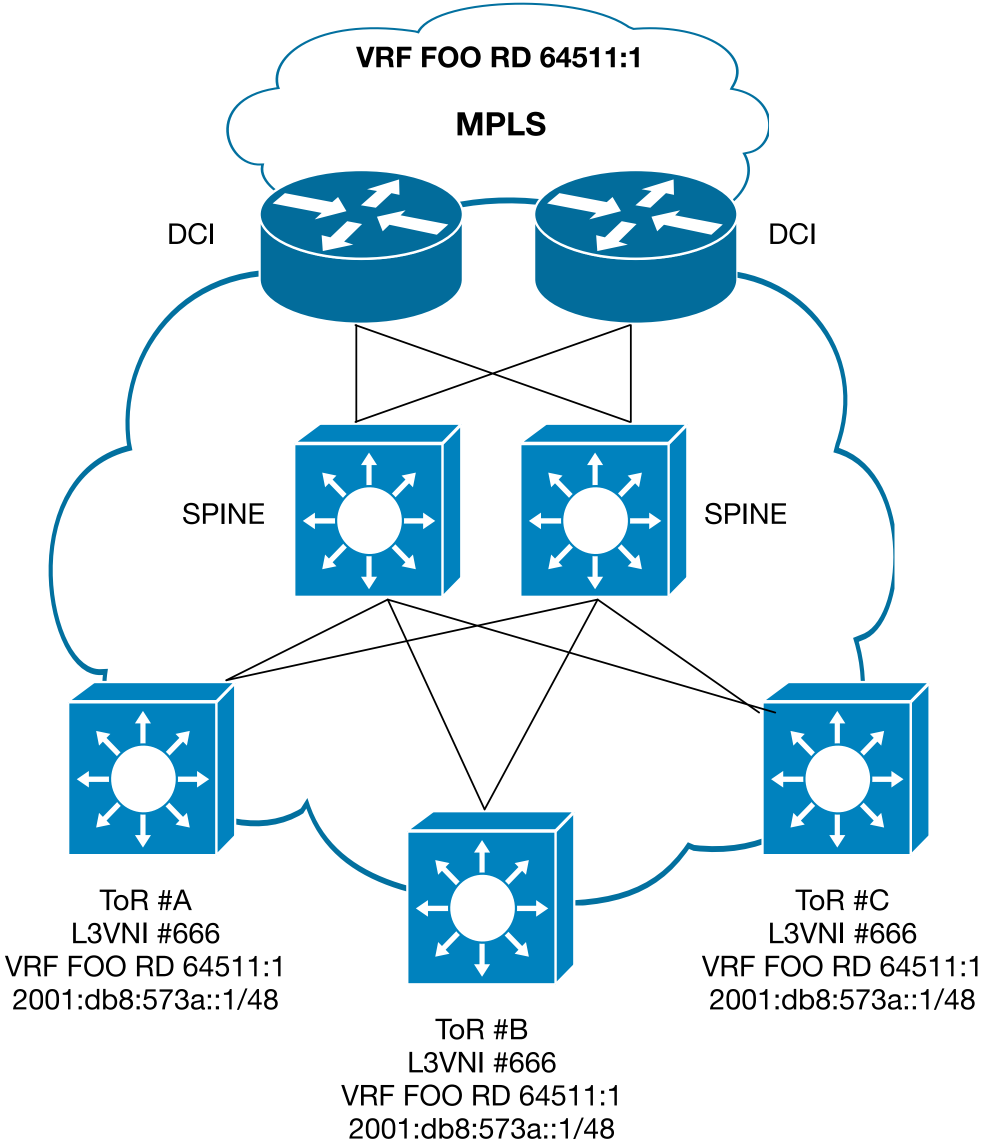
I would request you to allow me sometime to research on it and get back to you.

Thanks and have a great day ahead!

