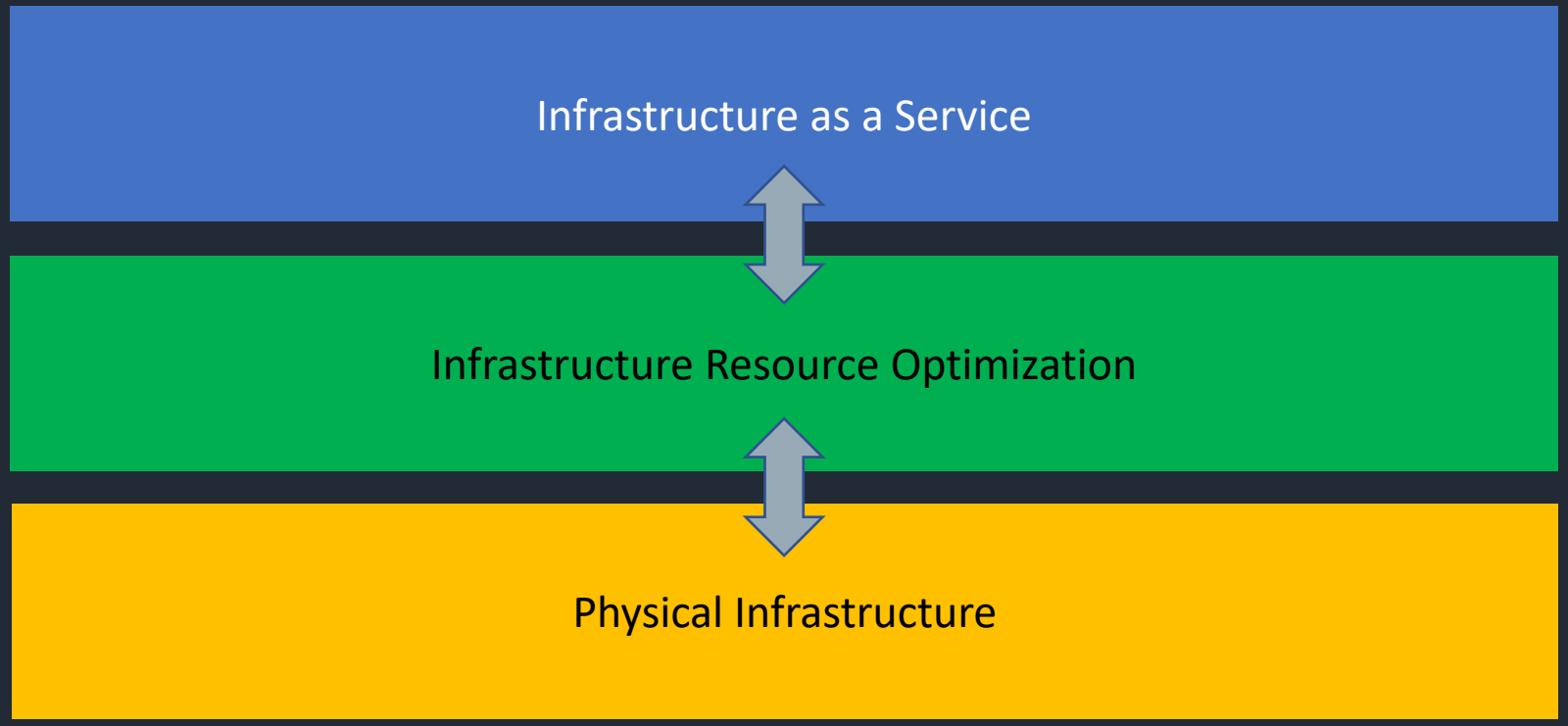
EVPN VXLAN Demystified

Aldrin Isaac

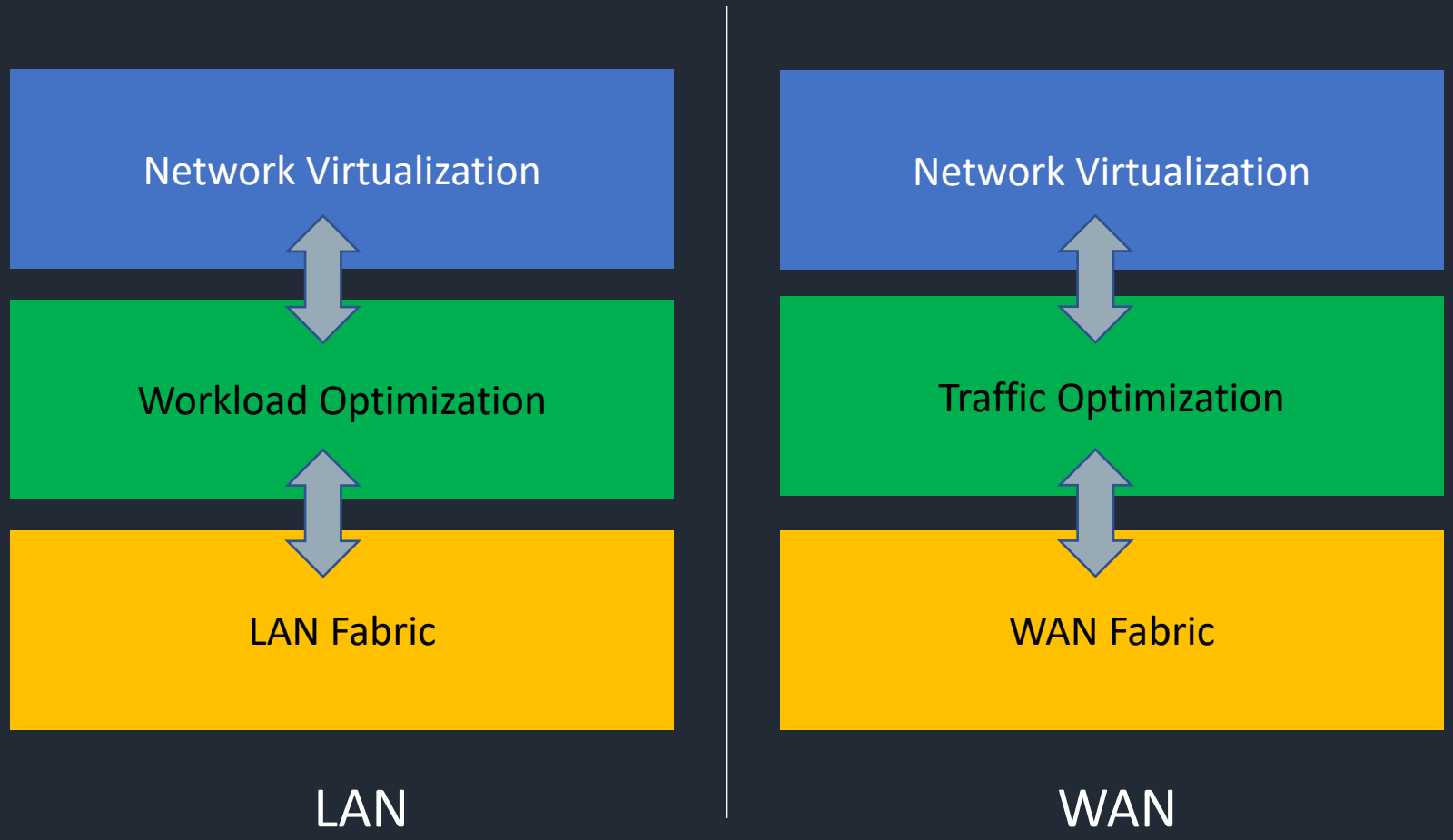
Co-author RFC7432

Juniper Networks

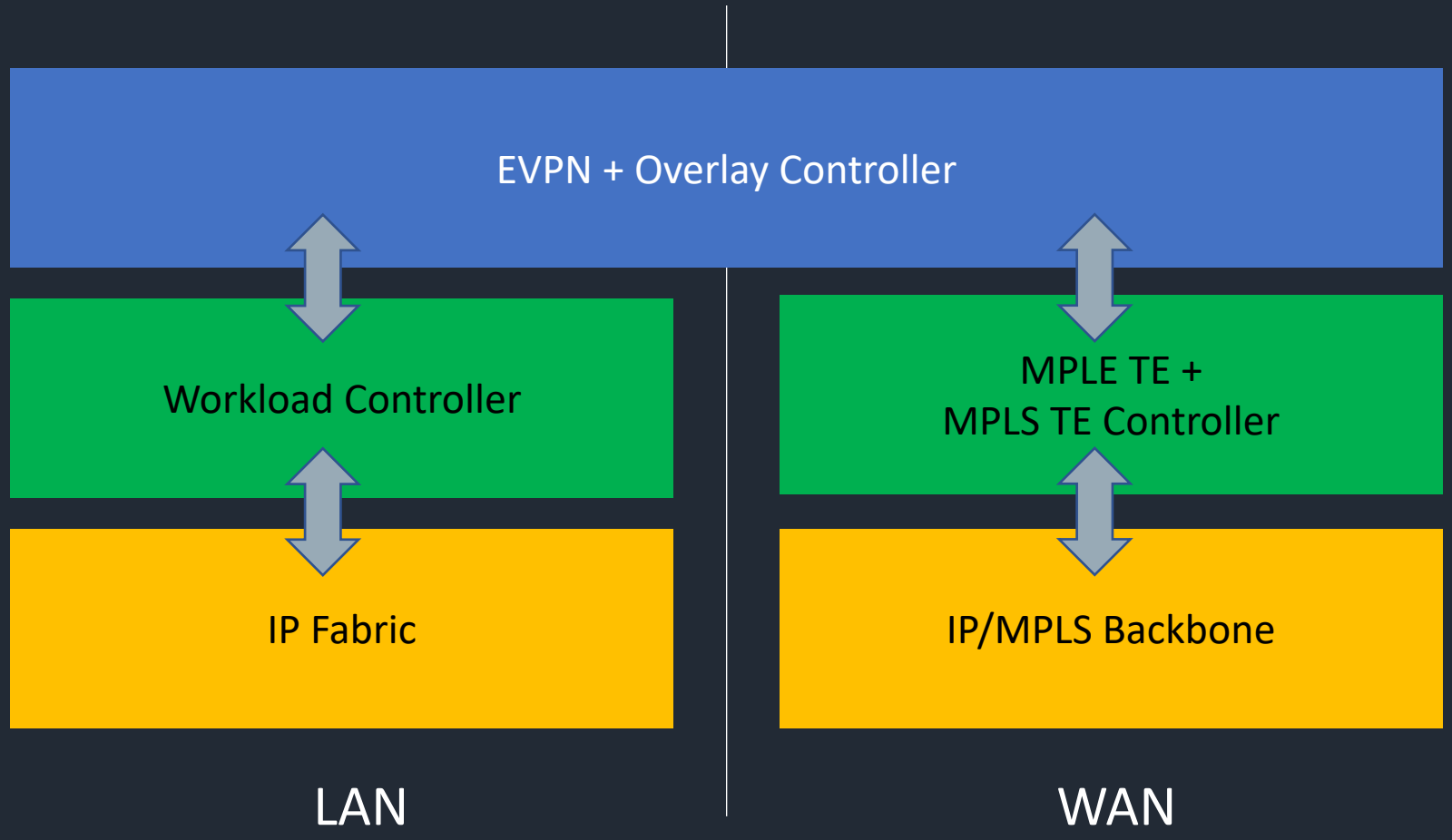
Infrastructure Subsystems -- Requirements



