Virtual Subnet (VS): A Scalable Data Center Interconnection Solution

draft-xu-virtual-subnet-05

Xiaohu Xu (xuxh@huawei.com)

NANOG52, Denver

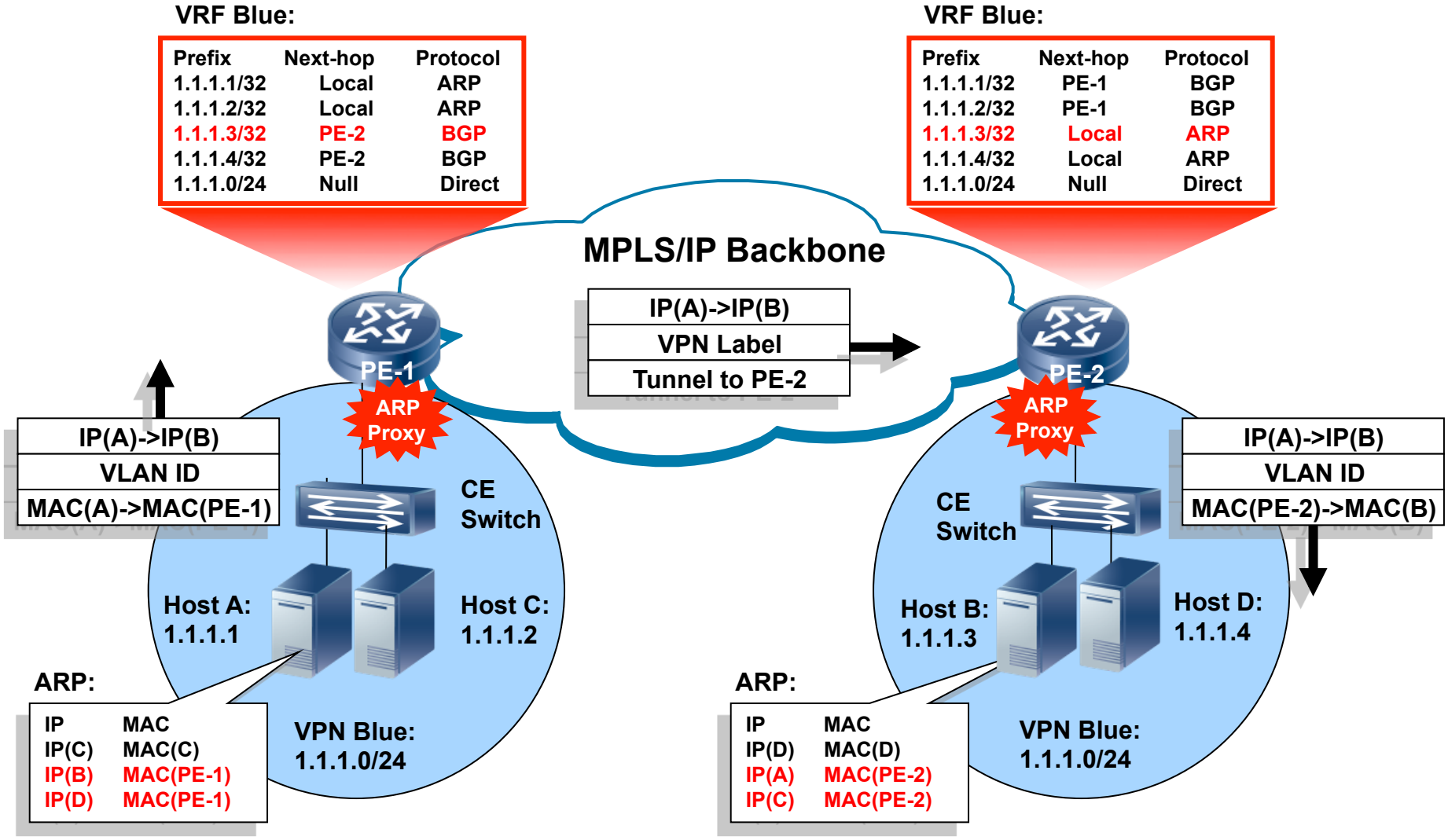
Requirements for Data Center Interconnection

