Designing Multi-Tenant Data Centers using EVPN-IRB

Neeraj Malhotra, Principal Engineer, Cisco

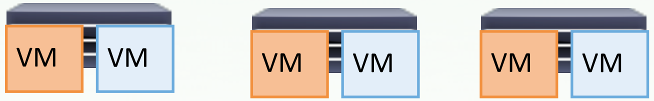
Ahmed Abeer, Technical Marketing Engineer, Cisco

Where We Are

Compute and NF Virtualization

SDN trying to achieve end to end Operational simplicity and programmability

Tenant Workloads, NFVs Spawned Anywhere



Workloads are Mobile

Situation

L2 Switched DC Fabrics Designed for Physical Compute

Disjoint control planes and data planes Across L2, L3, DC and WAN

Workload location determined by VLAN location

Immobile Workloads

Centralized east-west routing
Scale Bottleneck, single point of failure

No Traffic Steering, ECMP, FRR

Flood and Learn is sub-optimal

Complication

Network Fabric becomes the bottleneck

End to End Operational Simplicity and Programmability cannot be achieved

Sub-optimal BW and compute usage

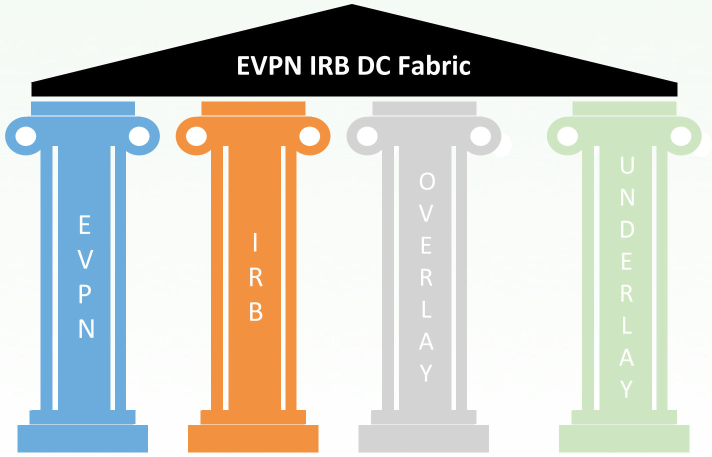
No flexible workload placement, mobility

Loss of Competitive Advantage

Implication

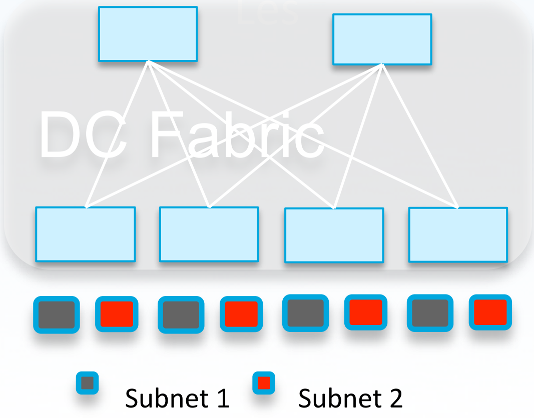
Where We Must Go

- L3 Underlay DC Fabric
- VPN Overlay based on EVPN-IRB
- Distributed any-cast routing architecture

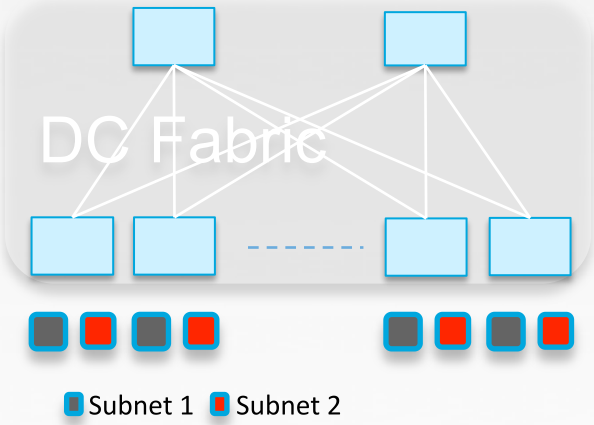


Proposal

1. Learn and evaluate the solution by starting with a small DC
2. Scale up horizontally



Action



- Unified control plane and data plane across DC and WAN
- End to End Operational Simplicity and Programmability

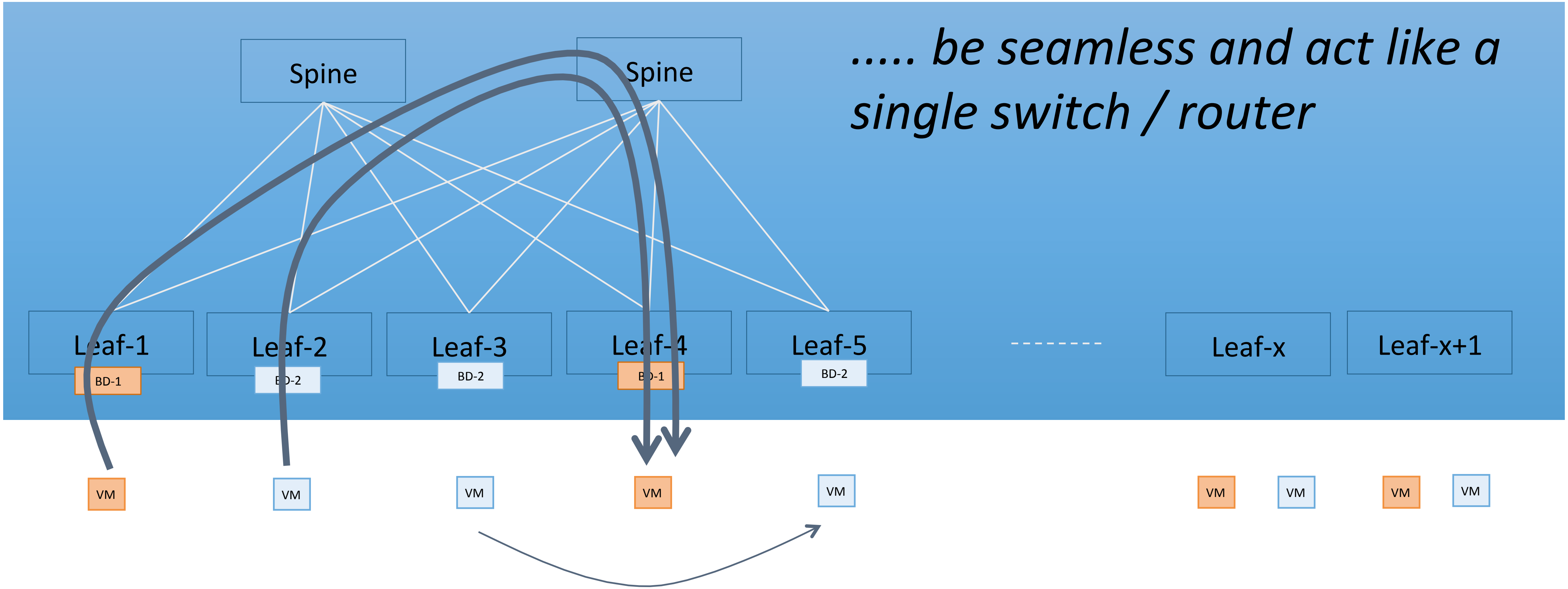
Benefit

Objectives

Architecture Objectives – Evolving DC Requirements

- Operational simplicity via uniform control, data plane across L2, L3, DC, WAN
- Flexible workload placement and mobility within DC and across DCs
- Efficient bandwidth utilization within DC – no flood and learn, ECMP
- Traffic engineering - traffic steering, ECMP, FRR
- Horizontal Scaling
- Multi-tenancy with L2 and L3 VPN in DC
- Interworking with Legacy L3VPN / L2VPN WAN

A DC network fabric must .....



Why not VPLS?

