cisco Live!

Let's go

#CiscoLiveAPJC

The bridge to possible

Architecting L4-L7 Network Services in a Multi-tenant Data Center with VXLAN EVPN

Matthias Wessendorf, Principal Engineer
@matteq4er

BRKDCN-2974

Cisco Webex App

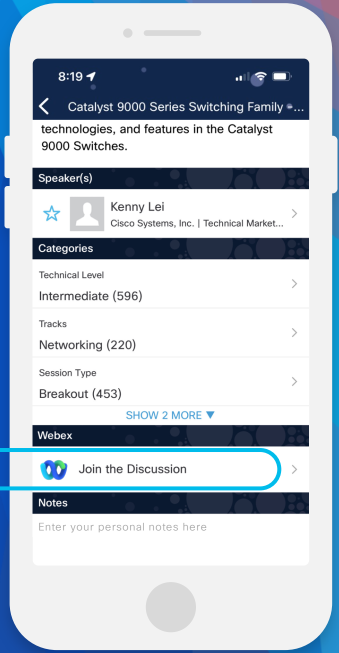
Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click "Join the Discussion"
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated
by the speaker until December 22, 2023.



8:19

Catalyst 9000 Series Switching Family ...

technologies, and features in the Catalyst 9000 Switches.

Speaker(s)

Kenny Lei
Cisco Systems, Inc. | Technical Market...

Categories

Technical Level
Intermediate (596)

Tracks
Networking (220)

Session Type
Breakout (453)

SHOW 2 MORE ▼

Webex

Join the Discussion

Notes

Enter your personal notes here

https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-2974

Session Objectives

- **At the end of the session, the participants should be able to:**
  - ✓ Articulate the different deployment options and integration considerations for service nodes in a VXLAN EVPN Fabric
  - ✓ Understand the supported deployed model to integrate services in a Multi-DC VXLAN EVPN deployment based on the VXLAN Multi-Site architecture
- **Initial assumption:**
  - ✓ The audience already has a good knowledge of the VXLAN EVPN technology (underlay, overlay, control and data plane, etc.)
  - ✓ This is not a deep dive on service nodes functionalities or configuration

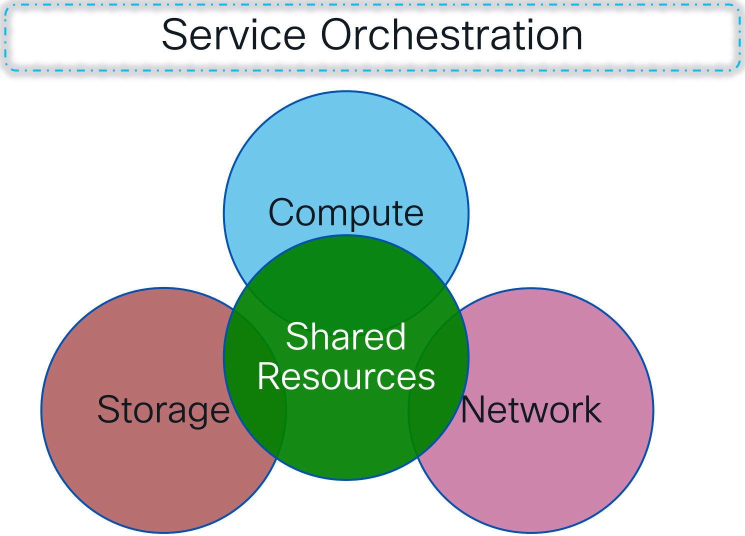
Agenda

- Multi-Tenancy for the DC Infrastructure
- Layer 4-7 Services Integration in a VXLAN EVPN Fabric
- Types of Network Services Deployments
- How to Attach Service Nodes
- Tenant Edge Firewall
- Intra-Tenant Firewall
- Layer 4-7 Services Integration in a VXLAN Multi-Site Architecture

Multi-Tenancy Functionality in Enterprise Data Centers

What is Multi-Tenancy for the DC Infrastructure?

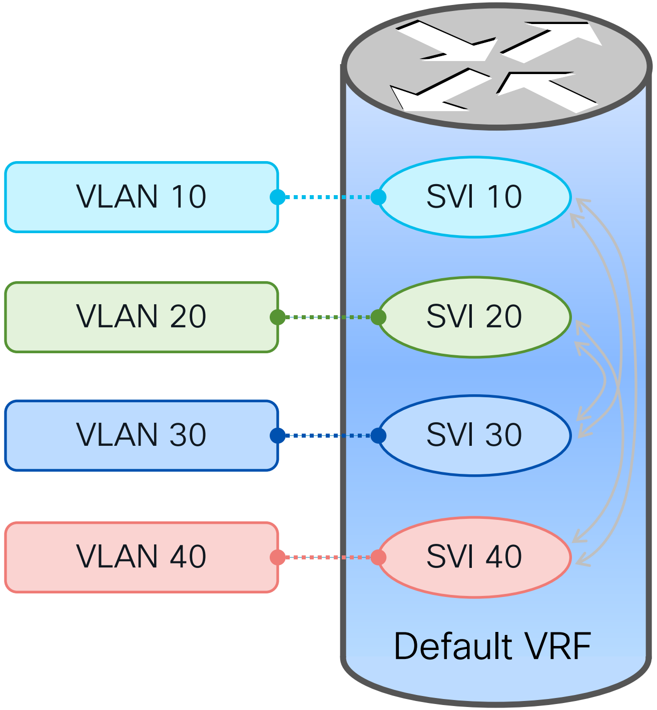
- Process of creating an environment where resources are split and combined, based on consumption, demand, supply and policies



Layer-2 Network Segmentation

- Prevents hosts in a given Layer-2 segment, from observing traffic of hosts in a different segment
  - Separation of Broadcast/Flood domains into bridge domains/segments
  - Splitting IP networks in smaller subnets
  - Containment of the Fault domain to a given Layer-2 bridge domain
  - VLAN is an overloaded notion ~ Layer-2 segment, Bridge-domain, Broadcast Domain, Flood Domain

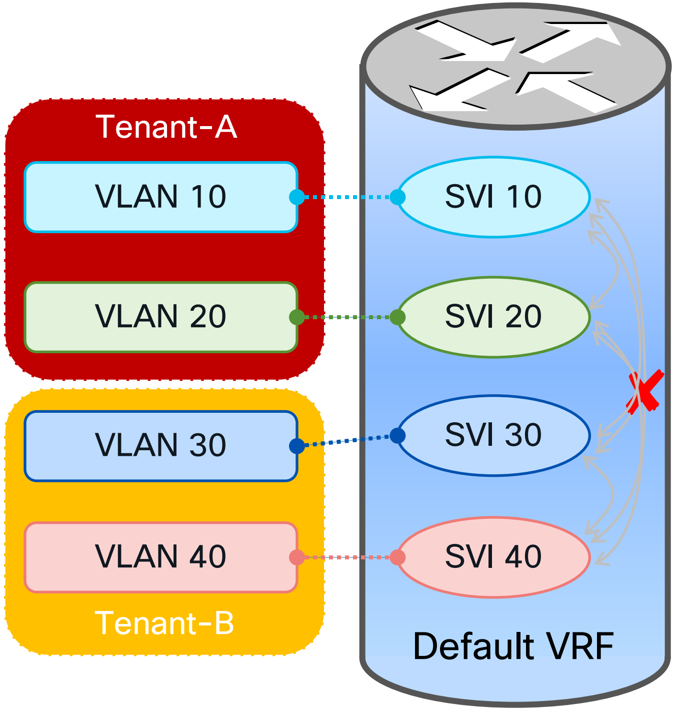
Layer-2 Segment Termination



- SVI – Layer-2 segment termination mechanism
- SVI (Switch Virtual Interface) terminates a VLAN and is assigned an IP address
- Multiple VLANs can terminate on a single device
- FHRP is typically used to provide HA
- SVI is a member of “Default VRF” by default
- Data traffic can be routed within a given VRF without restrictions

Restricting Forwarding between Segments

Use of ACLs

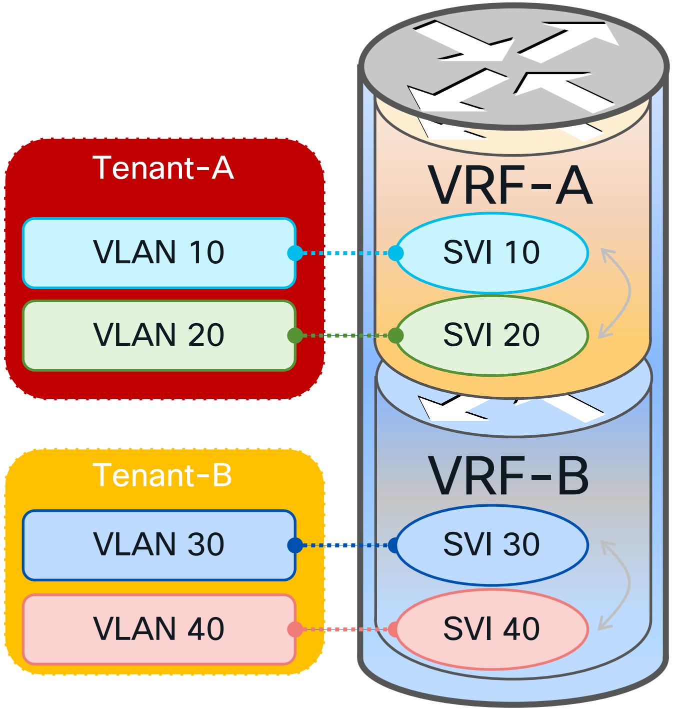


- Access Control Lists (ACL) between VLANs

Source \ Destination	VLAN 10	VLAN 20	VLAN 30	VLAN 40
VLAN 10	✓	✓	✗	✗
VLAN 20	✓	✓	✗	✗
VLAN 30	✗	✗	✓	✓
VLAN 40	✗	✗	✓	✓

- Number and complexity of ACLs becomes too high
- No overlapping IP subnets between tenants

Routing Domain – VRF



- Virtual Routing and Forwarding (VRF)
- Independent IPv4 and IPv6 address spaces
- Full unicast and multicast routing protocol support
- Two VRFs by default: Mgmt VRF and Default VRF
- All IP-based features in NX-OS are VRF aware
- Non-default VRFs are locally-significant on a router
- Data traffic is not routed across VRFs with the default configuration

Layer 4-7 Services Integration in a VXLAN EVPN Fabric

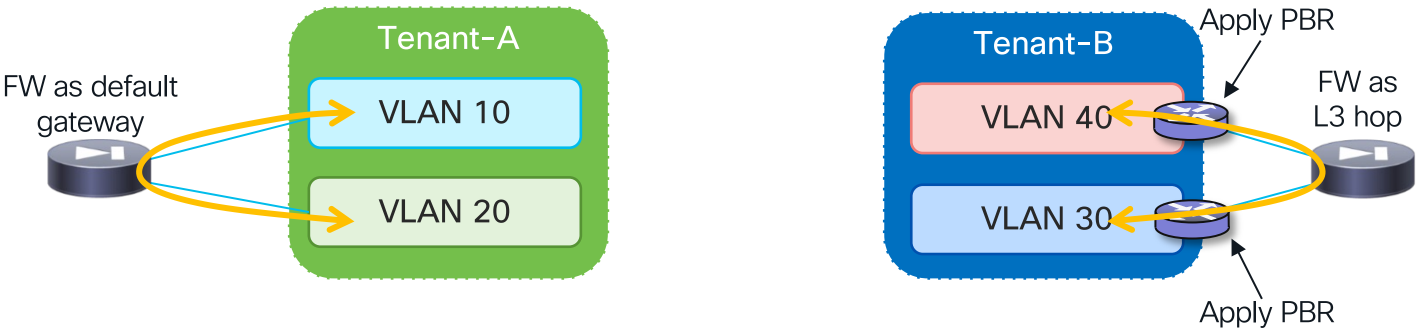
Types of Service Deployment

Prerequisites for Connecting Services

- In DC environments, Services may typically work in one of two modes:
  - Transparent, also called Layer 2 (also known as GO THROUGH)
  - Routed, also called Layer 3 (also known as GO TO)
    - Subnet default gateway configured on the firewall (most popular option)
    - Subnet default gateway configured in the network and firewall is the routed next hop (or PBR is used to steer traffic to the firewall)
- This will affect what network configurations are deployed in the fabric
- Be sure to define upfront the role of the service node (policy enforcement intra-tenant, inter-tenant, etc.)

Intra-Tenant (Intra-VRF) Services

- Filtering/policy enforcement between segments of the same Tenant
  - ➢ Intra-VRF, inter-subnets



Option 1 : FW as default GW

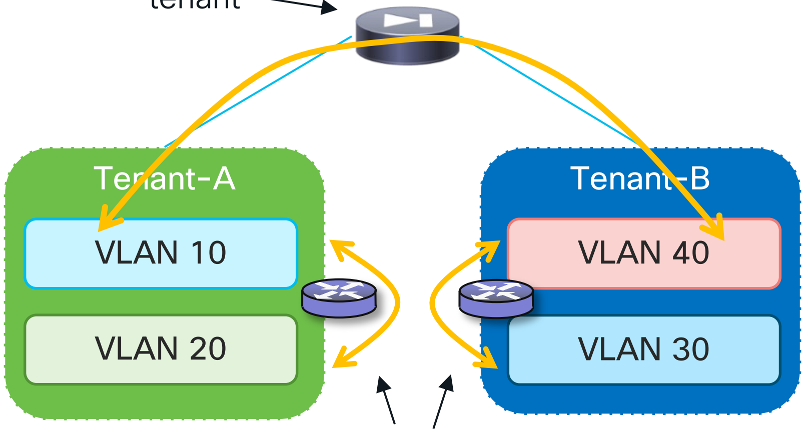
Option 2 : PBR with FW as L3 hop

Option 3 : FW in transparent (less common)

Tenant Edge Services

- Filtering/policy enforcement between Tenants (FW function front-ending each tenant domain)
  - ➢ Inter-VRF

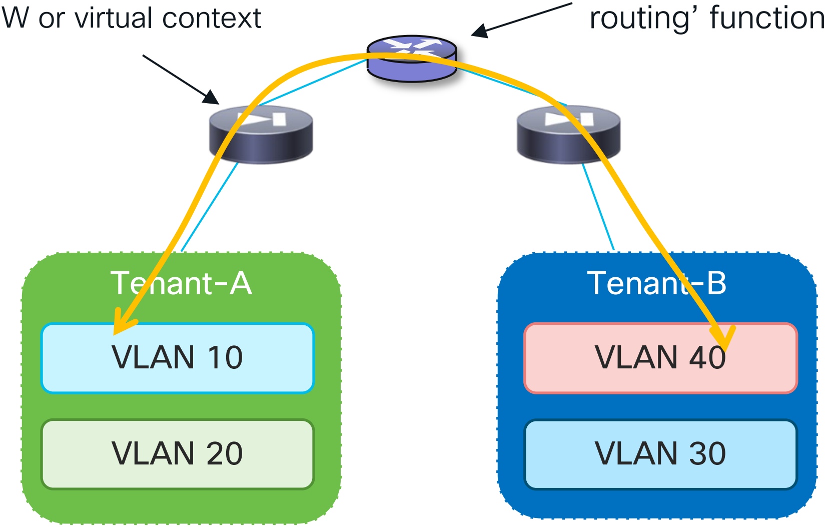
FW as 'fusion router', interface dedicated per tenant



