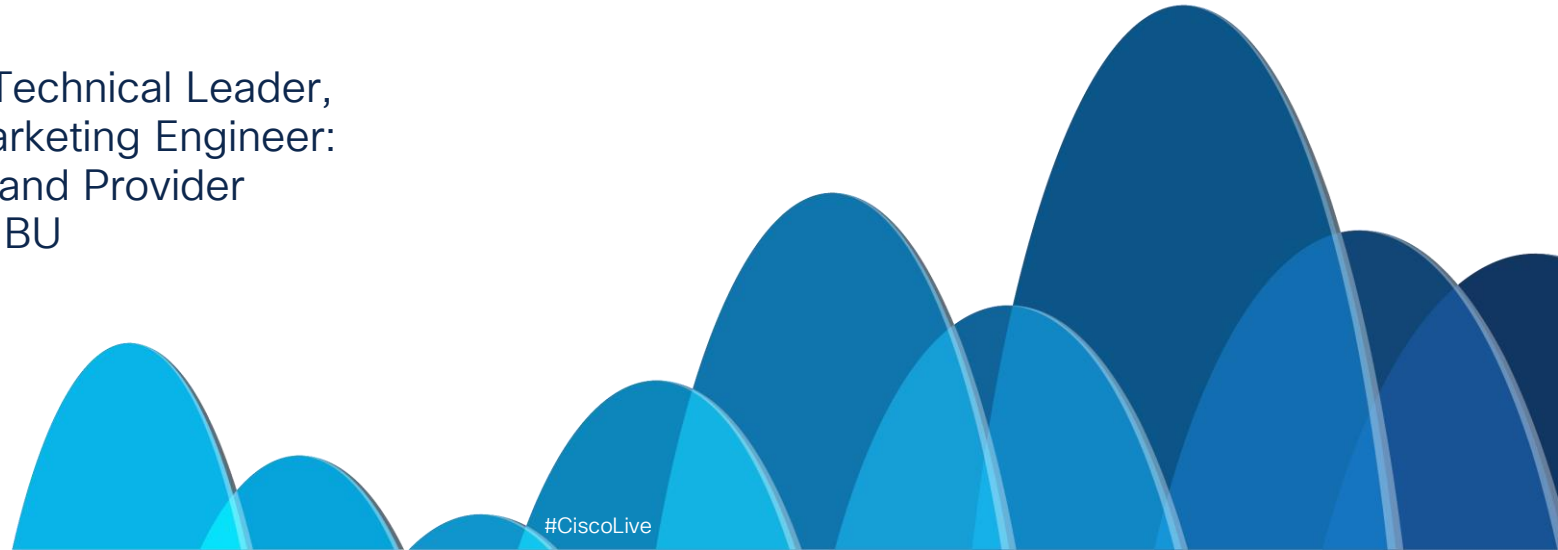
The bridge to possible

Design and Automate VXLAN Multi-Site with NDFC

Parth Patel, Technical Leader,
Technical Marketing Engineer:
Data Center and Provider
Connectivity BU

BRKDCN-2988



cisco Live!

Cisco Webex App

https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-2988

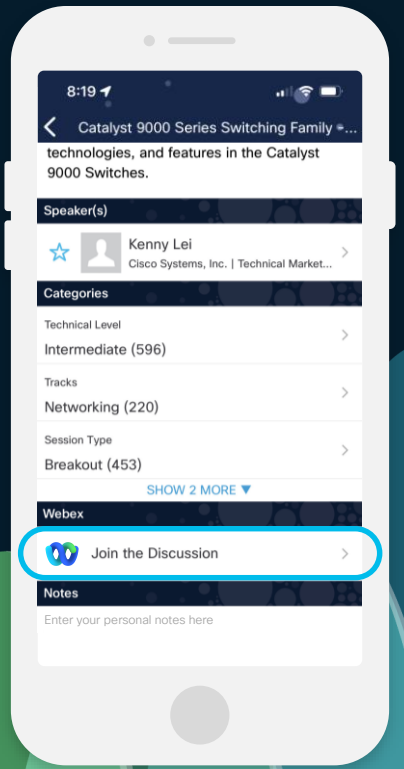
Questions?

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

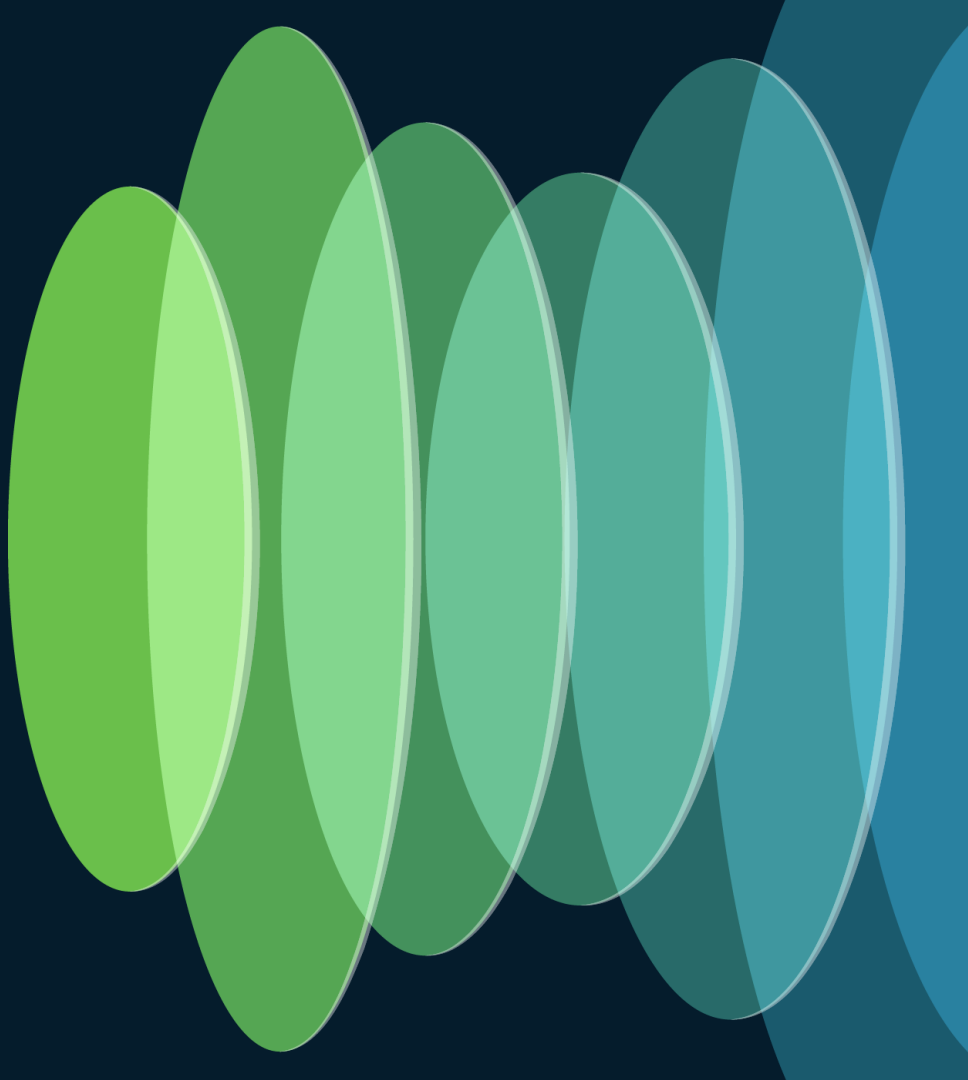
Webex spaces will be moderated by the speaker until June 7, 2024.



Agenda

- VXLAN EVPN Multi-Site Overview
- Design VXLAN EVPN Multi-Site with NDFC
- Introduction to NDFC
- Automate VXLAN EVPN Multi-Site with NDFC
- Conclusion

First Design
Highly Available Data Center

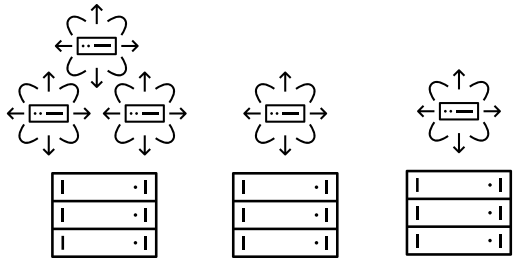


cisco Live!

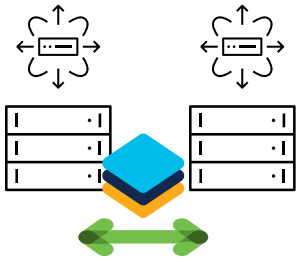
What defines a Highly Available Data Center?

Ask yourself

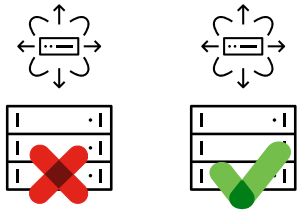
Do we have Network redundancy beyond a single rack?



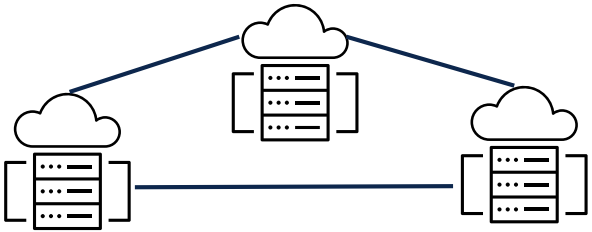
Do we have Application availability beyond a single rack?



Can we survive a single rack failure? How big is the change or failure domain?

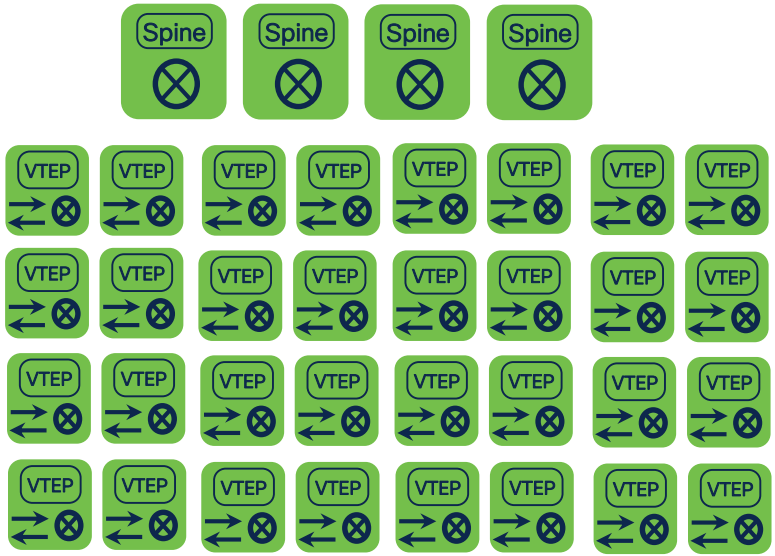


Do we have Network and Application availability beyond a single location?



Is your DC Highly Available Yet?

Where do you stand? “The Single” Or…

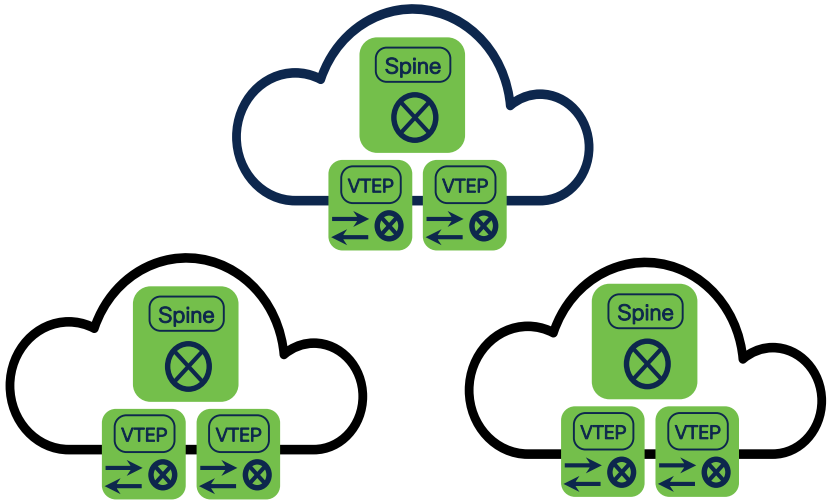


“The Data Center of Yesterday”

- Single Underlay Domain End-to-End
- Single Replication Domain for BUM
- Single Overlay Domain – End-to-End Encapsulation
- Single Overlay Control-Plane Domain – End-to-End EVPN Updates
- Single VNI Admin Domain

Is your DC Highly Available Yet?

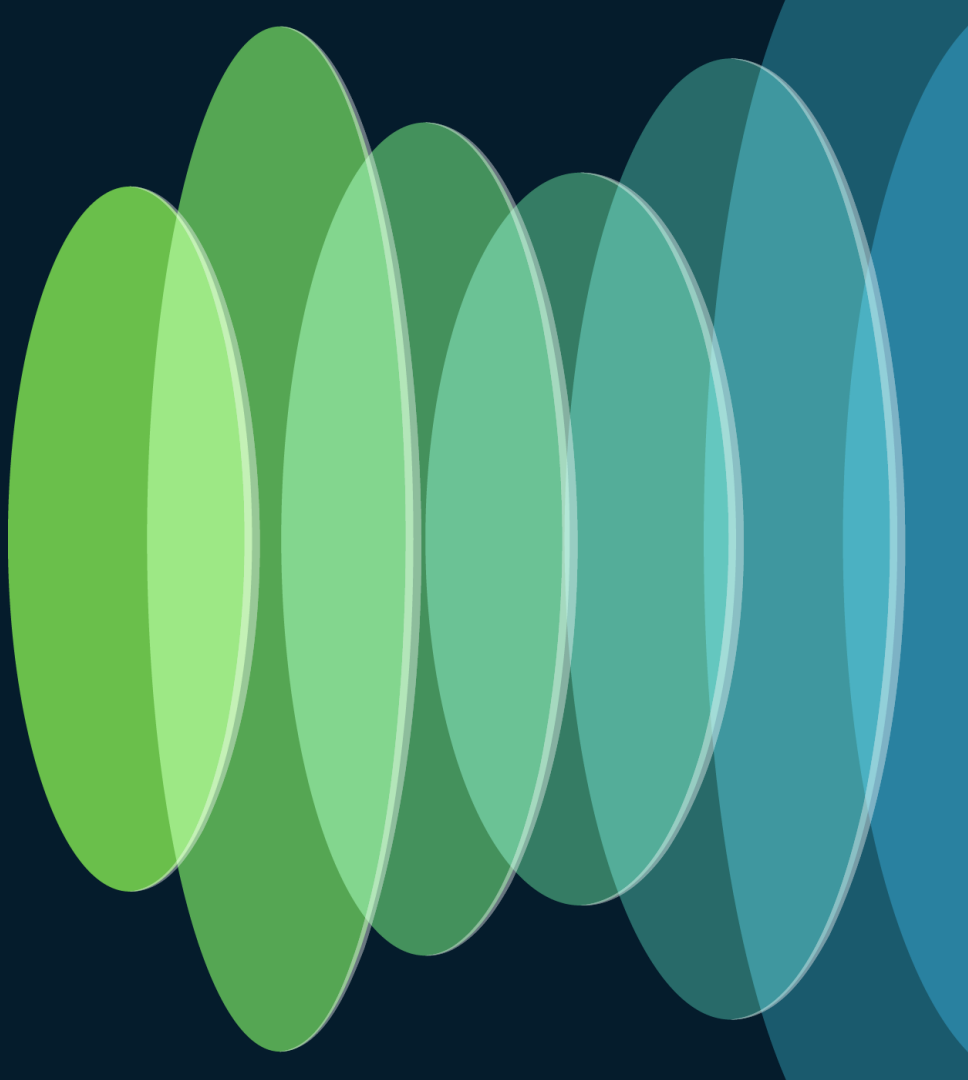
Where do you stand? “The Single” Or “The Multiple”



“The Data Center of Today”

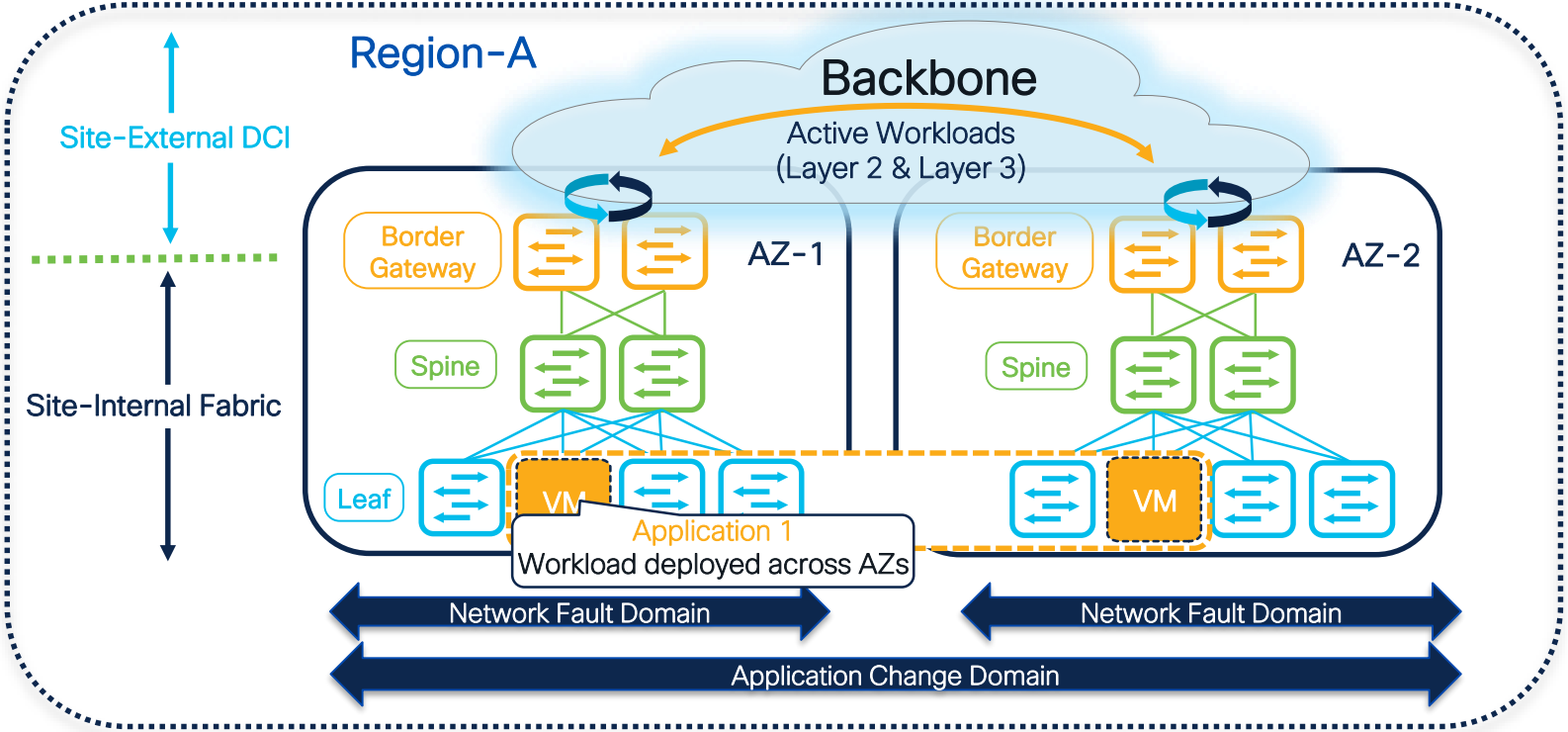
- Multiple Underlay Domains – Isolated
- Multiple Replication Domain for BUM – Interconnected and Controlled
- Multiple Overlay Domain – Interconnected and Controlled
- Multiple Overlay Control-Plane Domains – Interconnected and Controlled
- Multiple VNI Admin Domain – Downstream VNI