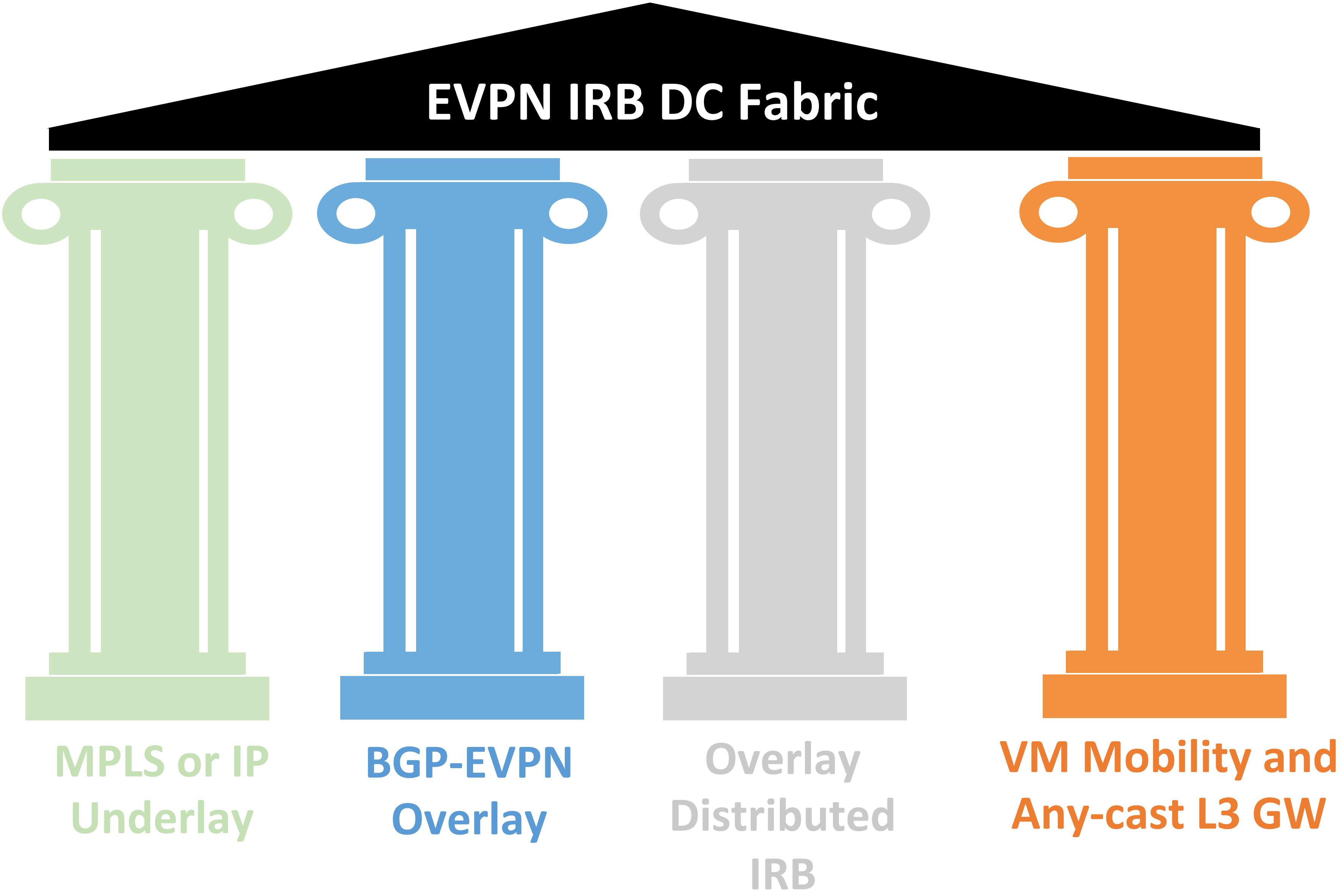
Why not use VPLS in DC?

Simply not designed for DC use-case

- L2 Only
- No All-Active Redundancy
- No per-flow ECMP Load-balancing
- Flood and Learn MAC learning Is Sub-optimal

What is the Solution?

Fabric Solution Components



IP or MPLS Underlay

Underlay vs. Overlay

Underlay = Transport

Physical Network
IP, MPLS / SR Transport

Traffic Steering, ECMP, FRR,.....

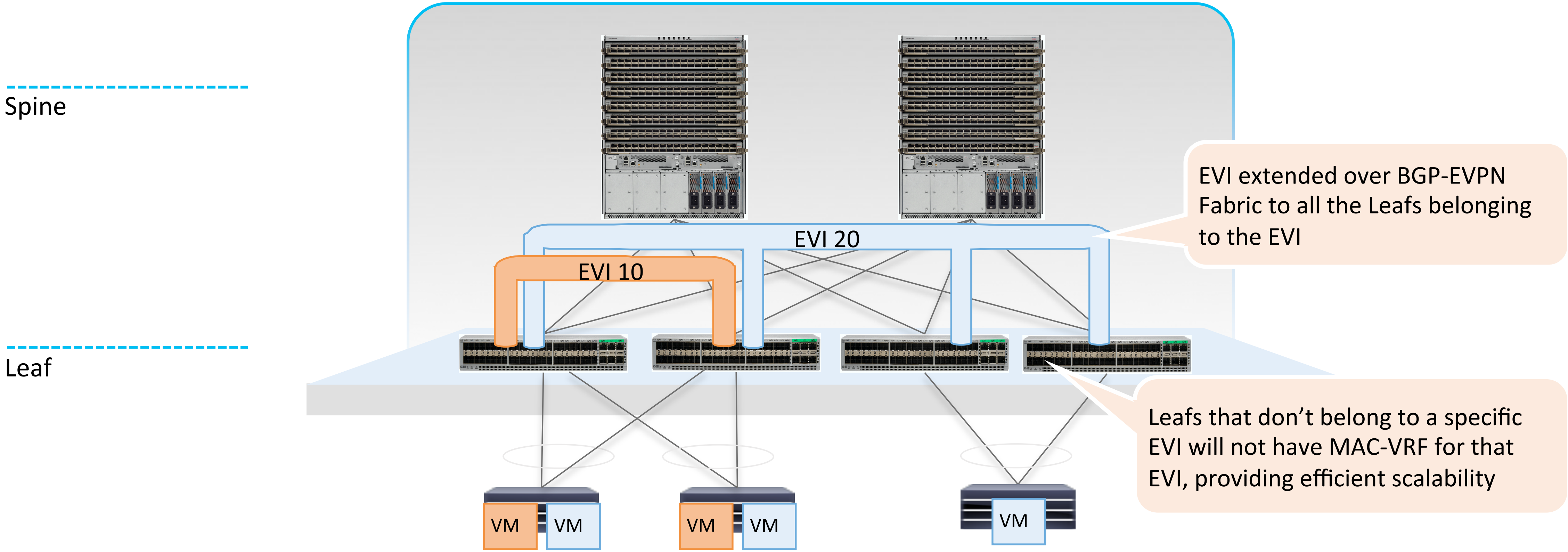
Overlay = VPN (L2+L3)

Control Plane – EVPN
Data Plane – MPLS, VXLAN,.....