Tenant as a security zone: allows intra-tenant communication

Per tenant physical FW or virtual context

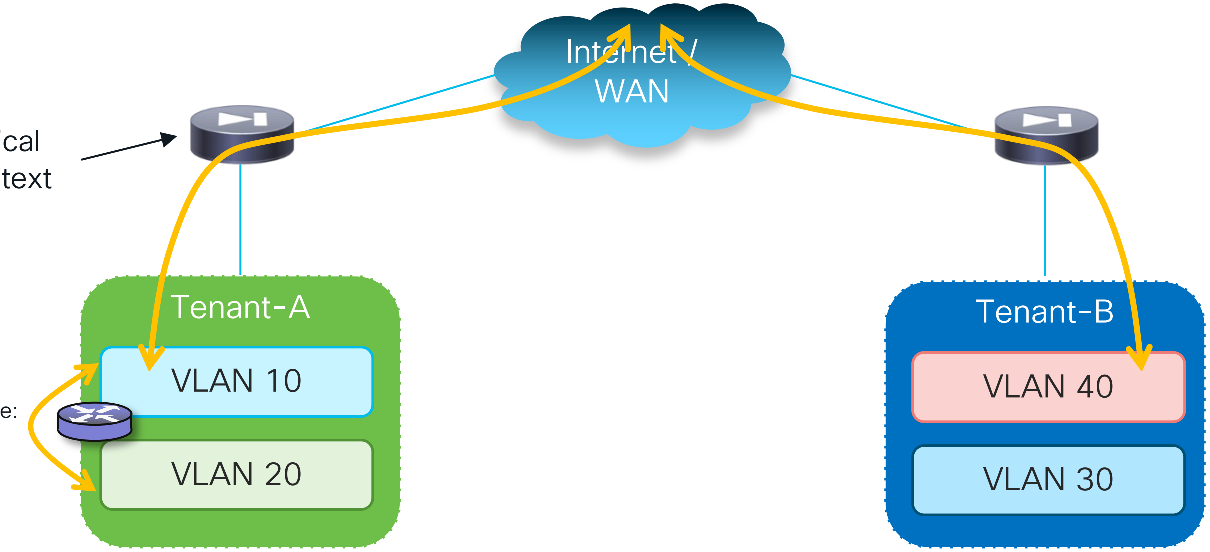
Separate 'fusion routing' function



Tenant Edge Services

Filtering for North-South Communication

- Filtering/policy enforcement between Tenants and the external world



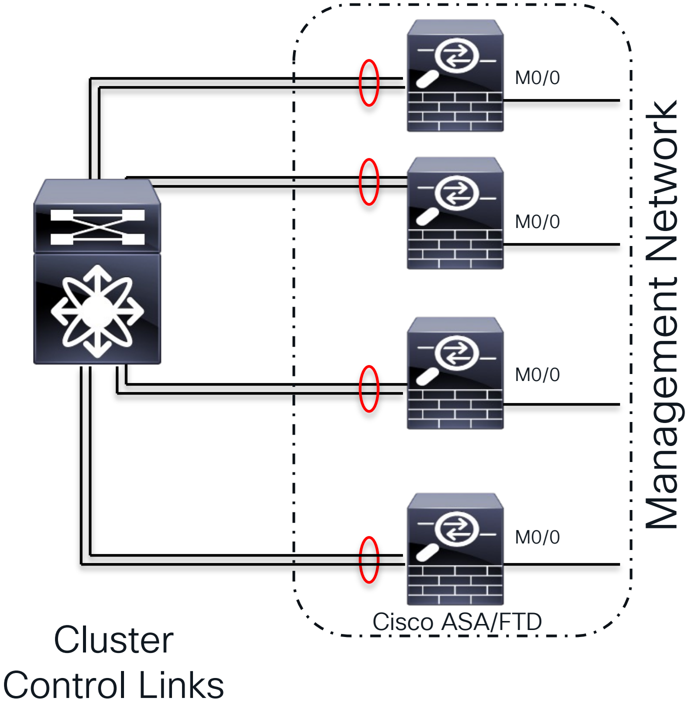
Per tenant physical FW or virtual context

Tenant as a security zone: allows intra-tenant communication

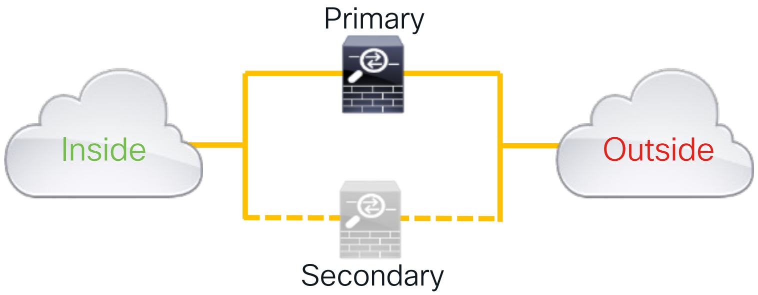
How to Attach Services Nodes?

Service Node Redundancy Models

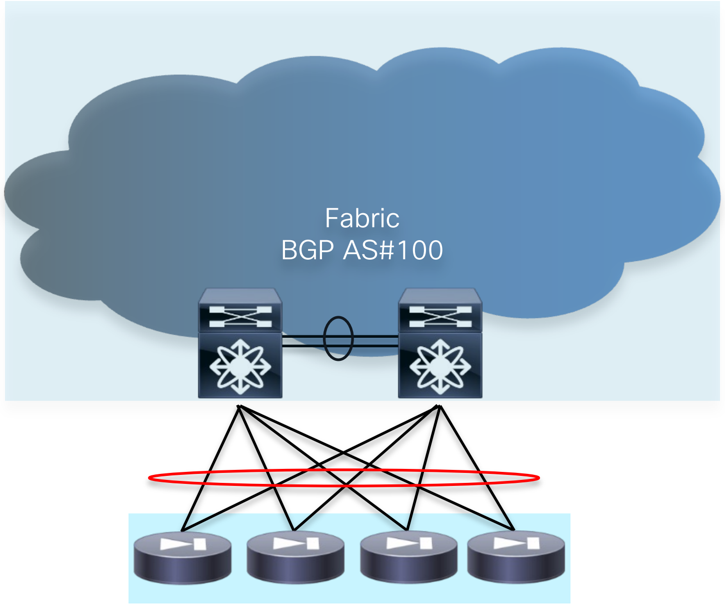
Active/Active Cluster



Active/Standby Pair



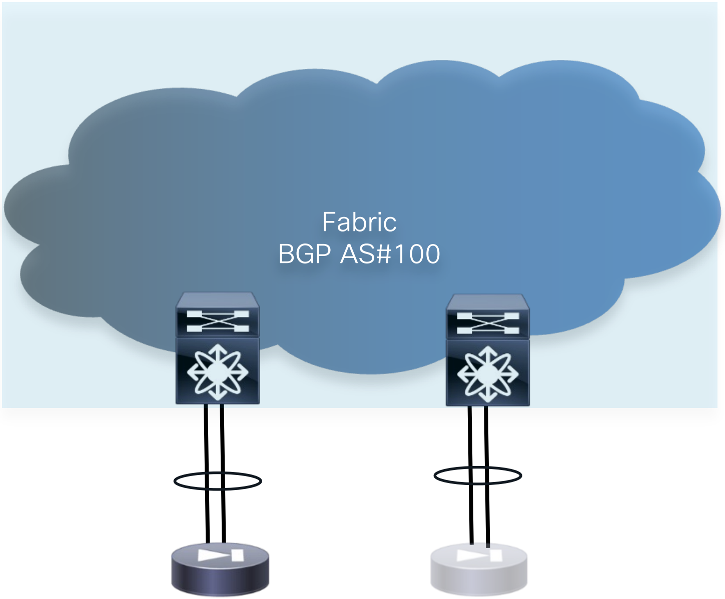
How to Physically Connect Service Nodes



Cluster

For clustered systems vPC is OK

(Cluster nodes need to be attached to the same vPC pair)

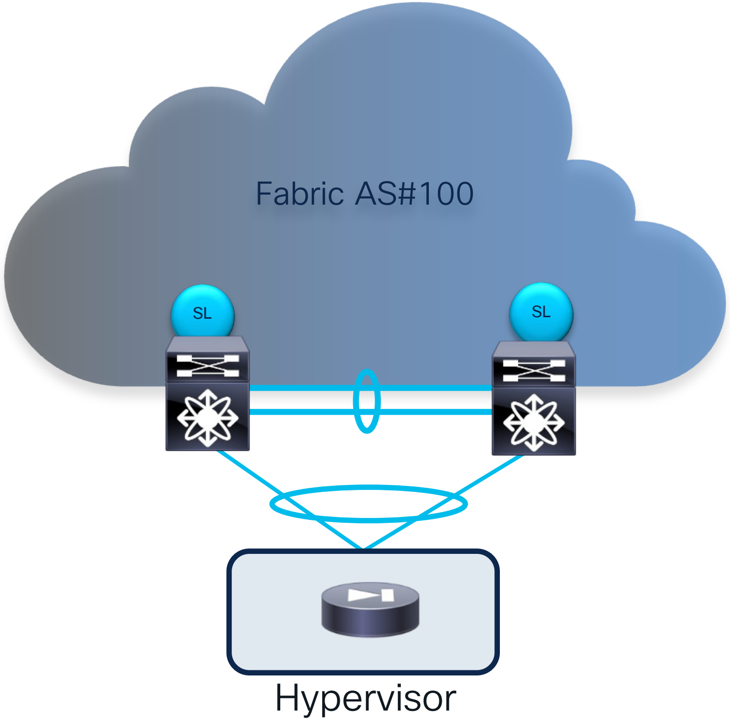


Active/Standby

For Active/Standby systems vPC is NOT a recommended choice

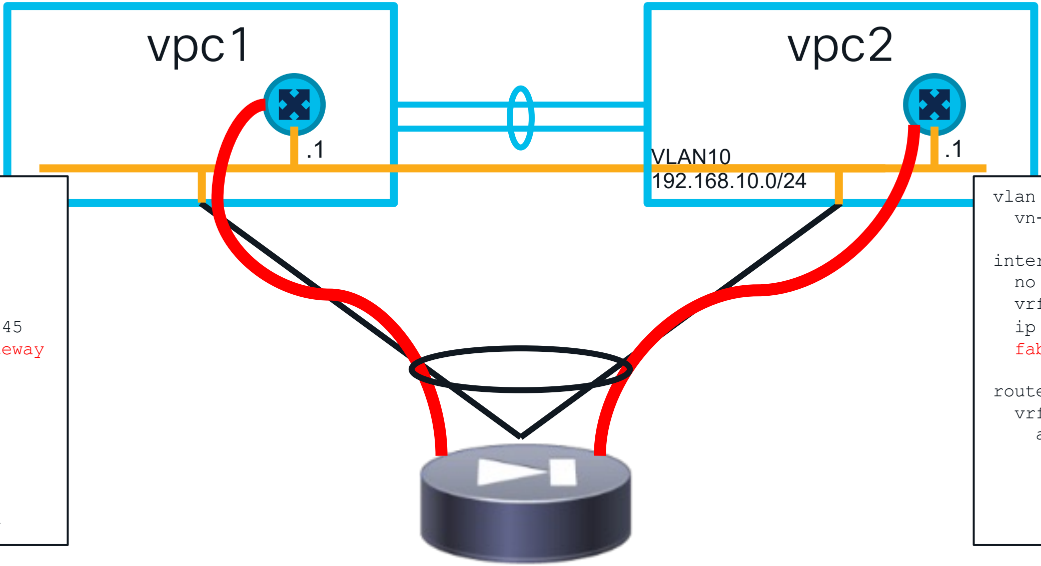
(no Multicast routing via vPC, consistent BW available, etc.)

Virtual Service Nodes Attachment to the Fabric



- Virtual service nodes deployed on a hypervisor may need to establish L3 peering with the fabric over vPC
- IPv4/IPv6 Layer 3 peering between leaf nodes and virtual service nodes is supported with the following considerations:
  - ➢ Peering can be established with unique SVI addresses on the leaf nodes only for non-VXLAN VLANs
  - ➢ For VXLAN VLANs, direct peering from the virtual router to to VTEPs' anycast GW IP address is not supported
  - ➢ The recommendation is to configure a loopback in tenant VRF on each VTEP for establishing the BGP peering with the virtual node

External Virtual Node Attachment to the Fabric



```
vlan 10
  vn-segment 30010

interface Vlan10
  no shutdown
  vrf member VRF-A
  ip address 192.168.10.1/24 tag 12345
  fabric forwarding mode anycast-gateway

router bgp 65501
  vrf VRF-A
    address-family ipv4 unicast
      neighbor 192.168.10.0/24
         remote-as 65502
         update-source VLAN 10
         address-family ipv4 unicast
```

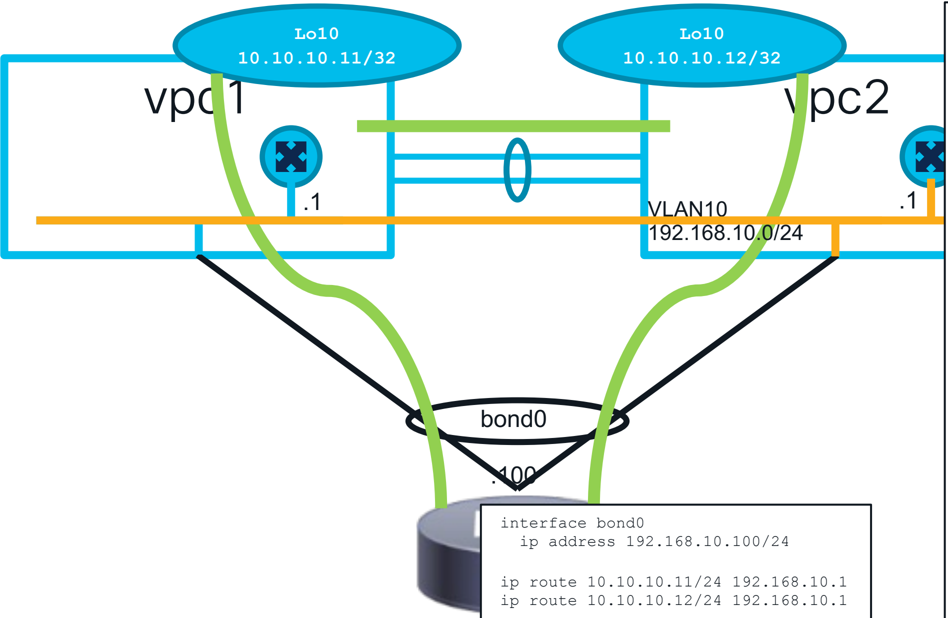
```
vlan 10
  vn-segment 30010

interface Vlan10
  no shutdown
  vrf member VRF-A
  ip address 192.168.10.1/24 tag 12345
  fabric forwarding mode anycast-gateway

router bgp 65501
  vrf VRF-A
    address-family ipv4 unicast
      neighbor 192.168.10.0/24
         remote-as 65502
         update-source VLAN 10
         address-family ipv4 unicast
```

```
interface bond0
  ip address 192.168.10.100/24

router bgp 65502
  address-family ipv4 unicast
    neighbor 192.168.10.1
      remote-as 65501
```

Not Supported

External Virtual Node Attachment to the Fabric

```
vlan 10
  vn-segment 30010
vlan 3967

system nve infra-vlans 3967

interface loopback10
  no shutdown
  vrf member VRF-A
  ip address 10.10.10.11/32 tag 12345

interface Vlan10
  no shutdown
  vrf member VRF-A
  ip address 192.168.10.1/24 tag 12345
  fabric forwarding mode anycast-gateway

interface vlan 3967
  no shutdown
  vrf member VRF-A
  ip address 10.10.0.1/30 tag 12345

router bgp 65501
  vrf VRF-A
    address-family ipv4 unicast
      neighbor 192.168.10.0/24
         remote-as 65502
         ebgp-multihop 5
         update-source loopback 10
         address-family ipv4 unicast
       neighbor 10.10.0.2
         remote-as 65501
         update-source VLAN 3967
         address-family ipv4 unicast
          next-hop-self
```



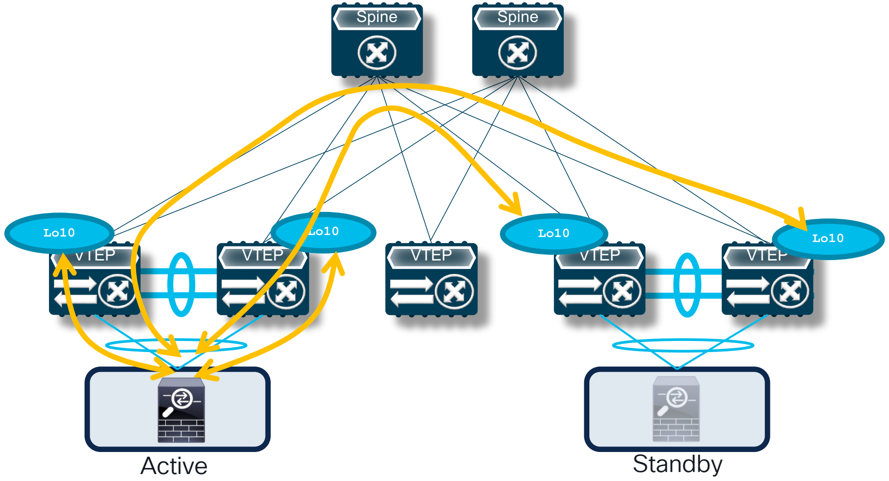
```
interface bond0
  ip address 192.168.10.100/24

ip route 10.10.10.11/24 192.168.10.1
ip route 10.10.10.12/24 192.168.10.1

router bgp 65502
  address-family ipv4 unicast
    neighbor 10.10.10.11
      remote-as 65501
      ebgp-multihop 5
      update-source loopback 10
  address-family ipv4 unicast
     neighbor 10.10.10.12
       remote-as 65501
       ebgp-multihop 5
       update-source loopback 10
```

```
vlan 10
  vn-segment 30010
vlan 3967

system nve infra-vlans 3967

interface loopback10
  no shutdown
  vrf member VRF-A
  ip address 10.10.10.12/32 tag 12345

interface Vlan10
  no shutdown
  vrf member VRF-A
  ip address 192.168.10.1/24 tag 12345
  fabric forwarding mode anycast-gateway

interface vlan 3967
  no shutdown
  vrf member VRF-A
  ip address 10.10.0.2/30 tag 12345

router bgp 65501
  vrf VRF-A
    address-family ipv4 unicast
      neighbor 192.168.10.0/24
         remote-as 65502
         ebgp-multihop 5
         update-source loopback 10
         address-family ipv4 unicast
       neighbor 10.10.0.1
         remote-as 65501
         update-source VLAN 3967
         address-family ipv4 unicast
          next-hop-self
```