VXLAN Multi-Site
Design Options



VXLAN EVPN Multi-Site Overview

Functional Components



Border Gateway

Deployment Considerations

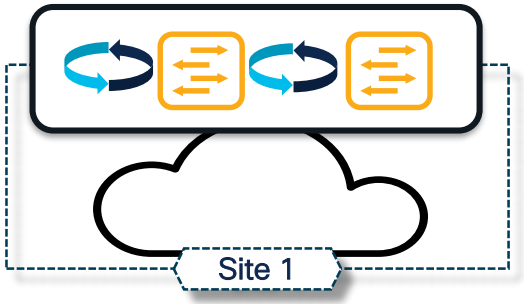
Border Gateways main functions and use-cases:

- Packet Re-Origination (L2 and L3)
- Inter-Site DCI (East-West)

- L3 Extension (North-South) IP Handoff
- Connect L4-L7 services and EPs Load Balancer ... Firewall
- Integration with Legacy Networks (Co-existence and/or Migration) VXLAN EVPN Greenfield ⟷ Classic Ethernet Brownfield
- VXLAN L3-extension to the Public Cloud aws Google Cloud Azure

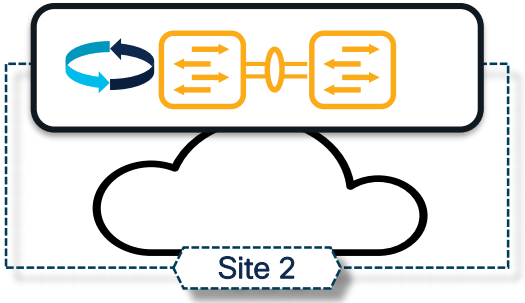
Border Gateway

When to use what



Anycast Border Gateway

- Up to 6 BGW
  - Simple Failure Scenarios
- Any Deployments
  - No End-Point or Network Services Connectivity on BGW
- Greenfield Deployments



VPC Border Gateway

- 2 BGW with Physical Peer-Link
- Small Deployments
  - End-Point or Network Services Connectivity on BGW
- Migration Use-Cases (Brownfield)
  - Classic Ethernet / FabricPath to VXLAN EVPN

BGW Node Placement Option

Flexible Design option-1

Dedicated Border Gateway

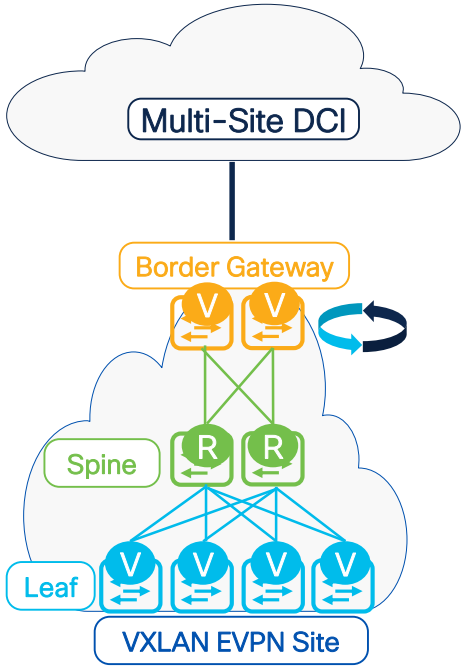
- Flexible scale-out approach for VXLAN EVPN Multi-Site DCI
- Flexible Anycast or VPC BGW models
- Capacity planning only for DCI traffic flows
- Clean role separation and uniform reachability from the entire fabric are the major advantages

Border Gateway **hosts:**
VTEP for:

- East-West (DCI Packet Re-Origination L2/L3)

V V = VTEP

R = RR/RP



BGW Node Placement Option

Flexible Design option-2

BGW on Spine nodes (**Border Gateway Spine**)

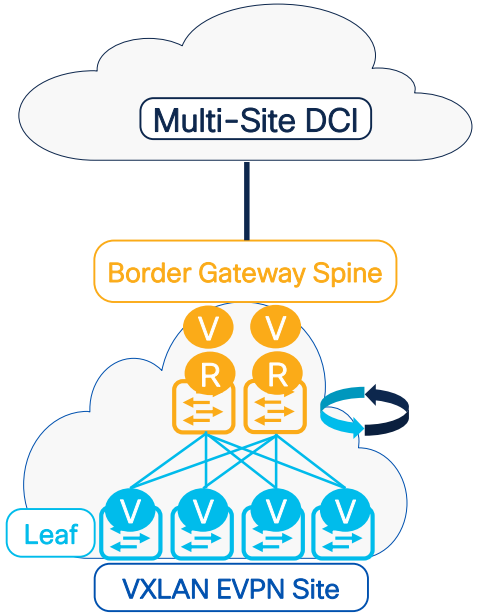
- ✔ Flexible scale-out approach for VXLAN EVPN Multi-Site DCI
- ✔ Anycast BGW only
- ⚠ Extra functional dependency (BGW + Spine)
- ⚠ Capacity planning needs to accommodate all flows

Border Gateway Spine hosts:
VTEP for:

- East-West (DCI Packet Re-Origination L2/L3)
- Route Reflector (RR) (Site-Internal iBGP EVPN)
- Rendezvous Point (RP) (Site-Internal Multicast Underlay BUM)

V V = VTEP

R = RR/RP



BGW Node Placement Option

Flexible Design option-3

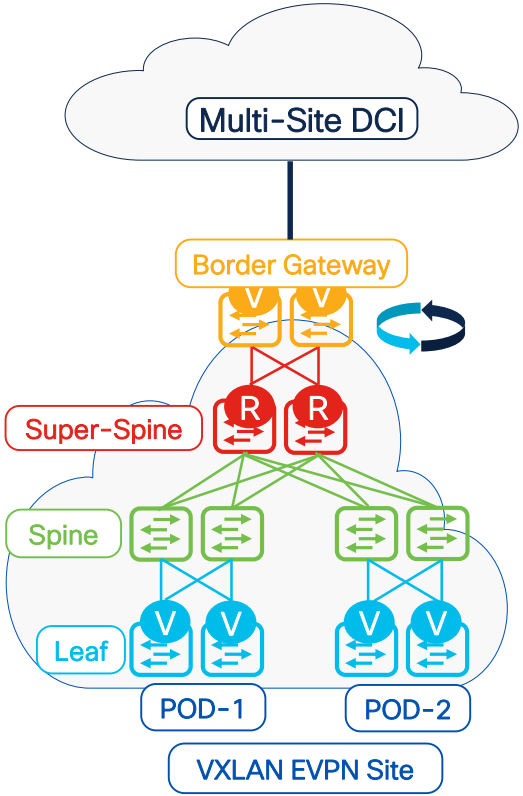
BGW on top of Super-Spine nodes

- Scale-out Multi-Clos Fabric to Interconnect the PODs using Super-Spine
- Architecture beyond a single server room. Simpler capacity planning
- Capacity planning only for DCI traffic flows
- Clean role separation and uniform reachability from the entire fabric are the major advantages

Border Gateway hosts:
VTEP for:

- East-West (DCI Packet Re-Origination L2/L3)

Super-Spine hosts:
- Route Reflector (RR) (iBGP EVPN)
- Rendezvous Point (RP) (Multicast Underlay BUM)



Border Node Placement Option

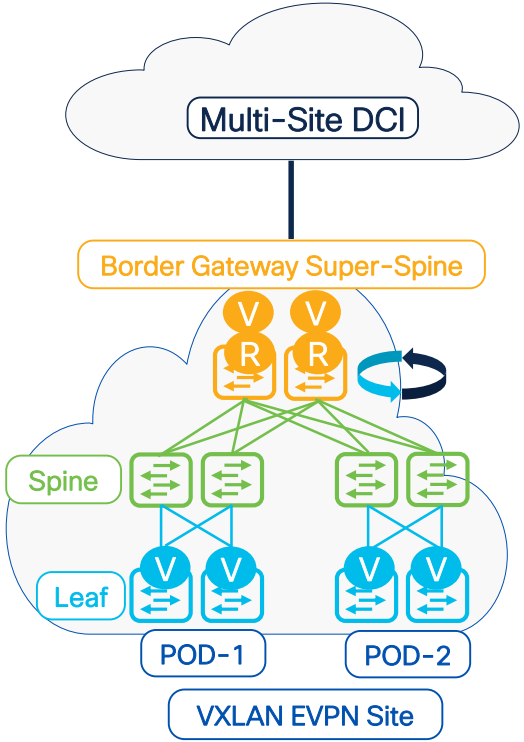
Flexible Design option-4

BGW on Super-Spine nodes (BGW Super-Spine)

- Scale-out Multi-Clos Fabric to Interconnect the PODs using Super-Spine
- Architecture beyond a single server room. Simpler capacity planning
- Capacity planning needs to accommodate all flows
- Extra functional dependency (BGW + Super-Spine). Not recommended due to Multi-POD failure dependency

Border Gateway Super-Spine hosts:
VTEP for:

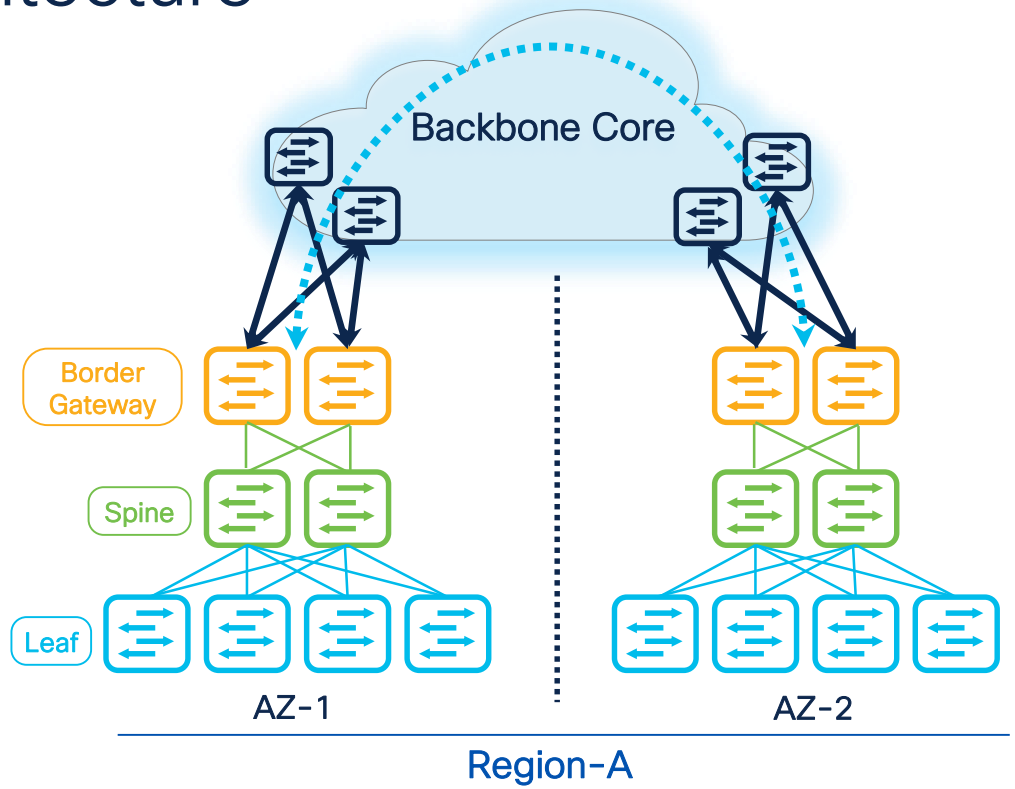
- East-West (DCI Packet Re-Origination L2/L3)
- Route Reflector (RR) (Site-Internal iBGP EVPN)
- Rendezvous Point (RP) (Site-Internal Multicast Underlay BUM)



VXLAN Multi-Site Architecture

DCI- BGW to Cloud

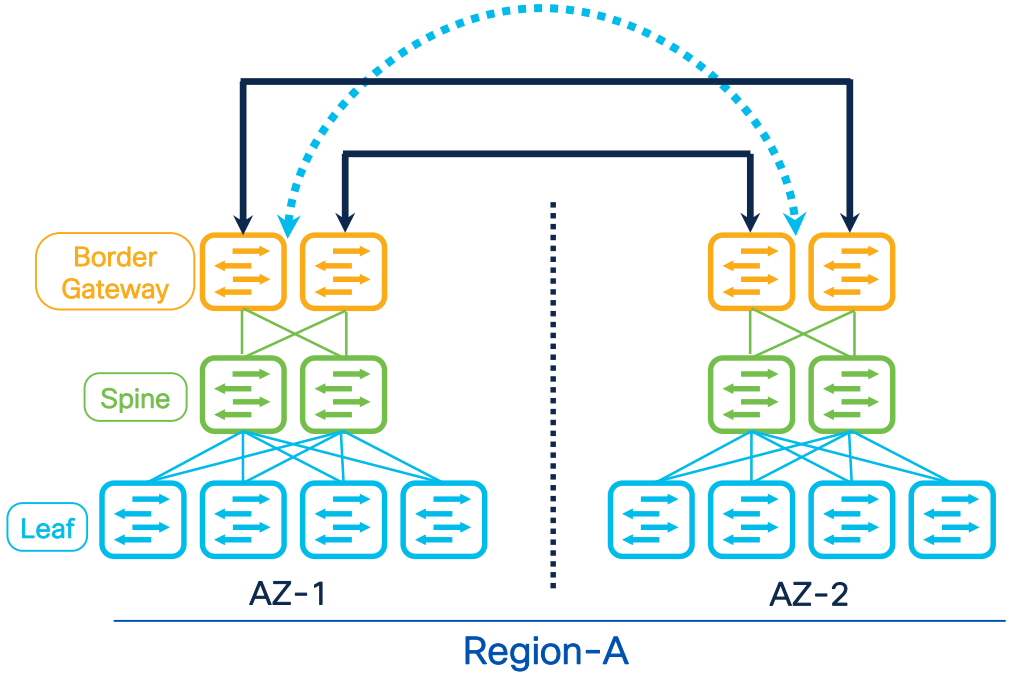
- Scalable design option with horizontal scale-out within and across multiple sites.
- Backbone/Cloud can be any routed service. (flat L3, MPLS-L3VPN)
- Multi-Site Underlay: eBGP IPv4 Unicast ⟷
  - Advertise Lo0 (Overlay Control Plane), Lo1 (BUM, External networks), and Lo100 (Multi-Site Inter-Site Transit communication)
  - Site-External DCI BUM: Ingress-Replication or Multicast supported. *Currently NDFC supports only Ingress-Replication.
  - Site-Internal Fabric BUM: Ingress-Replication or Multicast supported independently at each site.
- Multi-Site Overlay: eBGP EVPN Overlay ⟷
  - Full-mesh BGP EVPN peering across all BGWs.
- Ensure that PIP / VIP of all BGWs are known by every BGW and MTU must accommodate VXLAN encapsulated traffic.



VXLAN Multi-Site Architecture

DCI- BGW Back-to-Back

- Available option for connecting 2 or 3 sites.
- Minimum topology is the square. An enhanced option is to add links between site local BGWs for improved ECMP and Failure scenarios.
- Multi-Site Underlay: eBGP IPv4 Unicast ⟷
  - Advertise Lo0 (Overlay Control Plane), Lo1 (BUM, External networks), and Lo100 (Multi-Site Inter-Site Transit communication)
  - Site-External DCI BUM: Ingress-Replication or Multicast supported. *Currently NDFC supports only Ingress-Replication.
  - Site-Internal Fabric BUM: Ingress-Replication or Multicast supported independently at each site.
- Multi-Site Overlay: eBGP EVPN Overlay ⟷
  - Full-mesh BGP EVPN peering across all BGWs.
- Ensure that PIP / VIP of all BGWs are known by every BGW and MTU must accommodate VXLAN encapsulated traffic.

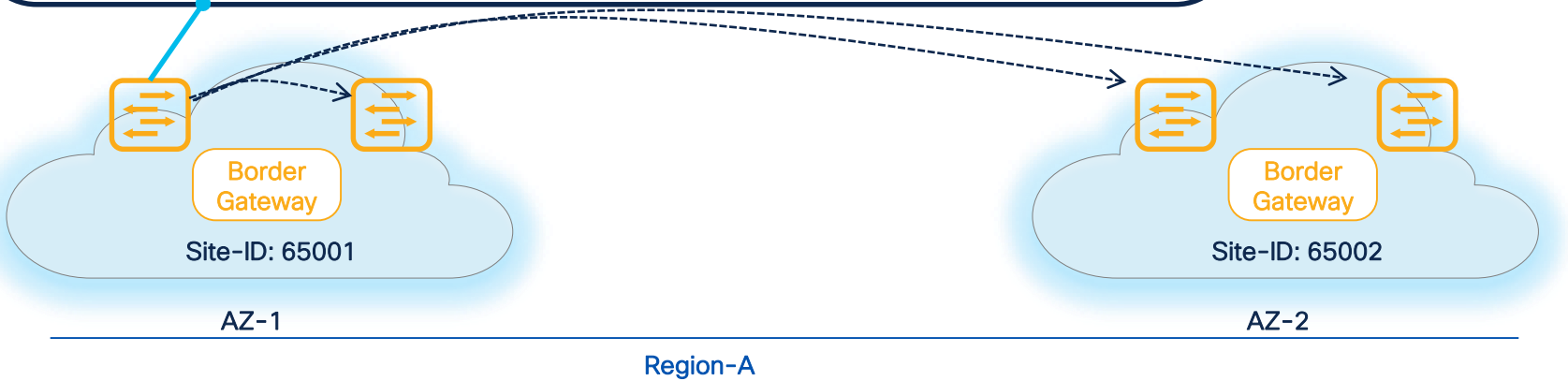


Multi-Destination Traffic Forwarding

EVPN Route-Type 4 (Ethernet Segment Route)

- BGW advertises BGP EVPN Type-4 update to all of their BGP EVPN peers (Site-Internal and Site-External)
- BGW will use this information for DF (Designated Forwarded) election on per VLAN/VNI
- EVPN Type-4 update will also carry ES Import Route-Target Extended Community
- Remote site BGWs will not install the route due to different RT and Site ID
- Command to check: **`show bgp l2vpn evpn route-type 4`**
  - `BGP routing table entry for [4]:[0300.0000.00fd.e900.0309]:[32]:[10.1.1.1]/136`
  - `[4] = EVPN Route-Type 4`
  - `[03]= ESI Type MAC-based`
  - `[0000.0000.fde9] = ESI system MAC for Site-ID and RT {fde9 = 65001}`
  - `[000309] = ESI local discriminator`
  - `[32] = IP Address Length`
  - `[10.1.1.1] = Sender (BGW Primary IP) used for DF election process`

```
evpn multisite border-gateway 65001
  delay-restore time 300
```



Multi-Destination Traffic Forwarding

BGW Designated Forwarded (DF) Election

- BGW learns about other BGWs originator IPs on the same site from exchanging EVPN Type 4 routes. The originator IP address used is the BGW NVE source interface's primary IP address.
- Each BGW creates an ordinal list of originator IP in numerical order from lowest to highest. Every BGW is then given an ordinal value based on its position in the ordinal list starting from position 0. The BGW with the lowest originator IP would get an ordinal value of 0. The ordinal value decides which BGW will be the DF for a VLAN/VNI.
- DF election Formula uses “mod” operation:
- `I = V mod N {I = ordinal value, V = VLAN #, N = # BGWs in a site}`
- If there are 4 BGWs in one site. The ordinal list of the BGWs will be arranged as shown below.
- Ordinal List = 10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4 {I=0, I=1, I=2, I=3}
- The DF for VLAN 48 will be: “I = 48 mod 4 = 0 = 10.1.1.1”
- The DF for VLAN 49 will be: “I = 49 mod 4 = 1 = 10.1.1.2”
- The DF for VLAN 50 will be: “I = 50 mod 4 = 2 = 10.1.1.3”
- The DF for VLAN 51 will be: “I = 51 mod 4 = 3 = 10.1.1.4”