Policy Driven

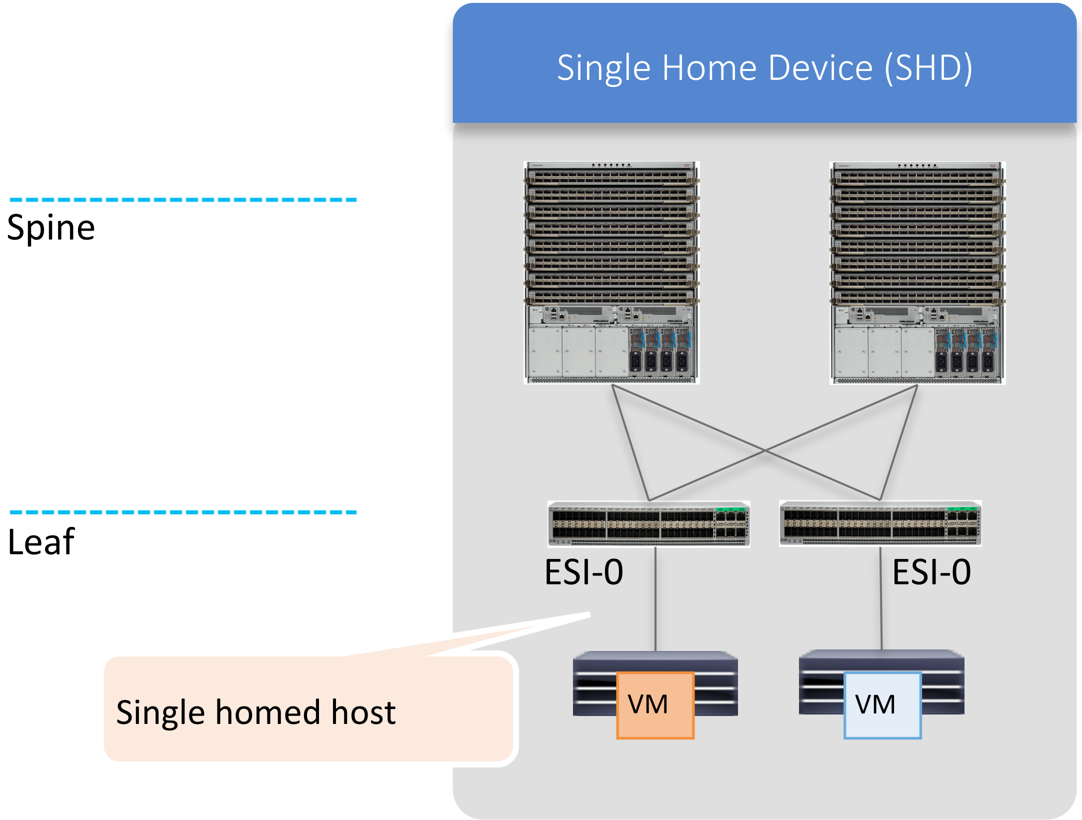
Overlay Control Plane – BGP EVPN

BGP EVPN – EVI

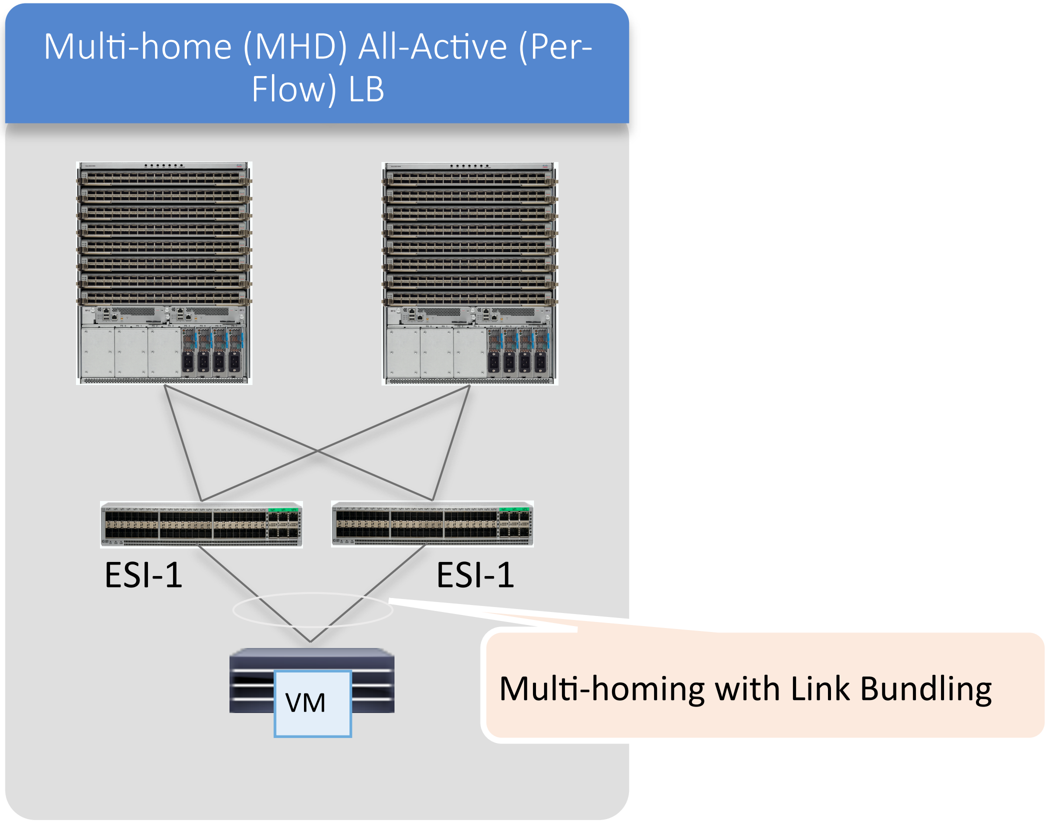


EVI: An EVPN instance extends Layer 2 between the Leafs

BGP EVPN – Host Connectivity Options, ESI



- Ethernet Segment Identifier (ESI) ‘0’
- No DF election

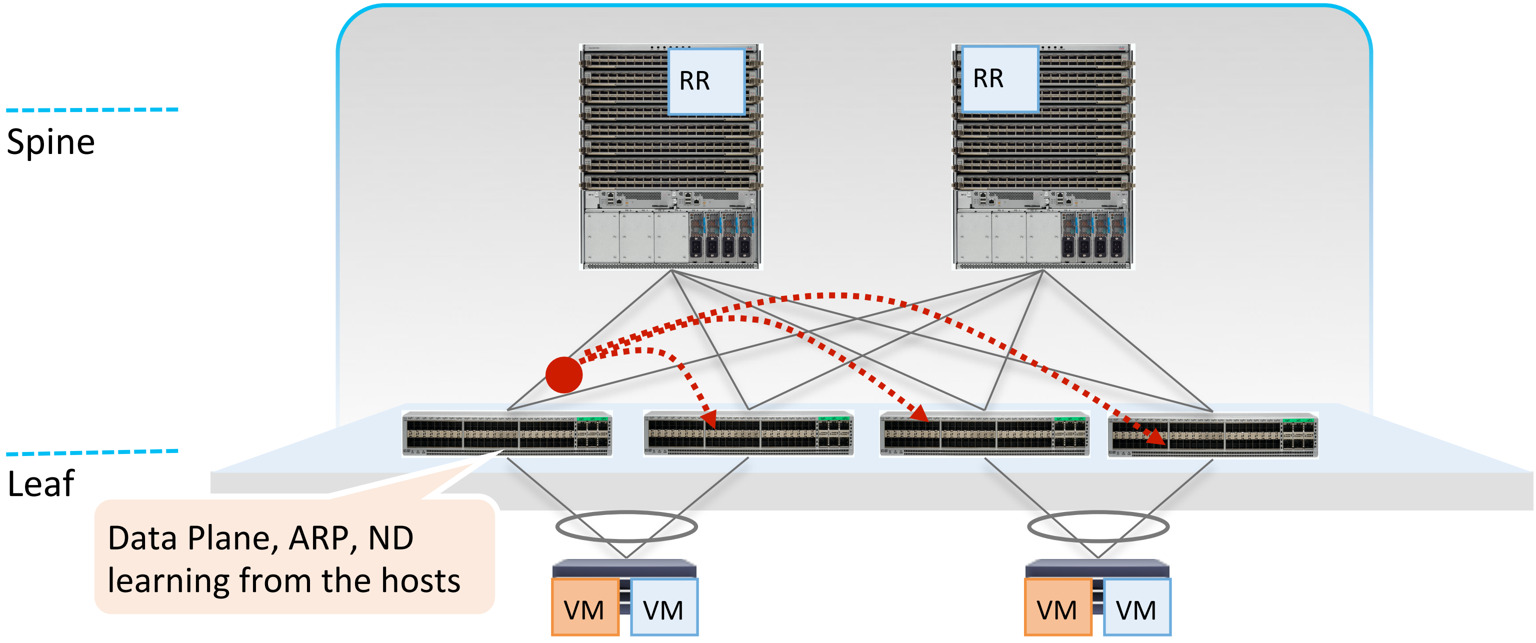


- Identical ESI on Leafs
- Per VLAN DF election

BGP EVPN – MAC and IP Learning

- MAC/IP addresses are advertised along with L2 and L3 VPN encap (MPLS label or VNID) to rest of Leafs via MAC+IP RT-2
- IP Prefix routes are advertised via BGP EVPN via RT-5

EVPN Route Type 2 carries MAC and IP reachability with L2+L3 VPN encapsulation, L2+L3 RTs