Supported

External Virtual Node Attachment to the Fabric

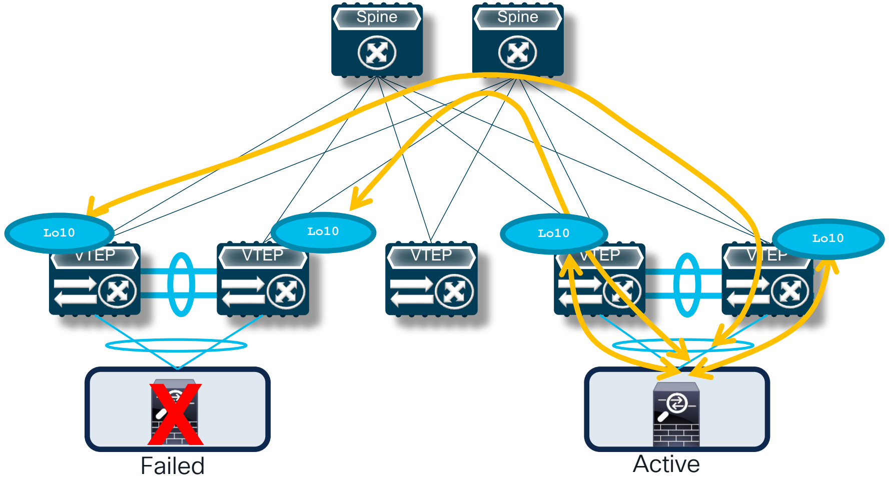
Virtual Nodes Connected to Separate Leaf Pairs



- Active/Standby virtual FW pair connected to separate leaf node pairs
- For minimizing the traffic outage after a FW failover event, the active virtual FW should peer with local and remote leaf nodes
  - Only possible in a VXLAN EVPN fabric when peering with loopbacks

External Virtual Node Attachment to the Fabric

Virtual Nodes Connected to Separate Leaf Pairs



- Active/Standby virtual FW pair connected to separate leaf node pairs
- For minimizing the traffic outage after a FW failover event, the active virtual FW should peer with local and remote leaf nodes
  - Only possible in a VXLAN EVPN fabric when peering with loopbacks
- After failover, there is no need to re-establish EBGP sessions between the virtual FW and the fabric
  - Leverages FW BGP graceful restart capabilities

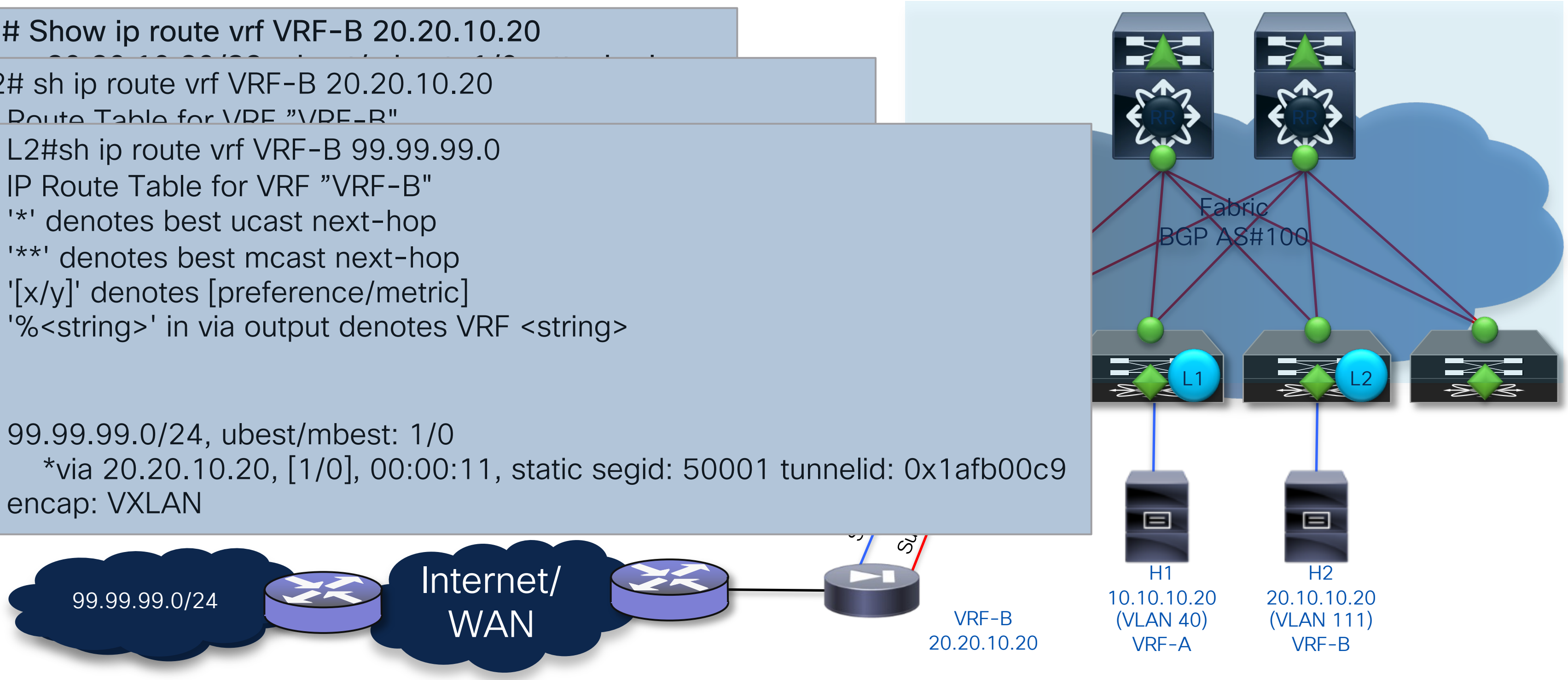
What about Static Routes

Check Availability of Static Routes Next Hop

- Problem with Redistributing Static Routes
  - What happens if the Next Hop goes down?
  - How to deploy this redundant?

- 2 Solutions
  - Recursive Next Hop (RNH)
  - Host Mobility Manager Tracking (HMM Tracking)

Recursive Next Hop (RNH)



```
BL1# Show ip route vrf VRF-B 20.20.10.20
```

```
L2# sh ip route vrf VRF-B 20.20.10.20
IP Route Table for VRF "VRF-B"
```

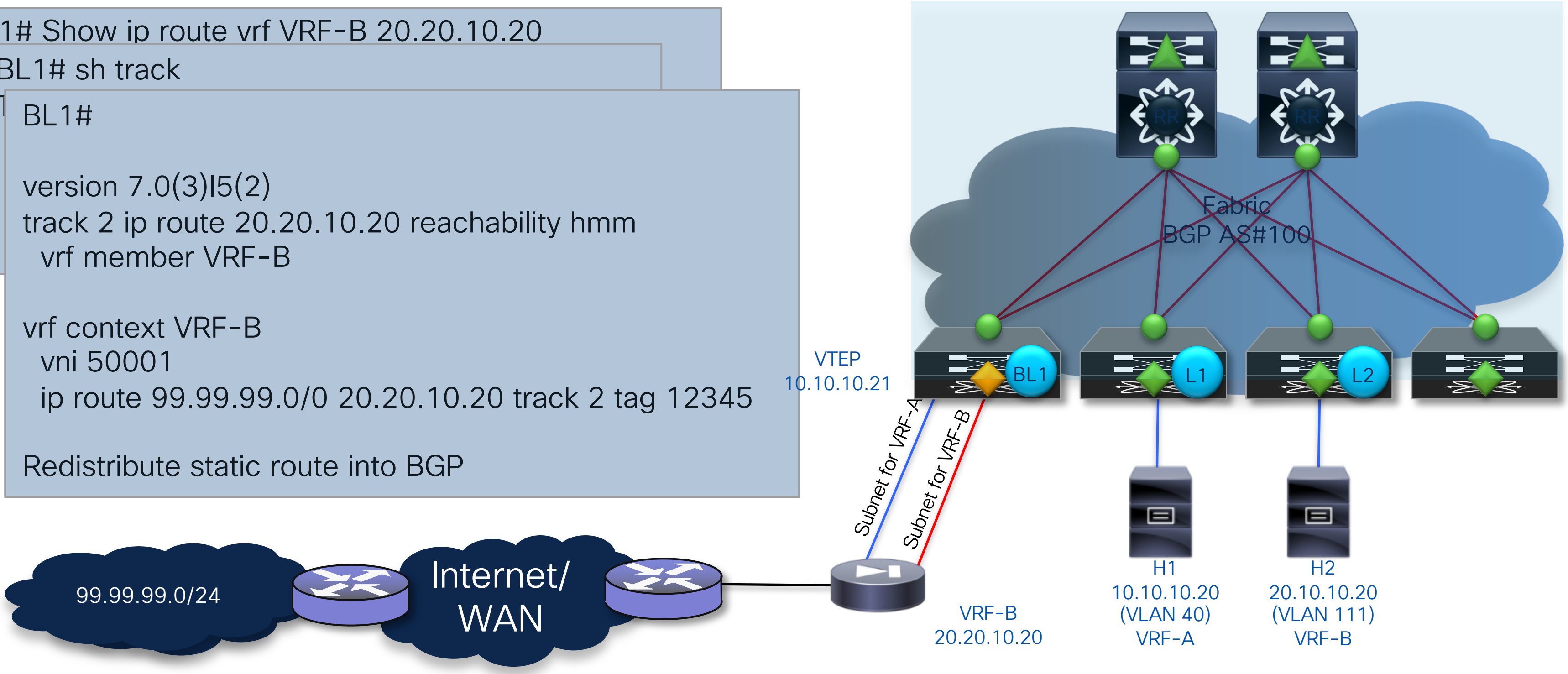
```
L2#sh ip route vrf VRF-B 99.99.99.0
IP Route Table for VRF ”VRF-B"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>


99.99.99.0/24, ubest/mbest: 1/0
    *via 20.20.10.20, [1/0], 00:00:11, static segid: 50001 tunnelid: 0x1afb00c9
encap: VXLAN
```

HMM Tracking

```
BL1# Show ip route vrf VRF-B 20.20.10.20

h
```

```
BL1# sh track
T
```

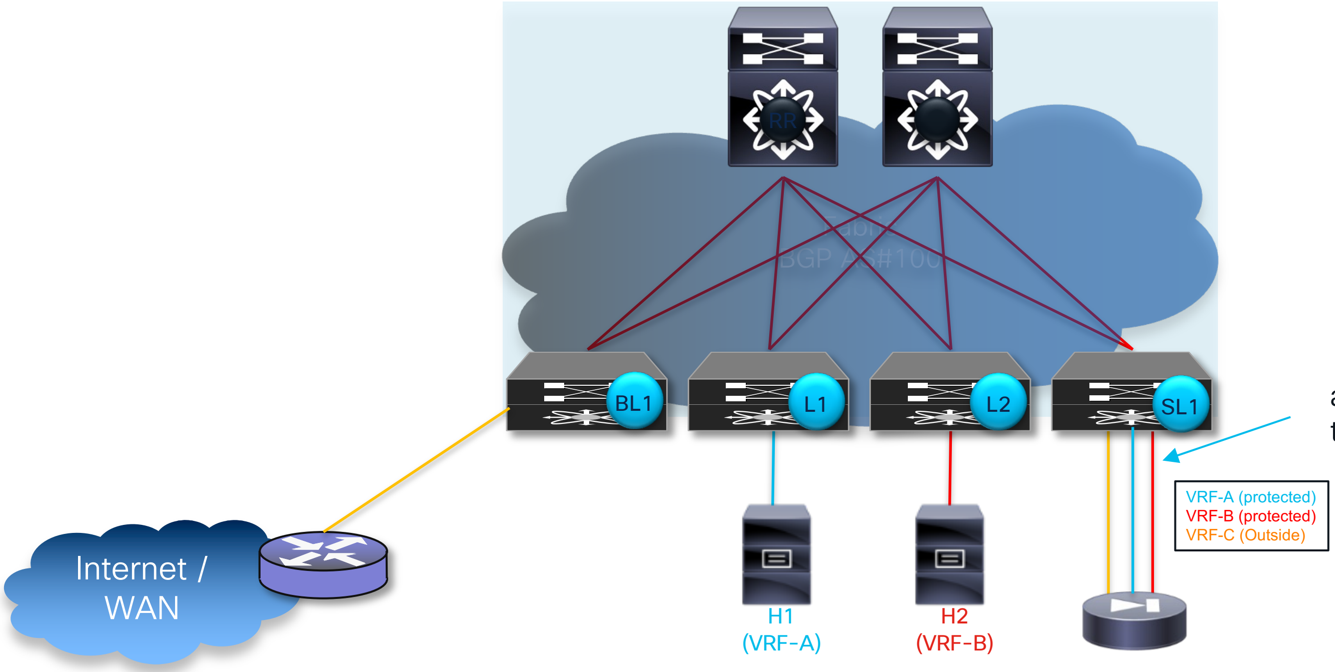
```
BL1#

version 7.0(3)I5(2)
track 2 ip route 20.20.10.20 reachability hmm
  vrf member VRF-B

vrf context VRF-B
  vni 50001
  ip route 99.99.99.0/0 20.20.10.20 track 2 tag 12345

Redistribute static route into BGP
```



Tenant Edge Firewall (Inter-VRF and North-South Flows)

Tenant Edge Firewall

Physical/Logical Topology

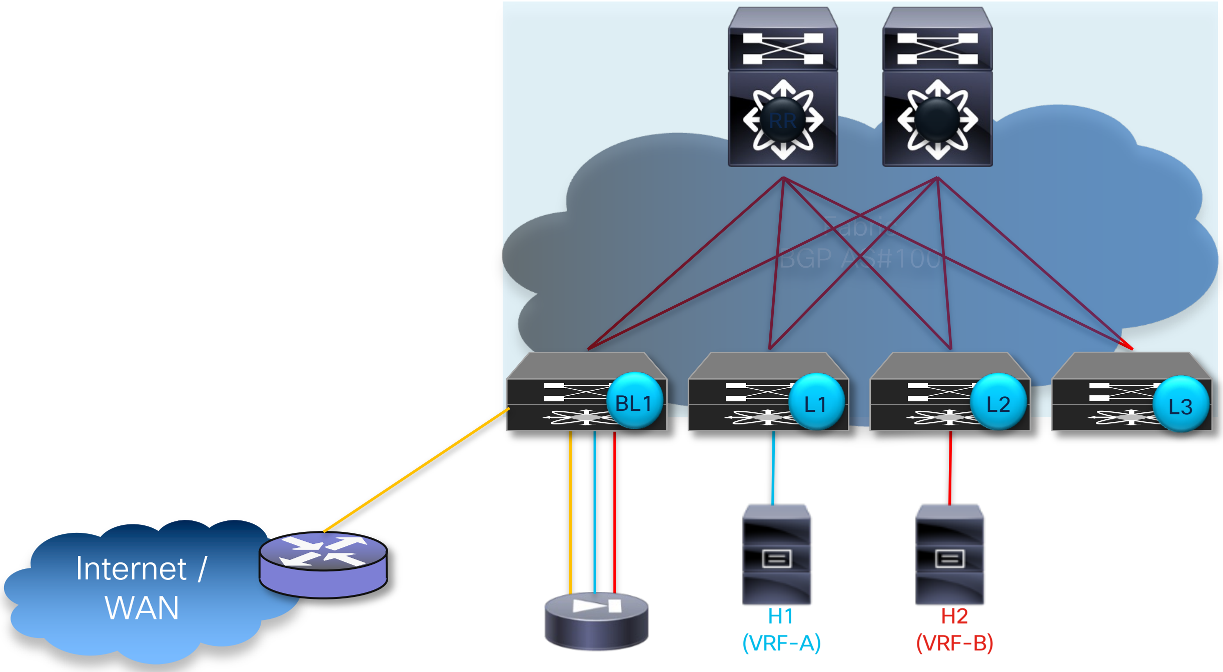


Usually those are logical interfaces associated to VLANs trunked between the FW and the SL