Network Subsystems -- Requirements

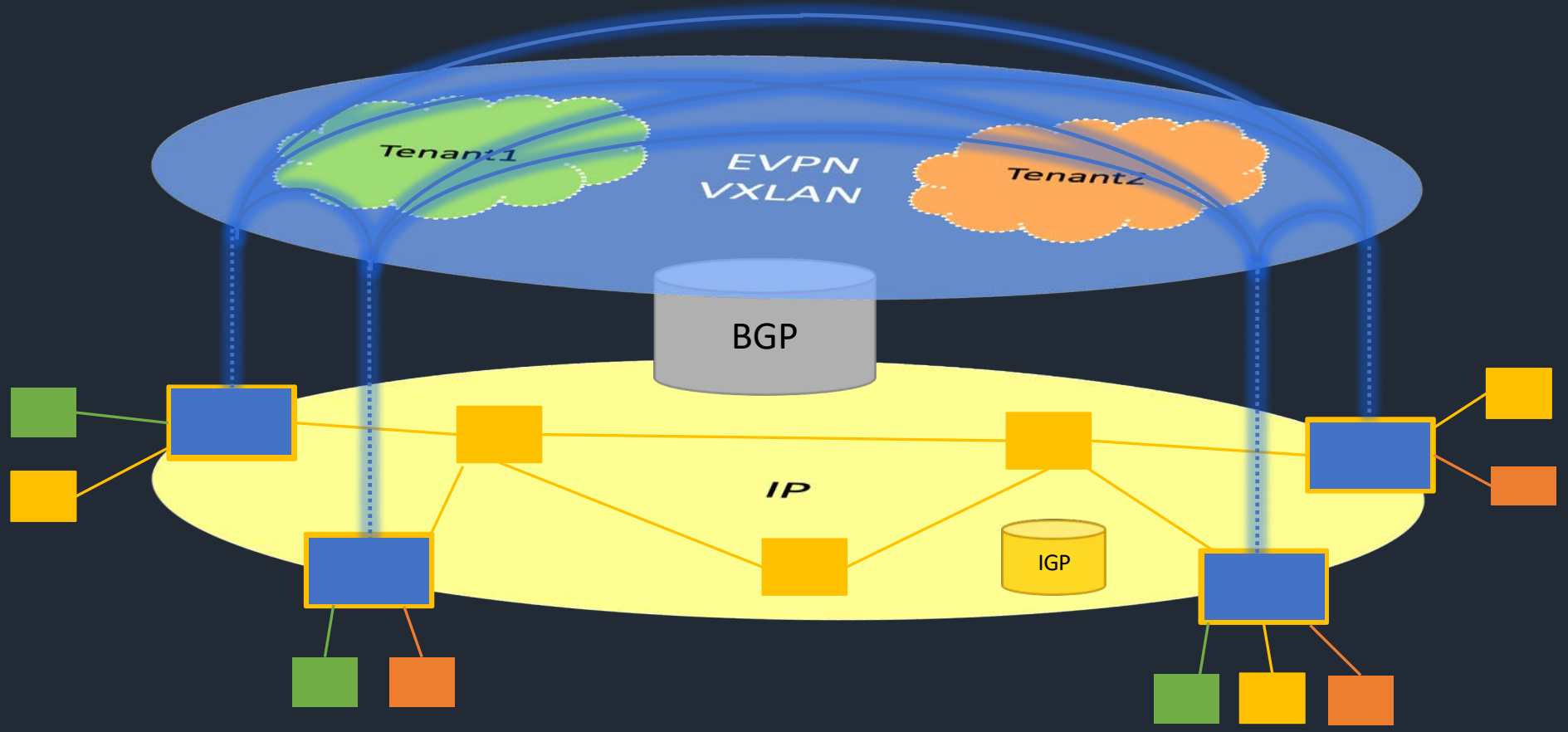


Network Subsystems -- Solutions



Network Modularity and EVPN

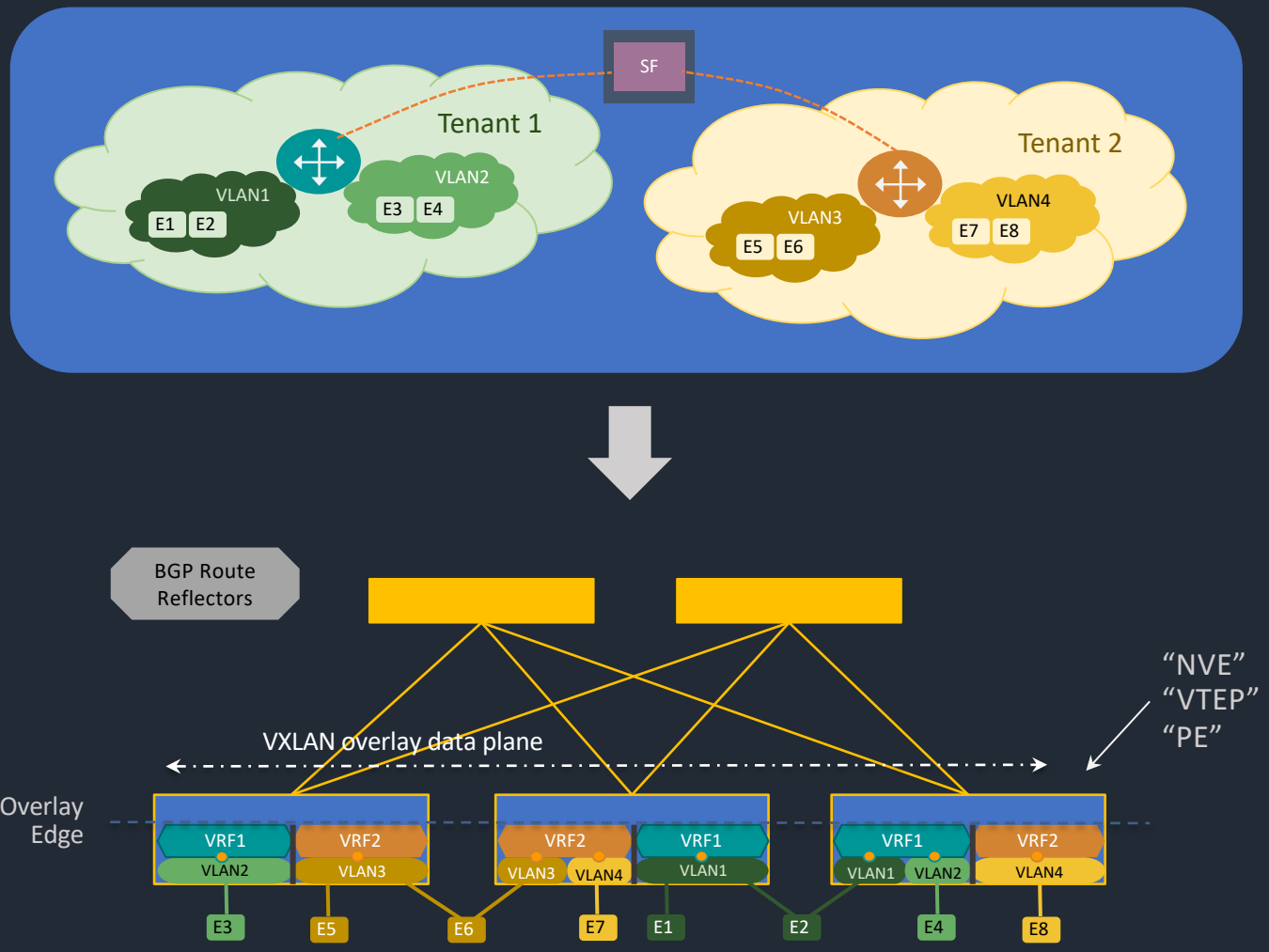
Network Modularity and EVPN



Why EVPN for Network Virtualization?

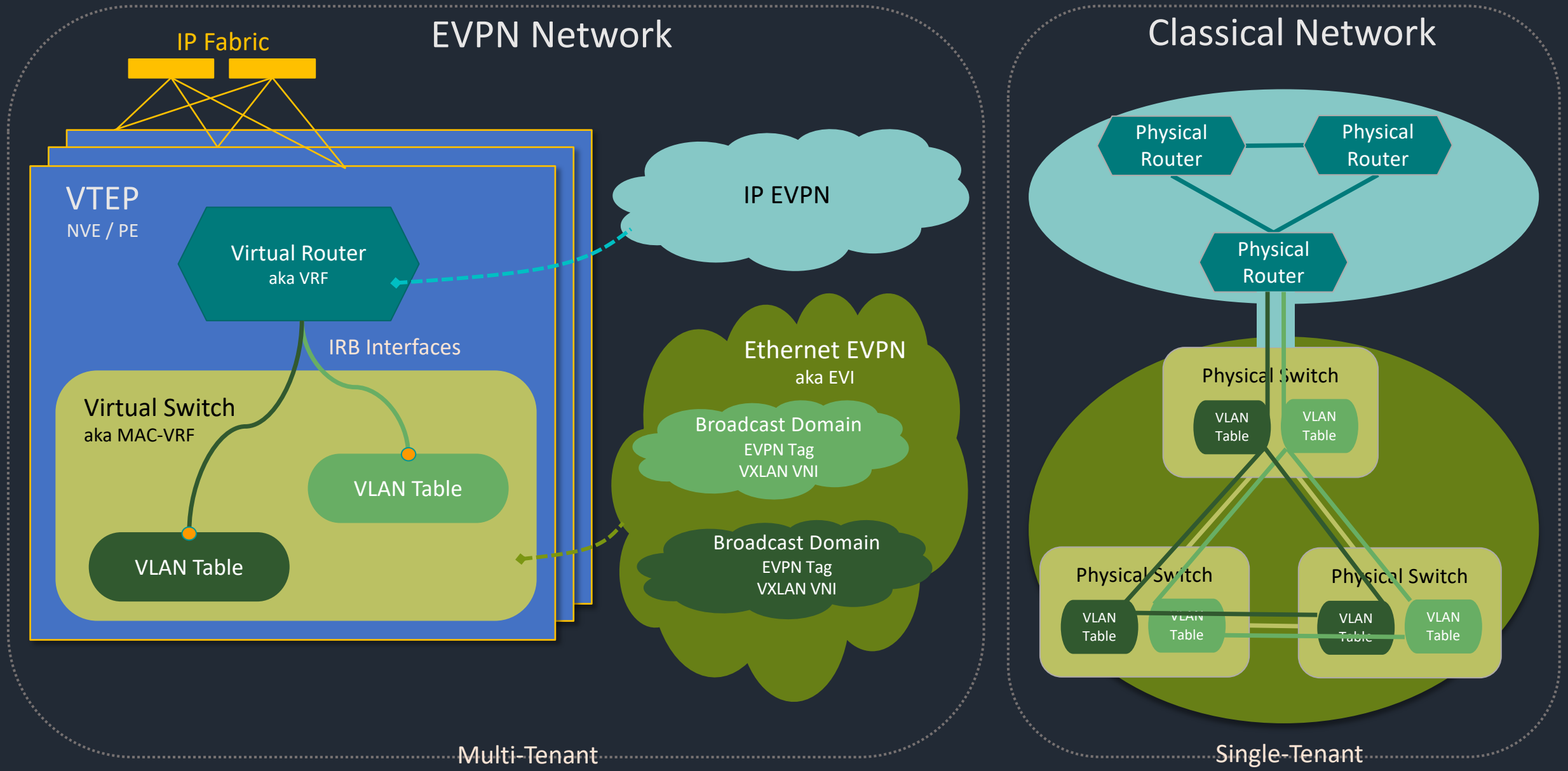
- Most pragmatic reason ⇒ there's no better open alternative at this time
- Gathering strong interest and adoption across multiple domains
- Built on top of trusted robust technologies – BGP and IP networks
- Simplified end-to-end network using a common technology across all multiple domains.
  - Translates to talent availability and fungibility, common tools, predictability, lower cost, etc
- Robust machinery for control and data plane scaling with minimum blast radius
- Seamless integration between IP and Ethernet forwarding paradigms with flexible overlay types and topologies
- All-active Ethernet multi-homing with end-to-end traffic load balancing
- Standards-based control plane based federation between autonomous systems
- Standards-based and operator-friendly network virtualization platform for SDN