Multi-Destination Traffic Forwarding

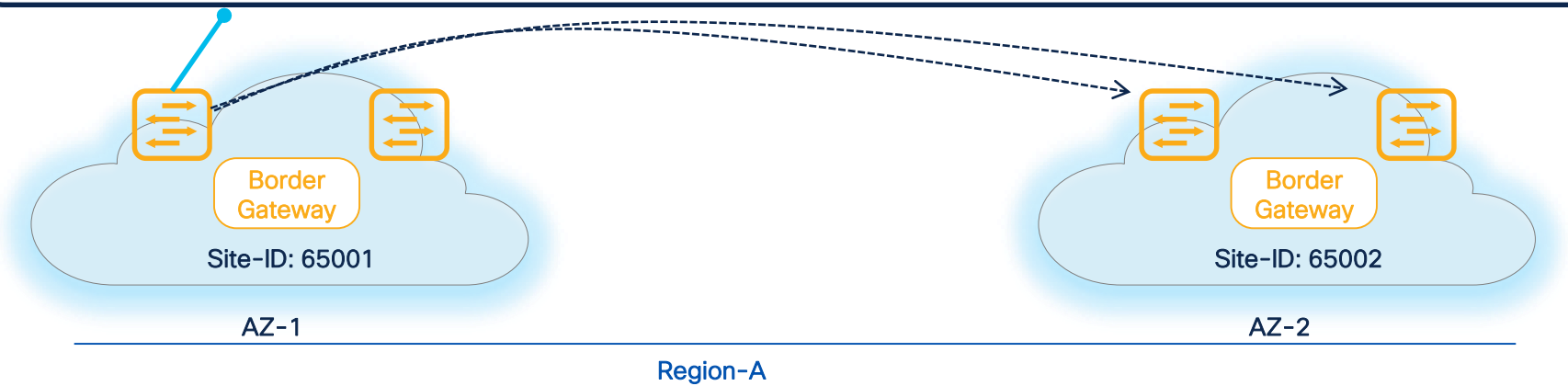
EVPN Route-Type 3 (Inclusive Multicast Ethernet Tag Route)

- BGW advertises BGP EVPN Type-3 only towards Site-External DCI
- Receiving BGW switch does not forward it to Site-Internal Fabric
- Command to check: **`show bgp l2vpn evpn route-type 3,`**
- **`show l2route evpn imet all detail`**
- **`Local: Advertising from Site 1 BGW to Site 2 BGW`**

```
BGP routing table entry for [3]:[0]:[32]:[10.3.0.4]/88
AS-Path: NONE, path locally originated
Extcommunity: RT:65001:10000 ENCAP:8
  PMSI Tunnel Attribute:
    flags: 0x00, Tunnel type: Ingress Replication
    Label: 10000, Tunnel Id: 10.3.0.4
Path-id 1 advertised to peers: 20.2.0.4
```

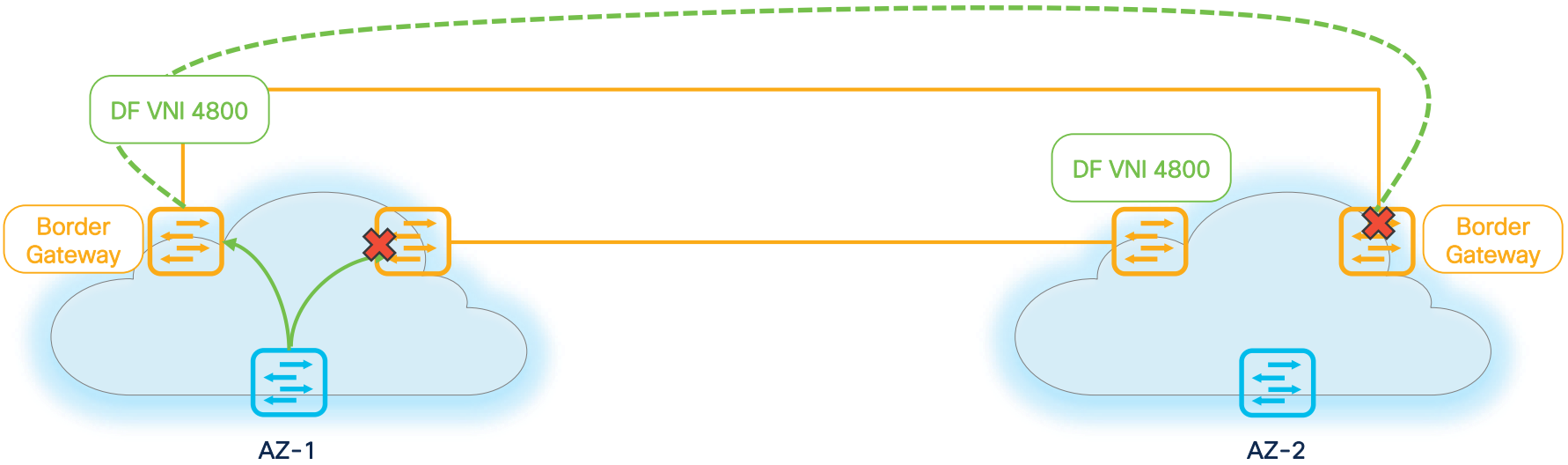
- **`Remote: Learning from Site 2 BGW to Site 1 BGW`**

```
BGP routing table entry for [3]:[0]:[32]:[20.3.0.4]/88
Path type: external, path is valid, is best path
 Imported to 1 destination(s)
 Imported paths list: L2-10000
AS-Path: 65002 , path sourced external to AS
Extcommunity: RT:65001:10000 ENCAP:8
  PMSI Tunnel Attribute:
    flags: 0x00, Tunnel type: Ingress Replication
    Label: 10000, Tunnel Id: 20.3.0.4
```



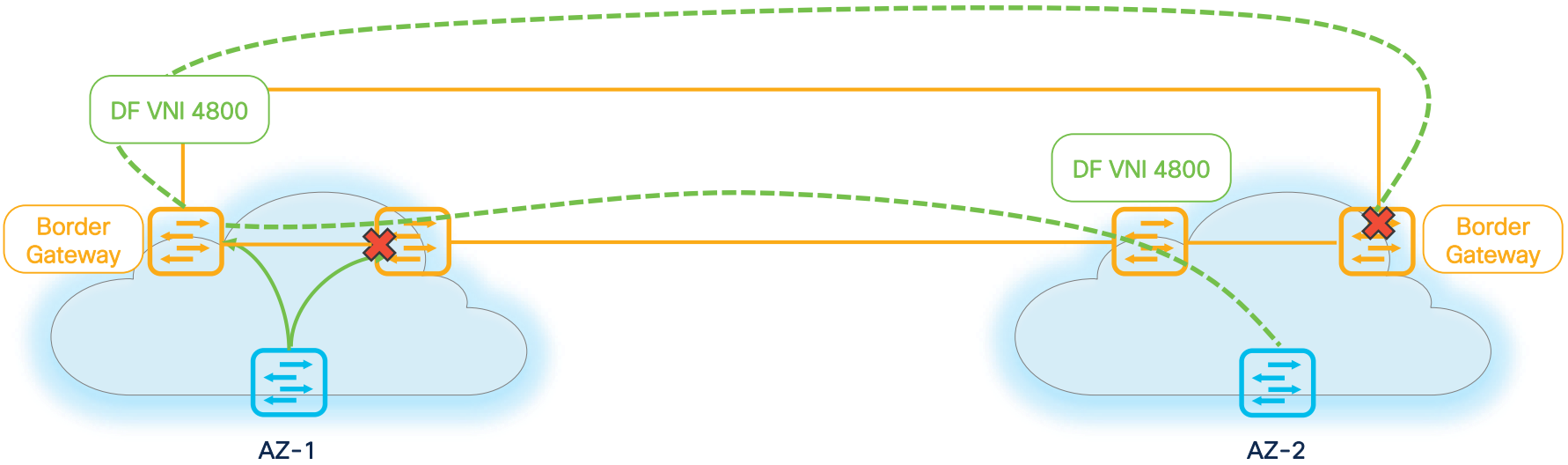
VXLAN Multi-Site Architecture

DCI- BGW Back-to-Back BUM (w/o Local L3 link)



VXLAN Multi-Site Architecture

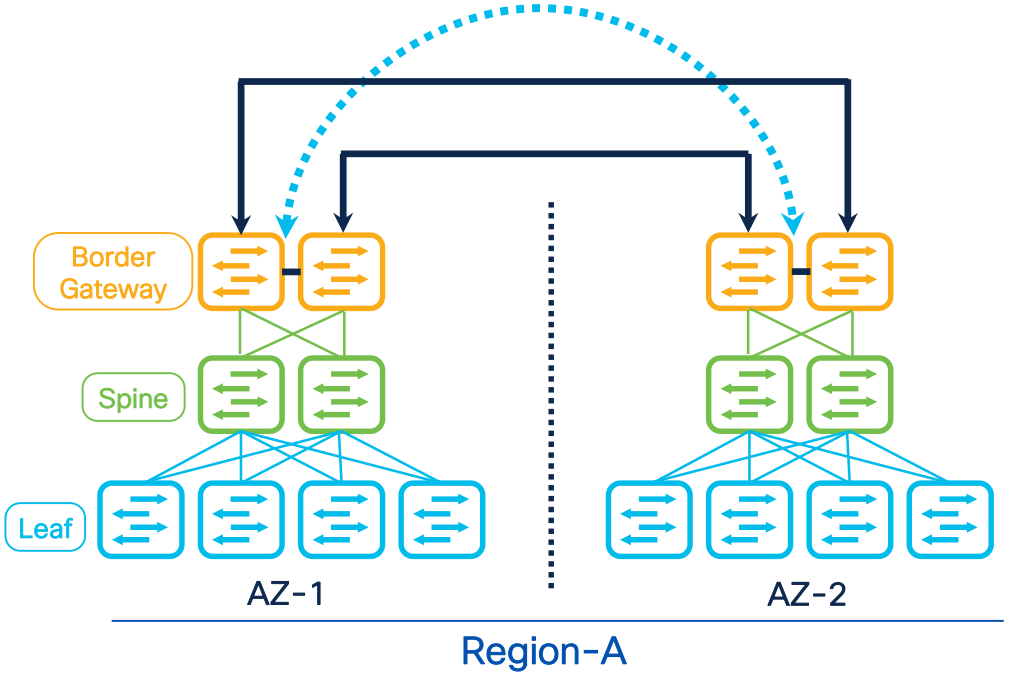
DCI- BGW Back-to-Back BUM (with Local L3 link)



VXLAN Multi-Site Architecture

DCI- BGW Back-to-Back (with Local L3 link)

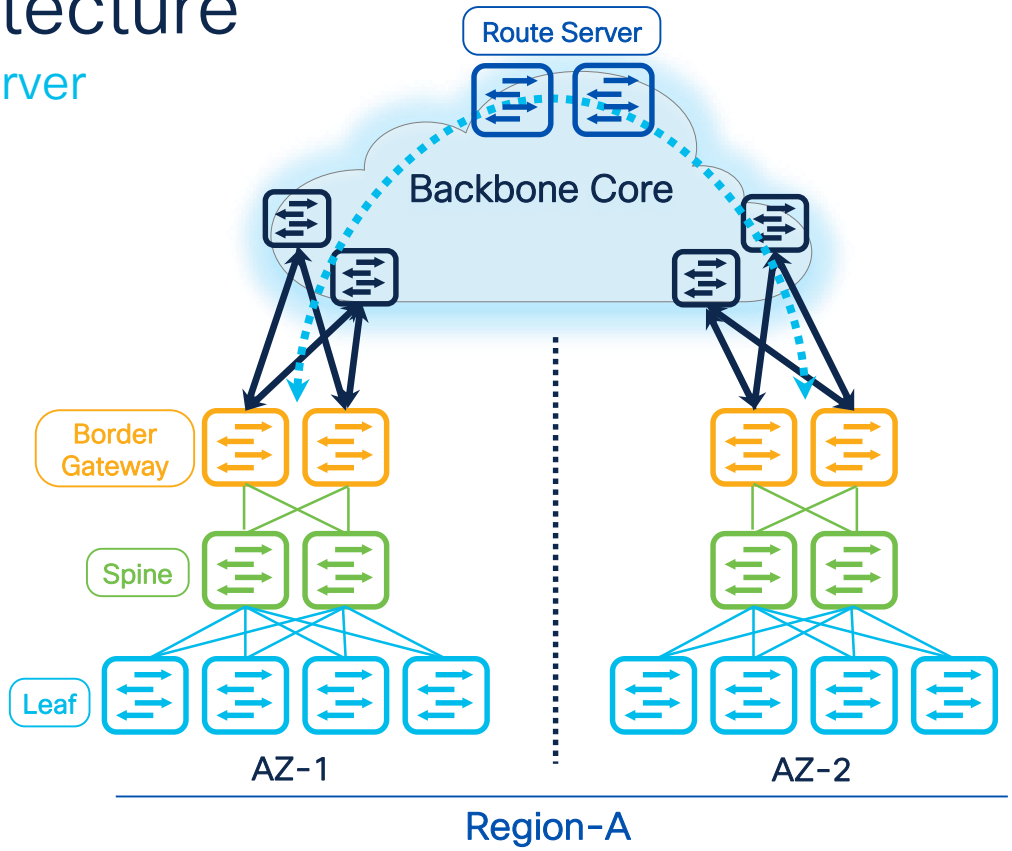
- Available option for connecting 2 or 3 sites.
- Minimum topology is the square. An enhanced option is to add links between site local BGWs for improved ECMP and Failure scenarios.
- **Multi-Site Underlay: eBGP IPv4 Unicast** ⟷
  - Site-External DCI BUM: Ingress-Replication or Multicast supported. *Currently NDFC supports only Ingress-Replication.
  - Site-Internal Fabric BUM: Ingress-Replication or Multicast supported independently at each site.
- **Multi-Site Overlay: eBGP EVPN Overlay** ⟷
  - Full-mesh BGP EVPN peering across all BGWs.
- Ensure that PIP / VIP of all BGWs are known by every BGW and MTU must accommodate VXLAN encapsulated traffic.



VXLAN Multi-Site Architecture

DCI- BGW to Centralized Router-Server

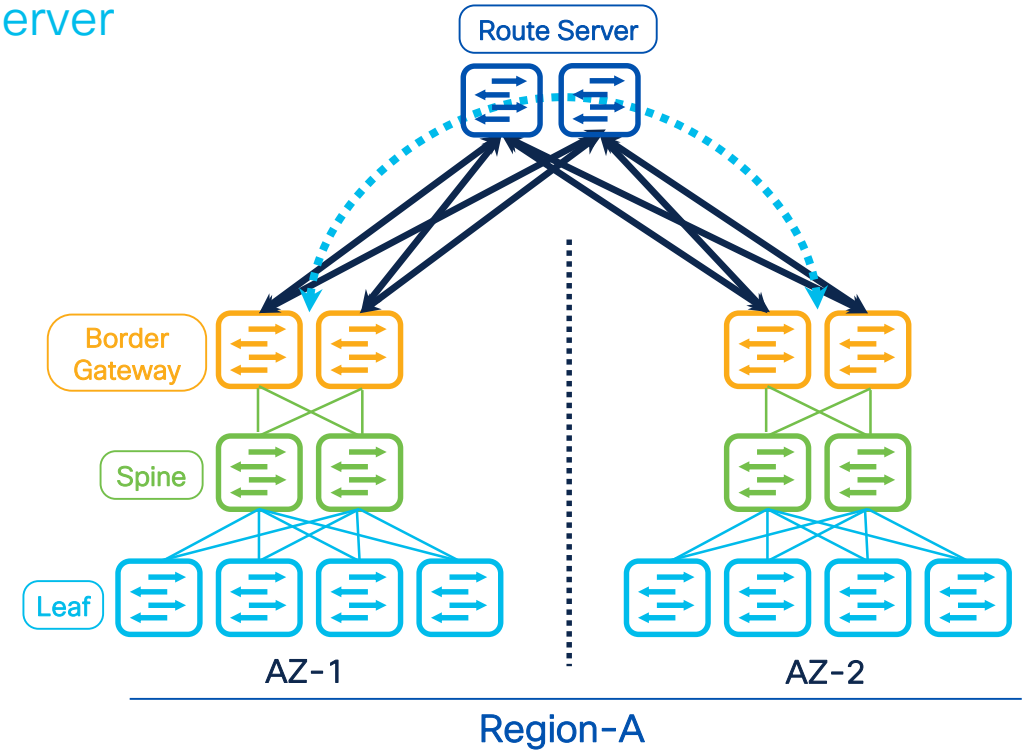
- BGP Route-Server (RS) function can be used in a highly scalable design to act as the central EVPN peering point for the Multi-Site.
- RS does not need to be on the data path.
- **Multi-Site Underlay: eBGP IPv4 Unicast** ⟷
  - **Advertise Lo0 (Overlay Control Plane), Lo1 (BUM, External networks), and Lo100 (Multi-Site Inter-Site Transit communication)**
  - Site-External DCI BUM: Ingress-Replication or Multicast supported. *Currently NDFC supports only Ingress-Replication.
  - Site-Internal Fabric BUM: Ingress-Replication or Multicast supported independently at each site.
- **Multi-Site Overlay: eBGP EVPN Overlay** ⟷
  - BGP EVPN peering only between BGWs and RS.
- **Route-Server (RS) must support...**
  - EVPN AFI and Router-Server function per RFC 7947
  - Next-hop-unchanged, Retain RT, and RT Rewrite function
- Ensure that PIP / VIP of all BGWs are known by every BGW and MTU must accommodate VXLAN encapsulated traffic.