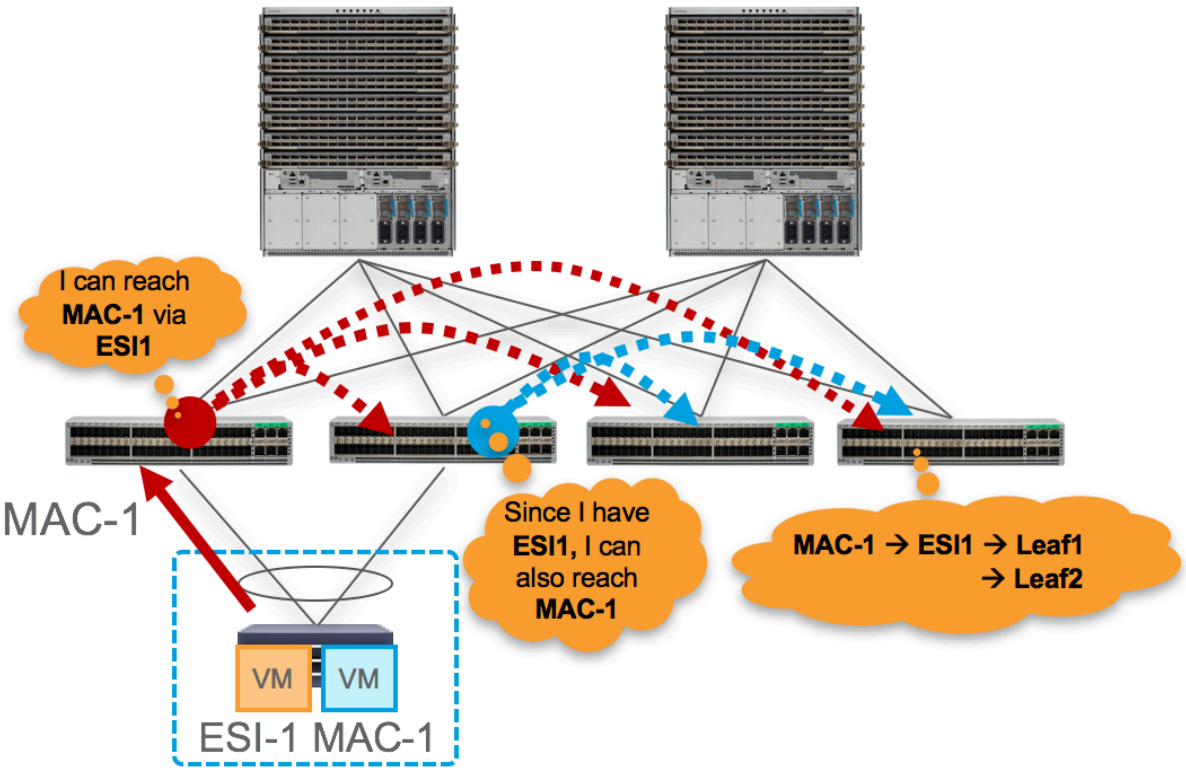


RD
Ethernet Segment Identifier
Ethernet Tag ID
MAC Address Length
MAC Address
IP Address Length
IP Address
MPLS Label1
MPLS Label2

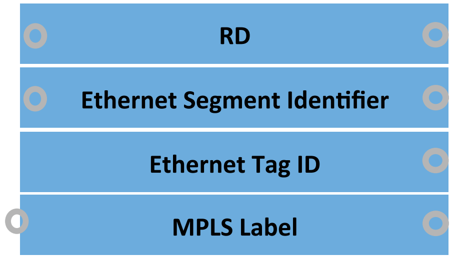
BGP EVPN – Load Balancing via Aliasing

Challenge:

How to load-balance traffic towards a multi-homed device across multiple Leafs when MAC addresses are learnt by only a single Leaf?



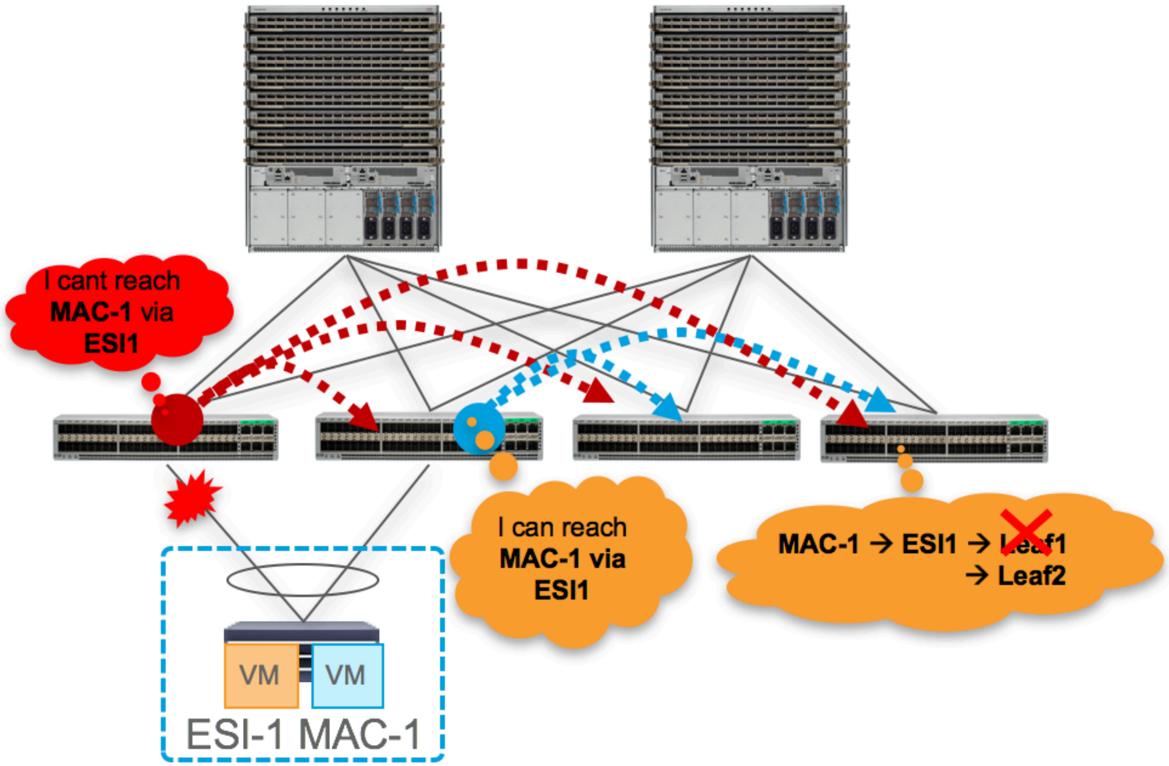
EVPN Route Type 1 advertises ESI reachability per-EVI to enable MAC ECMP without an explicit MAC route advertisement



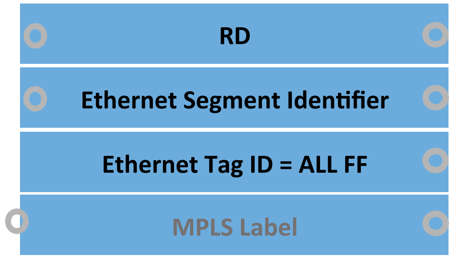
BGP EVPN – Fast Convergence via Mass-Withdraw

Challenge:

How to inform other Leafs of a failure affecting many MAC addresses quickly while the control-plane re-converges?

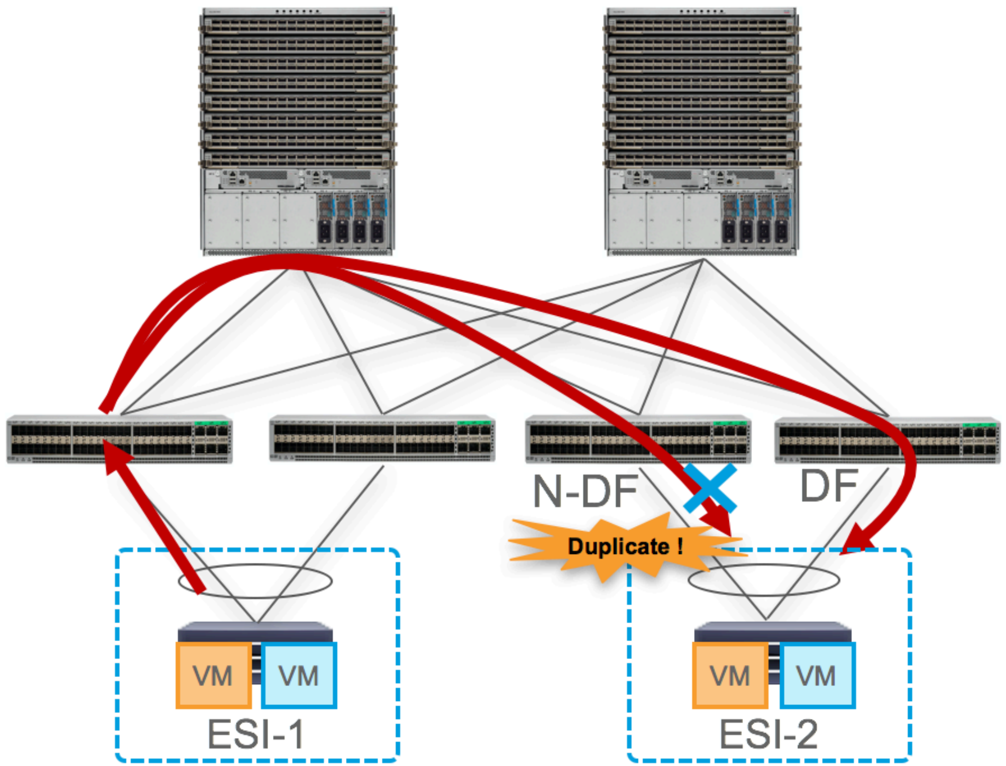


EVPN Route Type 1 also advertises ESI reachability globally for ALL EVIs to enable MAC independent convergence on ESI failure



BGP EVPN - Designated Forwarder (DF)

Challenge:
How to prevent duplicate copies of flooded traffic from being delivered to a multi-homed Ethernet Segment?