Anyway let's talk about EVPN / VXLAN

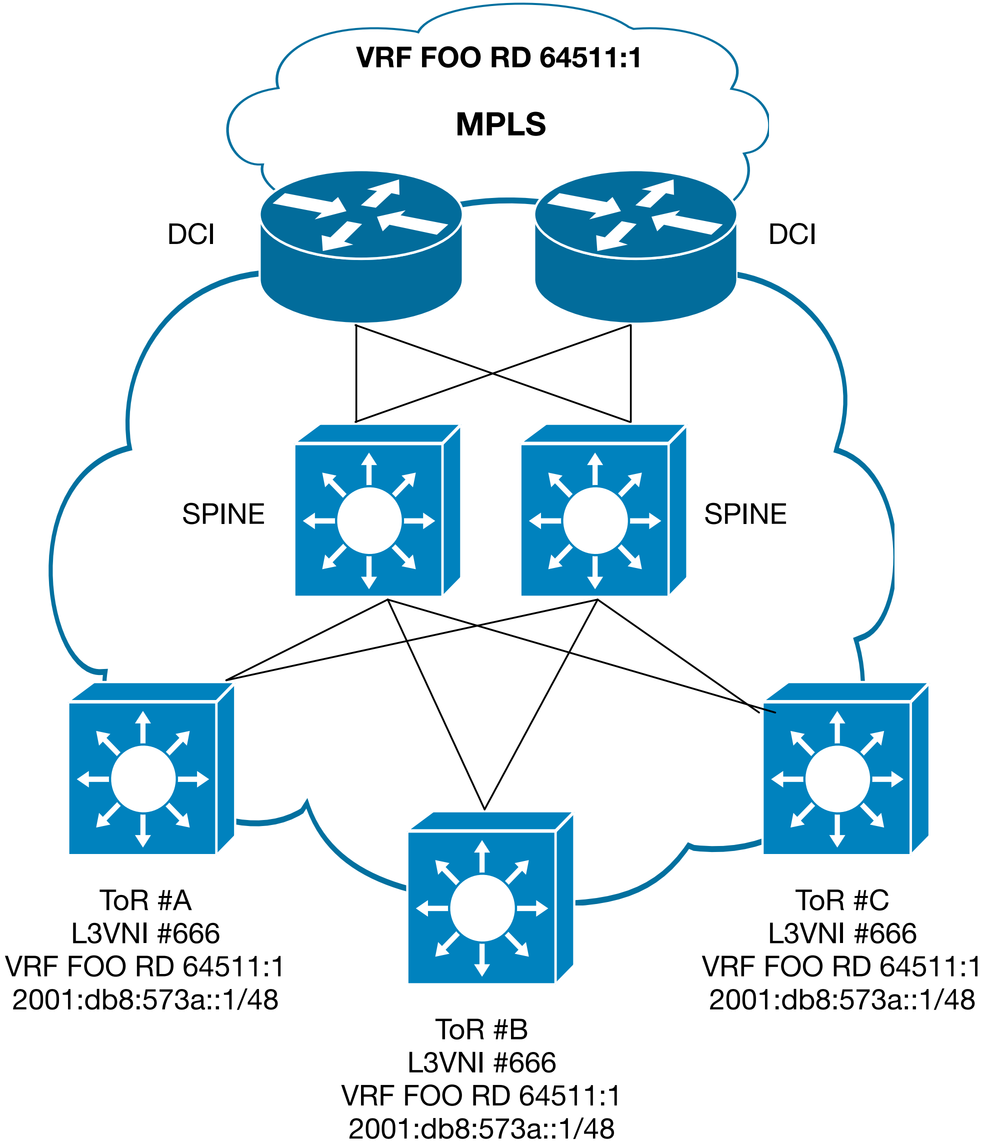
- **Here is a typical GOLF DC**
  - I'm sure you have seen our GOLF talk, if not, go watch it at **https://youtu.be/jXOrdHfBqb0**
- **In GOLF we use BGP EVPN as control plane and VxLAN as data plane (RFC8365)**
  - Top of Rack switches (ToR / Leaf) provide virtualised layer 3 services in the fabric in the form of L3VNI.
  - The L3VNI is signalled by BGP EVPN and implemented in the VxLAN encap to keep traffic segregated.
  - When the traffic is on-box, it is constrained to a VRF as usual, using the VRF RD.
  - DCI layer "stitches" L3VNI to MPLS VPN (6VPE).

claranet



EVPN / VXLAN (#2)