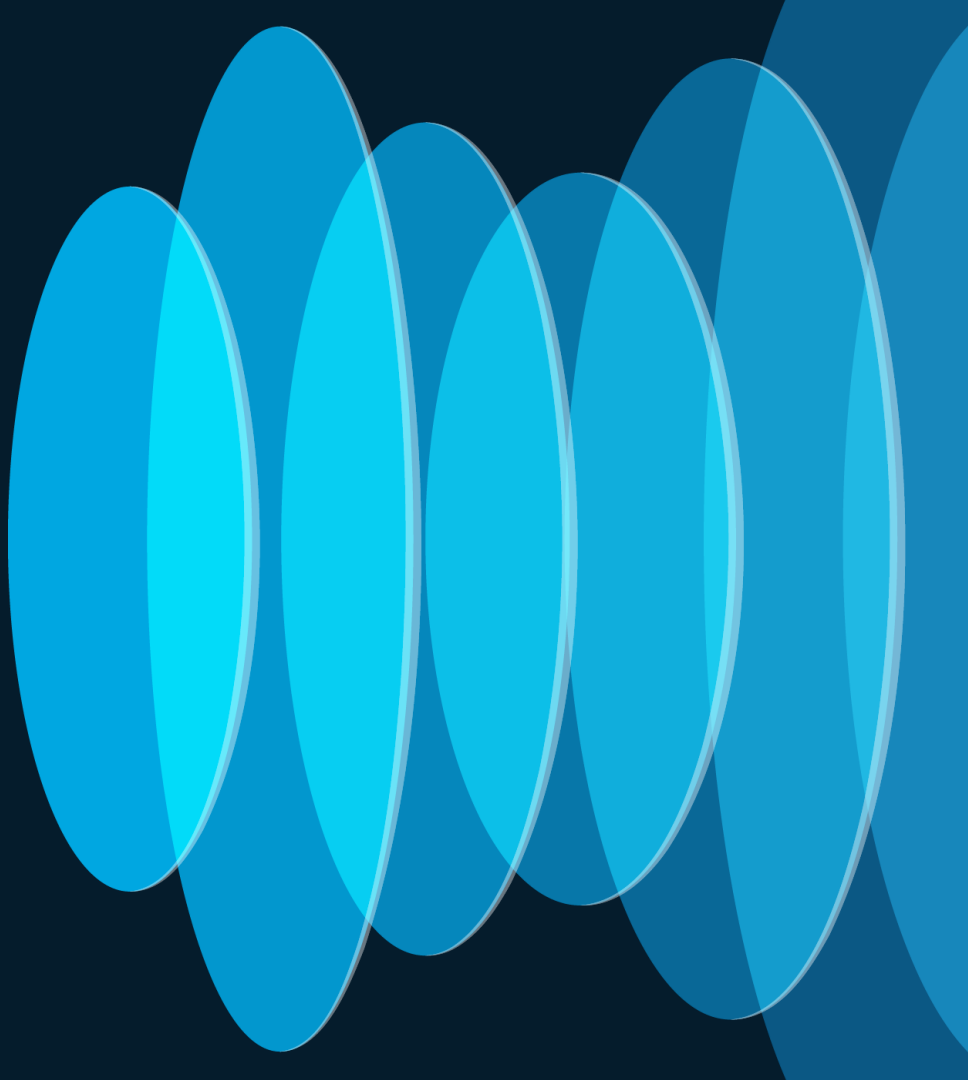
VXLAN Multi-Site Architecture

DCI- BGW to Centralized Router-Server

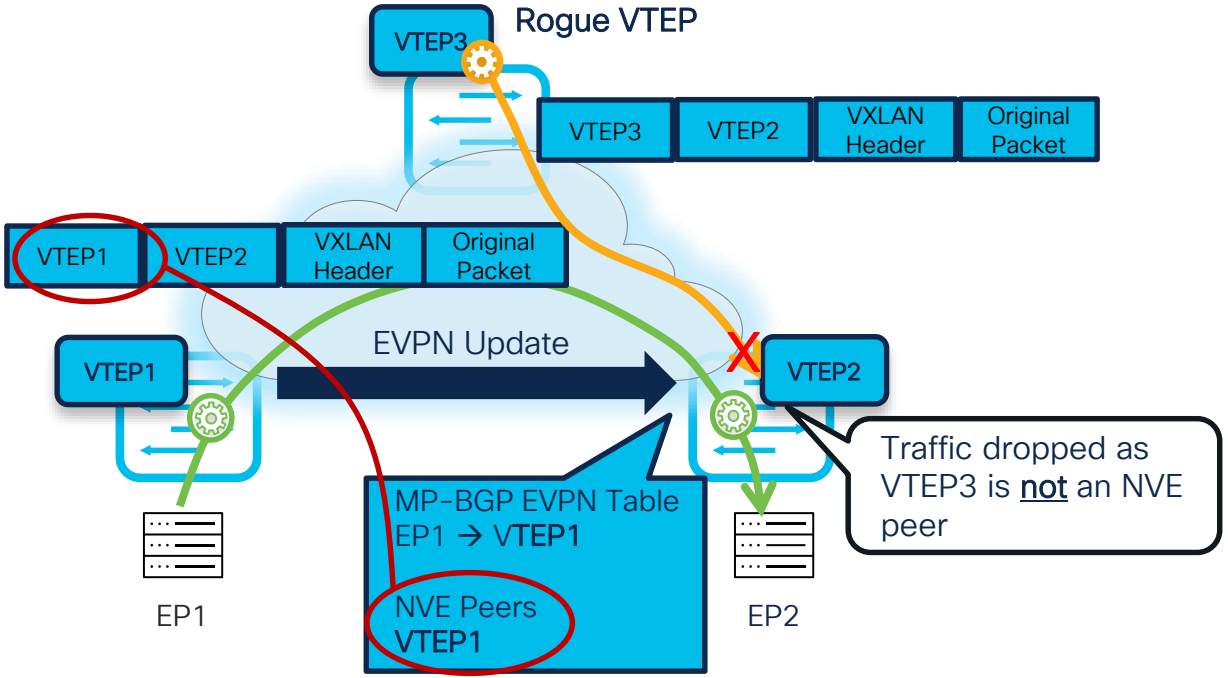
- BGP Route-Server (RS) function can be used in a highly scalable design to act as the central EVPN peering point for the Multi-Site.
- RS can also be on the data path.
- **Multi-Site Underlay: eBGP IPv4 Unicast** ⟷
  - Site-External DCI BUM: Ingress-Replication or Multicast supported. *Currently NDFC supports only Ingress-Replication.
  - Site-Internal Fabric BUM: Ingress-Replication or Multicast supported independently at each site.
- **Multi-Site Overlay: eBGP EVPN Overlay** ⟷
  - BGP EVPN peering only between BGWs and RS.
- **Route-Server (RS) must support…**
  - EVPN AFI and Router-Server function per RFC 7947
  - Next-hop-unchanged, Retain RT, and RT Rewrite function
- Ensure that PIP / VIP of all BGWs are known by every BGW and MTU must accommodate VXLAN encapsulated traffic.



Special Considerations for Layer-3 extension across Multi-Site



VXLAN Native Data Plane Security



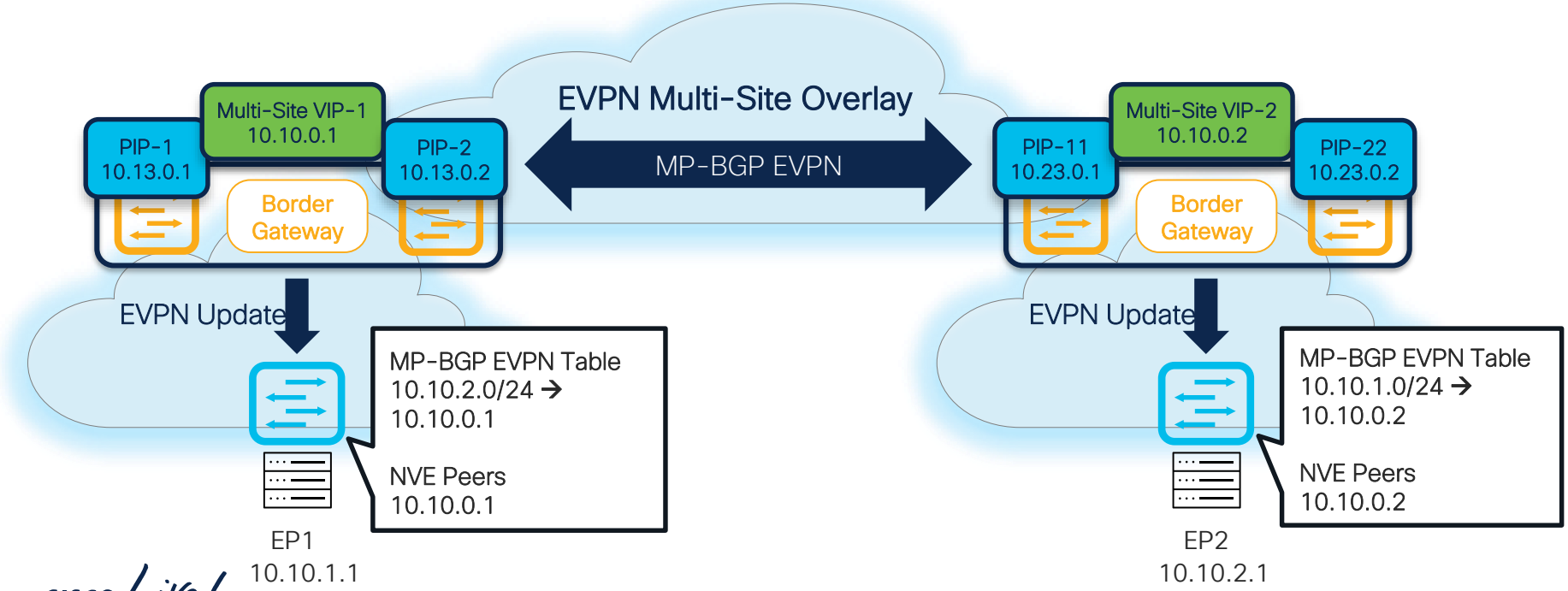
- VXLAN EVPN offers a native data plane security functionality
  - VXLAN traffic originated from a remote VTEP is only accepted when sourced from a TEP address that is an “NVE peer”
  - An NVE peer’s address is added to the local table based on the reception of MP-BGP EVPN updates carrying that specific address as next-hop
- Prevents the insertion of rogue VTEPs in a VXLAN EVPN fabric

Why can this become an issue for Layer-3 communication across sites?

Inter-Site Layer 3 Traffic – Control Plane

Inter-Site Type-2 and Type-5 EVPN updates always carry the local Multi-Site VIP as next-hop address

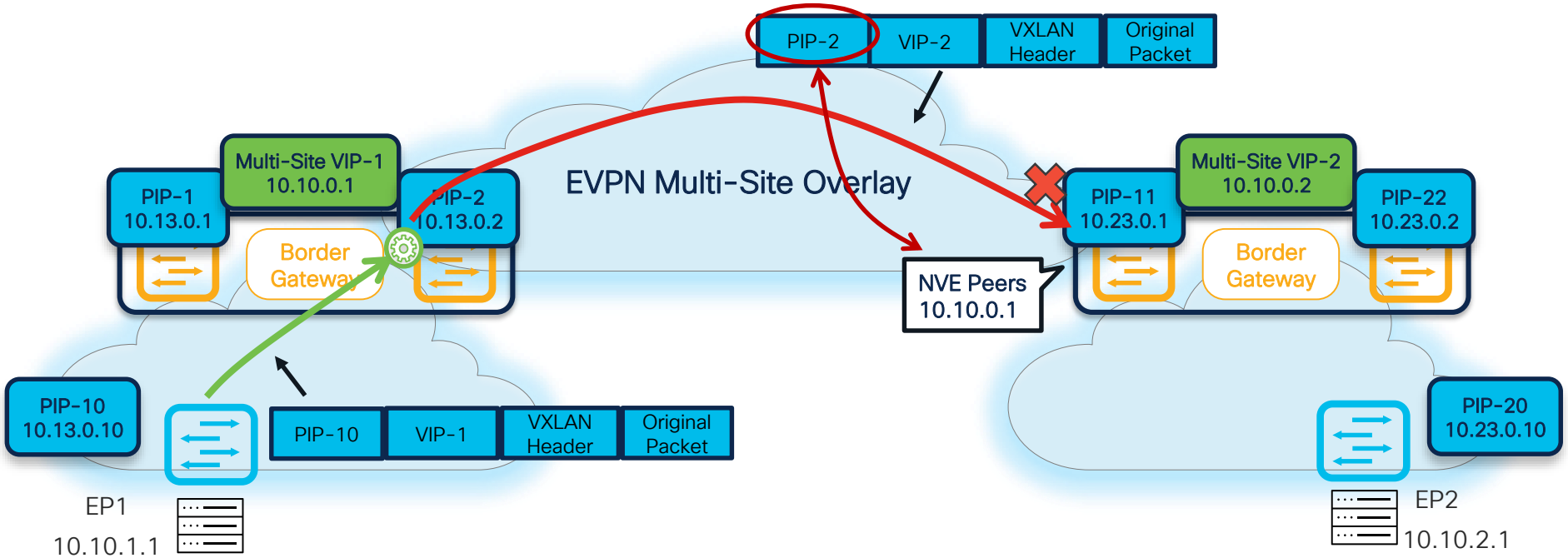
- Only exception are Type-5 updates for L3 networks locally connected to the BGWs



Inter-Site Layer 3 Traffic – Data Plane

Inter-Site traffic is always sourced by local BGWs from their specific PIP address

Same applies to Intra-Site traffic between local BGW and leaf nodes

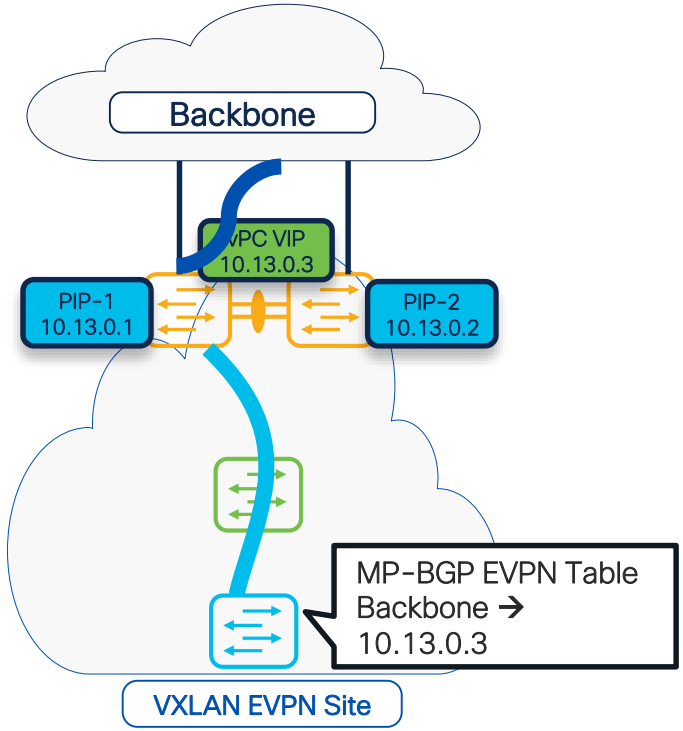


How to Prevent this issue?

Option-1: Advertise-PIP

1. External routes are injected into VXLAN fabric
2. Border advertises External routes as EVPN Type-5 with the BGP Next-Hop of vPC VIP (Anycast)
3. From Leaf perspective the Next-Hop to reach Backbone is Border Anycast vPC VIP

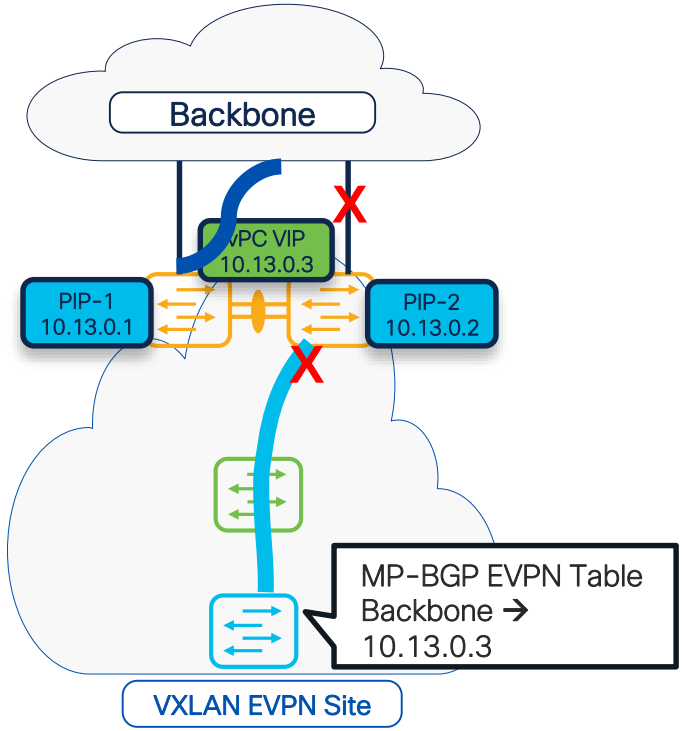
Problem Statement: While ARP/MAC/IPv6 ND entries are synced between the peers of a vPC pair, prefix routes belonging to an individual peer as well as external routes received by a peer are not synced between vPC peer switches. Using the VIP as the BGP next-hop for these routes can cause traffic to be forwarded to the wrong vPC peer and hence be black-holed.



Option-1: Advertise-PIP

1. Border-2 losses the link towards Backbone and Border-1 is the only available path towards the Fabric
2. Border-1 continues to advertise External routes as EVPN Type-5 with the BGP Next-Hop of vPC VIP (Anycast)
3. From Leaf perspective the Next-Hop to reach Backbone is Border Anycast vPC VIP. Hence, traffic can hash to either Border-1 or Border-2. If packet hits Border-2, it will drop the traffic!

Note: Border-1 still has an active link towards the Backbone and advertises the routes towards the Spine (RR). Later, the Spine will reflect the route to Border-2, but it will reject it due to Next-Hop being its own IP (VIP 10.13.0.3)



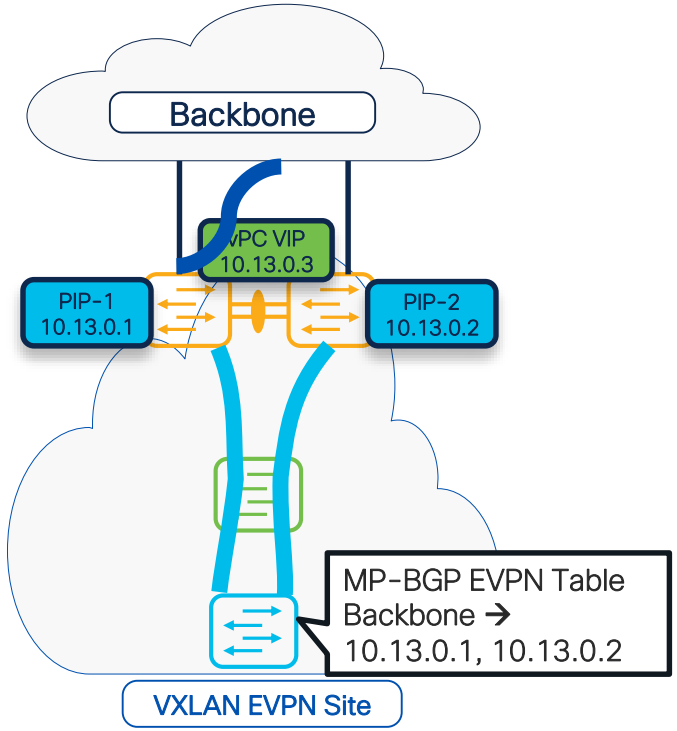
Option-1: Advertise-PIP

```
router bgp 65001
  address-family l2vpn evpn
    advertise-pip

Interface nve1
  advertise virtual-rmac
```

On both VPC peer Border Devices

The advertise-pip command lets BGP use the PIP as next-hop when advertising prefix routes or leaf-generated routes if vPC is enabled. With the advertise-pip and advertise virtual-rmac commands, EVPN Type-5 routes are advertised with PIP, and EVPN Type-2 routes are still advertised with VIP. In addition, a virtual MAC will be used with the VIP that is shared by both vPC peers, and individual peer specific system Router MAC will be used with PIP when the advertise-pip feature is enabled. In this way, the traffic will always be destined to the right vPC peer.