A Simple Network Virtualization Overlay Reference Model

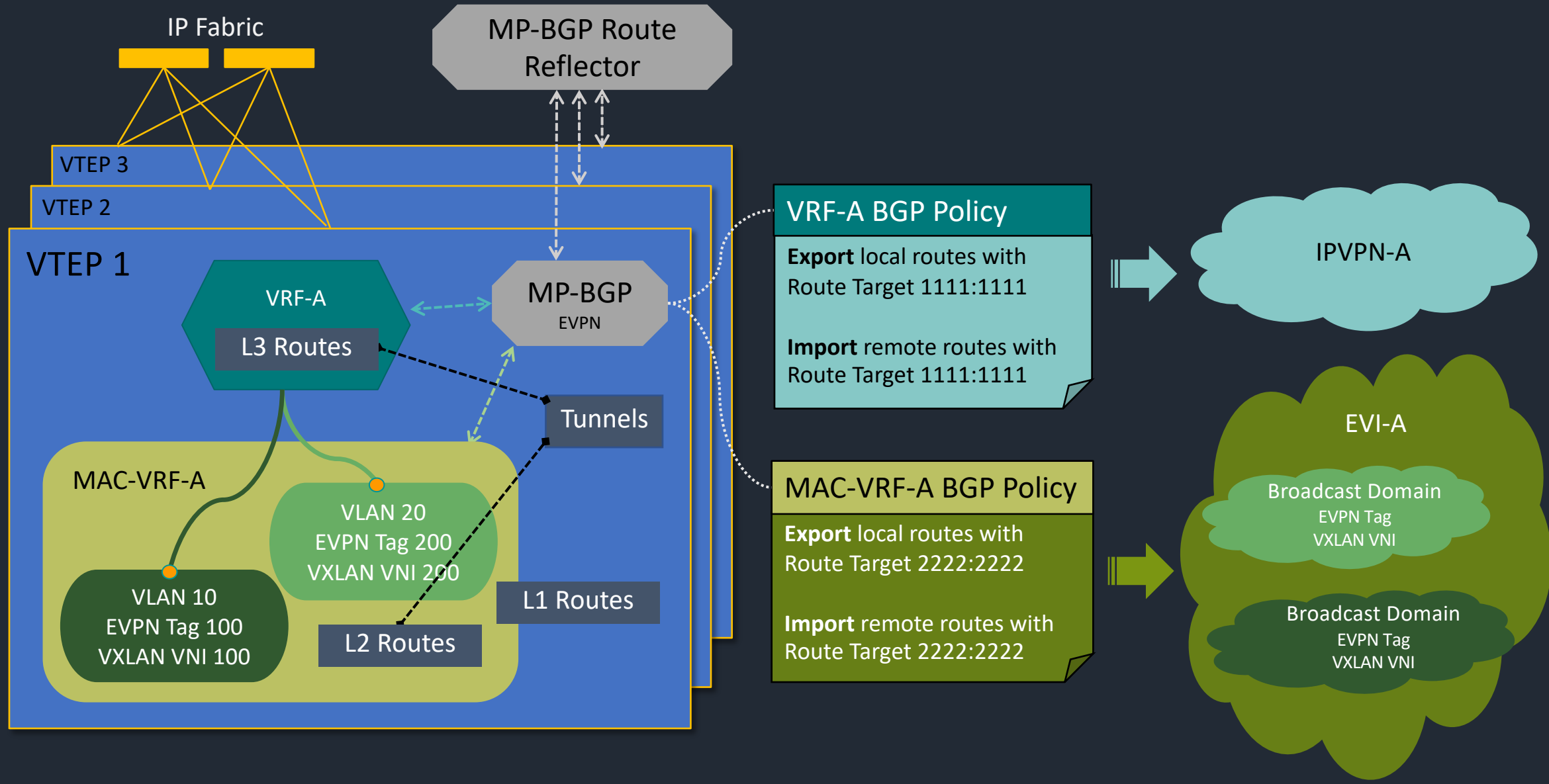


- For this talk, "tenants" are groups of location-independent endpoints where:
- Groups manifest as subnets that are routed to other groups of the same tenant (i.e. east-west) via a distributed routing function
- Tenants are routed to other tenants and to external destinations (i.e. north-south) through service function chains
- Tenants and groups are implemented as IP and Ethernet overlay virtual networks
- The network virtualization edge (NVE) function may be implemented on
  - ToR switch: to support physical end-systems
  - Virtual routers: to support virtual end-points
- Note: NVE are also referred to as PE in SP networks, or VTEP in VXLAN networks.

EVPN Parallels with Classical Networks



BGP-based VPNs Overview



EVPN Route Types – Modularity by OSI Layer

L3: IP Routing

- Type-5 IP Prefix Route
  - IP Unicast Forwarding

L2: Ethernet Bridging

- Type-2 MAC/IP route
  - MAC-Only
    - MAC unicast forwarding
  - MAC + IP
    - ARP Proxy
- Type-3 Inclusive Multicast Ethernet Tag (IMET) Route
  - BUM forwarding
- Type-6 Selective Multicast Ethernet Tag (SMET) Route
  - Selective IP multicast forwarding

L1: Ethernet Multi-Homing

- Type-4 Ethernet Segment (ES) Route
  - Designated Forwarder (DF) election
- Type-1 Ethernet A-D Route
  - Per ES
    - Split horizon, Fast convergence
  - Per EVI (ES:Tag)
    - Aliasing
- Type-7 Multicast Join Sync Route
  - Selective IP multicast support
- Type-8 Multicast Leave Sync Route
  - Selective IP multicast support

EVPN Route Types – Modularity by Unicast / Multicast

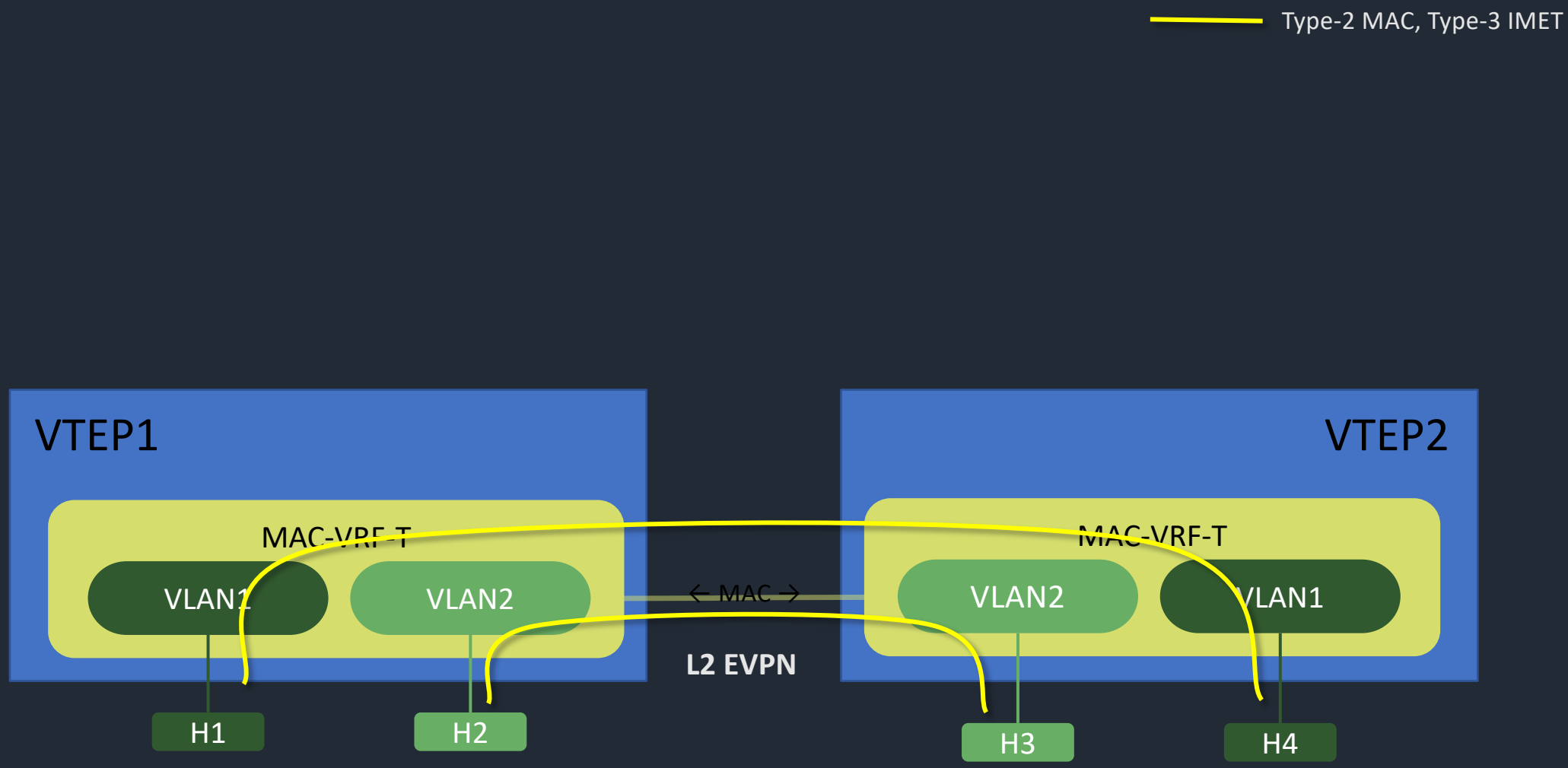
Unicast

- L1: Type-1 Ethernet A-D Route per ES
  - Fast convergence
- L1: Type-1 Ethernet A-D Route per EVI
  - Aliasing
- L2: Type-2 MAC/IP route
  - MAC unicast forwarding, ARP Proxy **
- L3: Type-5 Prefix Route Route
  - IP forwarding