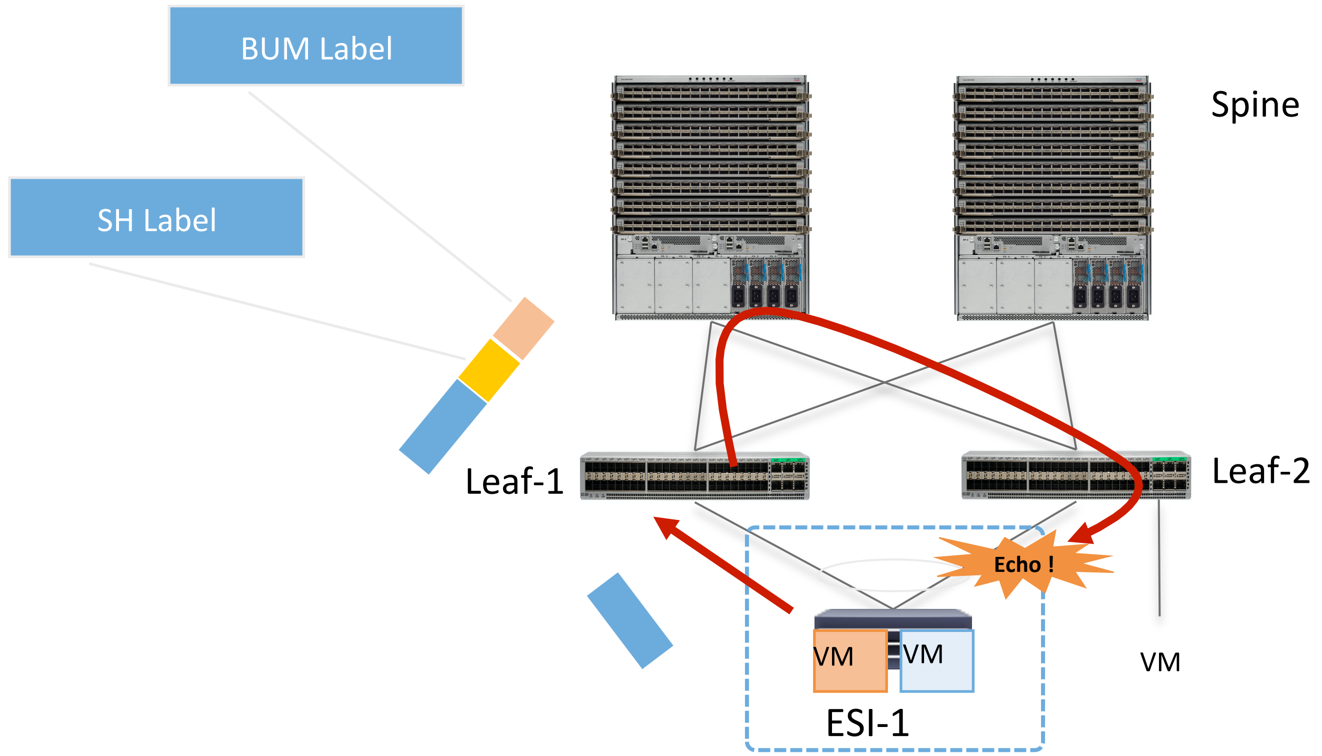


EVPN Route Type 4 enables ESI discovery and DF election

RD
Ethernet Segment Identifier
IP Address Length
Originating Router’s IP add.

BGP EVPN - Split Horizon Group Filtering

Challenge:
How to prevent flooded traffic from echoing back to a multi-homed Ethernet Segment?



Per- ESI SHG Label EXT-COMM with EVPN RT-1 enables SHG filtering to cut potential loops back to same ESI

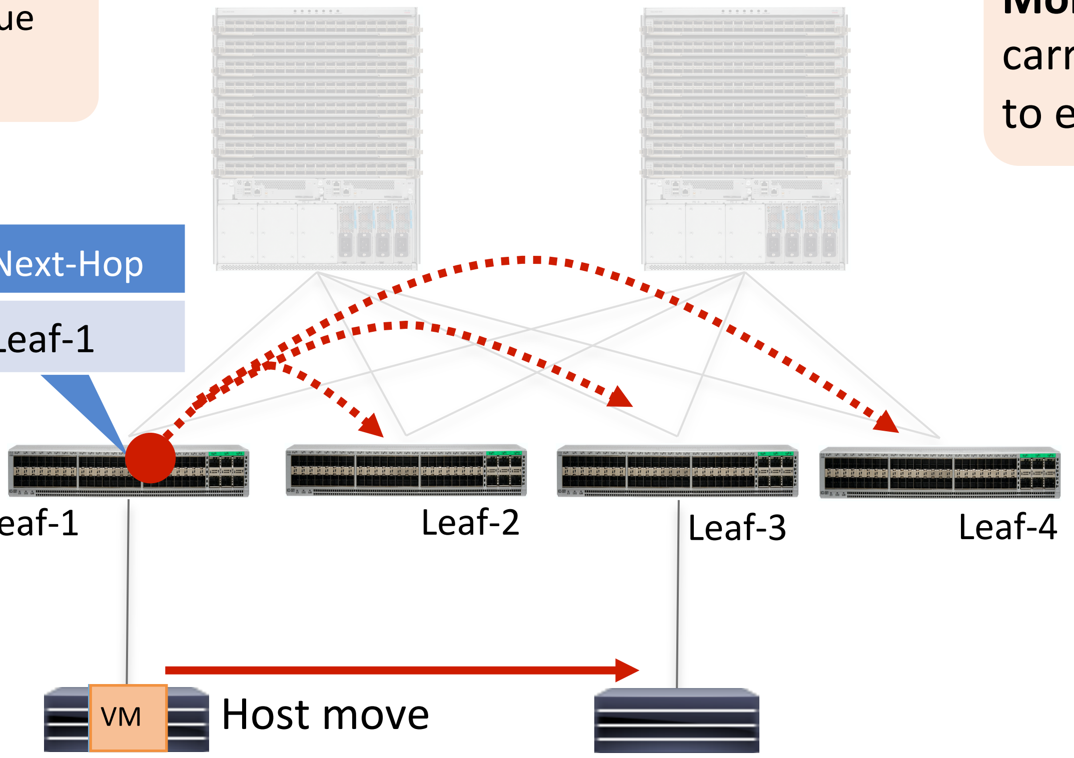
0x06
0x01
Flags
Reserved
ESI MPLS Label

VM Mobility – MAC + IP

Challenge:
How to detect the correct location of MAC after the movement of host from one Ethernet Segment to another also called "MAC move"?