Option-2: Per-VTEP, Per-VRF Loopback

- Define on all the BGW nodes a loopback interface (in a specific VRF) and advertise the information across site with a Type-5 EVPN update

```
route-map fabric-rmap-redist-subnet permit 10
  match tag 12345
!
interface loopback2
  vrf member t1-vrf1
  ip address 33.33.33.1/32 tag 12345
!
router bgp 65001
  vrf t1-vrf1
    address-family ipv4 unicast
      redistribute direct route-map fabric-rmap-redist-subnet
```

On local and remote BGW

Option-3: Extend one L2VNI End-to-End

- An L2VNI must be stretched between the local leaf nodes and BGW nodes. An L2VNI must be stretched across sites (i.e. defined on the local and remote BGW nodes). Does not necessarily need to be the same L2VNI

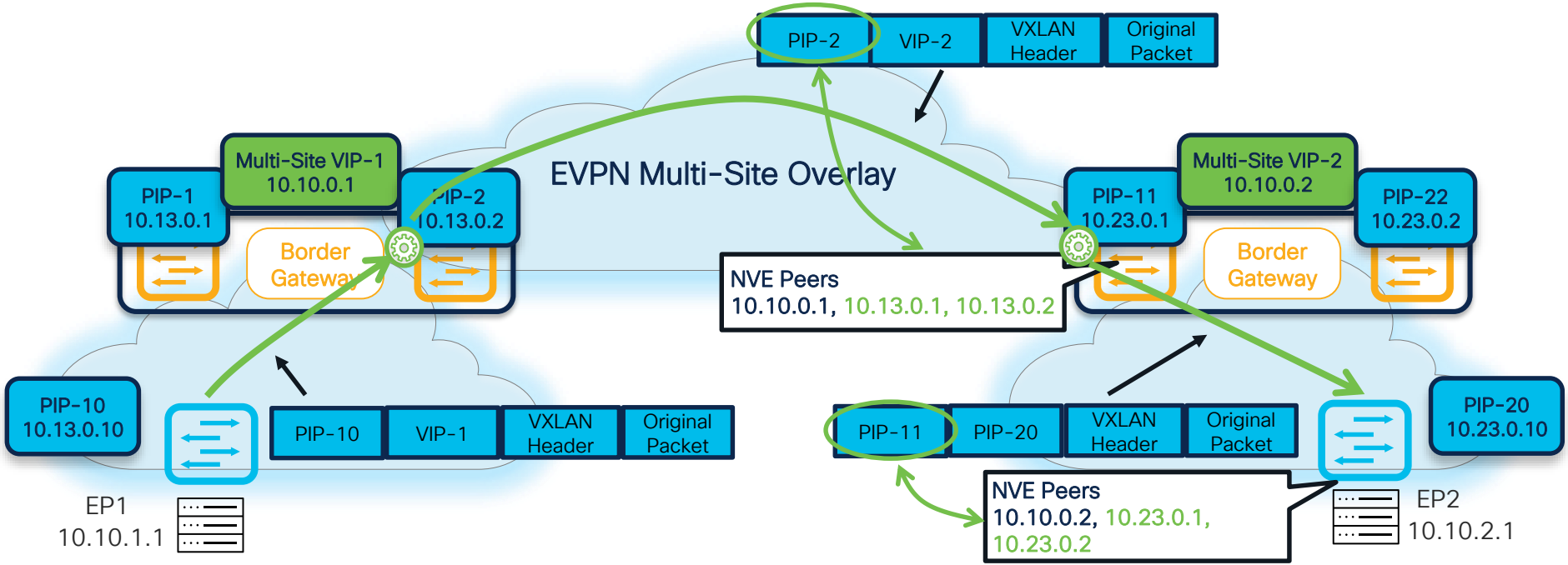
```
vlan 99
  vn-segment 9999
vlan 2300
  name Net-App
  vn-segment 30001

interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  multisite border-gateway interface loopback100
  member vni 9999 associate-vrf
  member vni 30001
    multisite ingress-replication
    mcast-group 239.1.1.1
```

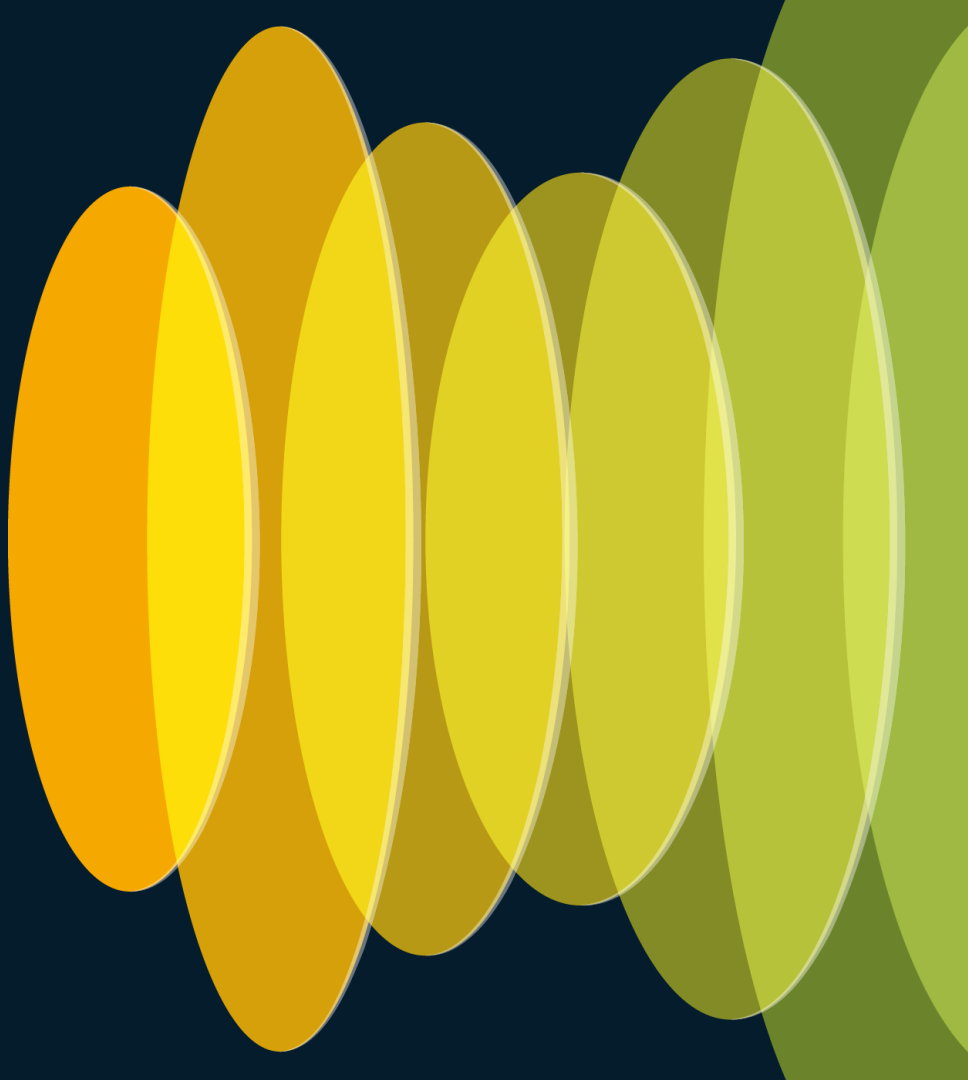
On local and remote BGW

Installing PIP Addresses as NVE Peers

Adopting any of the 3 solutions described in the previous slide ensures that the PIP addresses of the BGWs can be installed as NVE peers in the local leaf nodes and in the remote BGW nodes



Some more Important Requirements!



Multi-Site Route-Target Problem

```
vrf context myvrf_50000
  vni 50000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn {65001:50000}
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn {65001:50000}
evpn
  vni 30000 l2
    rd auto
    route-target import auto {65001:30000}
    route-target export auto {65001:30000}
```

```
vrf context myvrf_50000
  vni 50000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn {65002:50000}
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn {65002:50000}
evpn
  vni 30000 l2
    rd auto
    route-target import auto {65002:30000}
    route-target export auto {65002:30000}
```

EVPN Route-Target Re-Write

The ”rewrite-evpn-rt-asn” command modifies the incoming EVPN advertisements by swapping the remote AS portion in the RT with the local ASN, provided the update is coming from a neighbor that is locally configured.

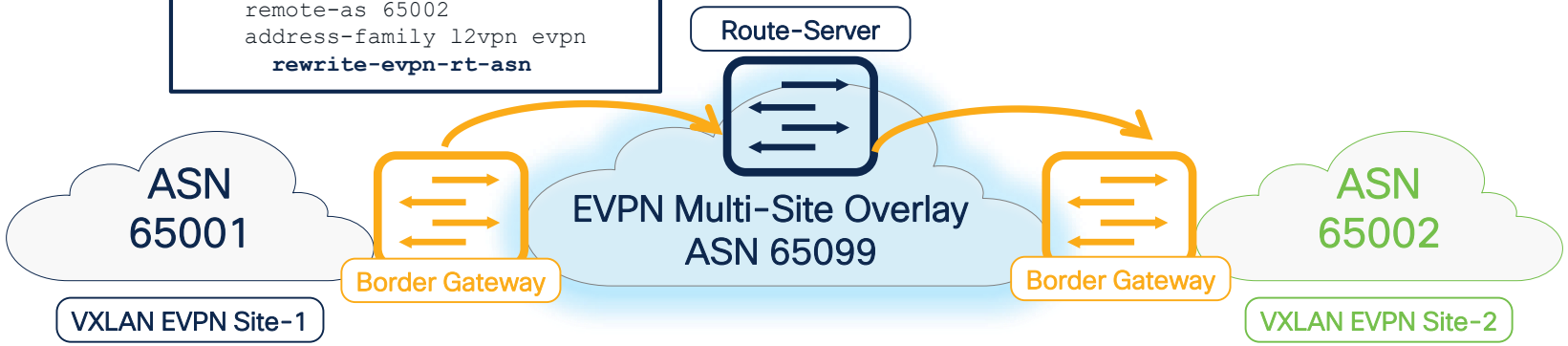
```
router bgp 65001
  neighbor 2.2.2.2
    remote-as 65002
    update-source loopback0
    ebgp-multihop 5
    peer-type fabric-external
    address-family l2vpn evpn
      send-community
      send-community extended
      rewrite-evpn-rt-asn
```

```
router bgp 65002
  neighbor 1.1.1.1
    remote-as 65001
    update-source loopback0
    ebgp-multihop 5
    peer-type fabric-external
    address-family l2vpn evpn
      send-community
      send-community extended
      rewrite-evpn-rt-asn
```

Route-Server in DCI

EVPN Route-Target Re-Write by Route-Server

```
router bgp 65099
  address-family l2vpn evpn
    retain route-target all
  template peer OVERLAY-PEERING
    update-source loopback0
    ebgp-multihop 5
    address-family l2vpn evpn
      send-community
      send-community extended
      route-map unchanged out
  neighbor 20.2.0.4
    inherit peer OVERLAY-PEERING
    remote-as 65002
    address-family l2vpn evpn
      rewrite-evpn-rt-asn
```



Peer-Type Fabric-External

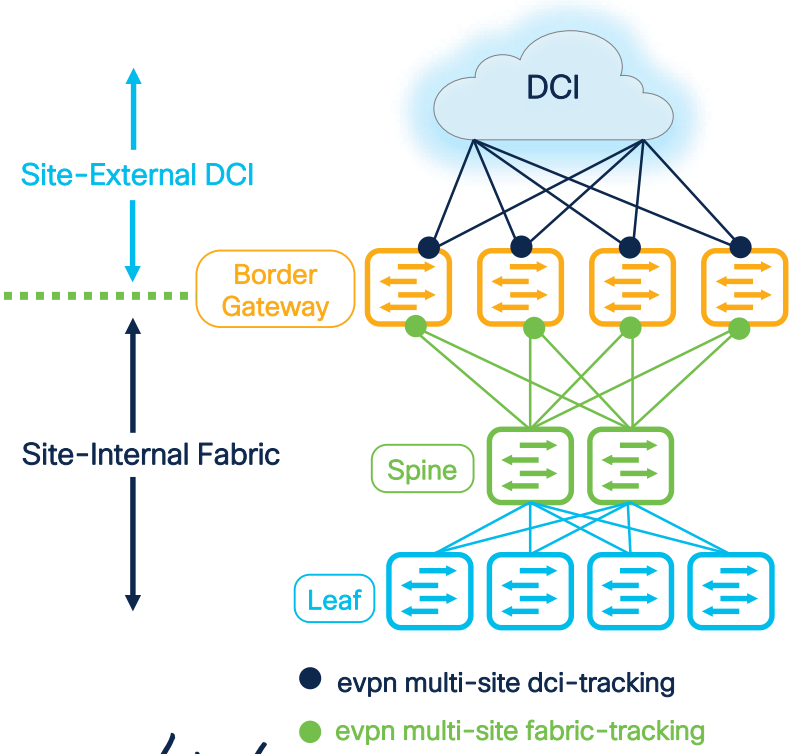
EVPN Split-Horizon and Route Re-Origination

```
router bgp 65001
  neighbor 2.2.2.2
    remote-as 65002
    update-source loopback0
    ebgp-multihop 5
    peer-type fabric-external
    address-family l2vpn evpn
      send-community
      send-community extended
      rewrite-evpn-rt-asn
```

- The command is defined towards Site-External EVPN peering. It provides the capability of VXLAN packet re-origination and implements VPN split horizon mechanism.
- EVPN route coming from Site-External peer-type must not be re-advertised back into the VPN. The route is only advertised towards Site-Internal VTEPs

Failure Detection on BGWs

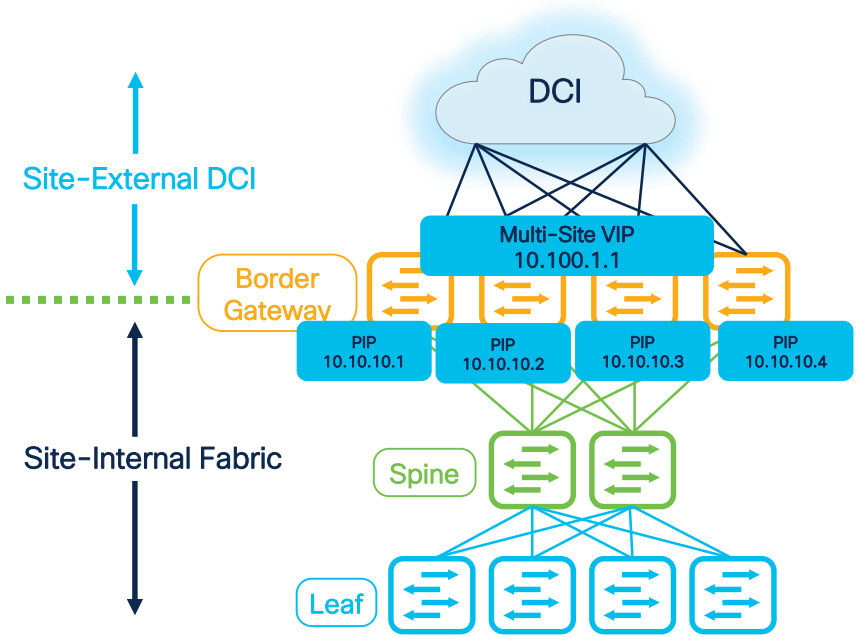
Use of Interface Tracking



- Because of the critical role played by the BGWs, it is critical to consider the required behavior during different failure scenarios.
- Tracking the state of the interfaces connecting the BGWs to the spines (“fabric-tracking”) and to the ISN (“dci-tracking”)
- The “dci-tracking” configuration is also required on Layer-3 interfaces locally connecting Anycast BGW nodes, needed for example, back-to-back topologies.
- Allows to define via configuration how the BGW is connected in the topology and to take proper action depending on the specific failure.
- The BGW node that gets isolated from the fabric or from the ISN needs to stop receiving traffic flows if it can not forward them to the destination.

Failure Detection on BGWs

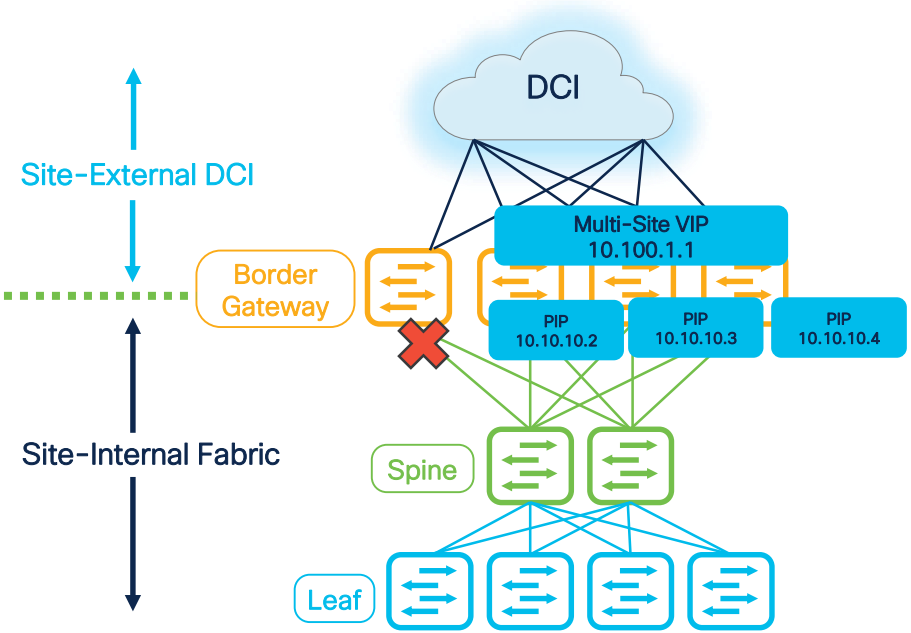
Fabric Isolation



- The Site-Internal interfaces on BGW nodes are constantly tracked to determine their status (‘evpn multisite fabric-tracking’ command)

Failure Detection on BGWs

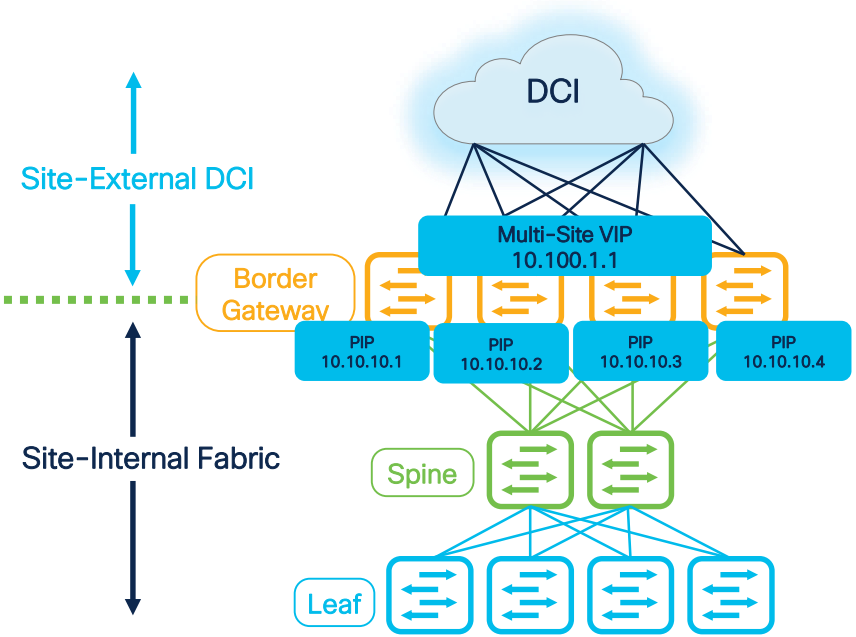
Fabric Isolation



- The Site-Internal interfaces on BGW nodes are constantly tracked to determine their status ('evpn multisite fabric-tracking' command)
- If all the Site-Internal interfaces are detected as down on a BGW:
  1) The BGW will isolate itself from Multi-Site traffic. Hence , it will shutdown Lo100 aka VIP and stop advertising to remote sites. So, the traffic from remote site will never come to this BGW as it will no longer be part of the ECMP.
  2) The BGW site external BGP session will be up and running but it will withdraw all EVPN routes (Type 2,3,4,5)
  3) The remaining BGWs withdraw all BGP EVPN Route Type 4 (Ethernet segment) routes received from the now isolated BGW because reachability is missing.
- As a result, the BGW becomes isolated from both the Site-Internal and Site-External networks.
- Seamless BGW node re-insertion using a "delay-restore" timer for the Multi-Site VIP address

Failure Detection on BGWs

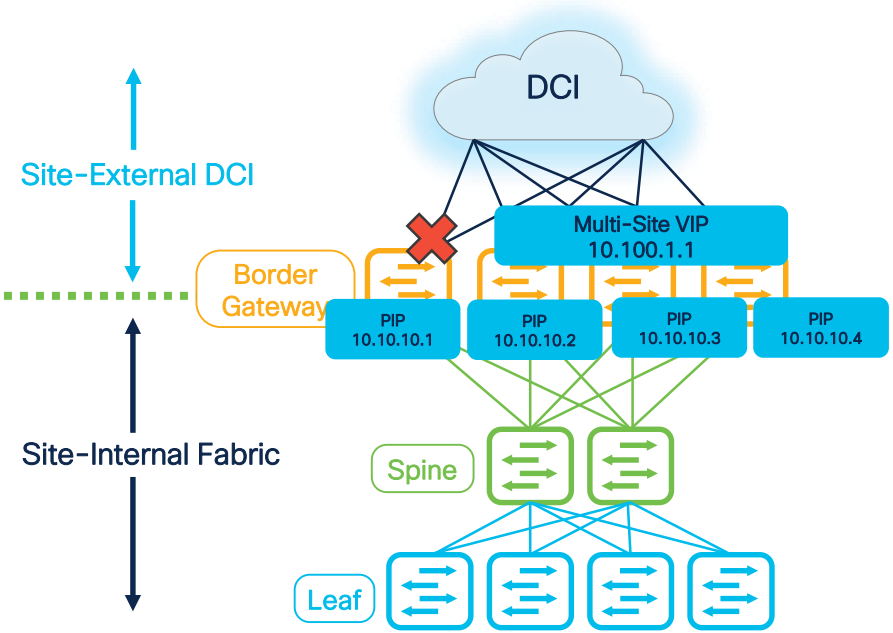
DCI Isolation



- The Site-External interfaces on BGW nodes are constantly tracked to determine their status (‘evpn multisite dci-tracking’ command)

Failure Detection on BGWs

DCI Isolation



- The Site-External interfaces on BGW nodes are constantly tracked to determine their status ('evpn multisite dci-tracking' command)
- If all the Site-External interfaces are detected as down on a BGW:
  1) The BGW will stop advertising Multi-Site VIP aka Lo100 towards Site-Internal network. Hence, all traffic destined to remote sites will be re-routed to remaining BGWs due to ECMP.
  2) The isolated BGW will also withdraw EVPN Type-4 as new DF election is triggered amongst the remaining BGWs.
  3) The isolated BGW do not need to withdraw Type-2,3, and 5 routes as DCI interfaces and remote BGP EVPN peering are down.
- As a result, the BGW continues to operate as a Site-Internal VTEP as it's PIP remains up. Hence, if any external network is connected to this BGW, traffic from local site can still be sent to this BGW.
- Seamless BGW node reinsertion using a "delay-restore" timer for the Multi-Site VIP address

Cool! But What's the Catch?