FW connected to a Service Leaf

Tenant Edge Firewall

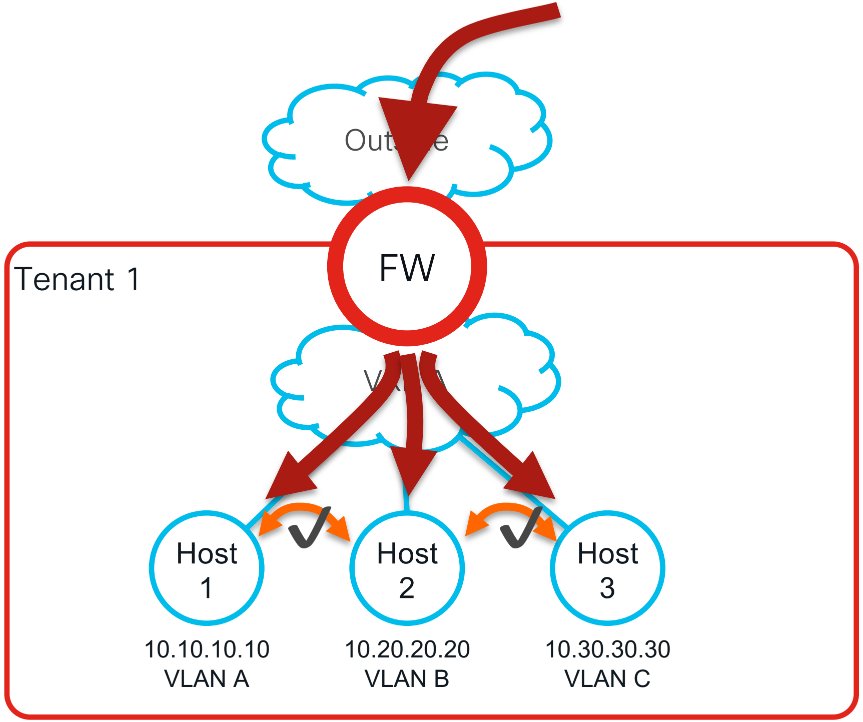
Physical/Logical Topology (Alternative Option)



FW connected to a Border Leaf

Tenant Edge Firewall

Terminology

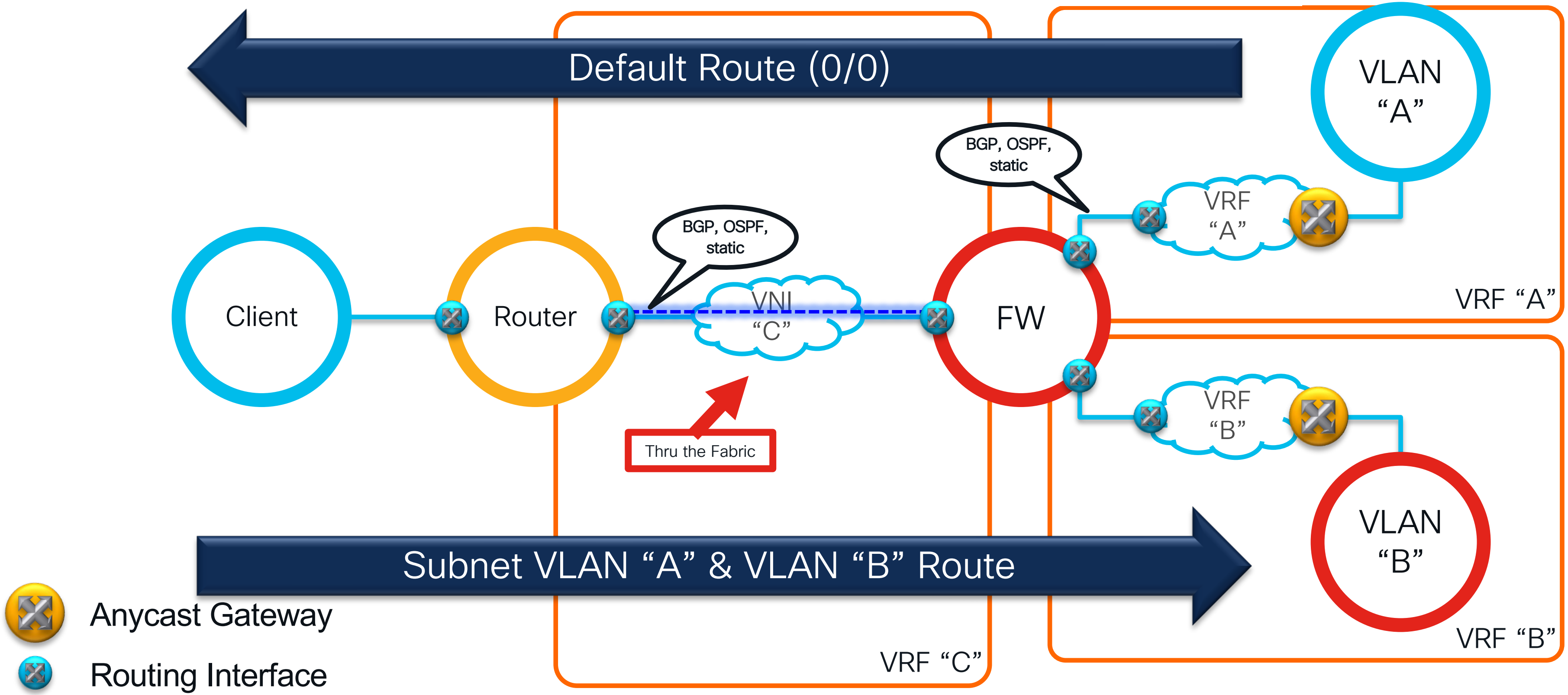


- Edge Firewall front-ends a Tenant (VRF) to control connectivity to another Tenant (VRF) or external network (North/South)
- All traffic is permitted / denied based on Services-Node policy

Tenant Edge Firewall

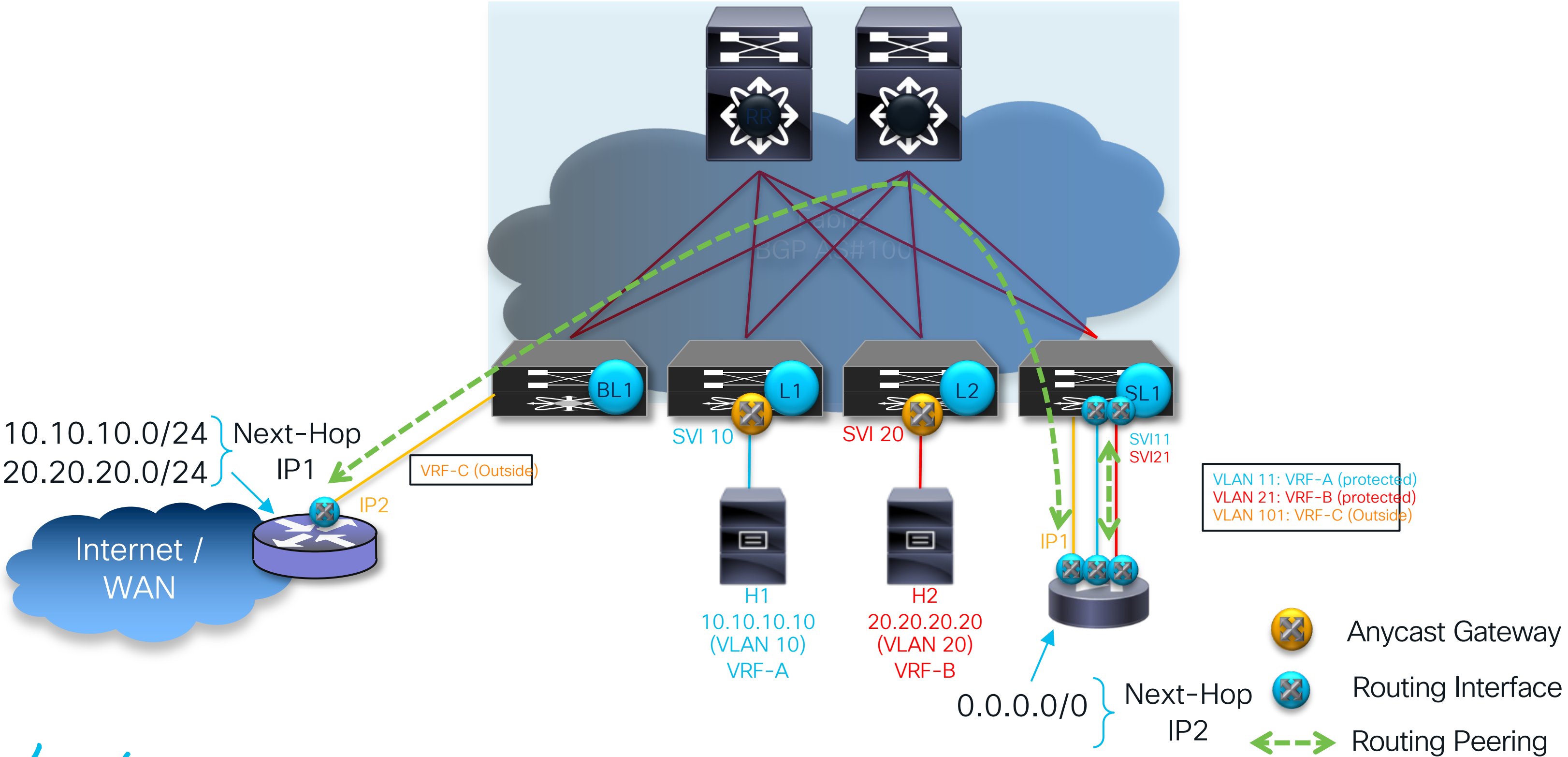
L3 FW Peering with the Fabric and the External Router

1



Tenant Edge Firewall

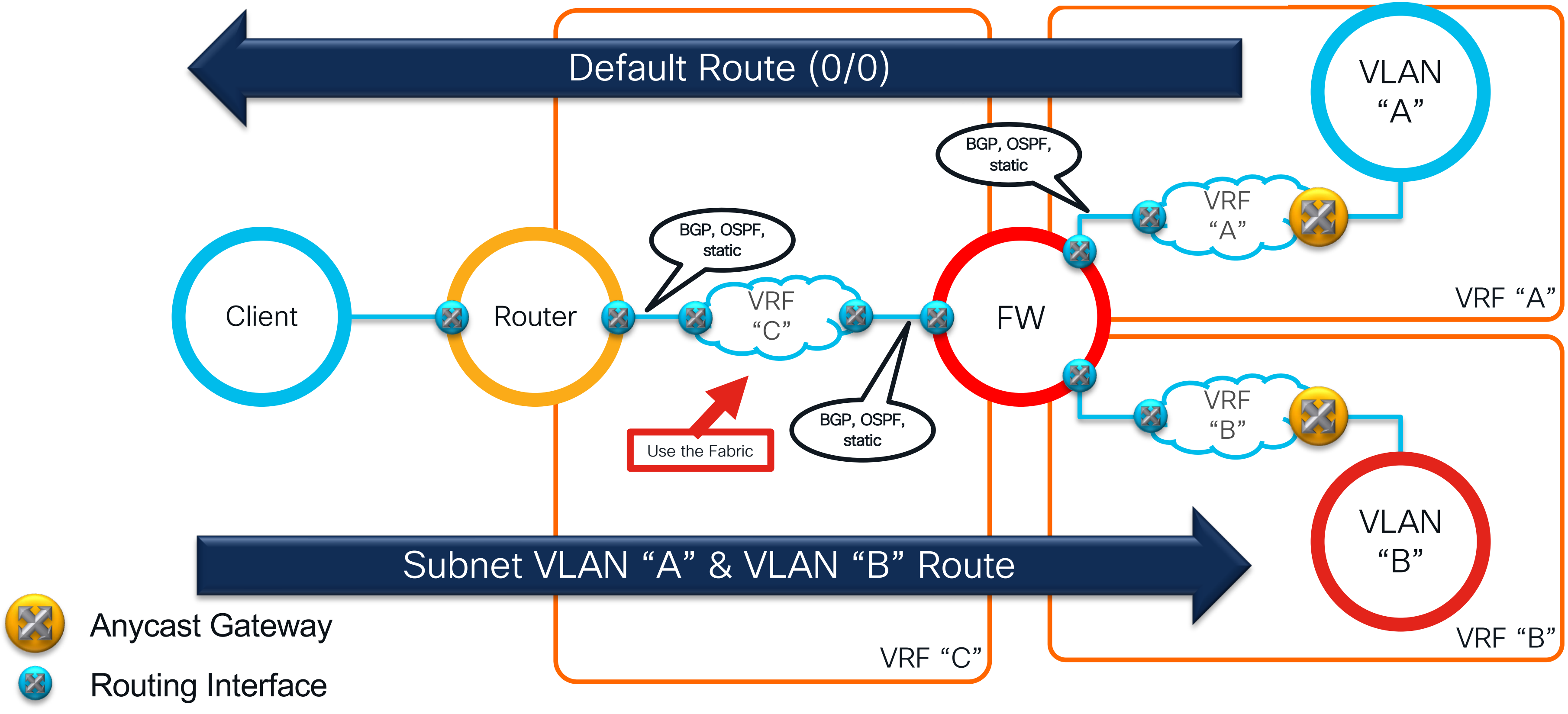
L3 FW Peering with the Fabric and the External Router