Sequence number and Next-Hop value will be changed after the host move

MAC	IP	ESI	Seq.	Next-Hop
MAC-1	IP-1	0	0	Leaf-1

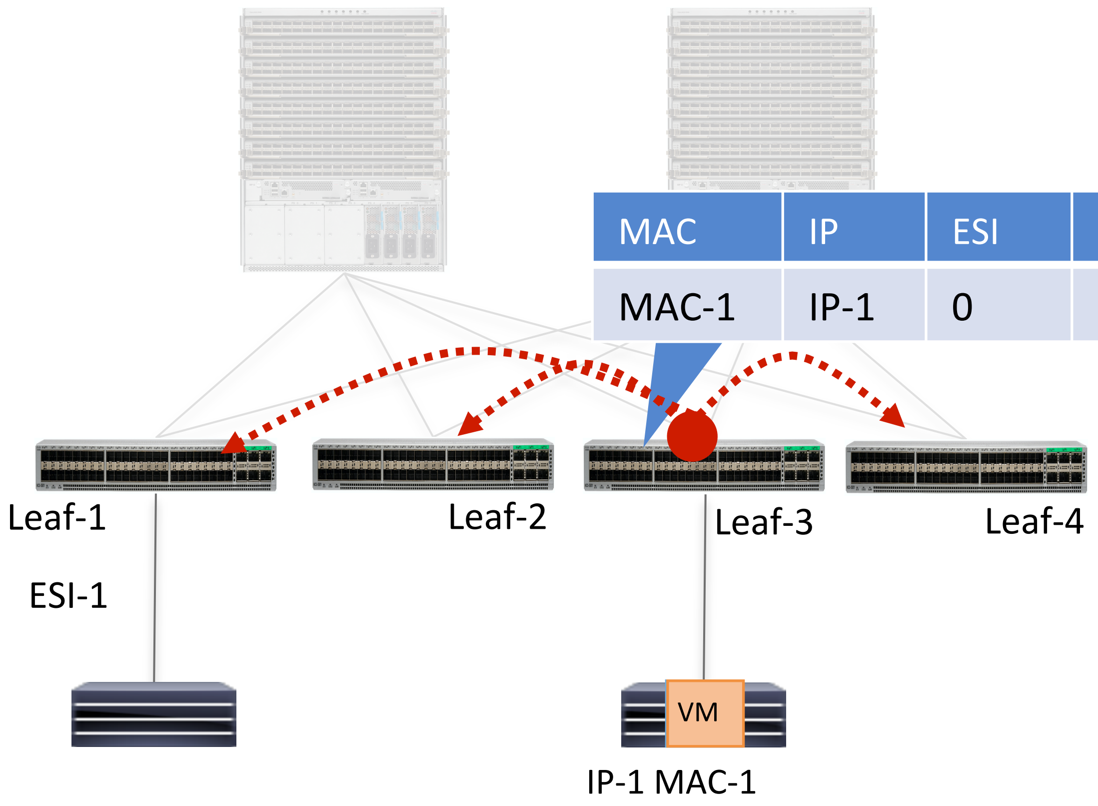


Mobility EXT-COMM with EVPN RT-2 carries MAC+IP route sequence number to enable MAC mobility

0x06
0x00
Reserved
Sequence Number

VM Mobility, continued

MAC	IP	ESI	Seq.	Next-Hop
MAC-1	IP-1	0	**1**	**Leaf-3**



Sequence number is incremented and Next-hop is changed to Leaf-3

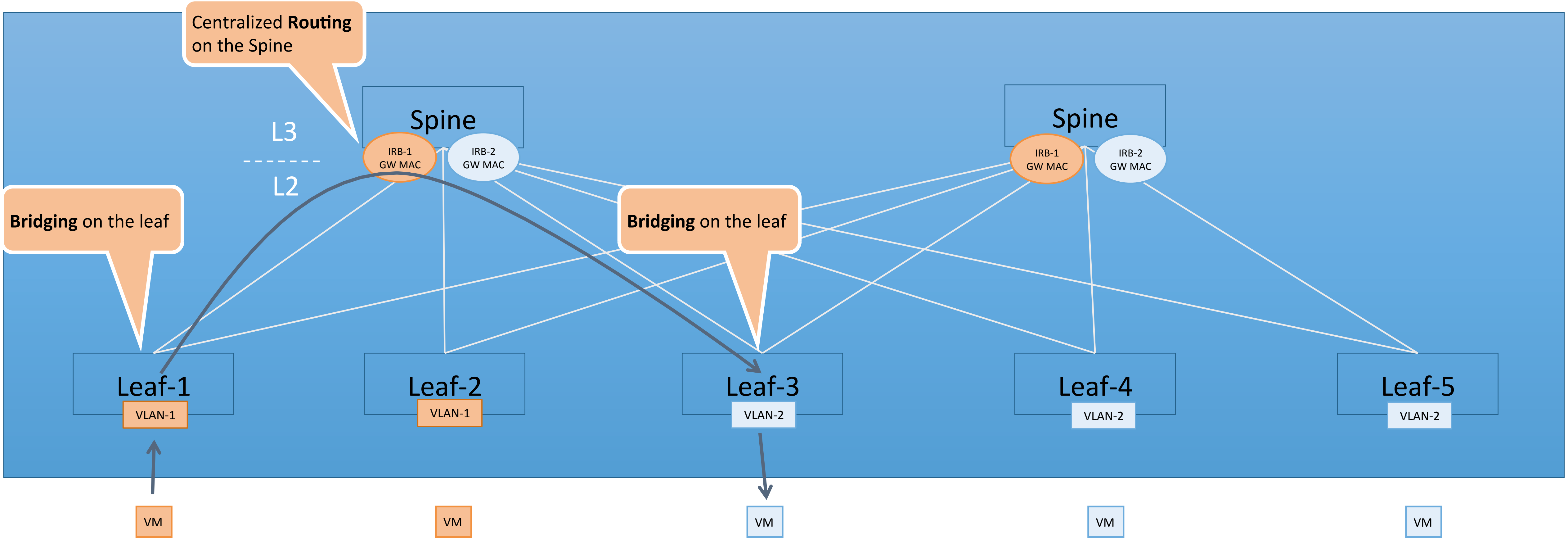
Overlay Integrated Routing and Bridging (IRB)

How do we do inter-subnet routing?

Overlay Routing Architectures

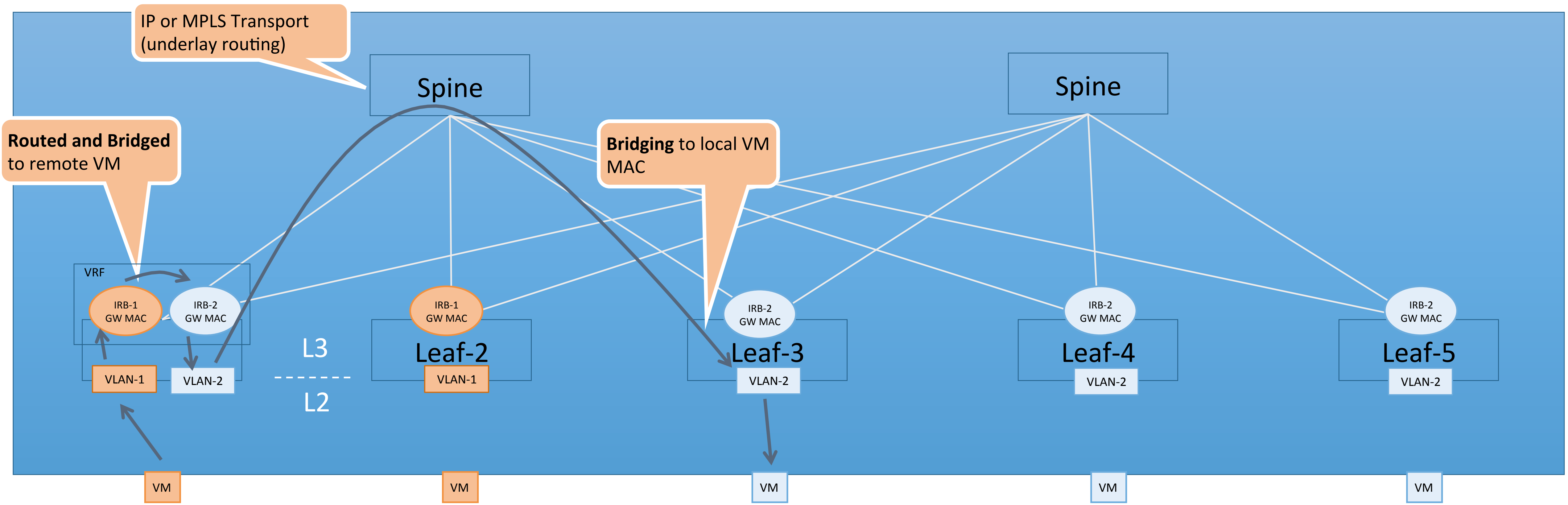
- Centralized Routing
- Distributed Routing – Asymmetric IRB
- Distributed Routing – Symmetric IRB