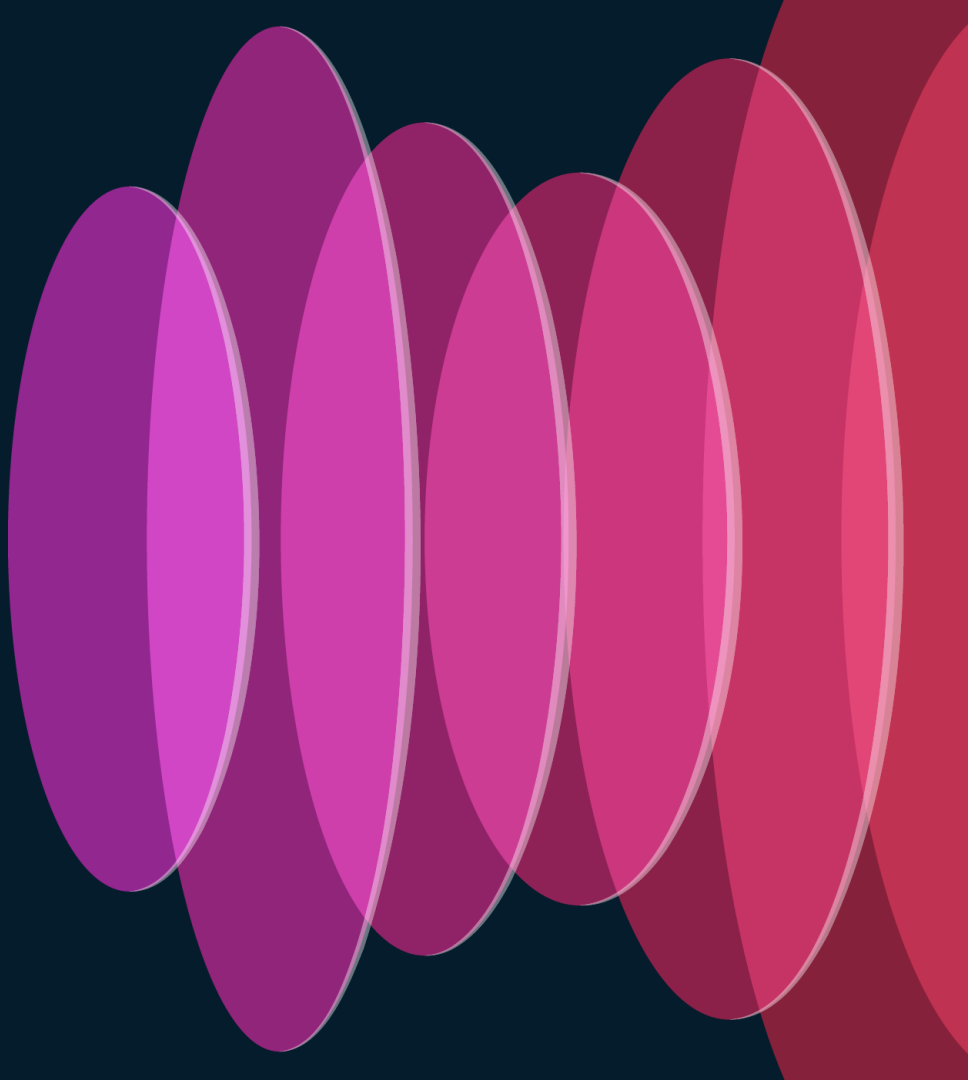
There is always a catch ☺

1. This seems really cool! Is it easy to configure?
   There are lots of moving parts: OSPF/BGP/VXLAN. Manual configuration can be challenging.

2. How easy is it to make changes?
   You still rely on traditional SSH based management to each device, which can be cumbersome and error prone.

3. How much Visibility do I have into the network?
   Visibility and Troubleshooting is still performed on a "switch-by-switch" basis.

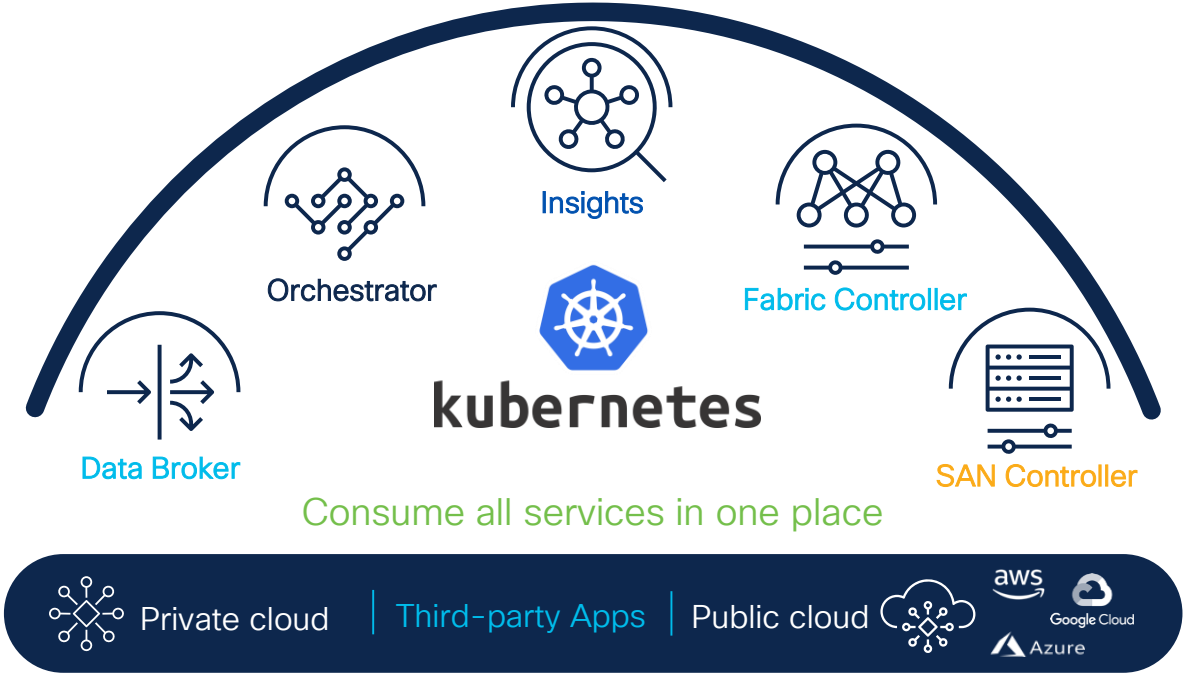
What is Nexus Dashboard Fabric Controller?



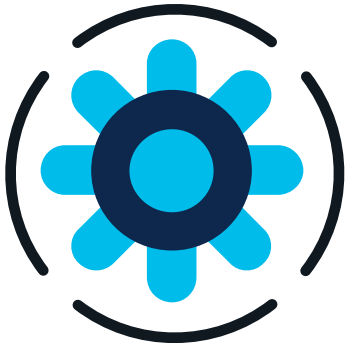
Cisco Nexus Dashboard

Simple to automate, simple to consume

Powering automation
Unified agile platform



Cisco Nexus Dashboard Fabric Controller



Automation

Accelerate provisioning and simplify deployments

Management

In depth Management and control for all network deployments

Visibility

Get Centralized Visibility and Monitoring views

Automation

Accelerate provisioning from days to minutes

Easy to understand approach to auto-bootstrapping of entire fabric

Rapid Deployment with Fabric Builder best practice templates for VXLAN-EVPN, BGP Routed, Campus, and More!

Optimized for both large deployments and traditional deployment models

Service Insertion and Layer-3 handoff

DevOps friendly

Benefits

Simplify fabric deployments	Developer agility	VXLAN EVPN Multi-Site

Management

Single point for management
for data center operations

Optimized for both large deployments
and traditional deployment models

Granular RBAC

Image management

RMA

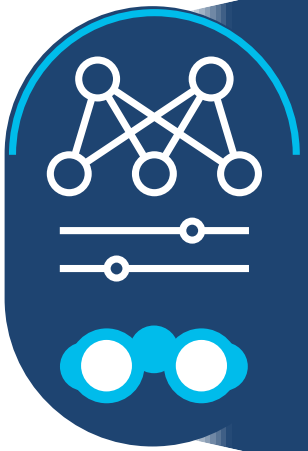
Scale within and across data centers
with One-Manage Federation

Management for non-Nexus platforms

Benefits

Reliability | Compliance | Secure

Visibility & Monitoring



Get comprehensive monitoring

Enhanced topology views

Compute and endpoint visibility

VXLAN OAM support with NDFC

Obtain detailed inventory, health, resource consumption information on devices

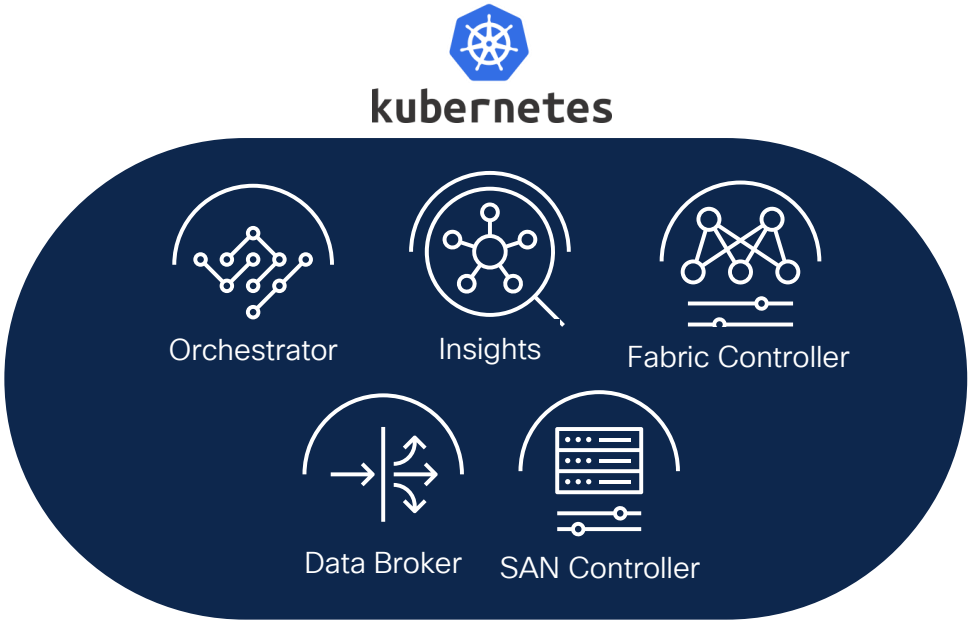
End-to-end visibility, monitoring and troubleshooting

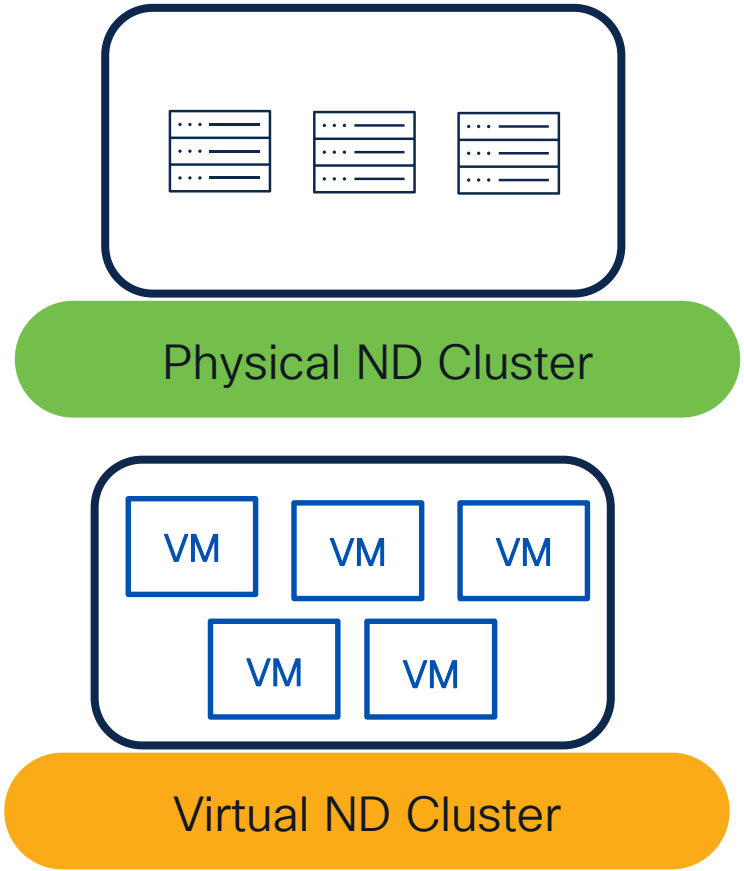
Integrate with NDI for Day 2 operations

Benefits

Intuitive	Deep visibility	Enhanced monitoring

Cisco Nexus Dashboard





Cisco Nexus Dashboard Formats – NDFC

Physical ND Cluster

Each node is a UCS Server with:

2.8GHz AMD CPU
256G RAM
4x2.4TB HDD
960 GB SSD
1.6 TB NVMe drive

For the latest information check the specific scalability guide. 12.1.3b Verified Scalability

Virtual ND Cluster

For NDFC each vND VM must satisfy the following requirements:

Specs	APP
vCPU	16
RAM (GB)	64
SSD (GB)	550

Cisco Nexus Dashboard Scaling – NDFC

Physical ND Cluster

Full scale for NDFC can be achieved with 3 nodes

Managed mode (VXLAN and BGP fabrics): 500 switches

Managed/Monitor mode (External fabrics): 1000 switches

Overall fabric count: 50

Recommended

Virtual ND Cluster

Full scale for NDFC can be achieved with 5 nodes

Managed mode (VXLAN and BGP fabrics): 400 switches

Managed/Monitor mode (External fabrics): 1000 switches

Overall fabric count: 50

3x vND can support 100 switches in managed mode

VXLAN EVPN (Greenfield)

Switches per Fabric: 200

Overlays: 500 VRF and 2000 Layer-3 Networks OR 2500 Layer-2 Networks

Multi-Site Domain: 30 fabrics

ToR/Leaf: 40 Leaf (VTEP) and 320 ToRs in DC VXLAN EVPN fabric

VXLAN EVPN (Brownfield)

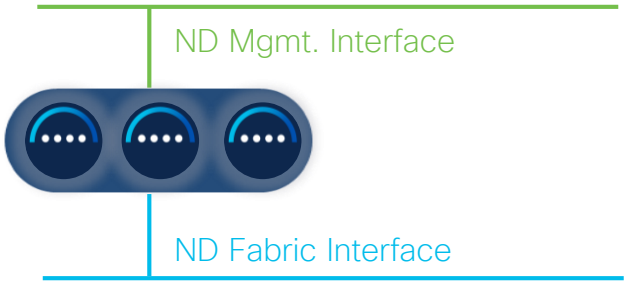
Switches per Fabric: 200

Overlays: 400 VRF and 1050 Layer-3/Layer-2 Networks

Multi-Site Domain: 30 fabrics

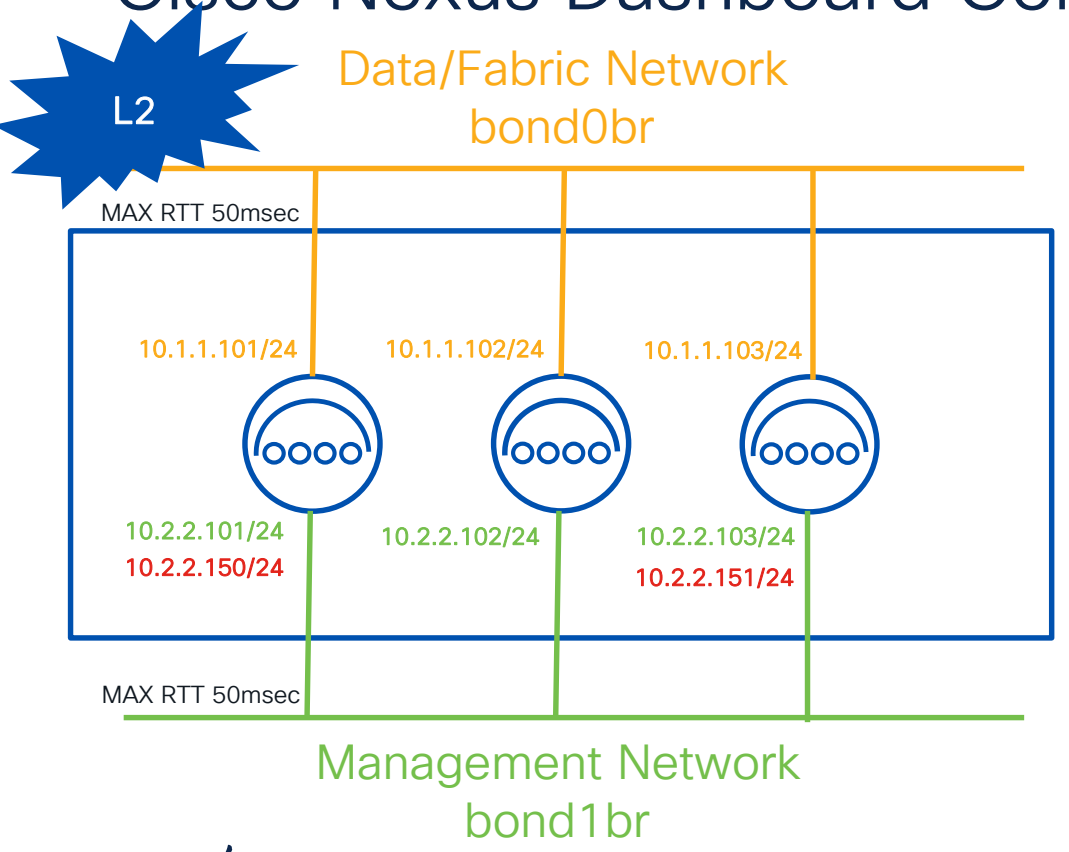
In any case at least 3 nodes must be deployed for proper redundancy. 1x vND also supported for Production