Tenant Edge Firewall

L3 FW Peering Only with the Fabric

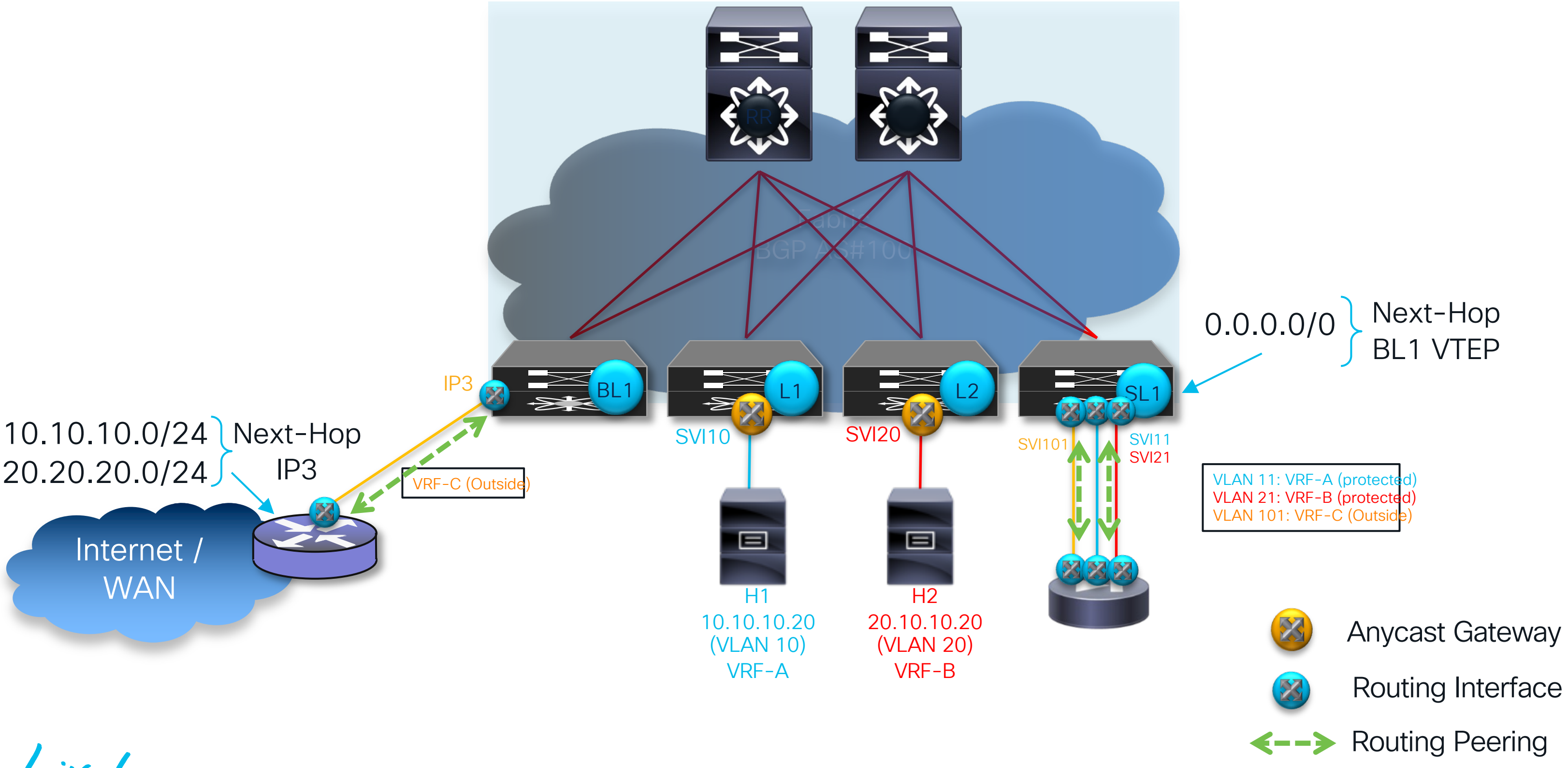
2



2

Tenant Edge Firewall

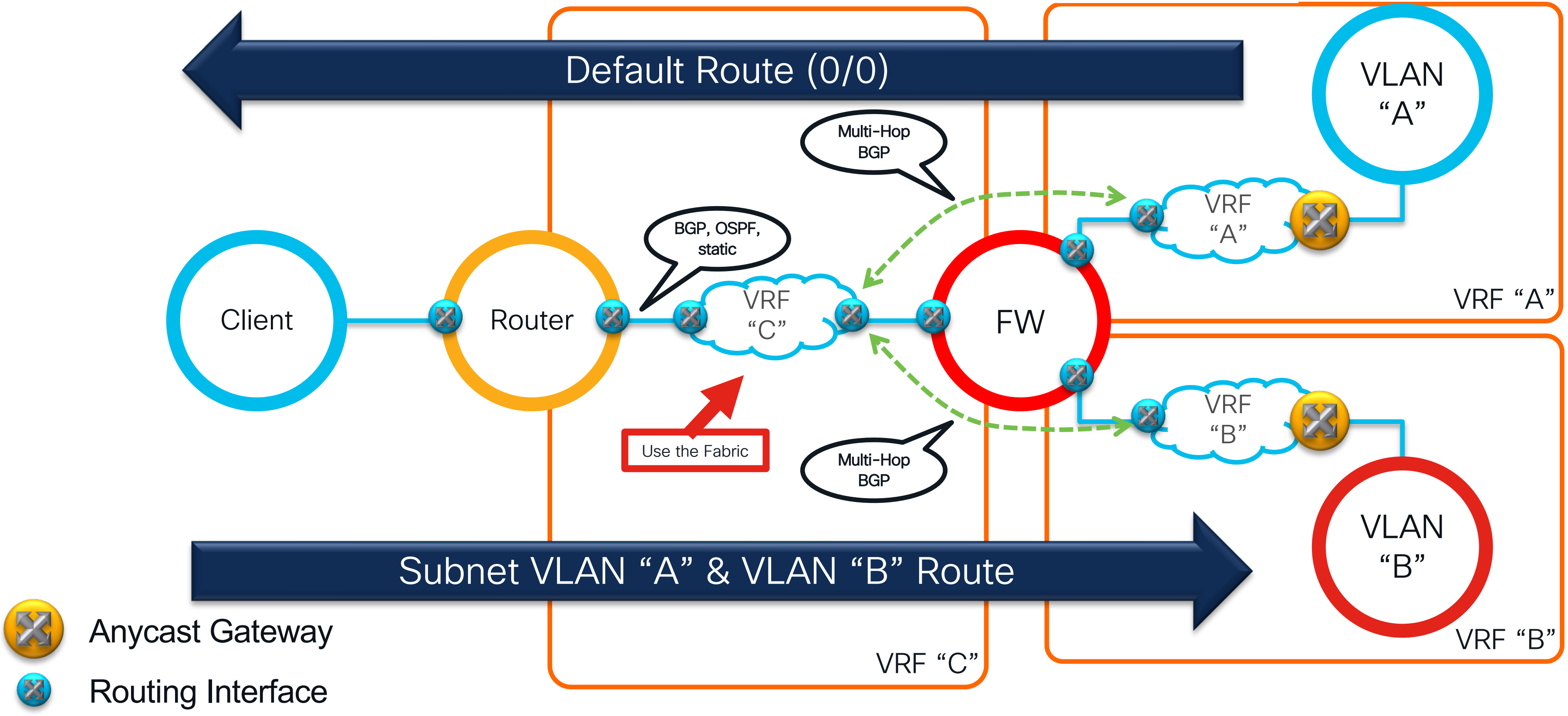
L3 FW Peering Only with the Fabric



Tenant Edge Firewall

Multi-Hop BGP Peering through the L3 FW

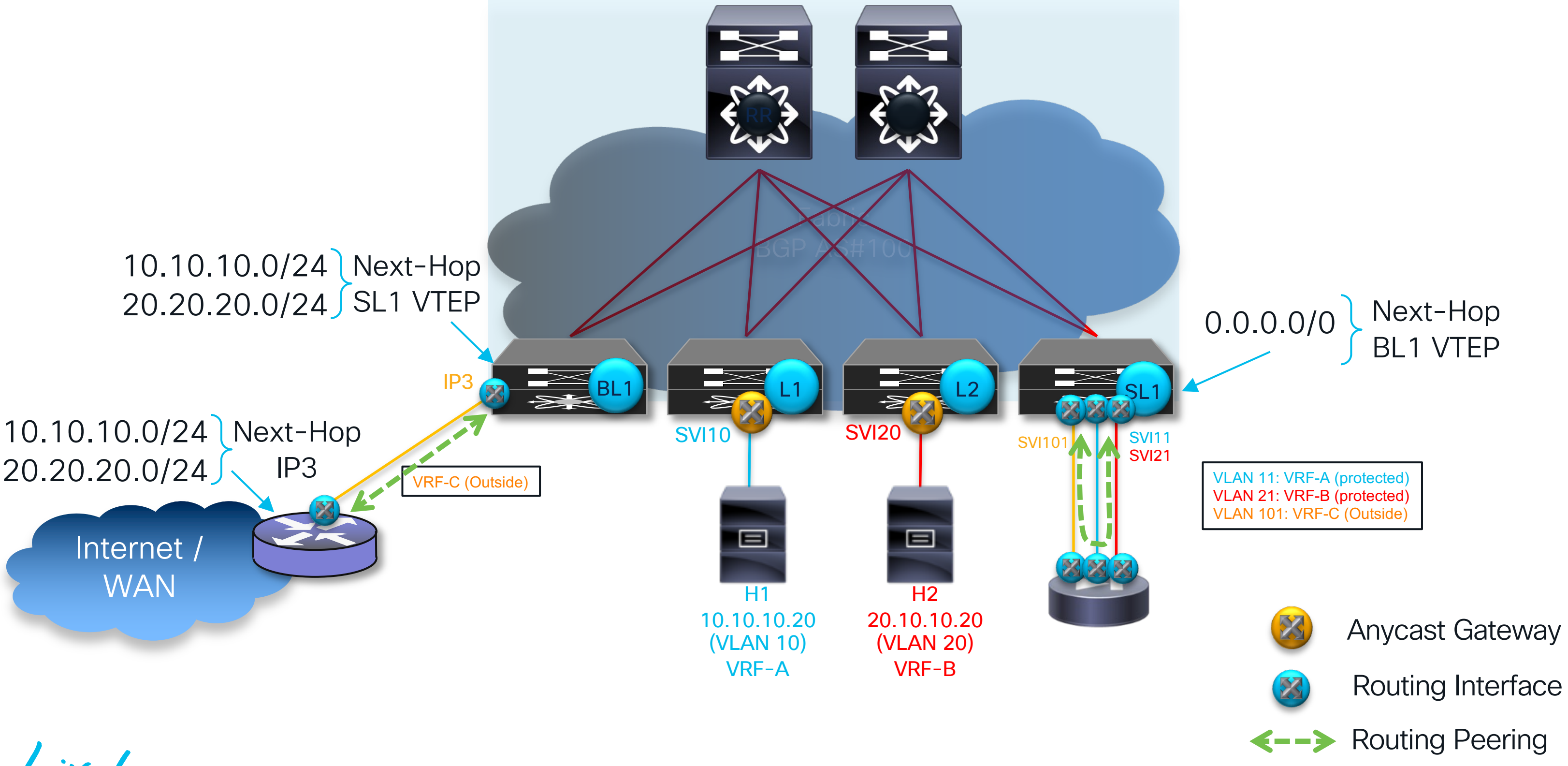
3



2

Tenant Edge Firewall

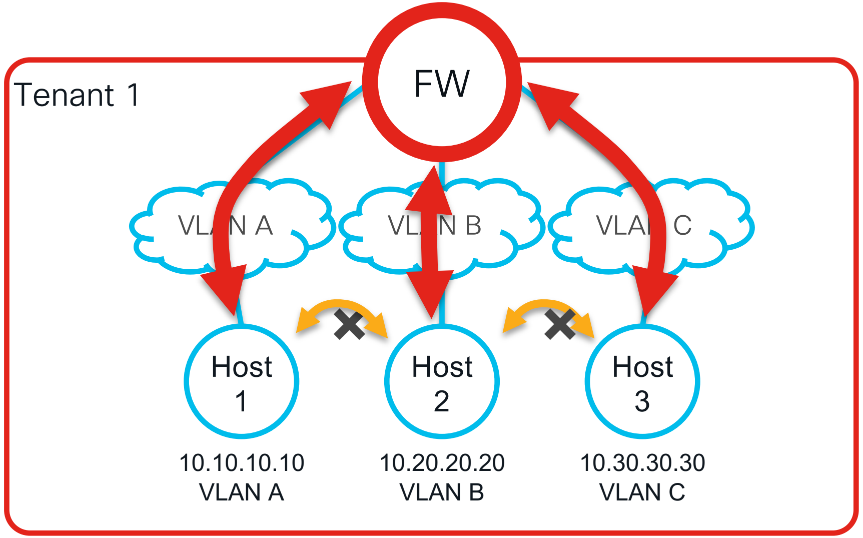
Multi-Hop BGP Peering through the L3 FW



Intra Tenant Firewall

Intra Tenant Firewall

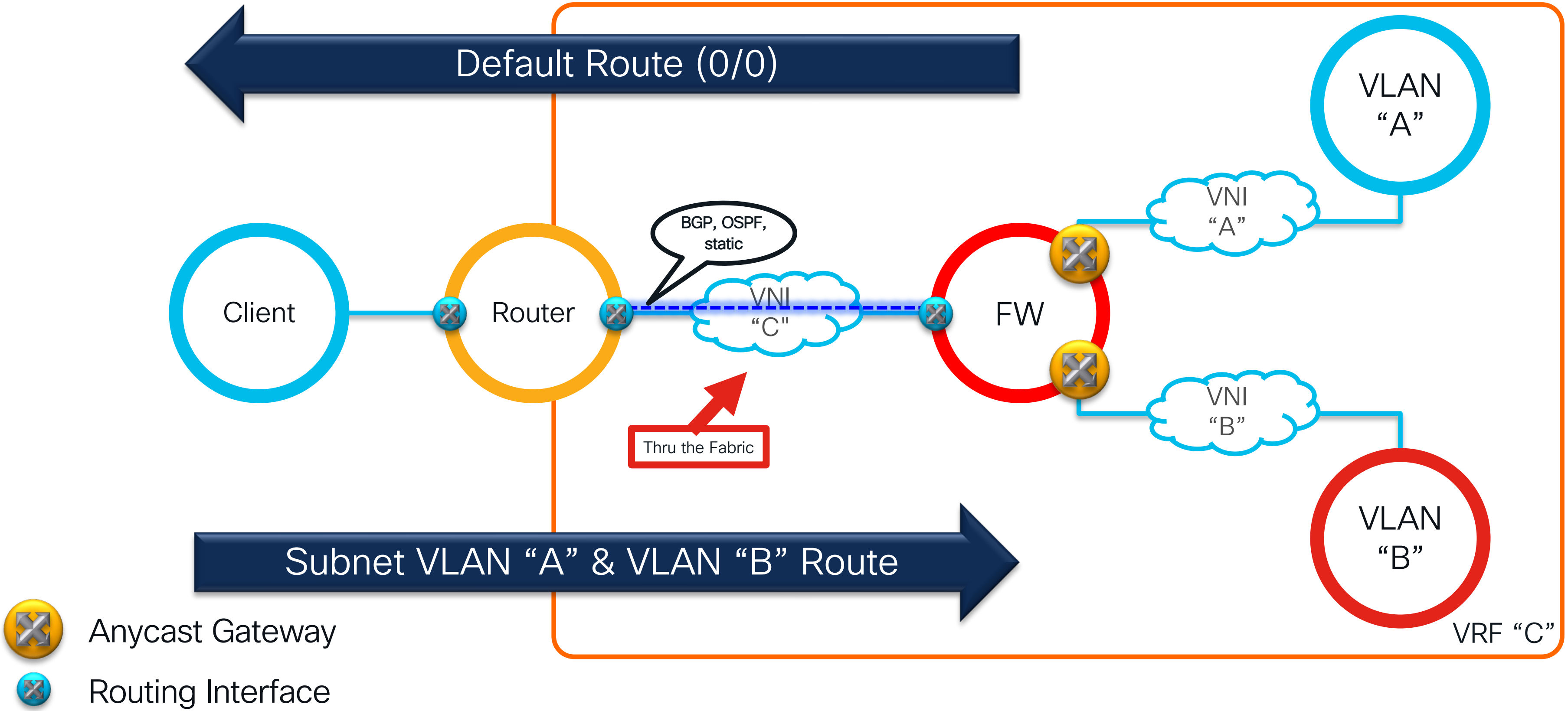
Terminology



- Edge Firewall that inspects traffic between endpoints within the same VRF (East/West)
- Follows traditional bridging towards endpoints with default gateway on the Service-Node
- Alternatively use EPBR if the default gateway is on the fabric
- All traffic is permitted / denied based on Services-Node policy

Intra Tenant Firewall

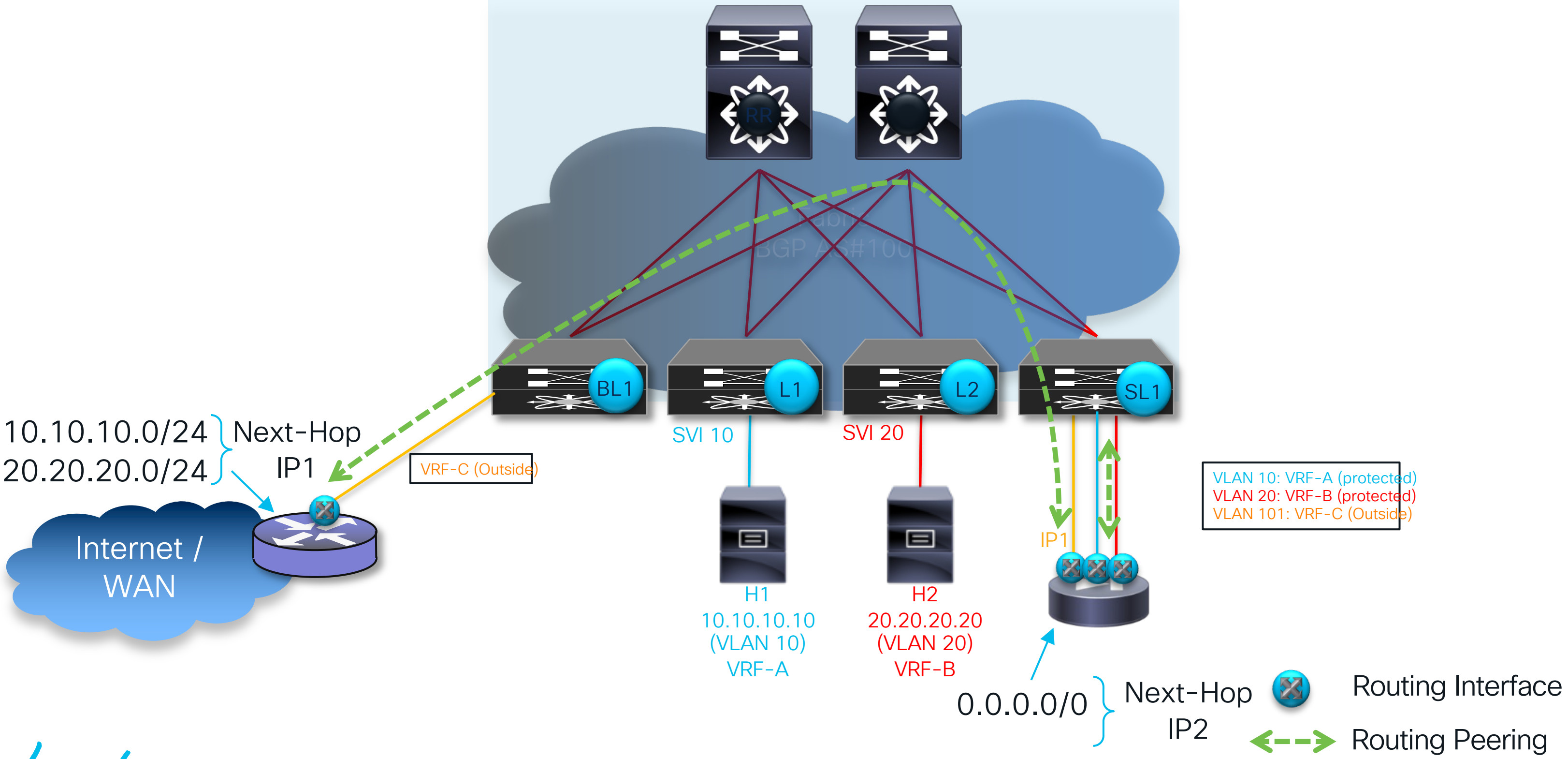
Firewall as Default Gateway Peering with the External Router