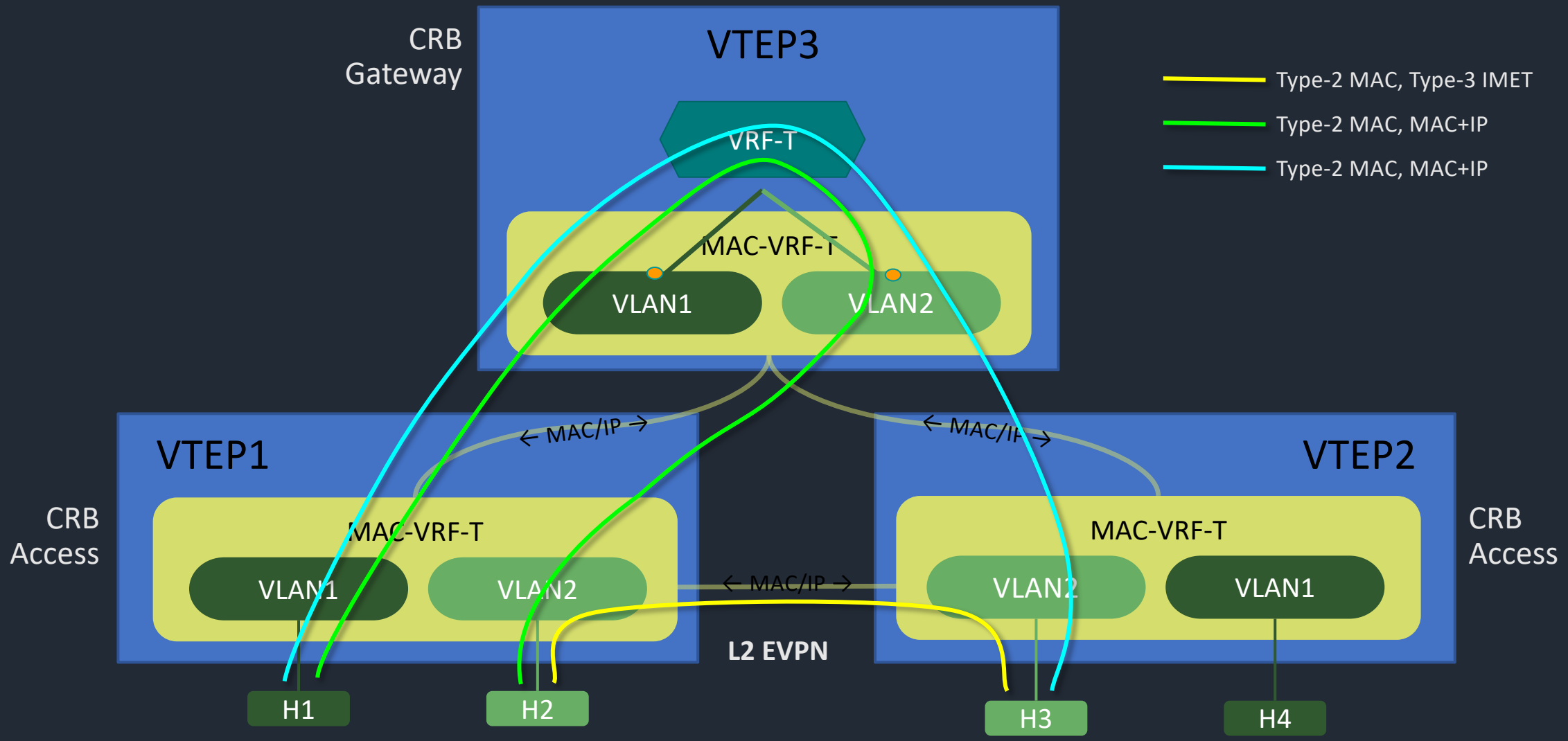
BUM and IP Multicast

- L1: Type-1 Ethernet A-D Route per ES
  - Split horizon
- L1: Type-4 Ethernet Segment (ES) Route
  - Designated Forwarder (DF) election
- L1: Type-7 Multicast Join Sync Route
  - Selective IP multicast support
- L1: Type-8 Multicast Leave Sync Route
  - Selective IP multicast support
- L2: Type-3 Inclusive Multicast Ethernet Tag (IMET) Route
  - BUM forwarding
- L2: Type-6 Selective Multicast Ethernet Tag (SMET) Route **
  - Selective IP multicast forwarding