Nexus Dashboard
Interface Types



- Each ND node has two interface types:
  - Management Interface: should be dedicated to the management of the ND cluster → connectivity to NTP and DC Proxy servers, Intersight, DNS, ND (and ND Apps) UI access and to perform firmware upgrade (for ND or Apps)
  - Fabric Interface: used for the bring up of the ND cluster (node to node communication) and application to application (NDO, NDI, NDFC, etc.) communication

Cisco Nexus Dashboard Connectivity – NDFC



TIP The two interfaces cannot share the same subnet

Intra/Inter APP PTP

ND Clustering

Switch Access*

SNMP TRAPS

DNS

POAP DHCP

NTP

GUI Access

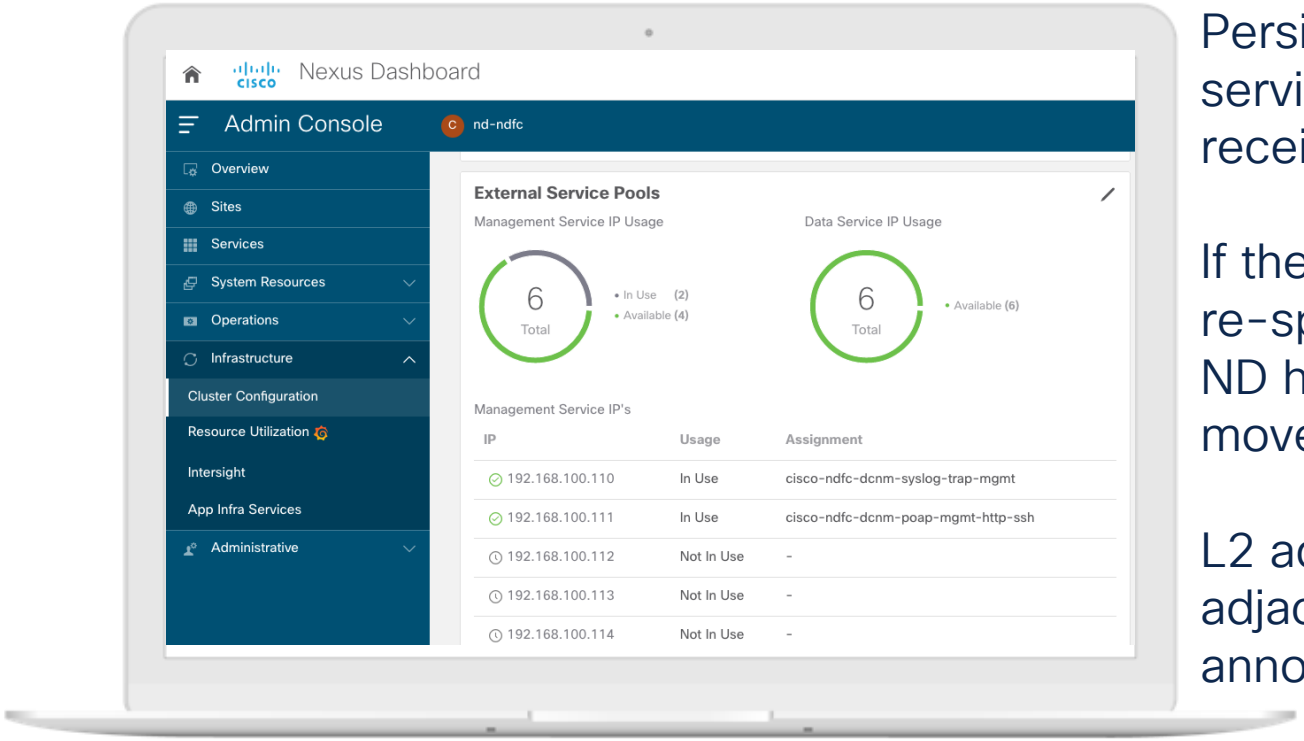
DC App Center

CLI via SSH

Intersight

* by default

NDFC Persistent IPs

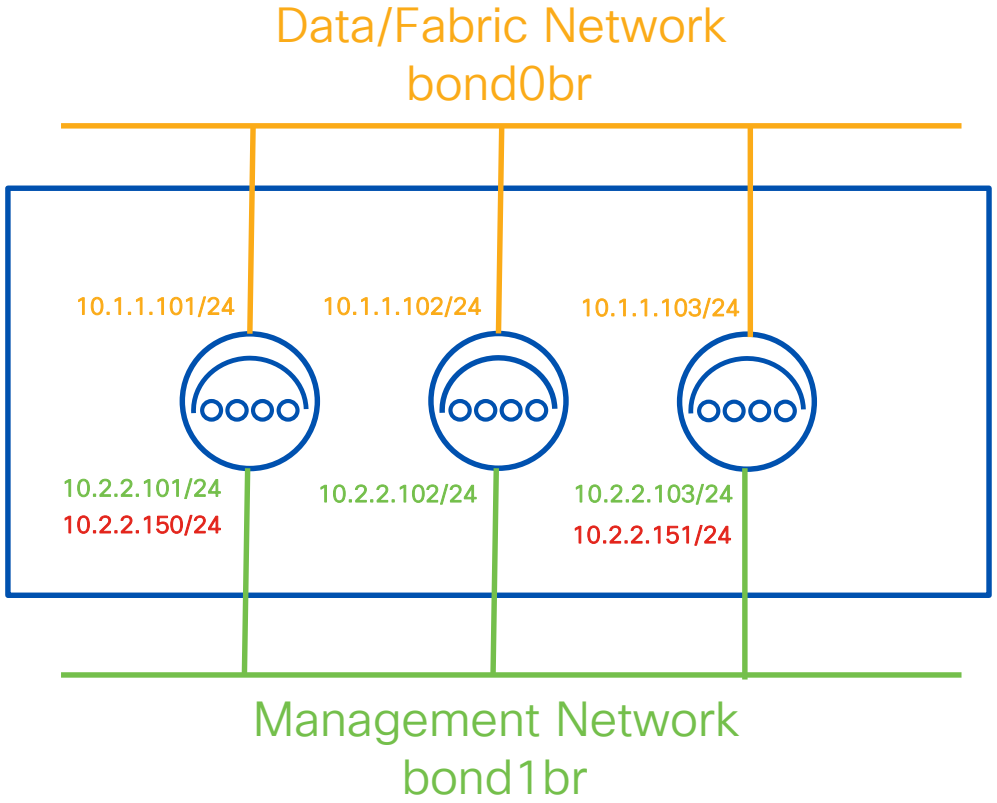


Persistent IPs are tied to a service, like the SNMP trap receiver

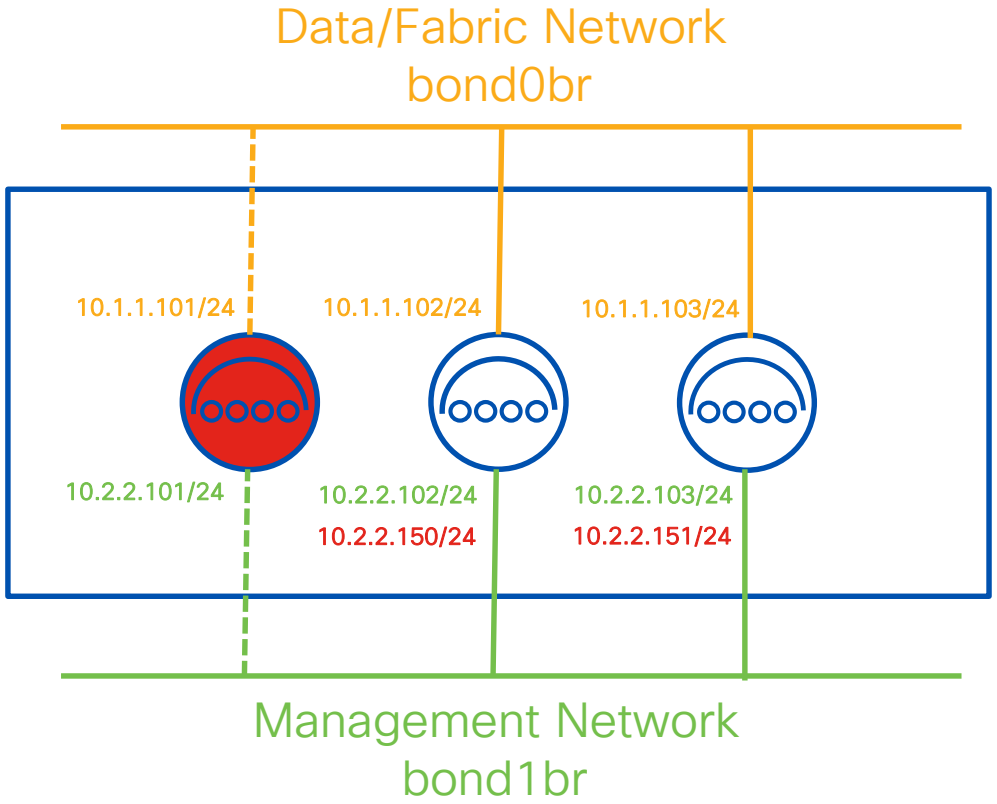
If the SNMP trap POD gets re-spawned into a different ND host the sticky IP will be moved there

L2 adjacency uses ARP, L3 adjacency BGP announcements

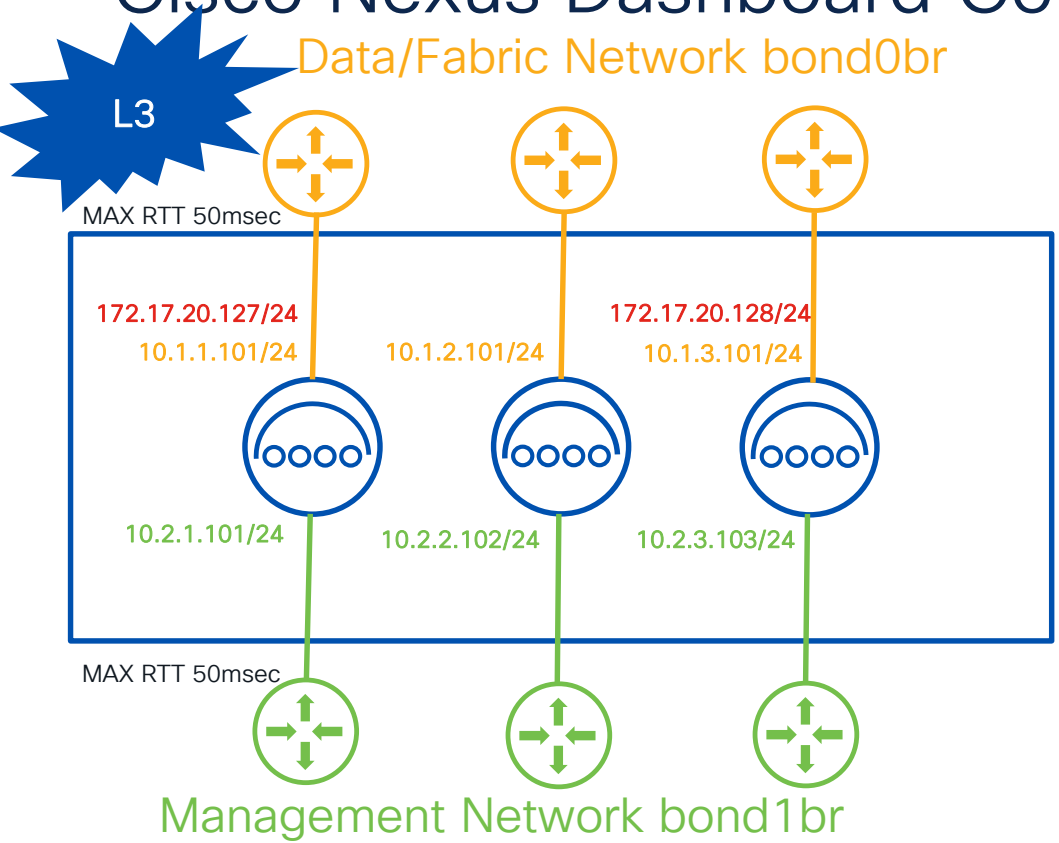
NDFC Persistent IPs – Normal conditions



NDFC Persistent IPs – Failover



Cisco Nexus Dashboard Connectivity – NDFC



This L3 options is valid since 12.1.1e

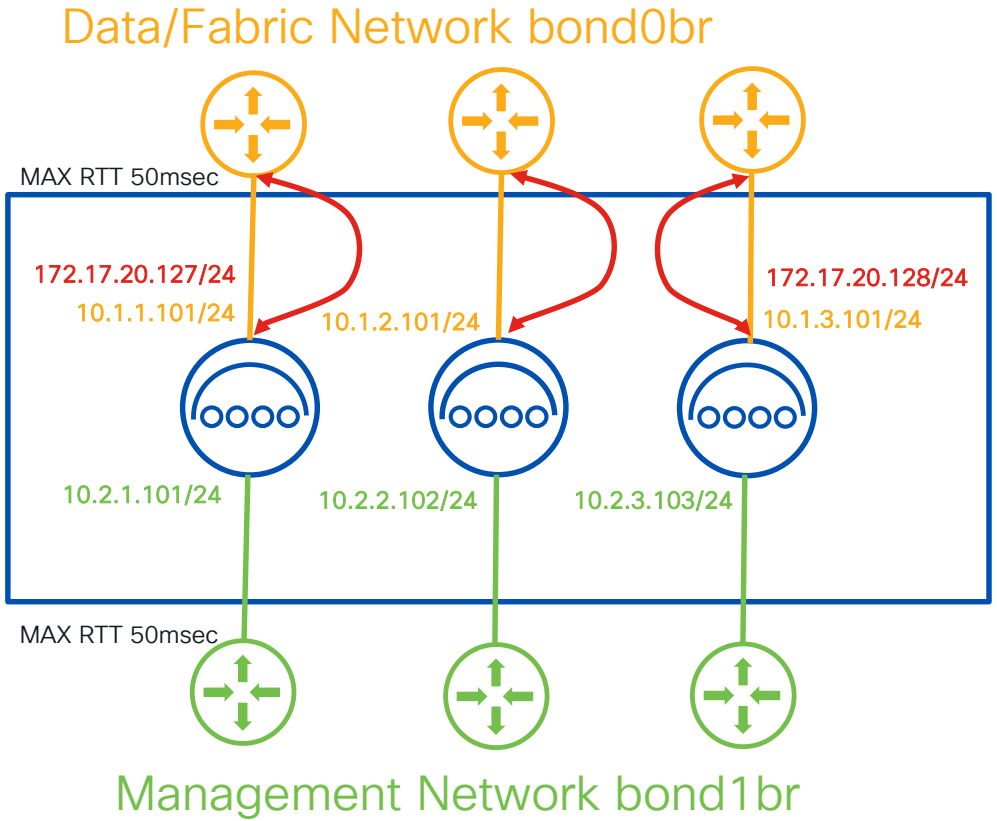
Each ND node on a different Subnet

BGP Sessions are established for Persistent IP advertisement (no multi-hop)

Persistent IPs must not overlap with ND subnets

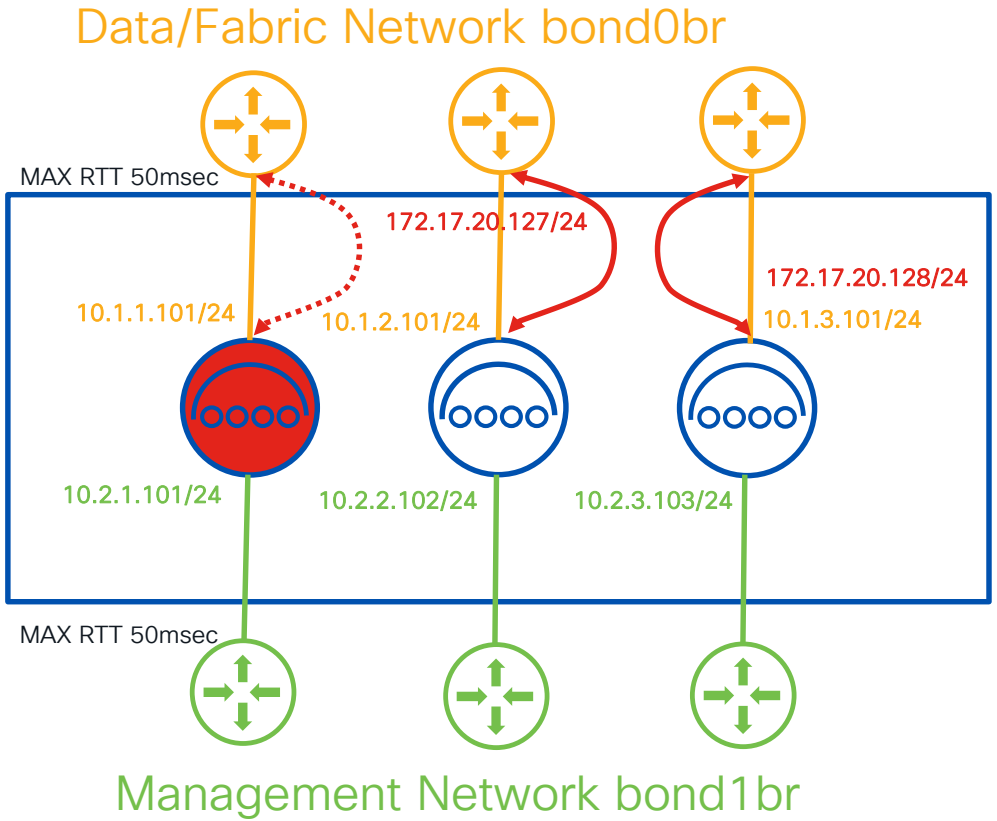
Cisco Nexus Dashboard Connectivity – NDFC

L3



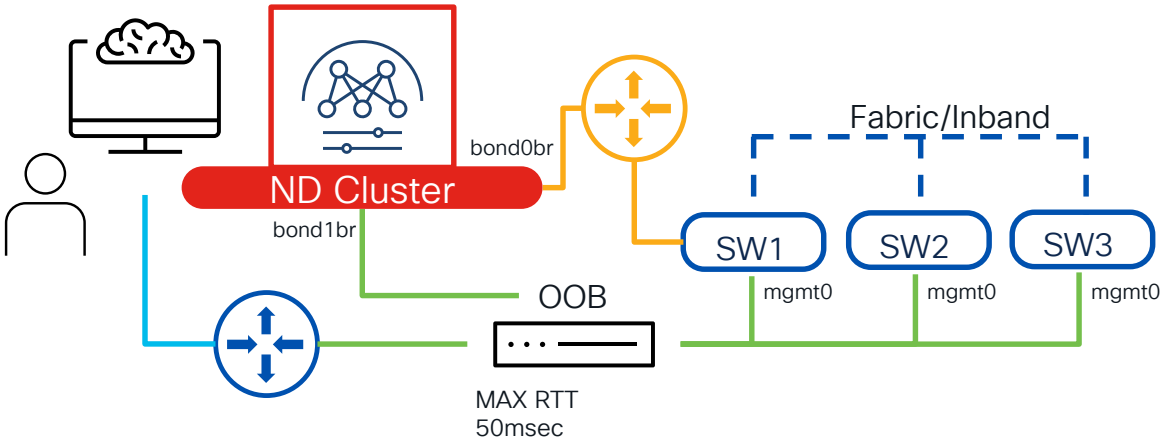
Cisco Nexus Dashboard Connectivity – NDFC

L3



Cisco NDFC Connectivity to the Switches

Use case #1



Discovery and Deployment happen via ND Management Interface as that subnet is directly connected

ND Data Interface eventually used for Endpoint Locator Feature (BGP towards Spine RR)

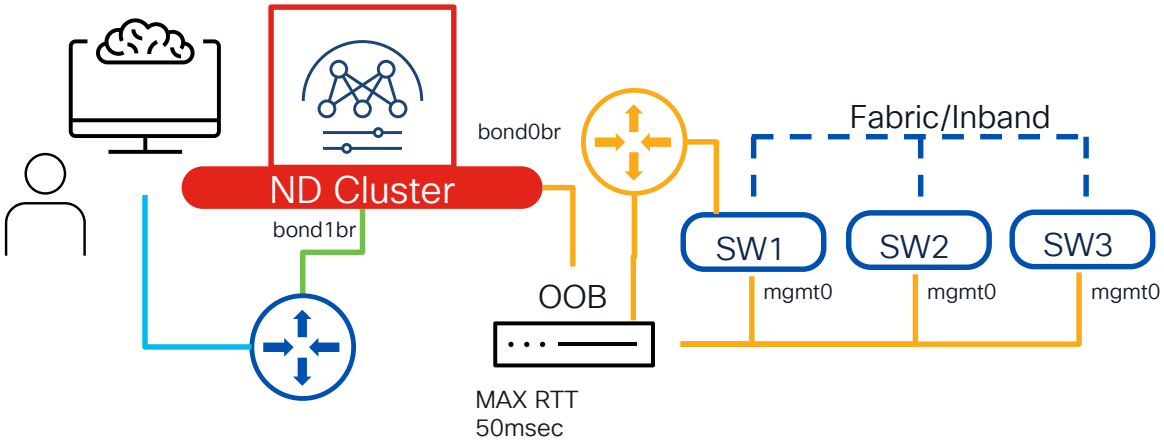
Persistent IPs are allocated on the Management Subnet

ND Mgmt Subnet	ND Data Subnet	Fabric Mgmt 0 Subnet	Fabric Inband Subnet
10.2.2.0/24	10.1.1.0/24	10.2.2.0/24	10.3.3.0/24

Works by default!