- To interconnect geographically dispersed cloud data centers, a highly reliable and scalable L2VPN solution across the WAN is essential.
- Detailed requirements include but not limited to the following items:
  - VLAN scalability (beyond 4K VLANs)
  - MAC table scalability (especially critical for commodity CE switches)
  - Broadcast storm reduction
  - Multi-homing and load-balancing

Virtual Subnet(VS) Overview

- By reusing the proven BGP/MPLS IP VPN [RFC4364] and ARP proxy [RFC925] technologies, VS provides a scalable IP-only L2VPN service for data center interconnection.
- In contrast to the existing VPLS solution, VS has the following distinct benefits:
  - Suppressing unknown unicast and ARP broadcast traffic from propagating across sites by restricting the reach of the flood domain within a single site.
  - Reducing the MAC table size of CE switches by using ARP proxy on PE routers.
  - Achieving multi-homing and load-balancing by enabling VRRP on PE routers.

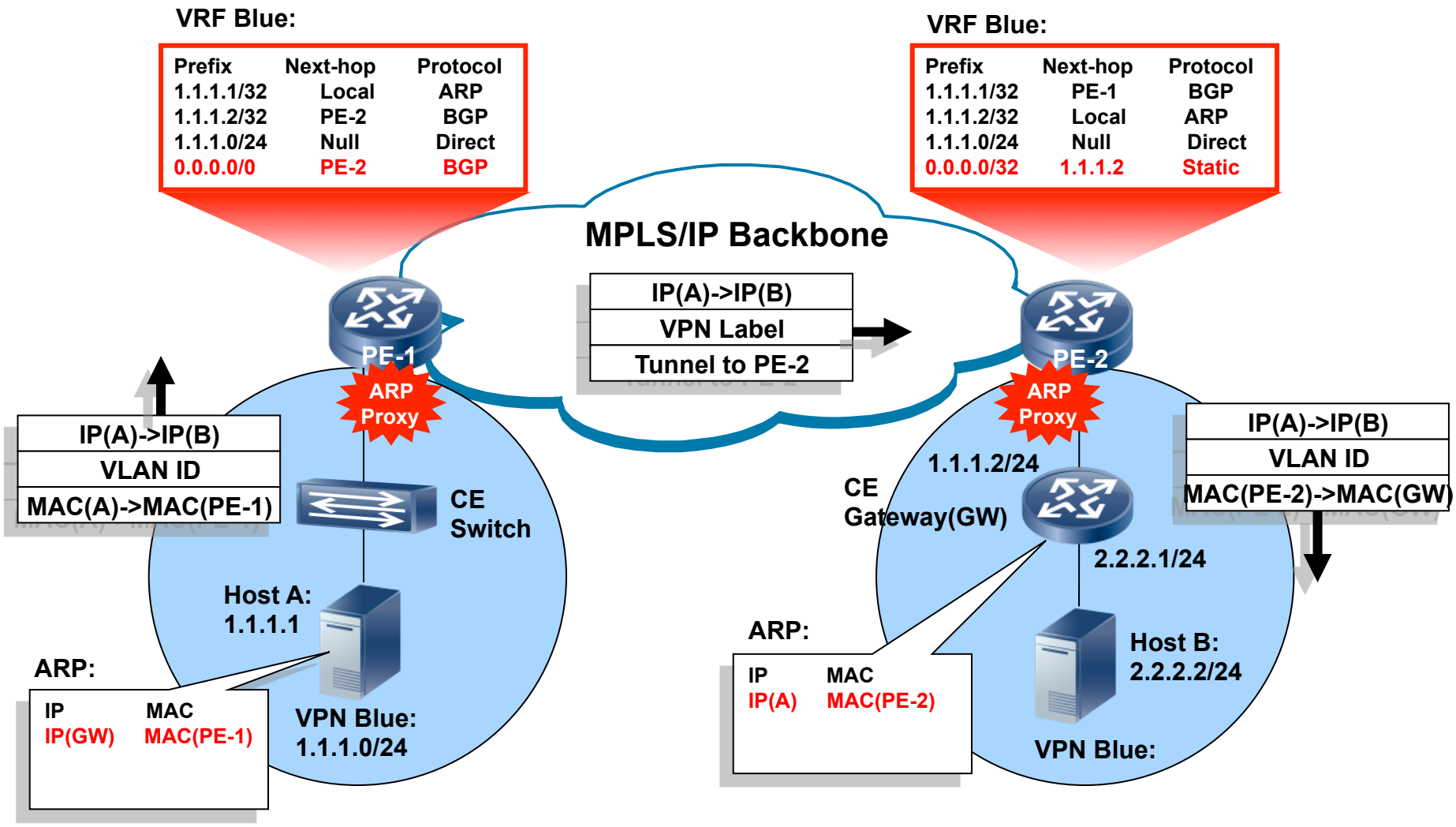
Intra-subnet Unicast



Intra-subnet Unicast

- Host routes (i.e., /32) for local CE hosts, which are generated automatically according to their corresponding ARP entries, are distributed among PE routers with the existing L3VPN signaling.
- Acting as an ARP proxy, each PE router returns its own MAC as a response to the ARP request for a remote CE host.
- Ingress PE routers tunnel received customer packets that are destined for remote CE hosts to the next-hop egress PE routers in accordance to the current L3VPN forwarding process.

Inter-subnet Unicast



Inter-subnet Unicast

- Only the PE router which is connected to a CE gateway router is entitled to announce a default route to other PE routers.
- Especially, in the CE gateway redundancy scenario where two CE routers of a given redundancy group are connected to two PE routers respectively, only the PE router which is connected to the CE router acting as the VRRP master is entitled to announce a default route.
  - The next-hop of the default route is set as the virtual router IP and that default route is not deemed as valid unless there is a directly connected host route for that next-hop address (i.e., the next-hop availability check).

CE Host Discovery

- Local CE hosts are automatically discovered by PE routers through ARP learning.
  - To keep the ARP cache entries from expiring, PE routers periodically send unicast ARP requests to the corresponding CE hosts.