Intra Tenant Firewall

Firewall as Default Gateway Peering with the External Router

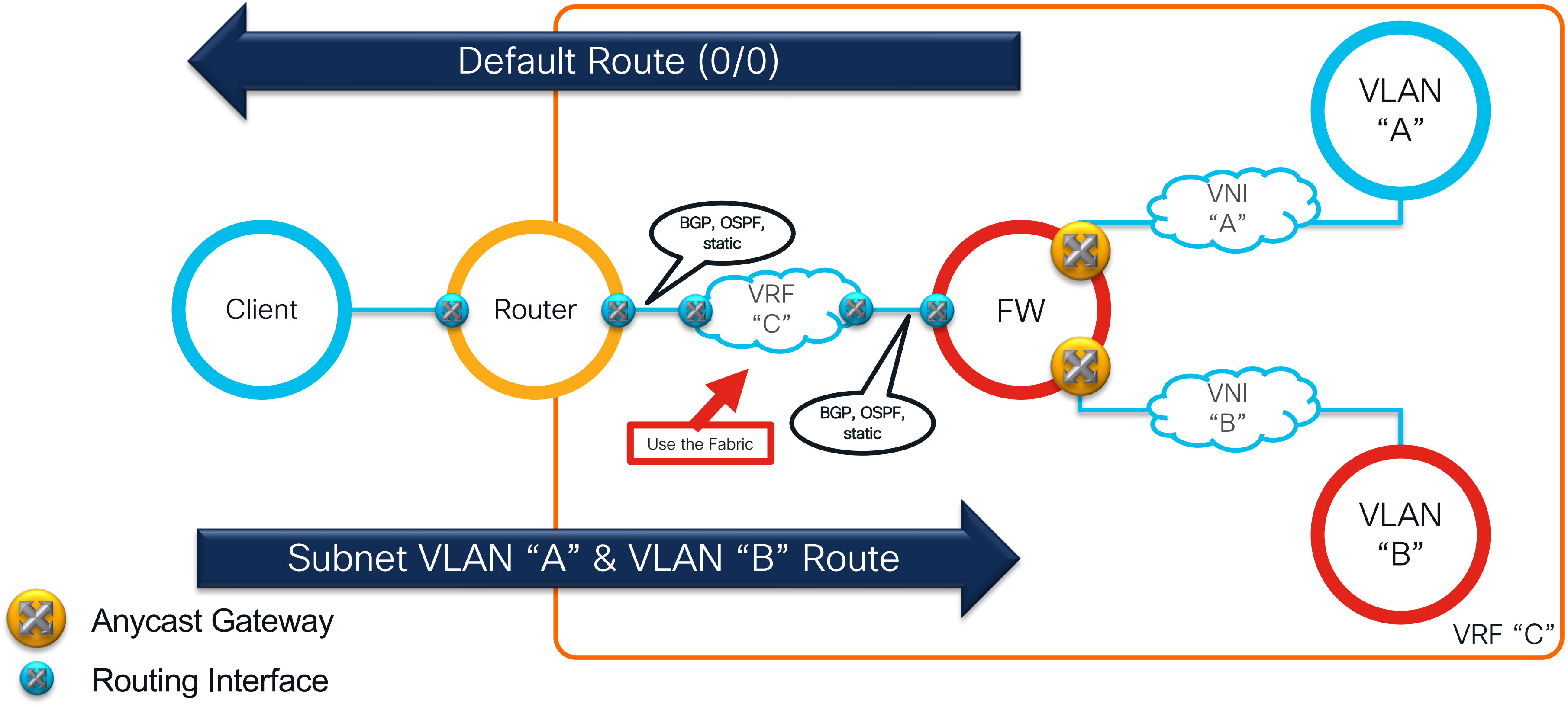
1



Intra Tenant Firewall

Firewall as Default Gateway Peering with the Fabric

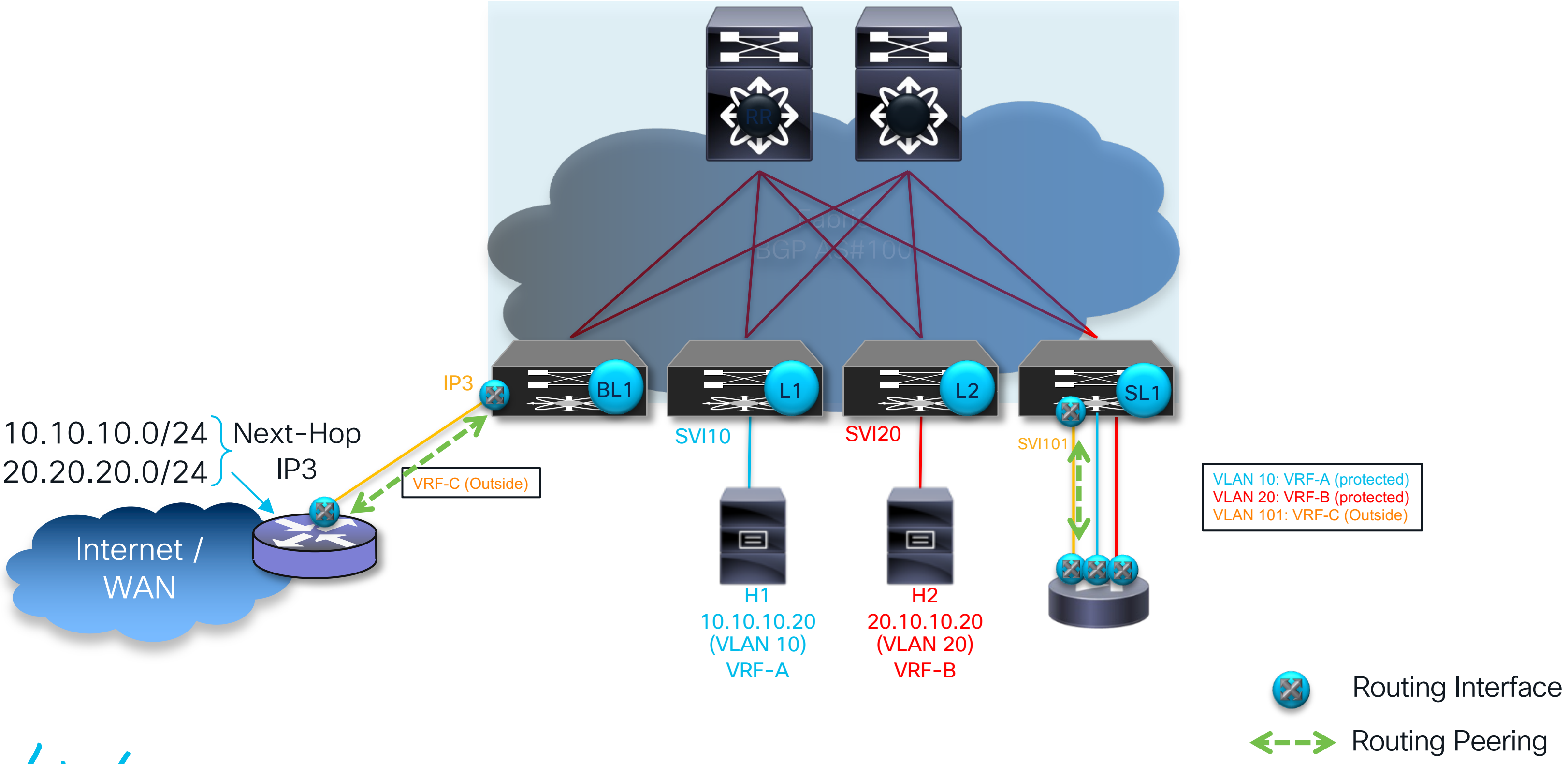
2



2

Intra Tenant Firewall

Firewall as Default Gateway Peering with the Fabric



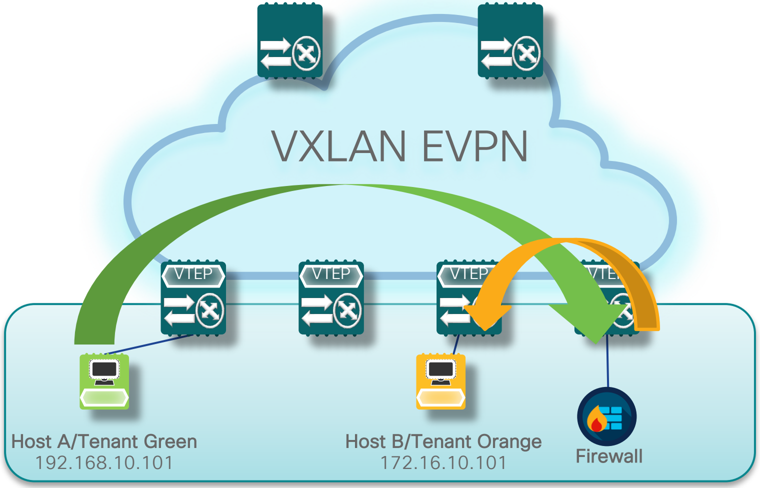
What if I don’t want to use the FW as Default Gateway?

What if I don’t want to Enhanced Policy-based Redirect (ePBR) the FW as Default Gateway?

Enhanced PBR

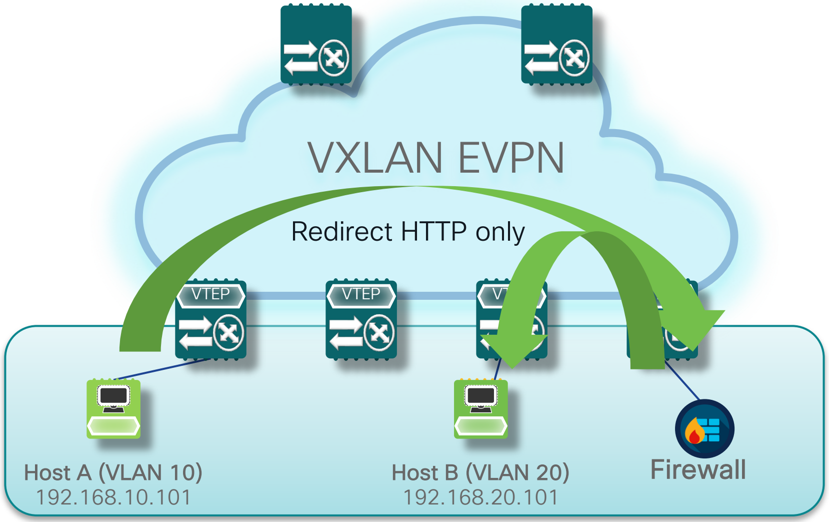
Enforcing Infra-VRF Policy Enforcement

Inter-VRF Enforcement (Routing Driven)



Routing rules reflect path via service devices

Intra-VRF Enforcement with EPBR



Selective Traffic Redirect using Policy Based Routing

Enhanced PBR

Solution Overview

1. Onboard Service Appliance

- Service IP address
- Forward and reverse attached interface (single/dual arm)
- Probes
- VRF membership
- Additional service end-points for creating appliance cluster

2. Define traffic redirect Policy

- Traffic Filtering or selection ACL
- Service-chain creation
- Load-balancing options (src/dst and buckets)
- Failover options (forward/bypass/drop)

3. Apply the ePBR Policy on relevant interfaces

- Apply policy on ingress interface where chaining needs to start
- VXLAN – Apply on L3 VNI interfaces on service leaf
- Apply policy with “reverse” keyword to maintain flow symmetry

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/93x/epbr/cisco-nexus-9000-series-nx-os-epbr-configuration-guide-93x/m-configuring-epbr.html

Enhanced PBR
Configuration Example

```
epbr service FW
   probe icmp source-interface loopback9
   vrf CustomerA-Service
   service-endpoint ip 193.40.1.1 interface VLAN401
           reverse ip 193.40.1.1 interface VLAN401

ip access-list WEB
10 permit tcp any any eq 80
20 permit tcp any any eq 443

epbr policy CustomerA-Redirect
   match ip address WEB
   load-balance method src-ip
    10 set service FW fail-action drop

interface vlan 2010
  !L3 VNI SVI
  epbr ip policy CustomerA-Redirect
  epbr ip policy CustomerA-Redirect reverse

interface vlan 301
  !SVI for tenant traffic/ingressing fabric
  epbr ip policy CustomerA-Redirect
```

Creates IP SLA and track

Set VRF for FW Needs to be deployed on every node doing redirect

Forward arm

Reverse arm

Single Armed FW

ACL matches web traffic

Define EPBR Policy

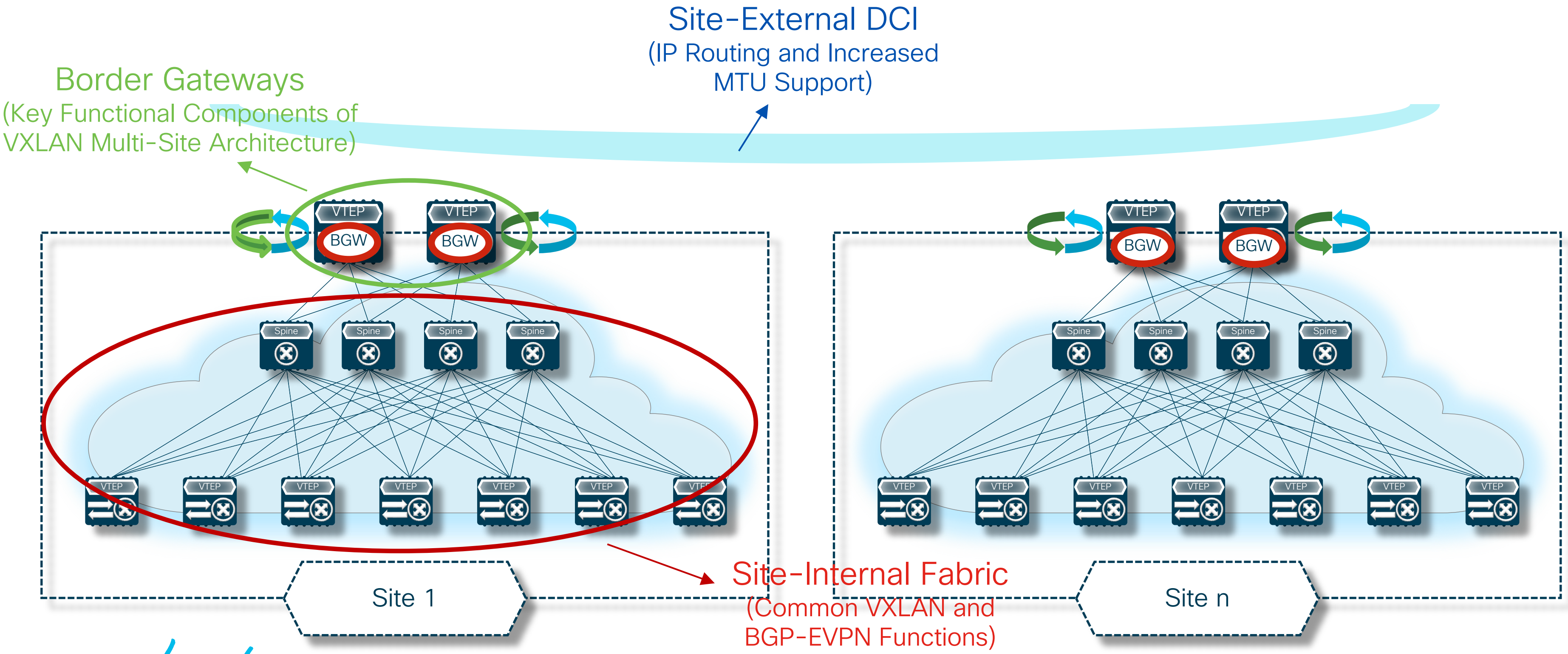
Policy needed on all interfaces where traffic can ingress

Layer 4-7 Services Integration in a VXLAN Multi-Site Architecture

VXLAN Multi-Site

Functional Components

https://tools.ietf.org/html/draft-sharma-multi-site-evpn

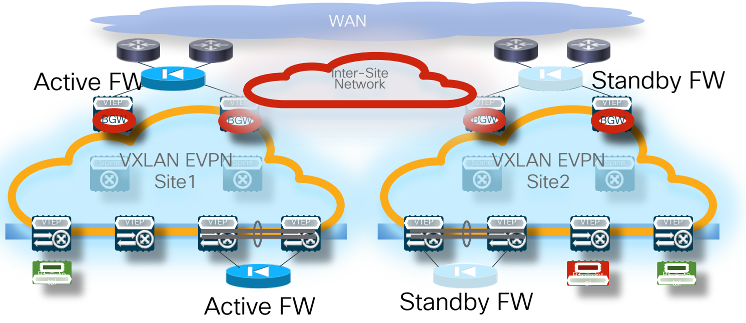


Site-External DCI
(IP Routing and Increased MTU Support)