Centralized Routing



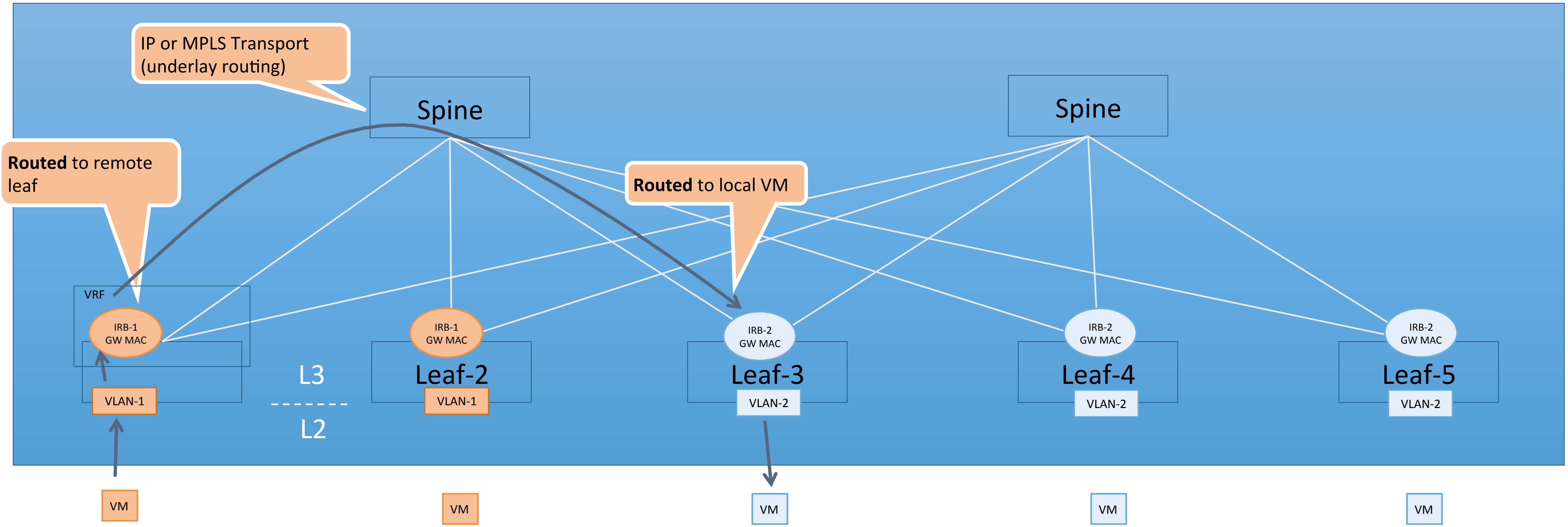
- east<->west routed traffic traverses to centralized L3 gateways
- ***Scale bottleneck:***
  - ***Centralized have full ARP/MAC state in the DC***
  - ***Centralized GW needs to host all DC subnets***

Distributed Routing – Asymmetric IRB



- Egress subnet is always local
- Inter-subnet packets routed directly to destination VM's DMAC
- ***Scale bottleneck:***
  - ***All egress subnets needs to be present on ingress leaf***
  - ***Ingress leaf maintains ARP/ND state every egress leaf***

Distributed Routing – Symmetric IRB

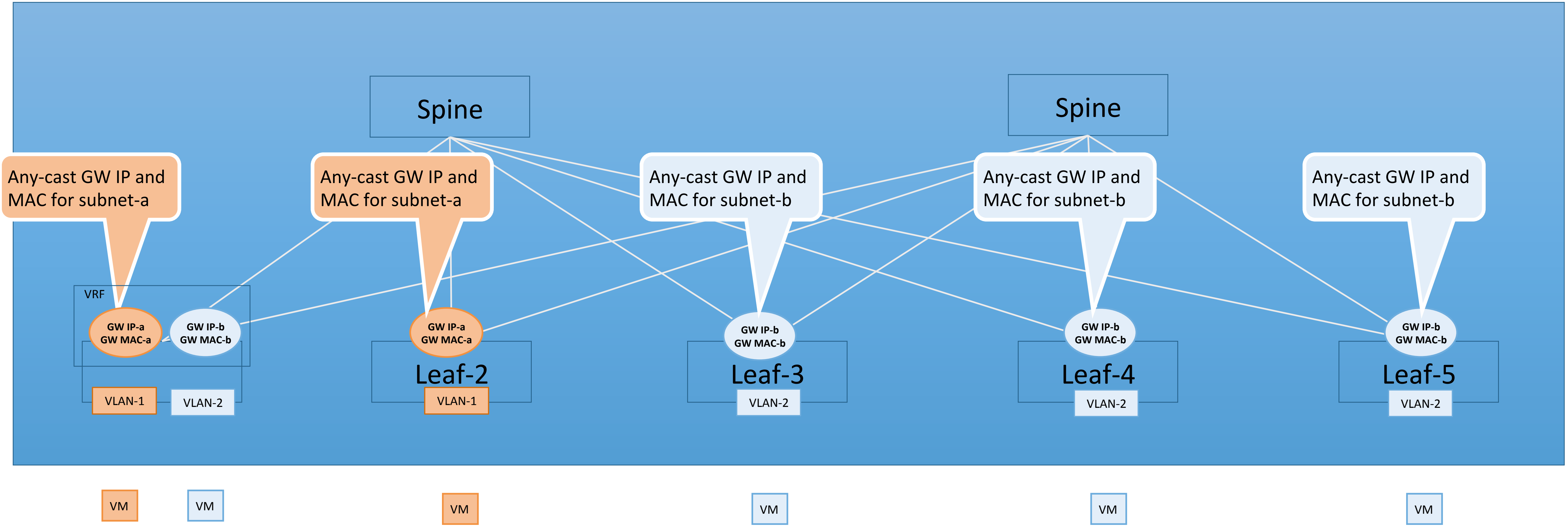


- *Remote VM IP is installed like a VPN IP route recursively over remote leaf next-hop*
- *No adjacencies to remote hosts even if the subnet is local*
- *Subnet does not need to be local on ingress leaf unless there are local hosts*

Overlay Distributed *Any-cast* GW

How do we let hosts move?

Symmetric IRB – Distributed **Any-cast** GW

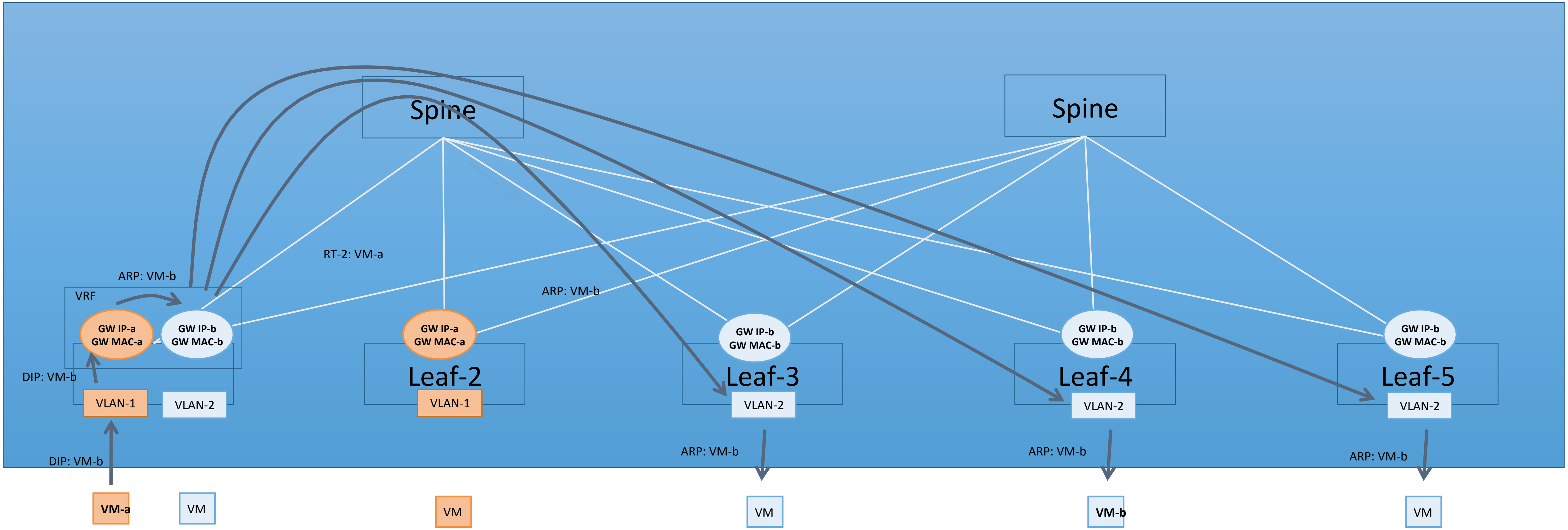


- Any-cast GW IP and Any-cast GW MAC configured on ALL leafs with local subnet
- Essentially, Subnet GW is distributed across ALL leafs with local subnet