- To be capable of learning all of the local CE hosts shortly after rebooting, PE routers should perform host scanning at least once :
  - E.g., PE routers could send an ICMP echo request to IP broadcast address over their VRF interfaces, every CE host receiving that ICMP request will respond with an ICMP echo reply. As a result, IP->MAC mappings (i.e., ARP entries) could be obtained.

ARP Reduction

- By enabling ARP proxy on PE routers, ARP broadcasts are strictly contained within a single site:
  - For an ARP request for a local CE host, discards it.
  - For an ARP request for a remote CE host, returns its own MAC as a response.
  - For an ARP request for an unknown CE host (i.e., no matching host route found), discards it.

CE Multi-homing

- VRRP is enabled among the PE routers of a multi-homed site and only the VRRP master is entitled to act as an ARP proxy.
- Active-active multi-homing is available for incoming traffic since all PE routers attached to a multi-homed site could advertise the corresponding host routes for their local CE hosts.

CE Mobility

- When a CE host moves from one VPN site to another,
  - The PE router attached to the current VPN site will advertise a CE host route upon receiving a gratuitous ARP request or reply from that CE host.
  - The PE router attached to the previous VPN site, upon receiving the above host route announcement, immediately sends an ARP request for that CE host to check whether that host is still connected to it.
    - If not, the PE router should delete the corresponding ARP entry and host route for that CE host, and accordingly withdrawn the corresponding BGP route advertised before.
    - Otherwise, it is judged as a case of CE multi-homing.

Multicast/Broadcast

- MVPN technologies including the ingress replication mechanism can be almost reused without any change to distribute customer multicast/broadcast traffic across sites.
  - Here the customer broadcast traffic is processed as the customer multicast traffic of a special group.

Comparison

	VPLS	VS
Unknown Unicast Flood Suppression	No	Yes
ARP Broadcast Reduction	No	Yes
MAC Table Reduction for CE Switches	No (CE switches need to learn MACs of both local and remote hosts).	Yes (CE switches only need to learn MACs of local hosts).
Active-active Multi-homing	No	Yes (for incoming traffic)
PE Failover without Performance Damage	No (triggered MAC withdraw causes unknown unicast traffic flood across sites for a short period of time)	Yes (will not cause traffic flood).

Thank You !