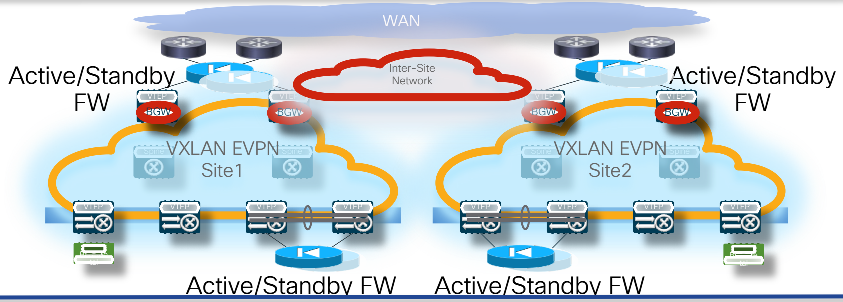
Border Gateways
(Key Functional Components of VXLAN Multi-Site Architecture)

Site-Internal Fabric
(Common VXLAN and BGP-EVPN Functions)

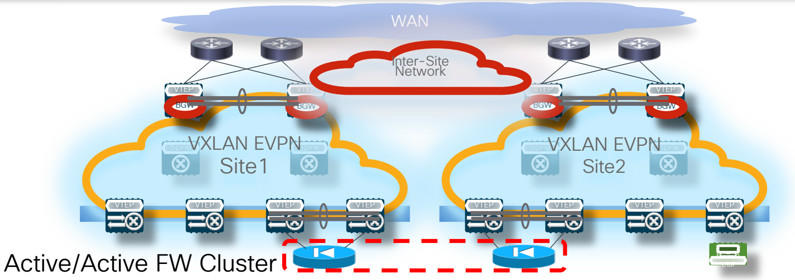
VXLAN Multi-Site and Network Services Integration



- Active and Standby pair deployed across Sites, enforcement for N-S and E-W flows
- No issues with asymmetric flows
- Various options possible (FW as endpoints gateway or fabric as endpoints gateway)



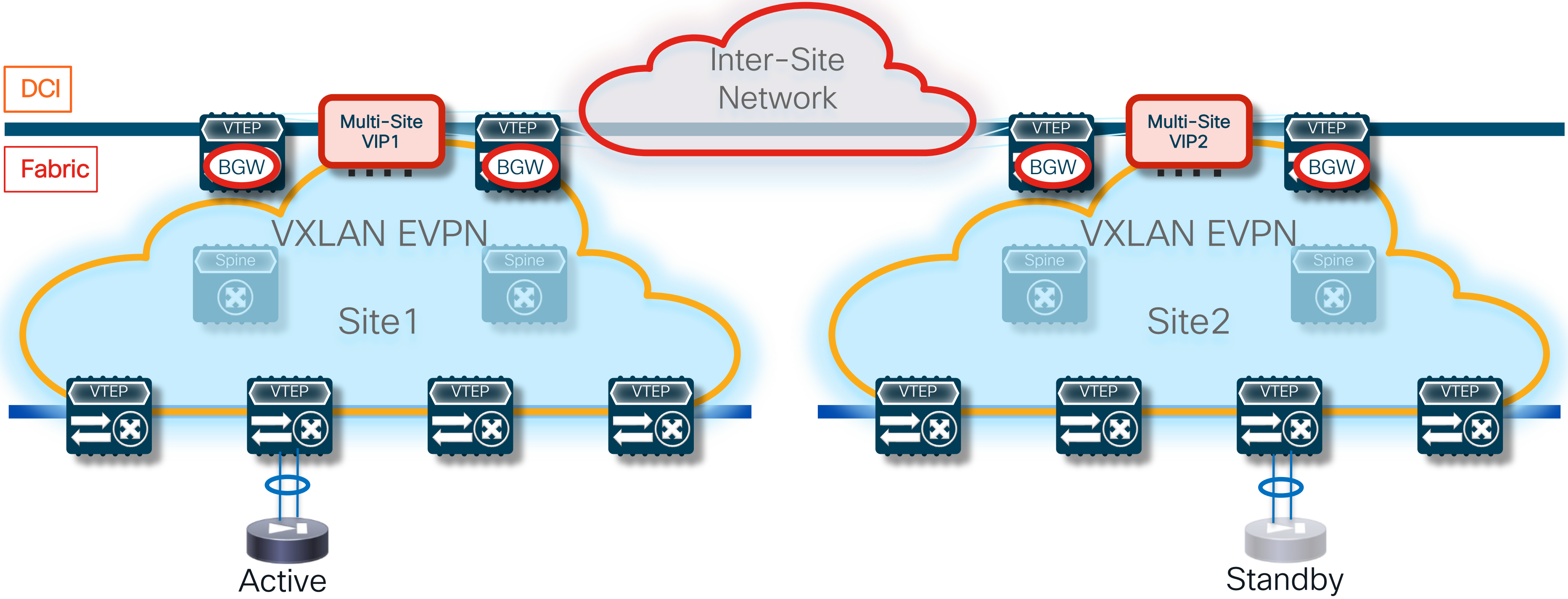
- Independent Active/Standby pairs deployed in separate Sites
- Need to avoid the creation of asymmetric paths crossing different active FW nodes
  - Only possible for N-S flows with perimeter FWs and host routes advertisement or with PBR



- Active/Active FW Cluster stretched across Sites
  - Split spanned ether-channel mode: supported with Cisco ASA/FTD from NX-OS release 10.2(2)
  - Individual mode: supported with Cisco ASA software for N-S and E-W flows

Active/Standby Pair Stretched across Sites

Active and Standby pair deployed across Sites



Active/Standby Pair across Sites

Deployment Considerations

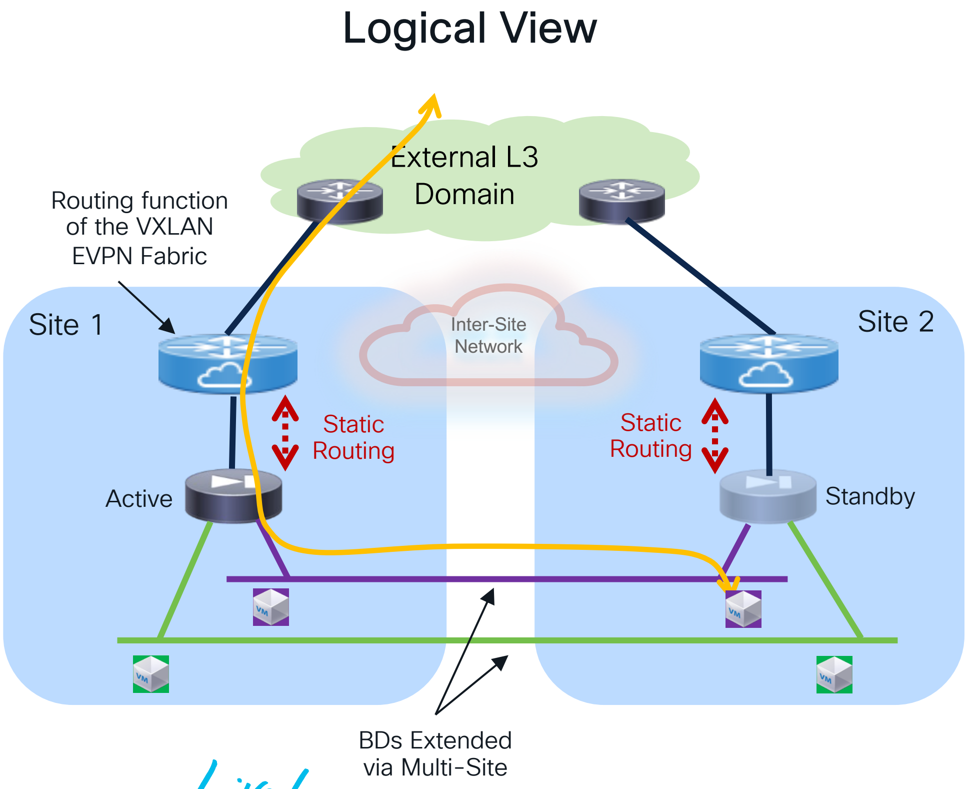
- Active/Standby model can be applied per context (i.e. can be deemed as 'active/active' support across contexts)
- Different deployment models
  - FW as default gateway for the endpoints using static routing
  - FW as default gateway for the endpoints peering with the fabric (via IGP or BGP)
  - FW as default gateway for the endpoints peering directly with the external routers (fabric as L2)
  - Fabric as default gateway and use of a perimeter FW

1. FW as Default Gateway Using Static Routing with the Fabric

1

Active/Standby Pair across Sites

FW as Default Gateway Using Static Routing with the Fabric

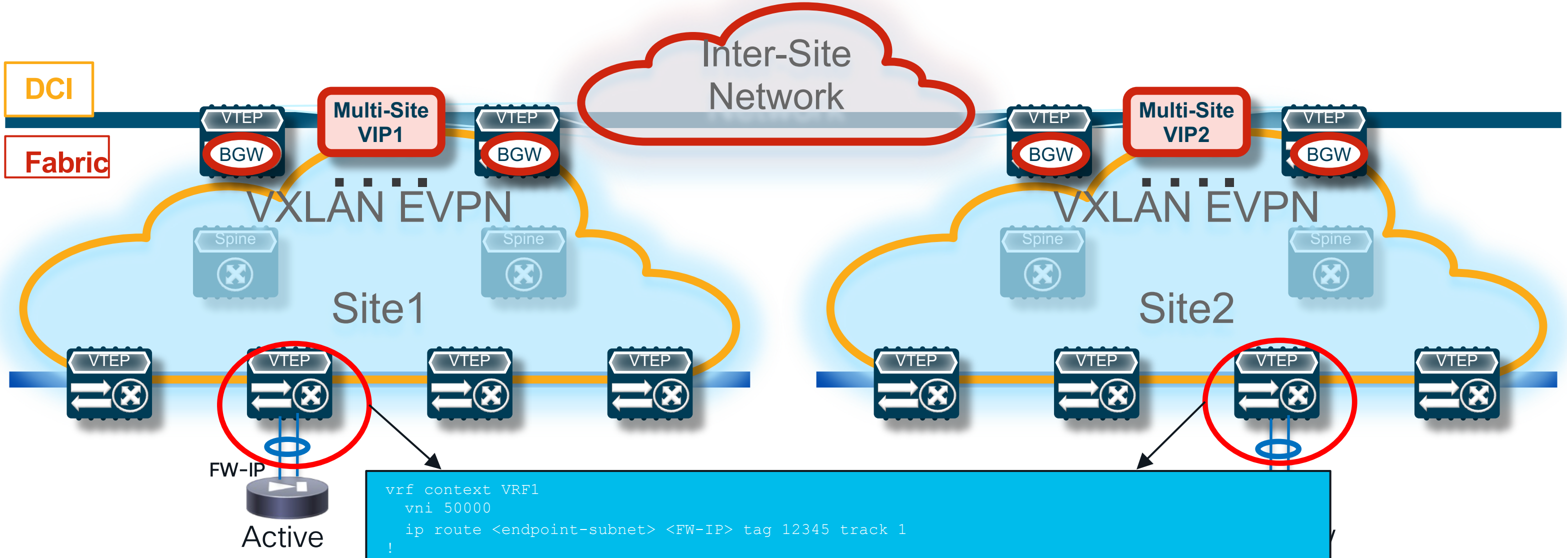


- FW allows to apply intra-tenant security policies (east-west) and between an internal subnet and the external L3 domain (north-south) or a subnet in a different tenant (inter-tenant)
- FW inside network(s) deployed as L2-only can be extended across sites to allow flexible deployment for endpoints
- Two deployment options:
  1. Centralized static routing with HMM tracking
  2. Distributed static routing with recursive next-hop

FW Using Static Routing with the Fabric

Centralized Static Routing with HMM Tracking (Configuration)

1



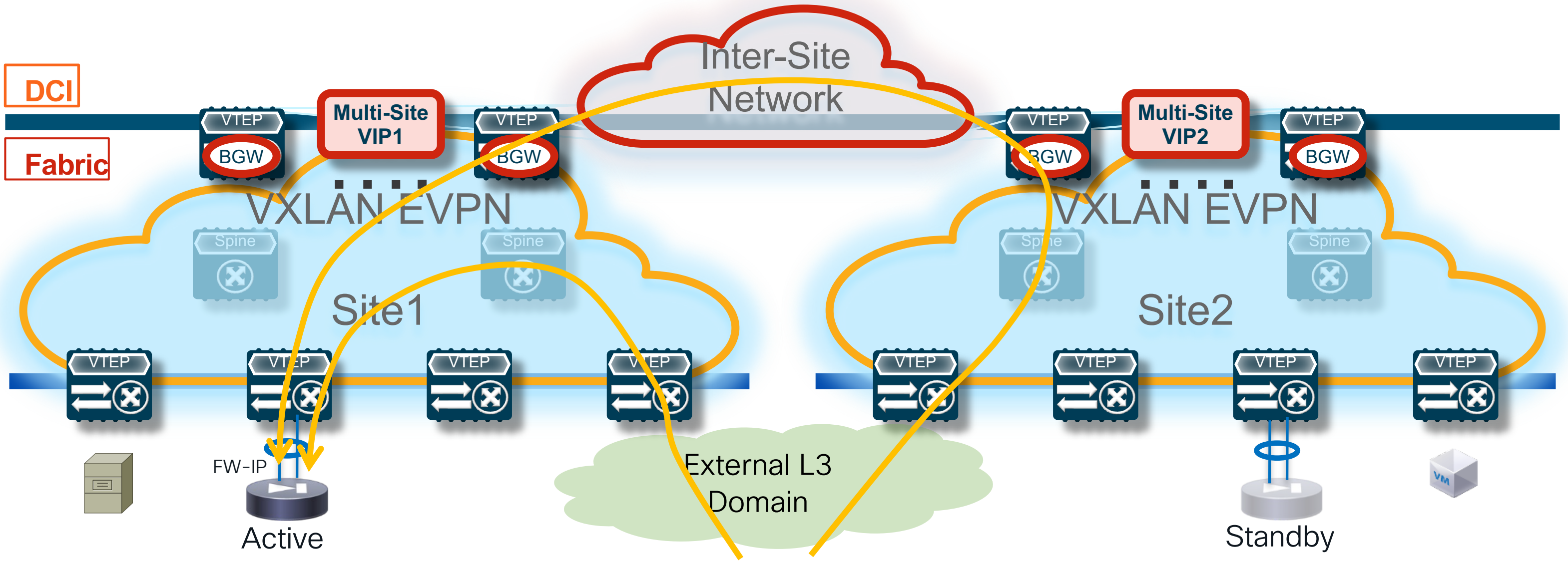
```
vrf context VRF1
  vni 50000
  ip route <endpoint-subnet> <FW-IP> tag 12345 track 1
!
track 1 ip route <FW-IP> reachability hmm
  vrf member VRF1
!


router bgp 65001
  vrf customera
    address-family ipv4 unicast
      advertise l2vpn evpn

      redistribute static route-map fabric-rmap-redist-subnet
```

Config applied only only on the leaf nodes connected to the Active and Standby FWs

1

FW Using Static Routing with the Fabric

Centralized Static Routing with HMM Tracking

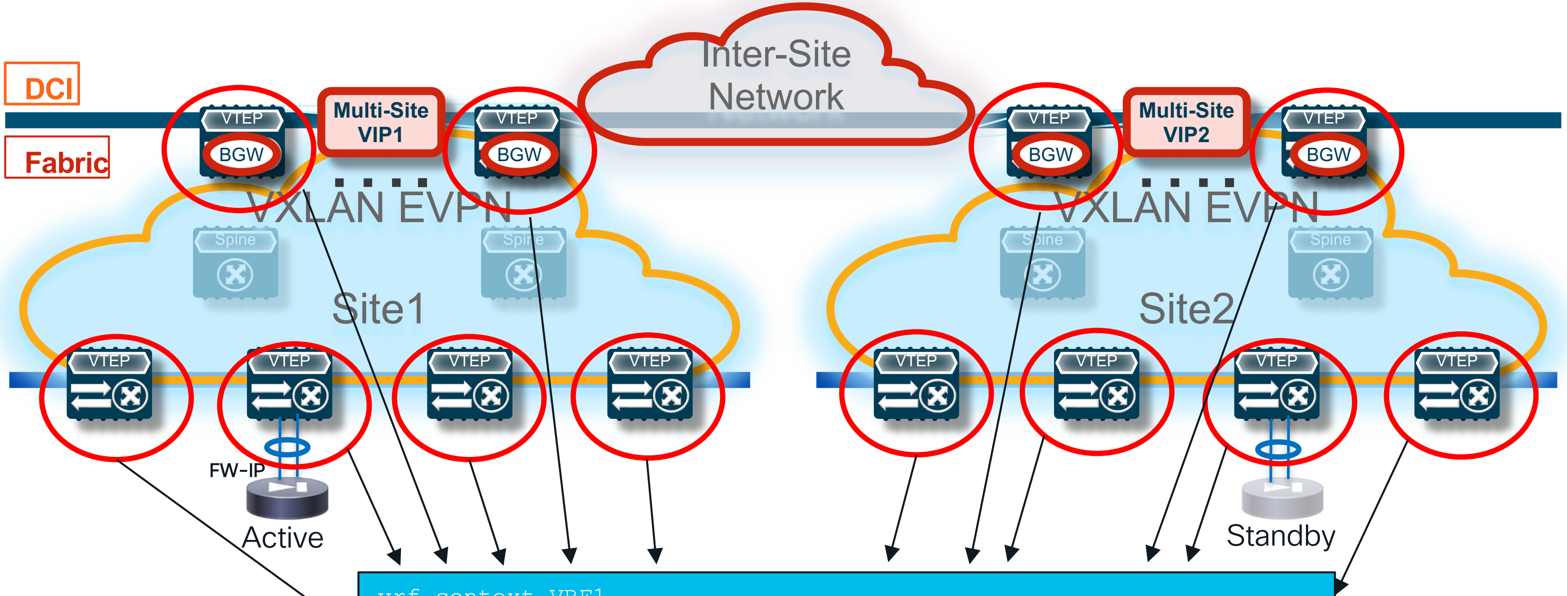


Traffic destined to endpoints behind the FW is always encapsulated toward the leaf node connected to the active FW