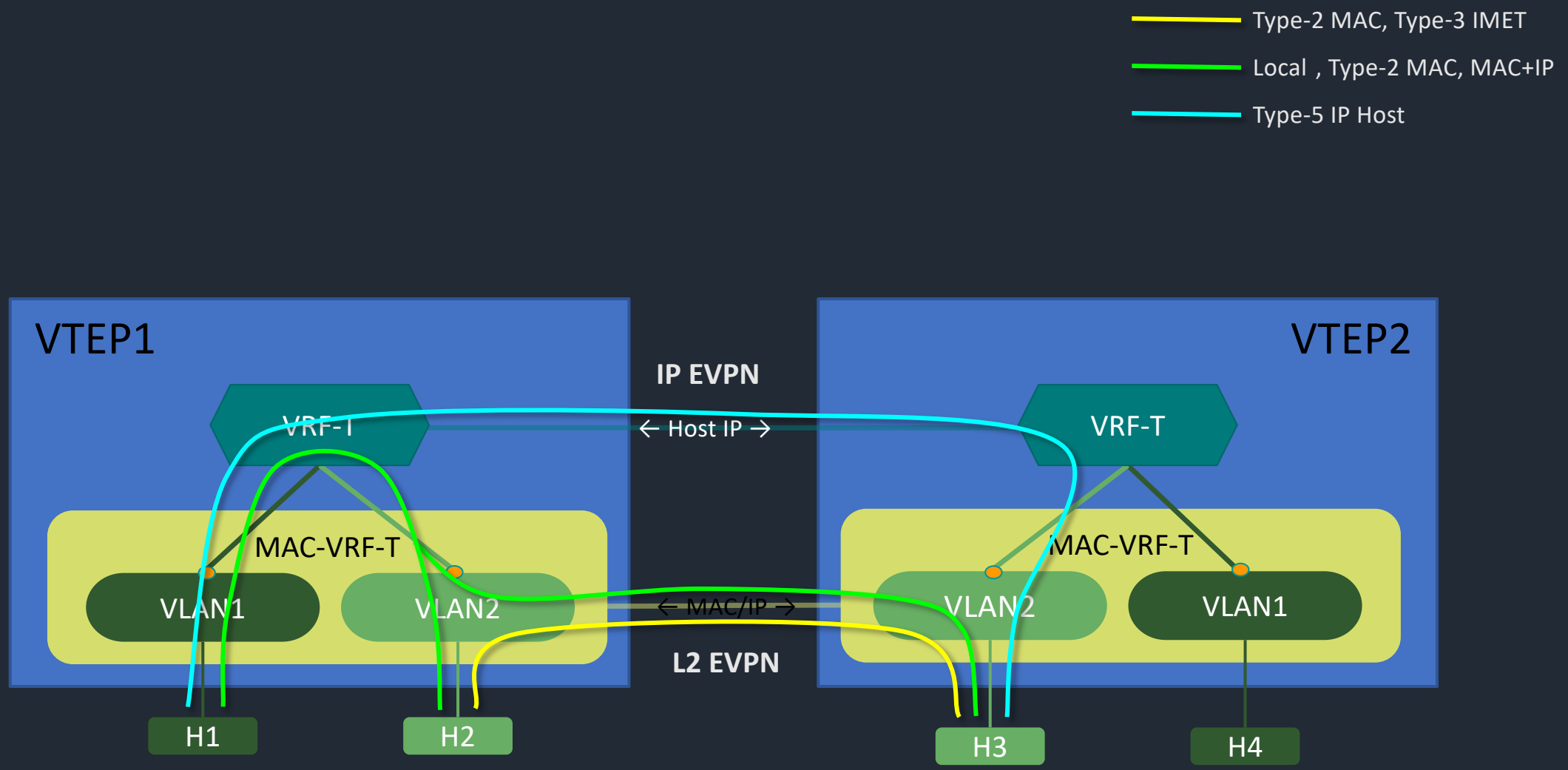
Pure Bridging Overlay



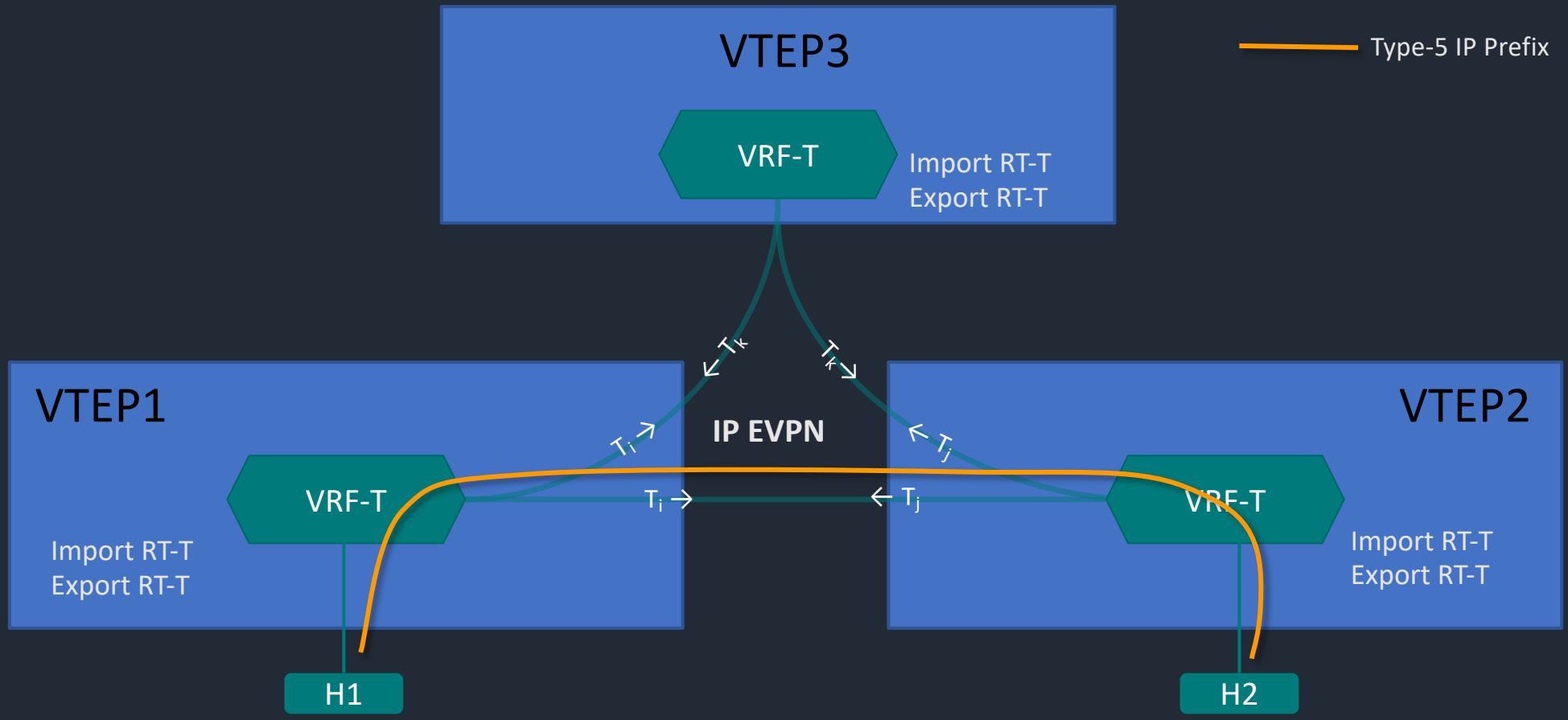
Centrally Routed Bridging Overlay (CRB)



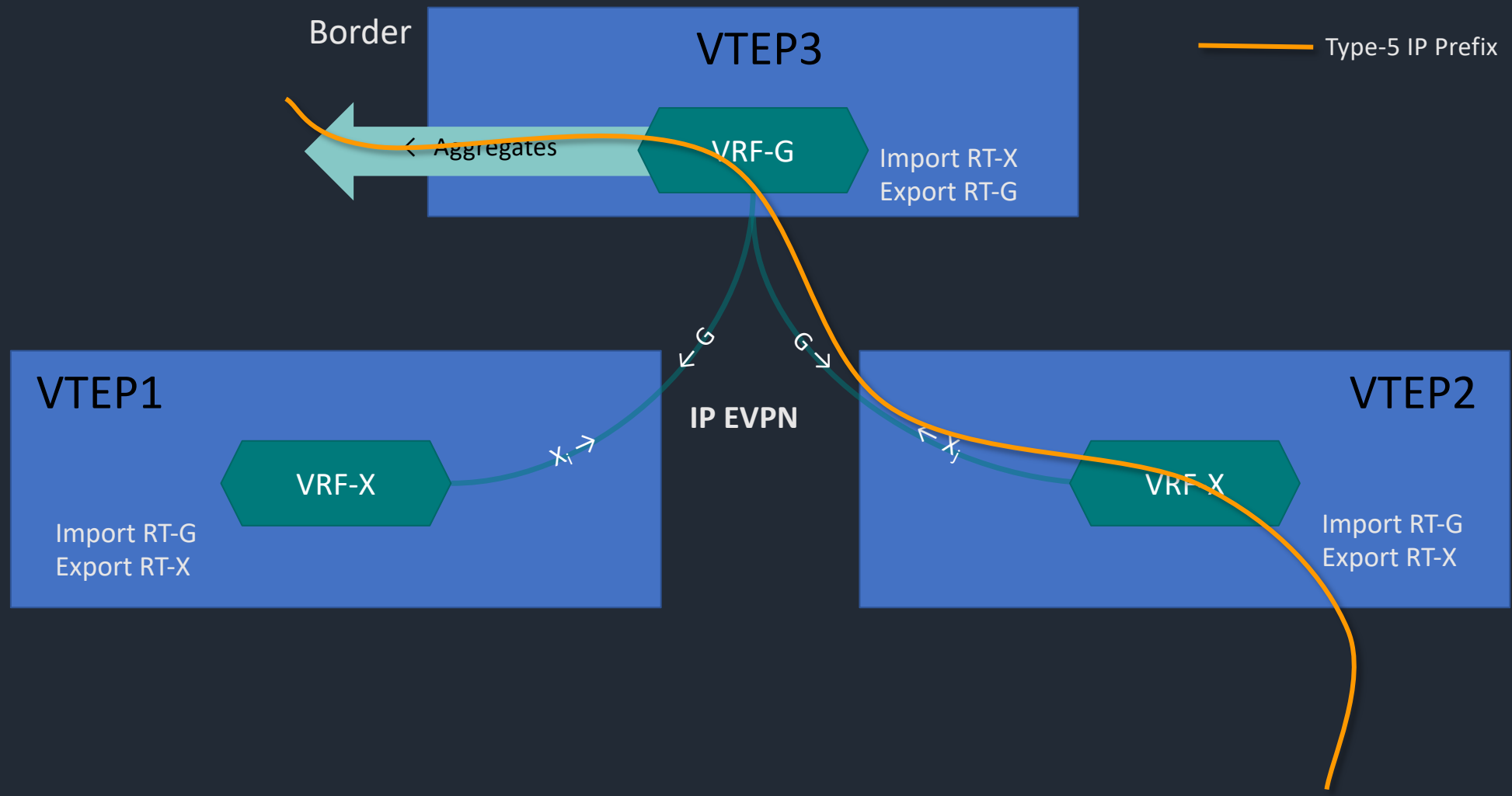
Edge Routed Bridging Overlay (ERB)



