RIB Reduction in Virtual Subnet

draft-xu-bess-virtual-subnet-rib-reduction-00

Xiaohu Xu (Huawei)

Susan Hares (Individual)

Yongbing Fan (China Telecom)

Christian Jacquenet (France Telecom)

Truman Boyes (Bloomberg)

Brendan Fee (Extreme Networks)

IETF92, Dallas

Motivation

- Virtual Subnet [draft-ietf-l3vpn-virtual-subnet] is intended for building L3 network virtualization overlays within or across data centers.
  - Since a subnet is extended across multiple PE routers, CE host routes need to be exchanged among PE routers. As a result, the resulting RIB/FIB size of PE routers may become a major concern in large-scale data center environments.
- [draft-ietf-bess-virtual-subnet-fib-reduction] introduces a method to reduce the FIB size of PE routers.
- This draft describes a method for further reducing the RIB size of PE routers, which is beneficial in the case where PE routers don't need to maintain all remote CE host routes on the control plane.
  - Remote CE host routes are learnt from RR on demand while remote subnet routes are learnt from RR as before.

Steps to Reduce RIBs in Virtual Subnet

- PE routers as RR clients advertise host routes for their local CE hosts to the RR. Meanwhile, PE routers notify the RR not to advertise any host route to them by using L3VPN Prefix ORF [draft-xu-bess-l3vpn-prefix-orf]
  - i.e., only requesting L3VPN routes with prefix length shorter than a /32 (VPNv4 case) or /128 (VPNv6 case).
- RR is configured with null routes for more specific subnets (e.g., 1.1.1.0/25 and 1.1.1.128/25) corresponding to the extended subnet (e.g., 1.1.1.0/24) and then redistributes these routes to BGP.
  - In the case where the RR is not available for transferring L3VPN traffic (e.g., the RR is running on a server), a PE router with a full routing table could advertise the above more specific subnet routes instead.
- Upon receiving a packet destined for a remote CE host, ingress PE router will forward the packet to the RR, which in turn forwards it to the egress PE.

On-demand Learning of Remote CE Host Route

- To avoid any potential path stretch penalty, PE routers could perform on-demand learning of remote CE host routes from the RR by using L3VPN Prefix ORF upon receiving:
  - An ARP request or Neighbor Solicitation (NS) message from a local CE host without matching CE host route for the target host.
  - A packet matching one of the more specific subnet routes (e.g., 1.1.1.0/25 and 1.1.1.128/25) learnt from the RR.
- Remote CE host routes would expire if they have not been used for packet forwarding for a certain period of time.
  - Once the expiration time for a given CE host route is approaching, PE routers would notify the RR to remove the corresponding L3VPN Prefix ORF entry.

RIB Reduction in Spine-Leaf Topology

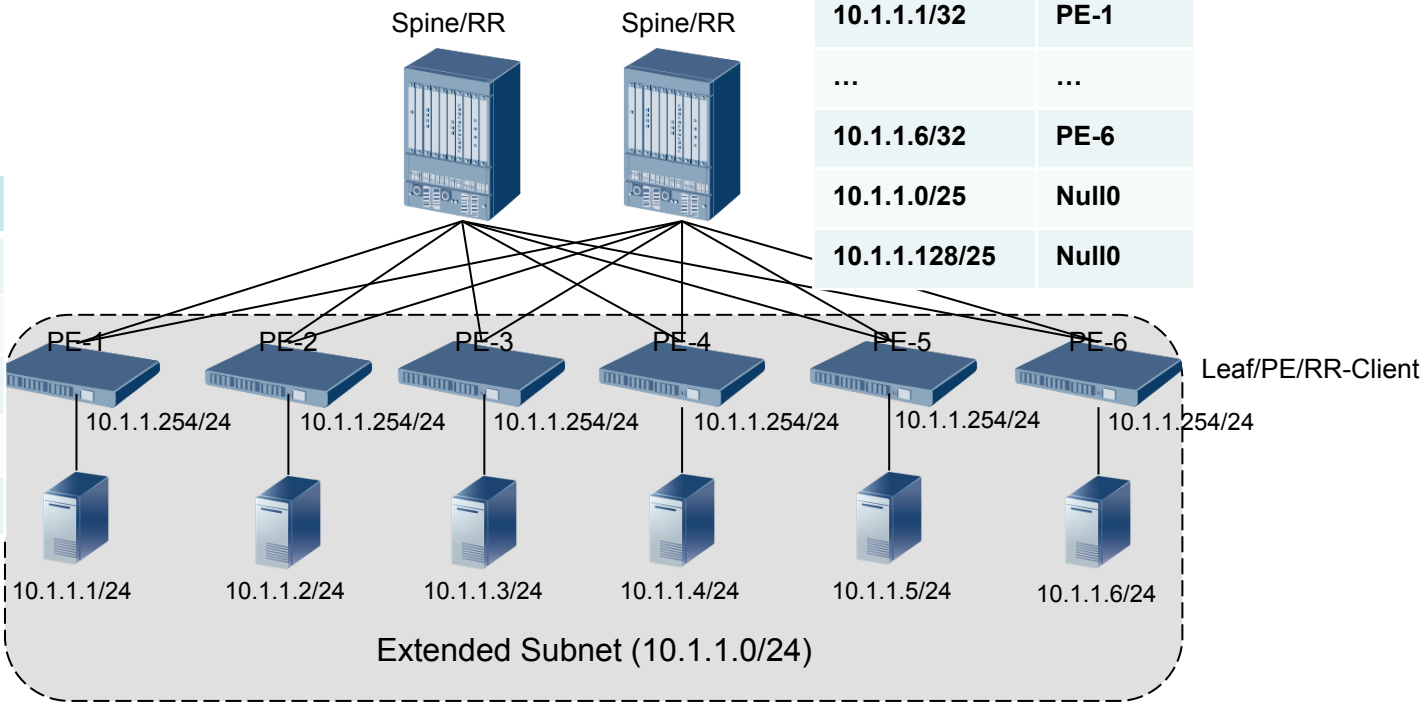
- In the spine-leaf topology, there is no need to enable the on-demand learning of remote CE host routes since those packets destined for remote CE hosts would have to traverse one of the spine nodes anyway.
  - PE routers just need to learn remote subnet routes, rather than remote host routes from RRs.

RIB of RR:

Prefix	Nexthop
10.1.1.1/32	PE-1
…	…
10.1.1.6/32	PE-6
10.1.1.0/25	Null0
10.1.1.128/25	Null0

RIB of PE-1

Prefix	Nexthop
10.1.1.1/32	10.1.1.1
10.1.1.254/32	127.0.0.1
10.1.1.0/24	10.1.1.254
10.1.1.0/25	RR
10.1.1.128/25	RR



Next Steps

- WG adoption as an informational draft?