1

FW Using Static Routing with the Fabric

Distributed Static Routing with Recursive Next-Hop (Configuration)



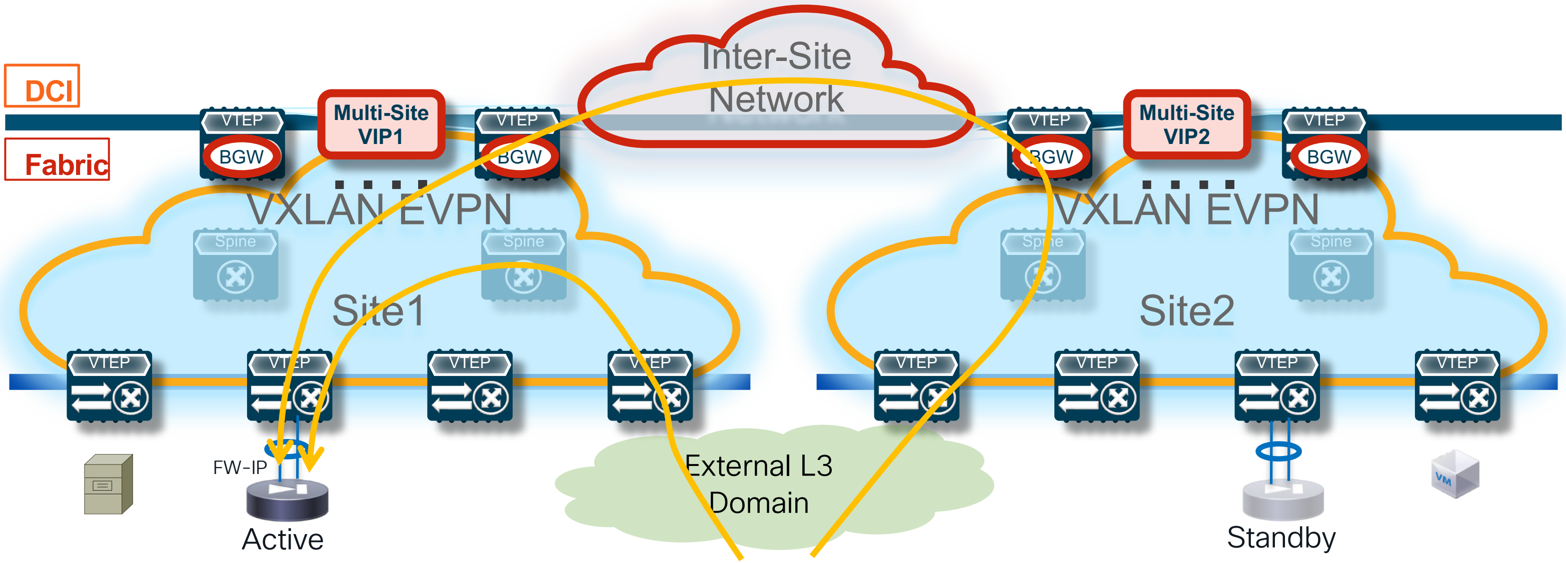
```
vrf context VRF1
  vni 50000
  ip route <endpoint-subnet> <FW-IP>
```

Config applied on all the leaf nodes and
also on the Border Gateways

FW Using Static Routing with the Fabric

Distributed Static Routing with Recursive Next-Hop

1



Traffic destined to endpoints behind the FW is always encapsulated toward the leaf node connected to the active FW

FW Using Static Routing with the Fabric

Centralized vs. Distributed Static Routing

1

Centralized Static Routing with HMM Tracking

- 👍 Centralized configuration (few touch points)
- 👎 Convergence depending on HMM tracking and static routing redistribution into EVPN
- 👎 Scalability dependent on the number of routes to redistribute

Distributed Static Routing with Recursive Next-Hop

- 👍 Simpler configuration
- 👍 Recursive Next-Hop functionality natively integrated into VXLAN EVPN
- 👍 Convergence only dependent on FW-IP discovery
- 👎 Distributed configuration (many touch points), can be simplified with a provisioning tool (NDFC)

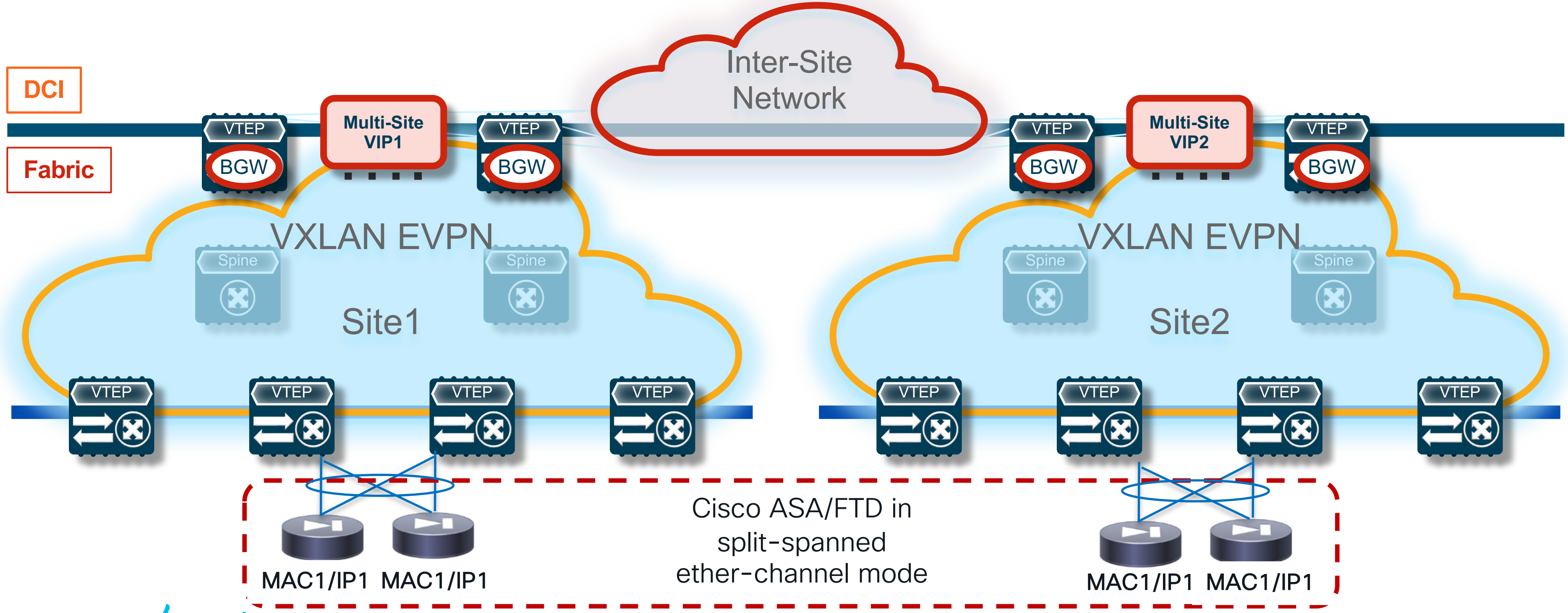
Active/Active FW Cluster across Sites

Active/Active FW Cluster across Sites

Split-Spanned Ether-Channel Mode

NX-OS 10.2(2)

Requires anycast IP service support across Sites



Cisco ASA/FTD in split-spanned ether-channel mode

NX-OS 10.2(2)

Active/Active FW Cluster across Sites

Overview

- FW cluster consists of multiple members, acting as a single device
- FW cluster is connected via L2 port-channel spanned across all cluster members (aka split spanned Etherchannel)
- Same cluster VIP/ cluster VMAC learnt across all instances
- BGP-EVPN VXLAN overlay per site, stitched at Border Gateway Nodes
- Each Site will have a <u>single VPC pair </u>connected to a part of cluster with a Port-channel interface that has an ESI assigned to it
- The cluster VIP and cluster VMAC will be advertised into the VXLAN/EVPN fabric as BGP EVPN RT-2 with the ESI set to the configured value VPC's Port-channel of each VPC pair. The next hop of the RT-2 will be the VPC pair's VTEP VIP address

Active/Active FW Cluster across Sites

Supported Deployment Models

NX-OS 10.2(2)

1. Firewall cluster as Default GW

   Static routing between the FW and the fabric

2. Default Gateway in the Fabric, Firewall at the perimeter

   Fabric peering multihop with external router (static routing on the FW)

3. FW one-arm mode and use of EPBR

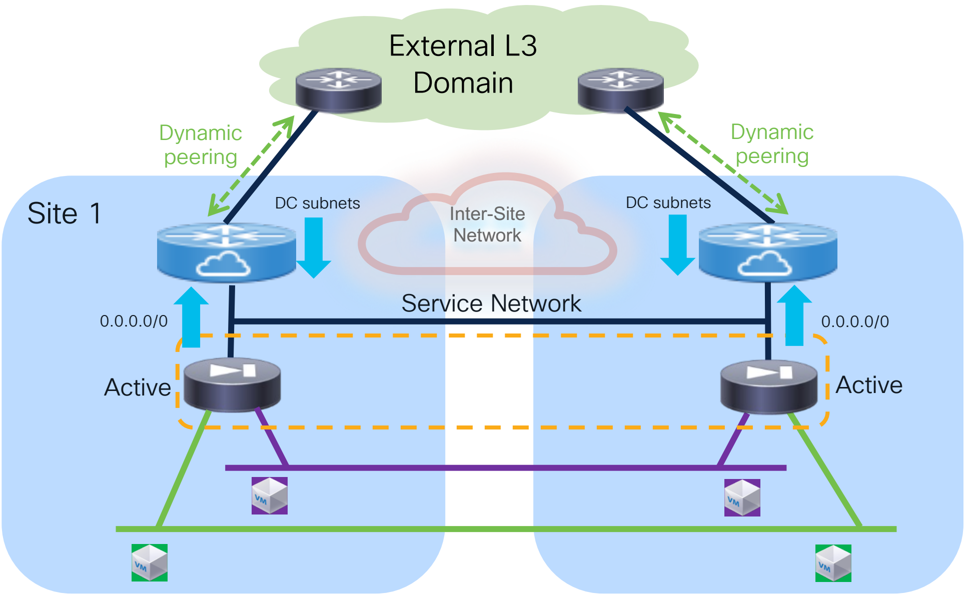
1. FW as Default Gateway

1

Active/Active Cluster across Sites

FW as Default GW and Static Routing between the FW and the Fabric

Logical View



- Service network defined to peer between FW and fabric
  - Must be stretched to ensure GARP can be sent across sites after a FW failover event
- Default gateway function on the FW distributed across sites
  - FW filtering function applied between subnets of the same VRF and for north-south communication
- Static routing between the FW and the fabric
- Dynamic peering between the fabric and the external routers

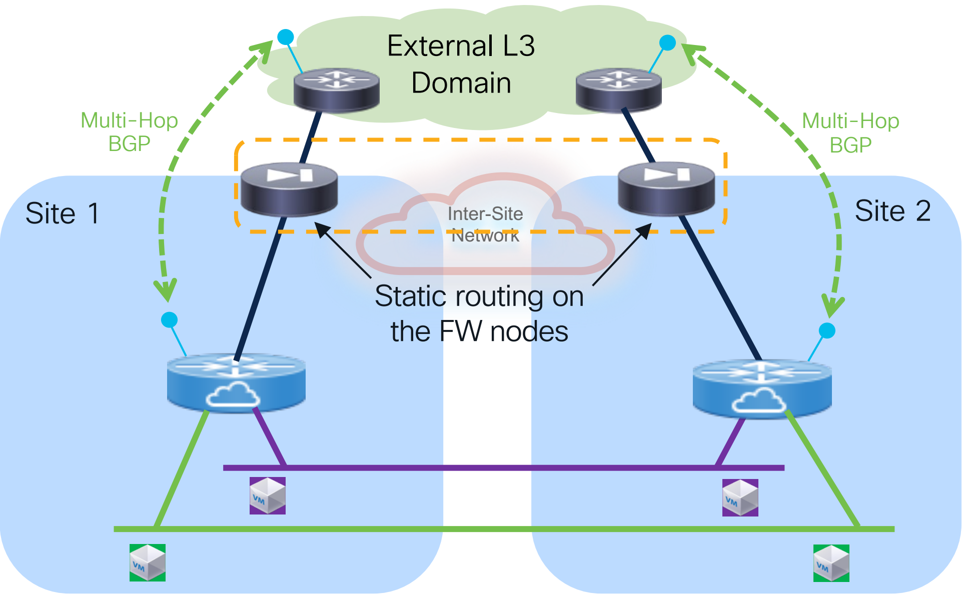
2. Tenant Edge Firewall

2

Active/Active Cluster across Sites

Use of Tenant Edge FW and HBR (North-South or inter-VRF)

Logical View



- Multi-Hop BGP session established between the fabric and the external routers through the FW cluster nodes
- Static routing only required on the FW nodes
- Host-routes for inbound traffic optimization not learned by the FW cluster
- FW enforcement applied to inter-VRF flows and to north-south communication

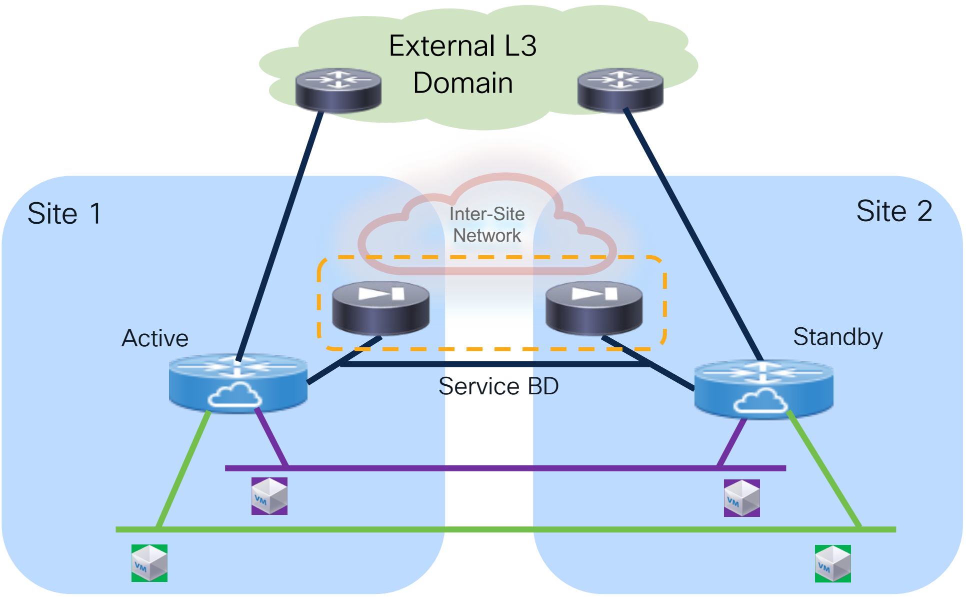
3. FW One-Arm Mode and use of EPBR

3

Active/Active Cluster across Sites

FW one-arm mode and use of EPBR

Logical View



- Service BD defined to connect the one-arm FW
  - Must be stretched to ensure reachability to the active FW for EPBR
  - Service BD must be part of a dedicated VRF (EPBR uses the “set VRF” option to redirect traffic to a service node in a remote site
  - 0.0.0.0/0 route only required on the FW
- FW enforcement for intra-VRF, inter-VRF and north-south flows

cisco Live!

Did you know?

You can have a one-on-one session with a technical expert!

Visit Meet the Expert in The HUB to meet, greet, whiteboard & gain insights about your unique questions with the best of the best.

Meet the Expert Opening Hours:

Tuesday	**3:00pm – 7:00pm**
Wednesday	**11:15am – 7:00pm**
Thursday	**9:30am - 4:00pm**
Friday	**10:30am - 1:30pm**

Session Surveys

We would love to know your feedback on this session!

- Complete a minimum of four session surveys and the overall event surveys to claim a Cisco Live T-Shirt

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Expert meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

The bridge to possible

Thank you

#CiscoLiveAPJC

cisco Live!

Let’s go

#CiscoLiveAPJC