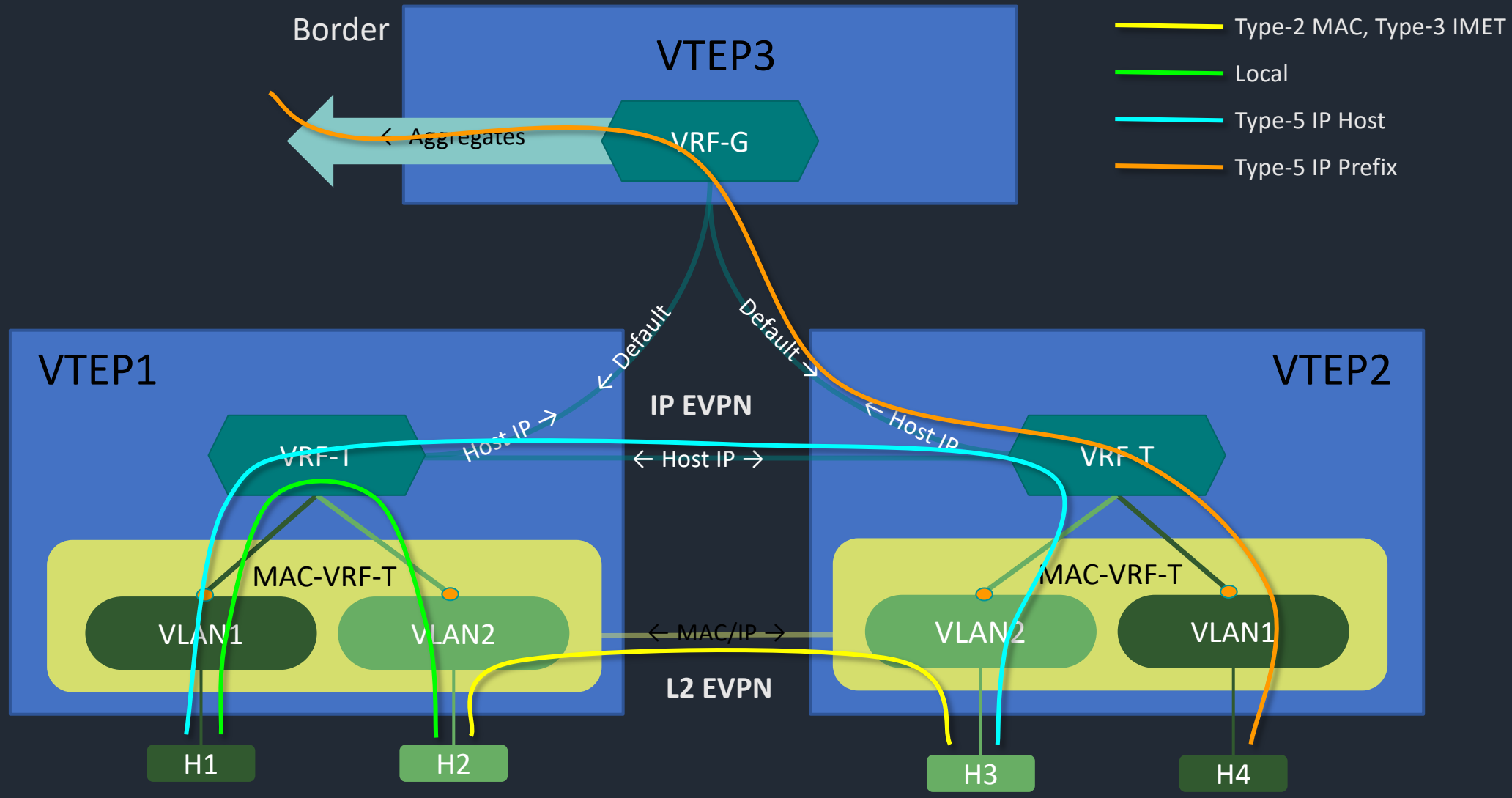
IP Routed Overlay -- Full Mesh Topology



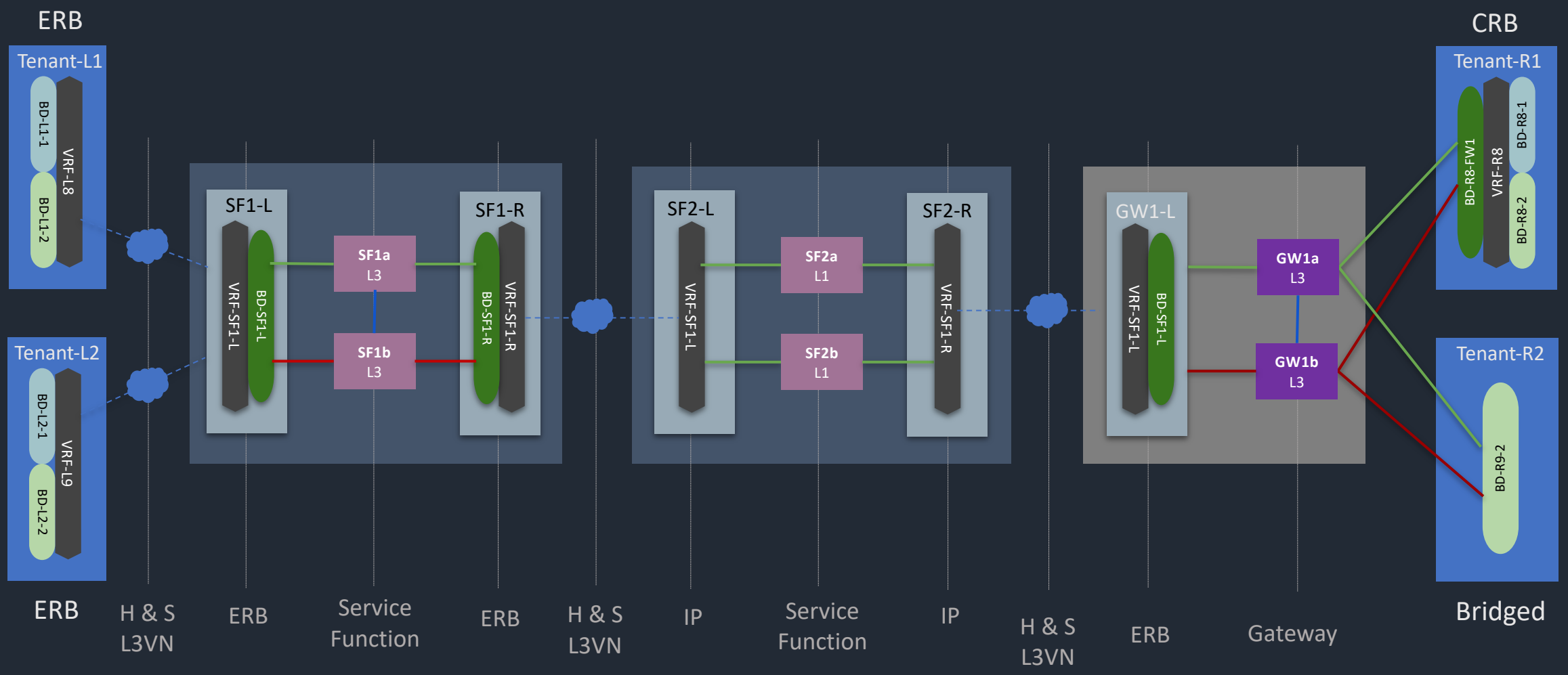
IP Routed Overlay -- Hub-and-spoke/Directional Topology