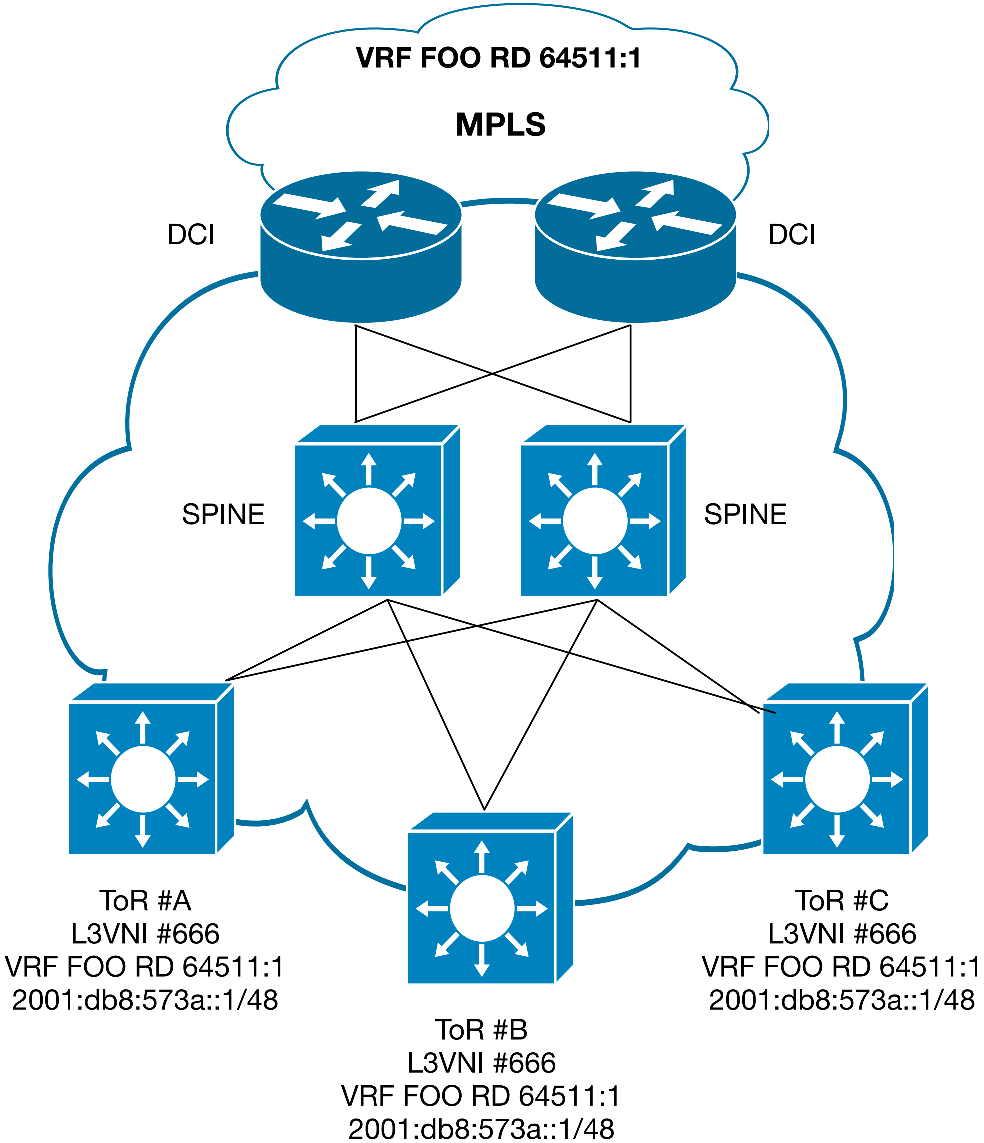
- **Delivers L2 and L3 services**
  - L3 provided at each ToR
  - Implemented as SVI (IRB)
- **ToR is anycasting**
  - Each Top of Rack has an anycast MAC
    - In our case `0000.1111.2222`
  - Anycast MAC is used to generate anycast link-local IPv6 (`fe80::11ff:fe11:2222`)
  - NDP exchanges between ToR and on-subnet neighbours (unfortunately, today this is not shared and thus not supressed)
  - Global scope address then added
  - Any other features on top of this (RA, ACLs etc..)
    - Assuming you have the TCAM that is…



claranet

EVPN / VXLAN (#3)