Control and Data Plane Call Flow

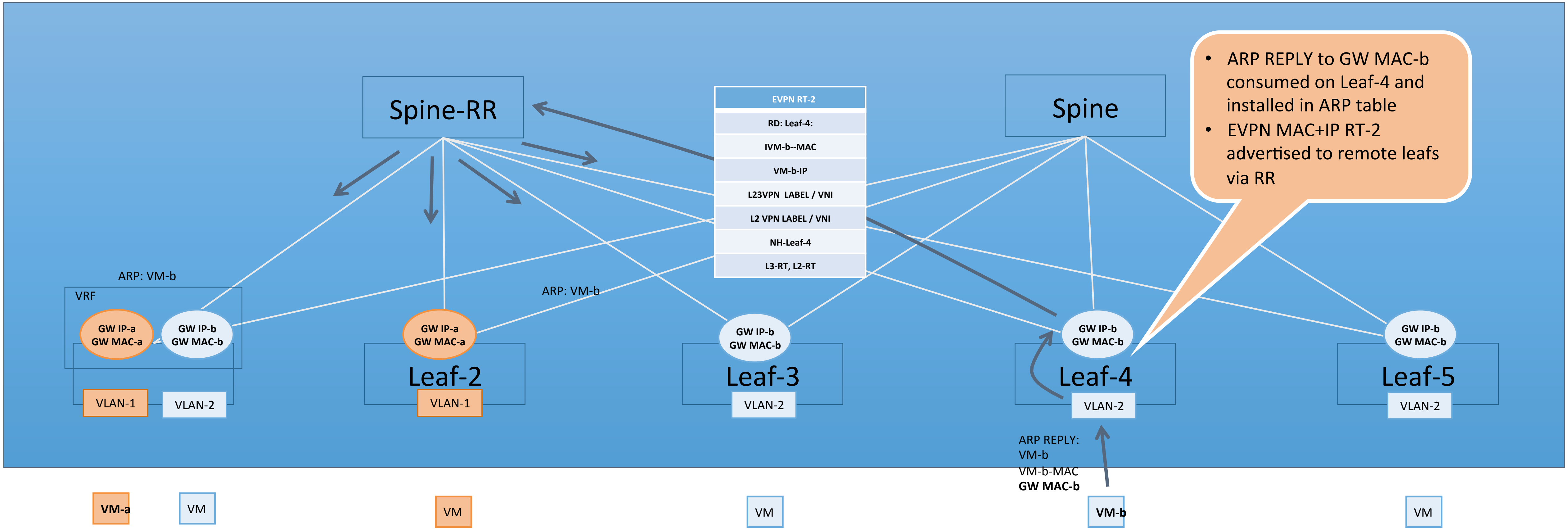
Putting it all together.....

Host Learning - ARP REQUEST contd.



1. IP packet destined to VM-b triggers ARP for VM-b on Leaf-1 from any-cast GW IP-b and any-cast GW MAC-b
2. ARP to VM-b flooded to all remote leafs where VLAN-b is stretched (via EVPN RT-3 enabled IR)
3. Leafs flood on local BD ports

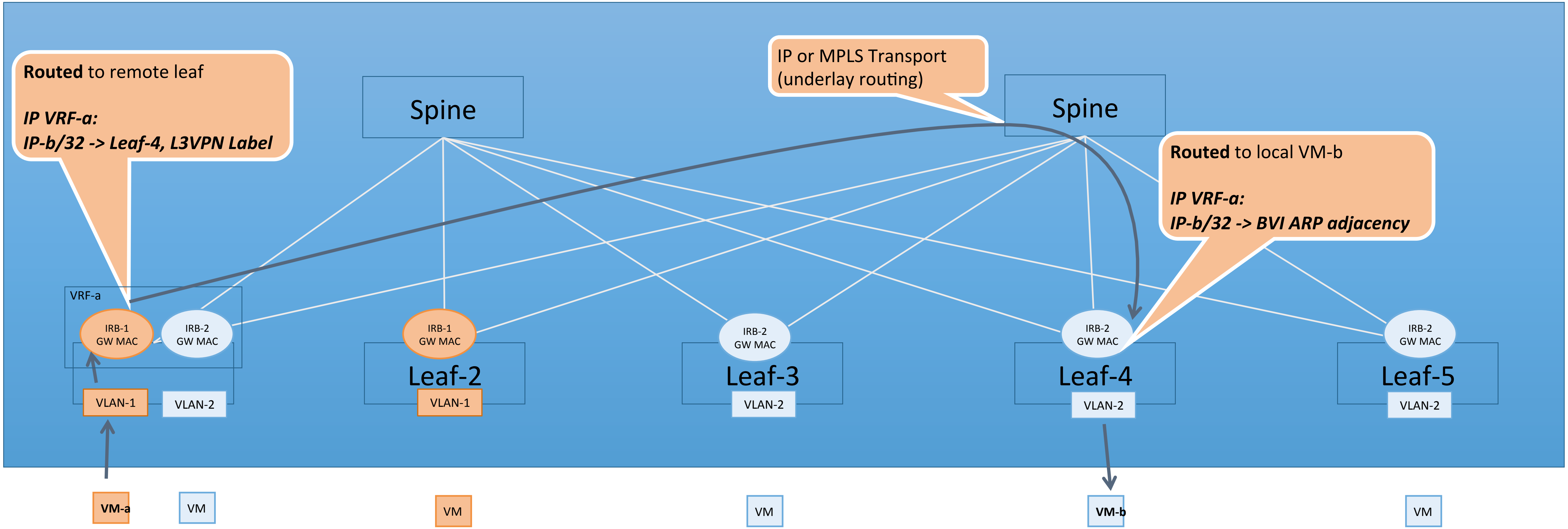
Host Learning – ARP REPLY, MAC+IP RT-2



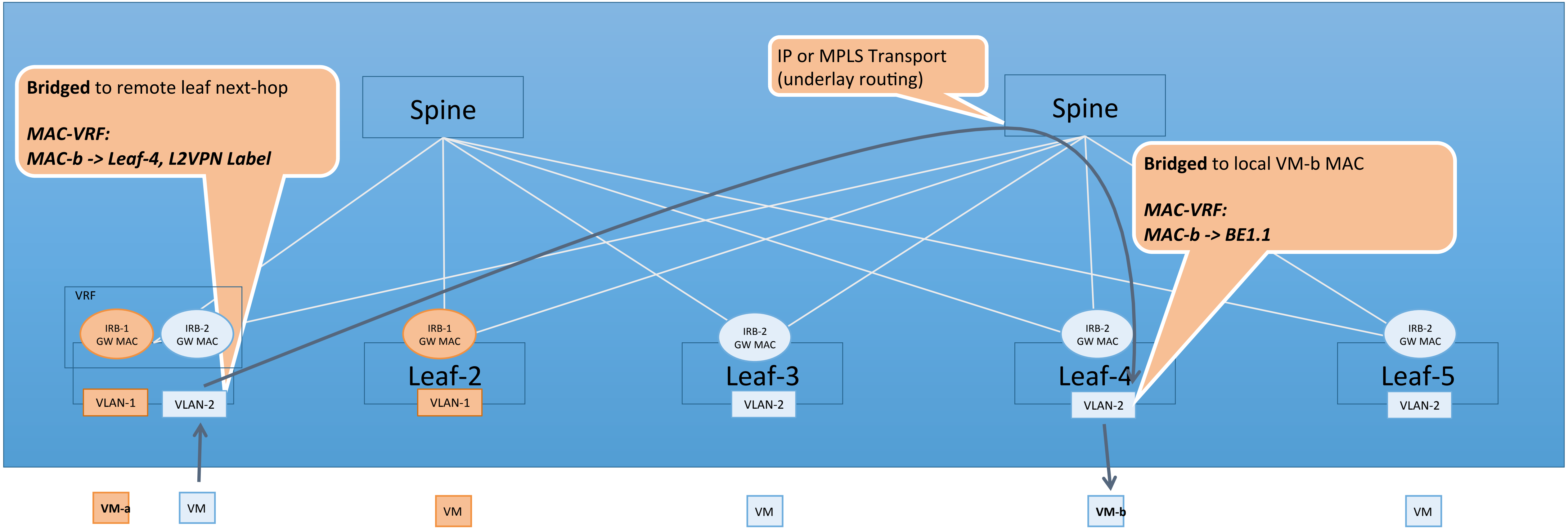
VM-b MAC Reachability installed in MAC-VRF across remote leafs

VM-b IP Reachability installed in IP-VRF across remote leafs as BGP L3VPN route independent of subnet being local or not

Inter-subnet traffic to VM-b



Intra-subnet traffic to VM-b



Summary

- Unified control, data plane across L2, L3, DC, WAN
- Flexible workload placement and mobility across L2 Overlay
- Optimal bandwidth utilization – no flood and learn, ECMP in overlay, underlay
- Traffic engineering with MPLS fabric - traffic steering, ECMP, FRR
- Horizontal Scaling with distributed symmetric IRB
- Multi-tenancy with L2 and L3 VPN
- Interworking with Legacy L3VPN / L2VPN WAN

Thank You

nmalhotr@cisco.com
aabeer@cisco.com