Edge Routed Bridging with IP Border Gateway

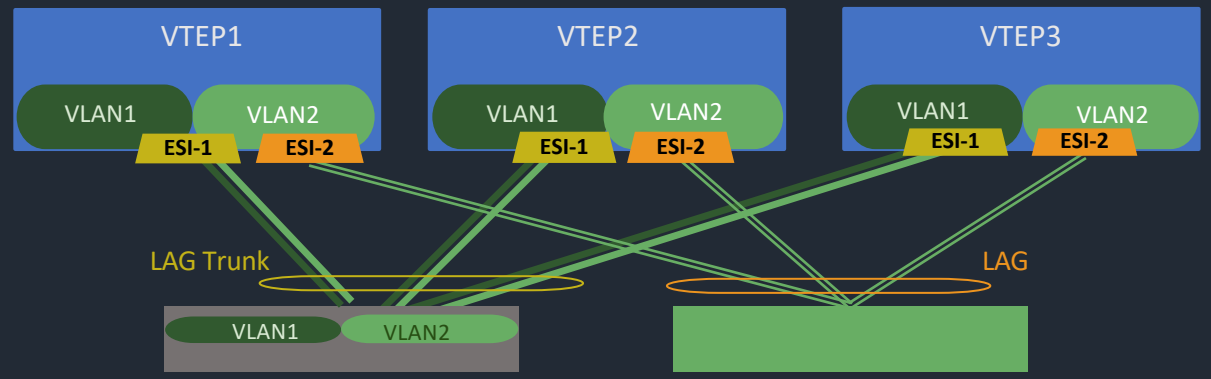


Service Chaining with these Building Blocks



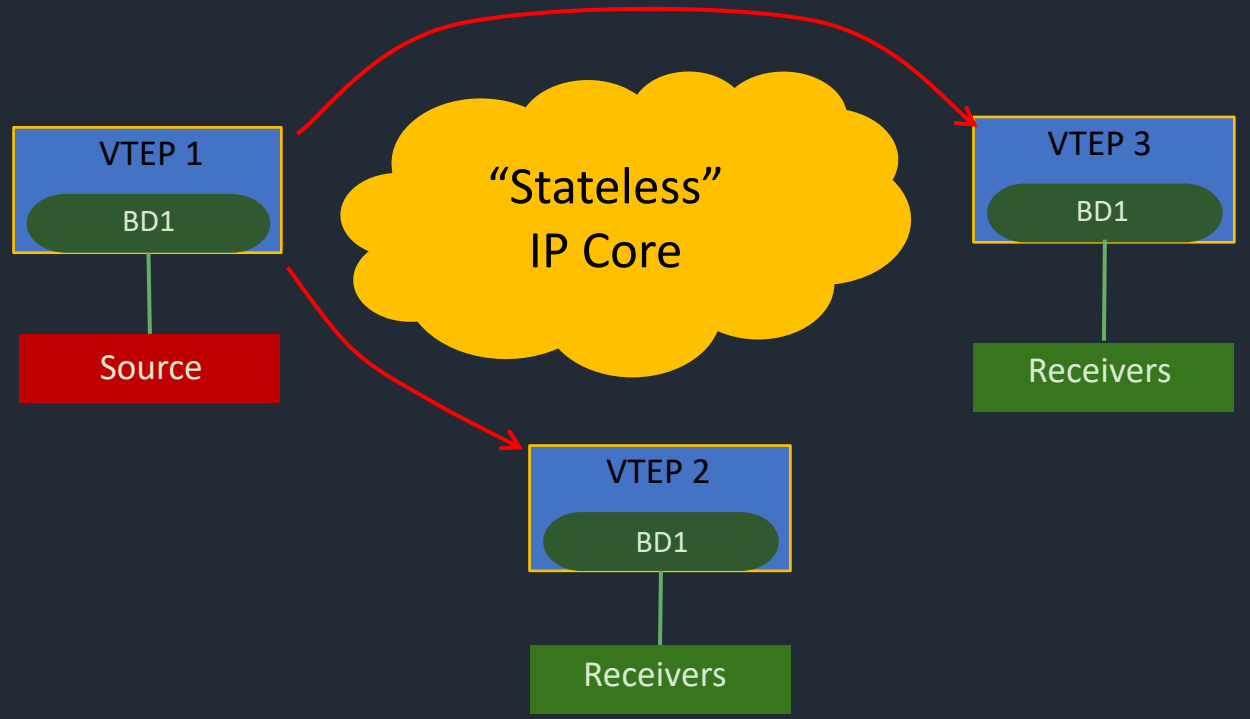
Ethernet Multihoming

- EVPN supports N-way Ethernet multihoming where N can be greater than 2
- No ICL link required
- Multi-homed end-systems are identified in the overlay by unique Ethernet Segment ID (ESI).
- EVPN Auto-ESI -- ESI generated automatically from LACP system-id or from BPDU root bridge snooping

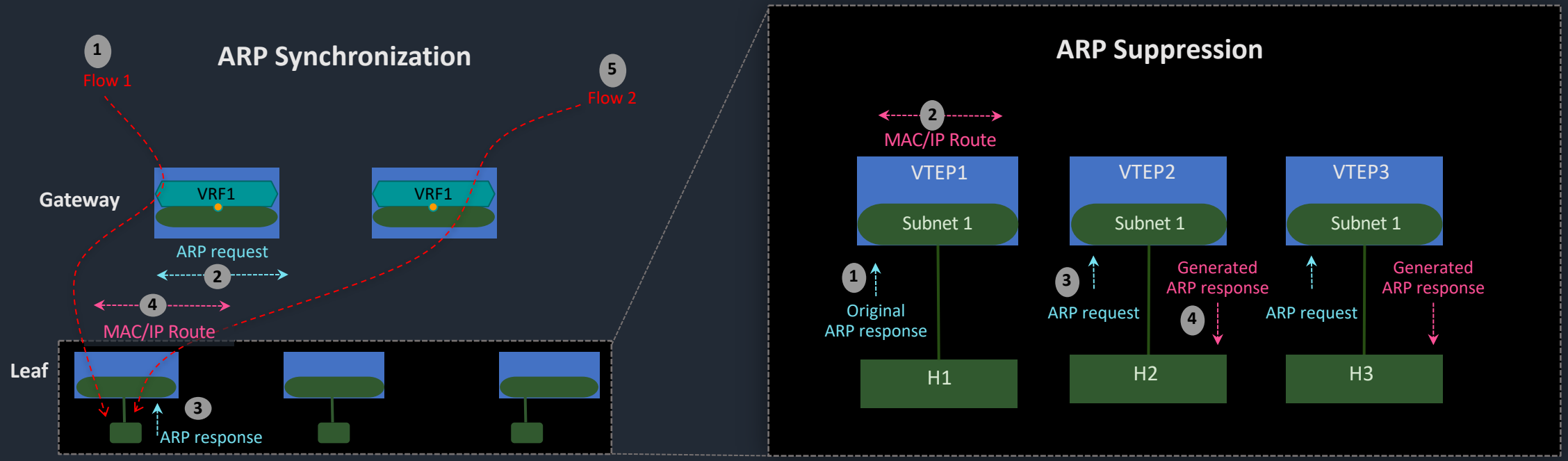


Pure Overlay BUM Replication

- Overlay replication uses “over-the-top” signaling
- No hop-by-hop per-flow or per-group multicast signaling or BUM state in underlay
- No traditional underlay multicast protocols translates to lean core network design
- Multicast convergence “same as” unicast convergence on transit link or node failure
- Selective IP multicast replication capability allows for IP multicast flow to only be replicated by an ingress VTEP only to egress VTEP that have at least one active receiver for that flow
- Further optimizations possible with Assisted Replication (AR) and Optimized Inter-Subnet Replication (OISM)



EVPN ARP Proxy -- Synchronization and Suppression



- ARP synchronization keeps the per-subnet ARP tables of tenant VRFs synchronized
- MAC-to-IP bindings are learned by Access VTEP from the Sender field of local ARP request and reply packets and advertised as Type-2 MAC+IP routes
- With distributed ARP broadcast suppression, Leaf VTEP will proxy respond to local ARP requests using the same synchronized MAC-to-IP bindings
- Reduces the impact of ARP broadcast on routers and hosts

Closing Thoughts

- EVPN based networks are only as complex as they need to be
  - Most use cases can be satisfied with only a few key building blocks
  - Complexity is proportional to the functionality required
- EVPN VXLAN is an open standard. Equivalent proprietary technology is not any simpler.

Thank You