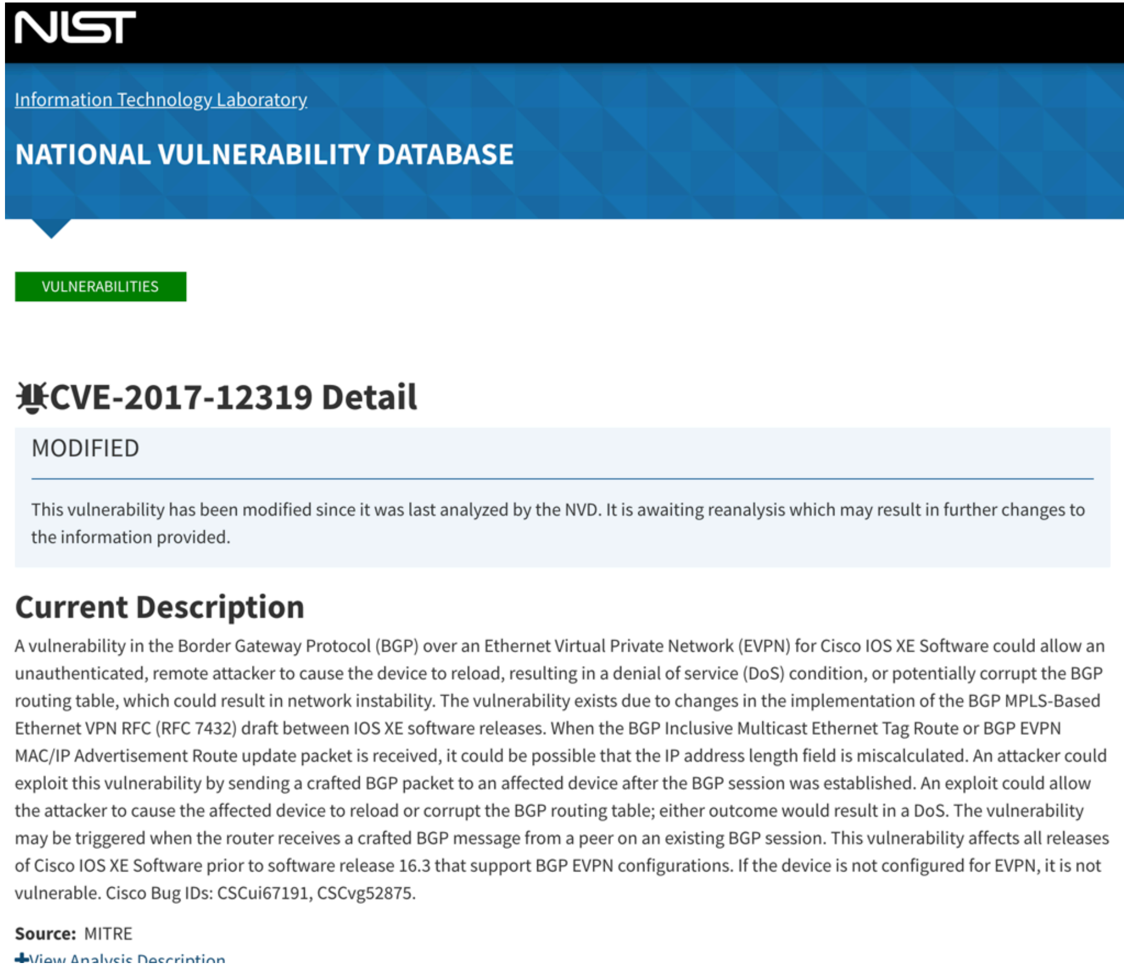
- **Address learning -> BGP**
  - IPv6 prefix carried in EVPN T2 & T5 advertisements
- **New IPv6 neighbours result in T2 activity**
  - Each time a neighbour is discovered, a T2 is “flooded”.
  - EVPN in our GOLF DCs based on hybrid of full-mesh & reflection.
    - Leaves are fully meshed with DCI
    - DCI peers with VPN RR
    - At the time our VPN RR serviced VPN IPv4 (RFC4364), VPN IPv6 (RFC 4659) and EVPN (RFC7432)
    - At the time, our RR implementation was IOS XE.
- **One day in 2017, everything broke**
  - RRs bounced sessions with DCI
  - Datacentre was EVPN isolated!
  - It did this repeatedly, again , and again…



claranet

CVE2017-12319 / CSCui67191

- **Remember that T2?**
  - T2 was sent to the DCI an onward to the RR
- **New IPv6 neighbours result in T2 activity**
  - T2 can signal address of variable lengths
  - RR (IOS-XE) EVPN implementation was different implementation of spec from DCI (IOS-XR) which was different implementation of spec from leaf (NX-OS)!
  - Two implementations validated lengths differently.
  - No proper error handling – sessions torn down!
  - When session came up again, T2 was back from the leaf – sessions torn down again!
  - Rinse and repeat…
- **We quickly took the leaf offline**
  - Valuable lesson – RR code diversity is an actual and important thing, this we have now fixed.



NIST

Information Technology Laboratory

NATIONAL VULNERABILITY DATABASE

VULNERABILITIES

CVE-2017-12319 Detail

MODIFIED

This vulnerability has been modified since it was last analyzed by the NVD. It is awaiting reanalysis which may result in further changes to the information provided.

Current Description

A vulnerability in the Border Gateway Protocol (BGP) over an Ethernet Virtual Private Network (EVPN) for Cisco IOS XE Software could allow an unauthenticated, remote attacker to cause the device to reload, resulting in a denial of service (DoS) condition, or potentially corrupt the BGP routing table, which could result in network instability. The vulnerability exists due to changes in the implementation of the BGP MPLS-Based Ethernet VPN RFC (RFC 7432) draft between IOS XE software releases. When the BGP Inclusive Multicast Ethernet Tag Route or BGP EVPN MAC/IP Advertisement Route update packet is received, it could be possible that the IP address length field is miscalculated. An attacker could exploit this vulnerability by sending a crafted BGP packet to an affected device after the BGP session was established. An exploit could allow the attacker to cause the affected device to reload or corrupt the BGP routing table; either outcome would result in a DoS. The vulnerability may be triggered when the router receives a crafted BGP message from a peer on an existing BGP session. This vulnerability affects all releases of Cisco IOS XE Software prior to software release 16.3 that support BGP EVPN configurations. If the device is not configured for EVPN, it is not vulnerable. Cisco Bug IDs: CSCui67191, CSCvg52875.

Source: MITRE

+View Analysis Description

claranet

Questions?

claranet | helping our customers do amazing things