Cisco NDFC Connectivity to the Switches

Use case #2



Everything is done over the ND Data Interface as that subnet is directly connected

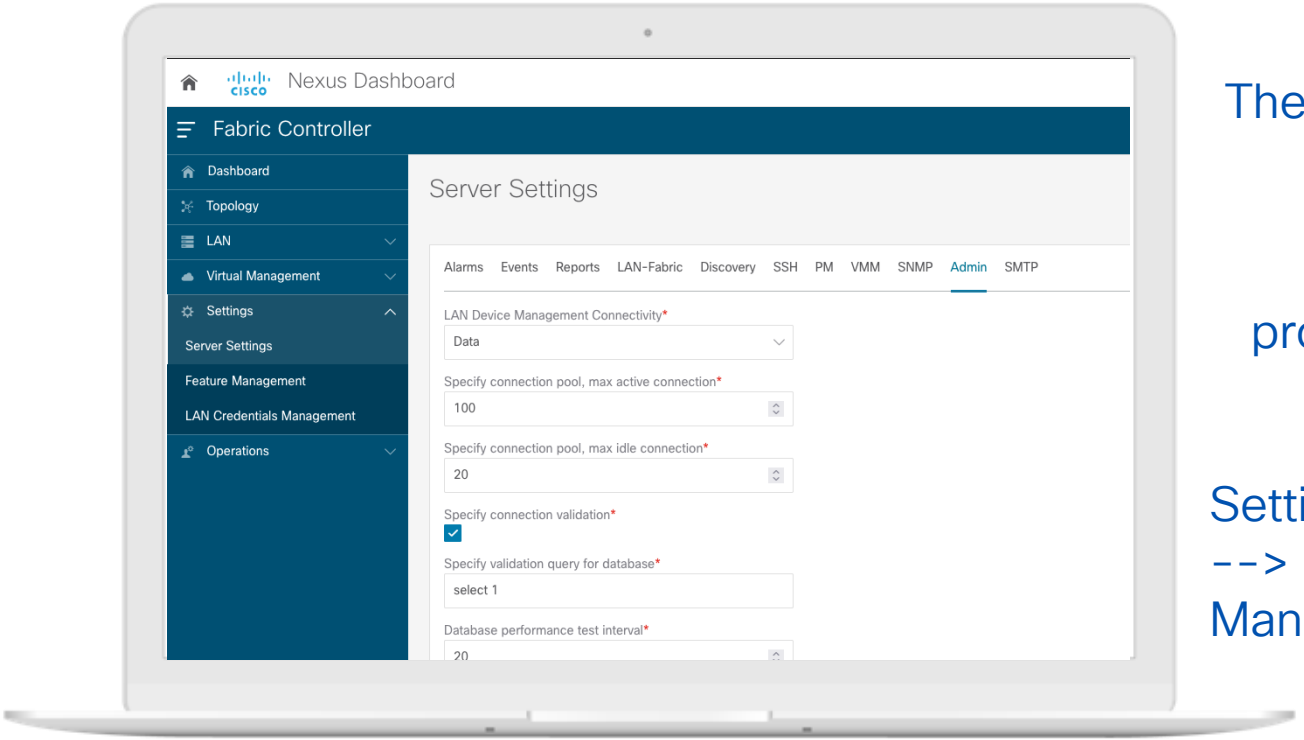
Persistent IPs are allocated on the Data Subnet

ND Mgmt Subnet	ND Data Subnet	Fabric Mgmt 0 Subnet	Fabric Inband Subnet
10.2.2.0/24	10.1.1.0/24	10.1.1.0/24	10.3.3.0/24

LAN Device Management Connectivity must be set to Data
(see next slide)

Cisco NDFC Connectivity to the Switches

Use case #2 continues



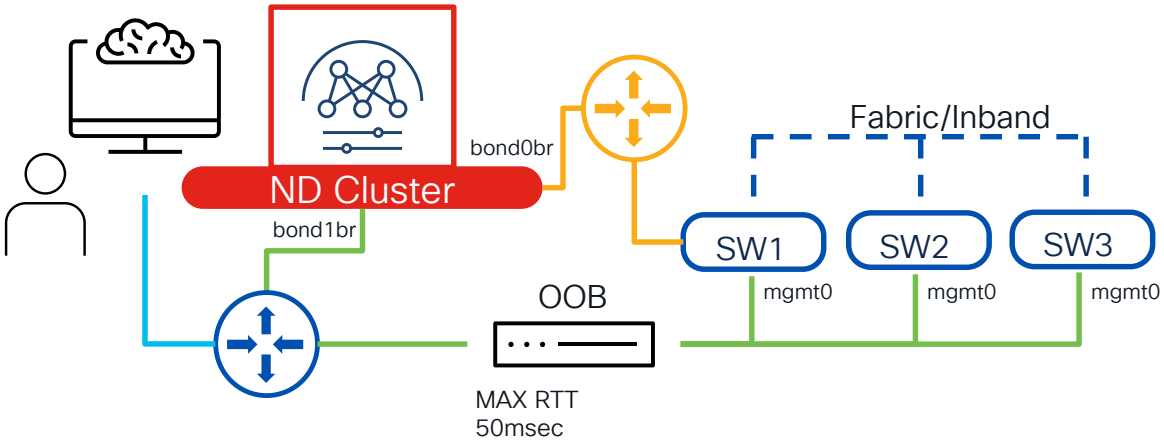
The change is global for the NDFC Instance

Persistent IPs will be provisioned over ND Data Interface

Settings --> Server Settings --> LAN Device Management Connectivity

Cisco NDFC Connectivity to the Switches

Use case #3



Discovery and Deployment happen via ND Management Interface

ND Data Interface eventually used for Endpoint Locator Feature (BGP towards Spine RR)

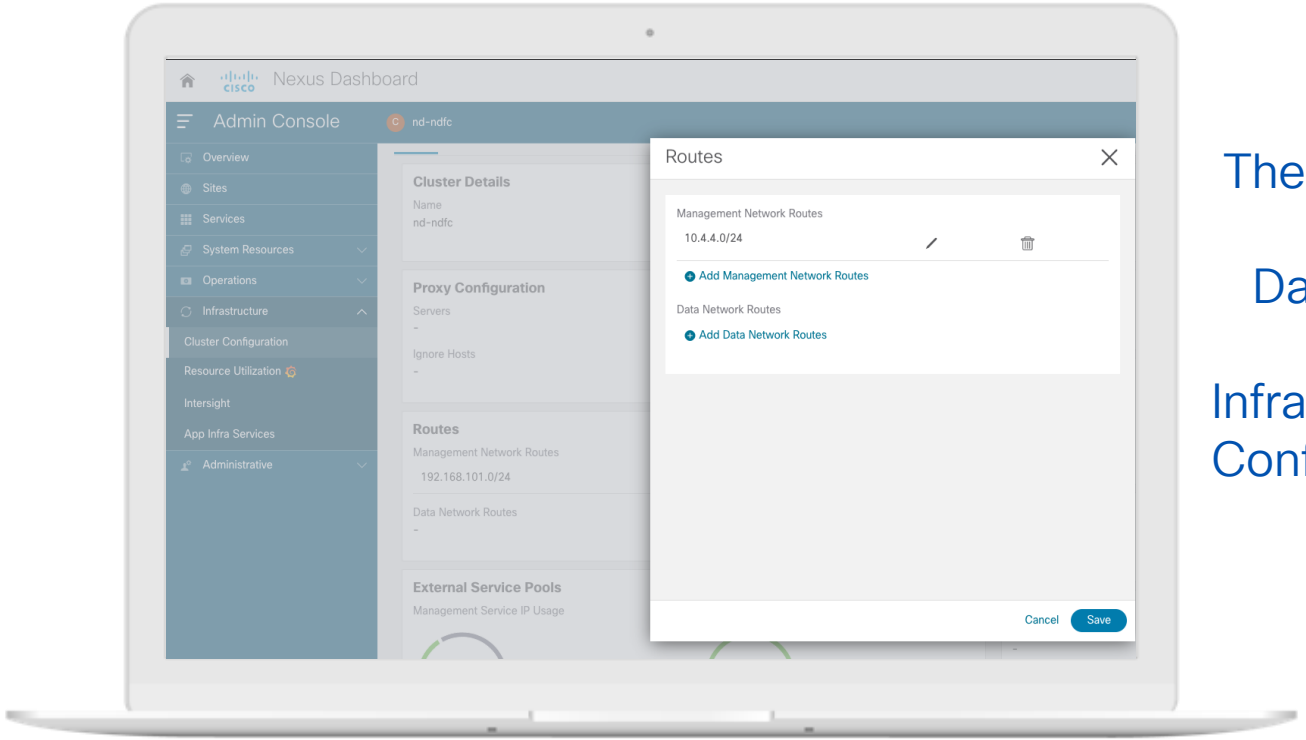
Persistent IPs are allocated on the Management Subnet

ND Mgmt Subnet	ND Data Subnet	Fabric Mgmt 0 Subnet	Fabric Inband Subnet
10.2.2.0/24	10.1.1.0/24	10.4.4.0/24	10.3.3.0/24

A static route to 10.4.4.0/24 must be added in ND Management Interface (see next slide)

Cisco NDFC Connectivity to the Switches

Use case #3 continues

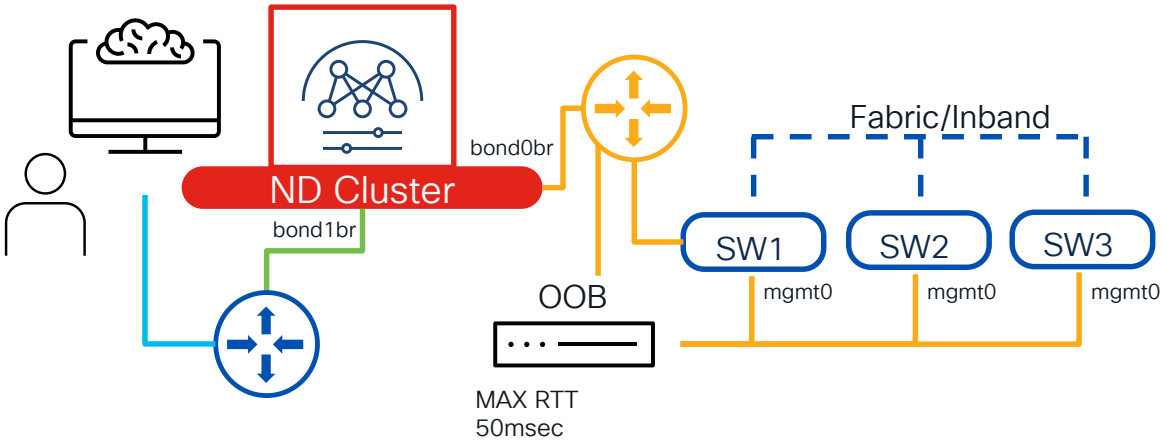


The static route needs to be added in the Nexus Dashboard Control Panel.

Infrastructure --> Cluster Configuration --> Routes

Cisco NDFC Connectivity to the Switches

Use case #4



ND Mgmt Subnet	ND Inband Subnet	Fabric Mgmt 0 Subnet	Fabric Inband Subnet
10.2.2.0/24	10.1.1.0/24	10.4.4.0/24	10.3.3.0/24

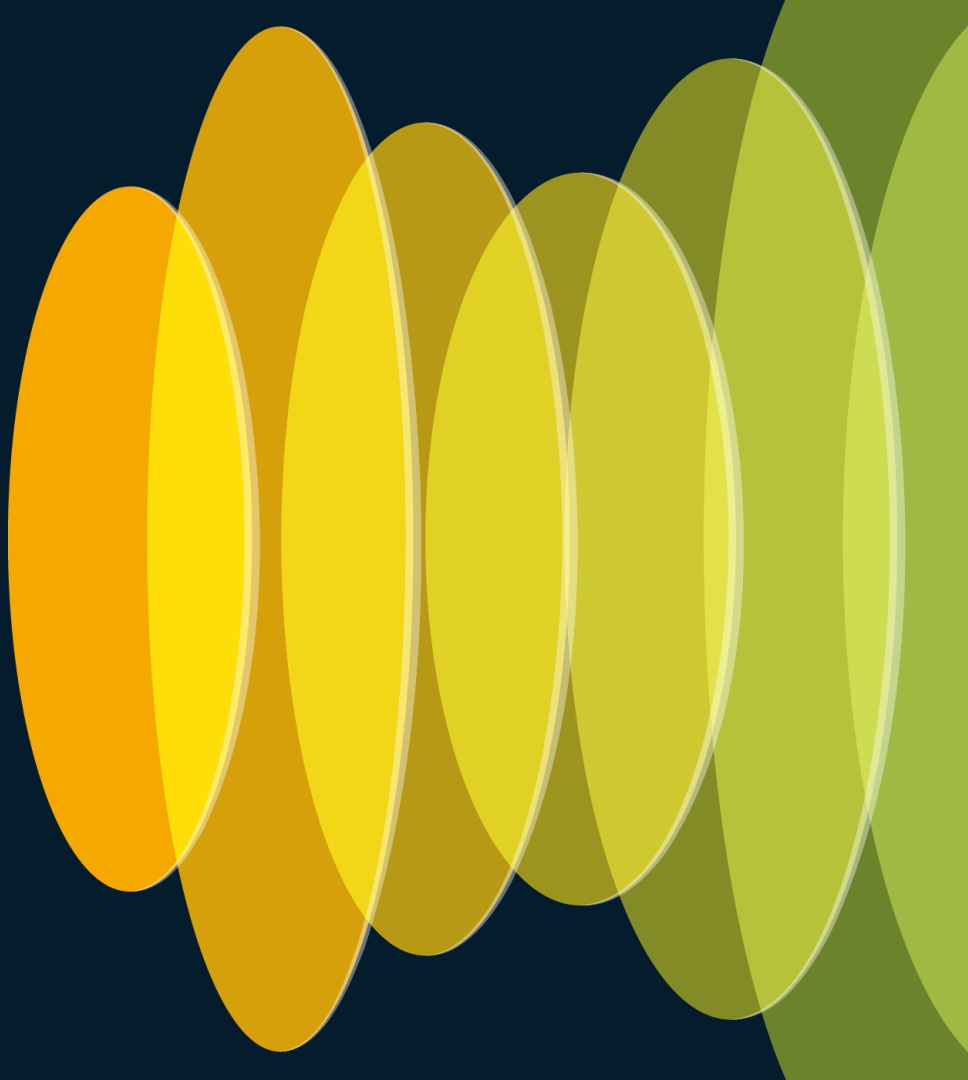
Everything is done over the ND Data Interface as that subnet is directly connected

Persistent IPs are allocated on the Data Subnet

LAN Device Management Connectivity must be set to Data

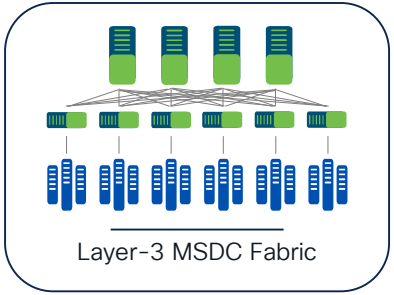
A static route to 10.4.4.0/24 must be added in ND Data Interface, not for routing but for POAP

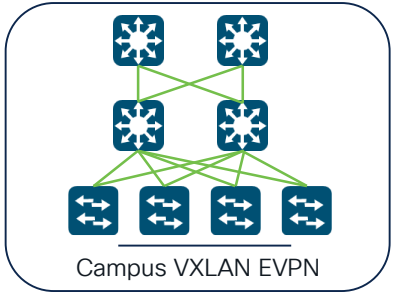
Why do YOU need NDFC?

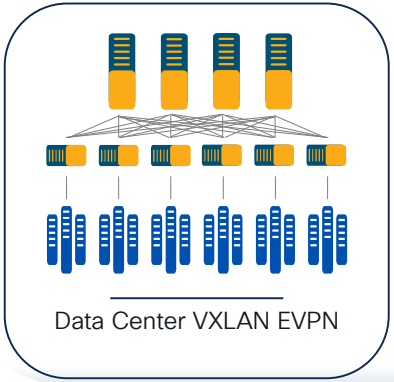


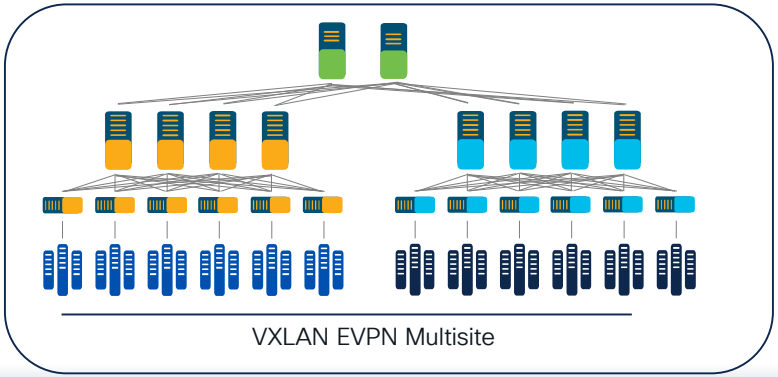
Why NDFC?

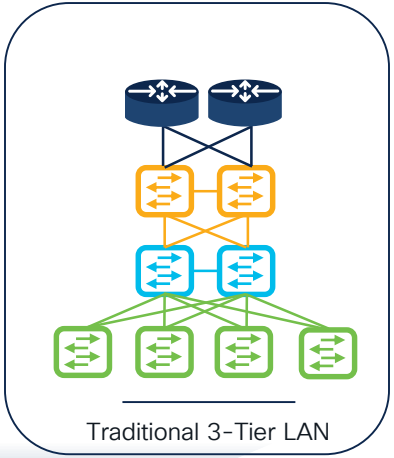
Multi-Architecture







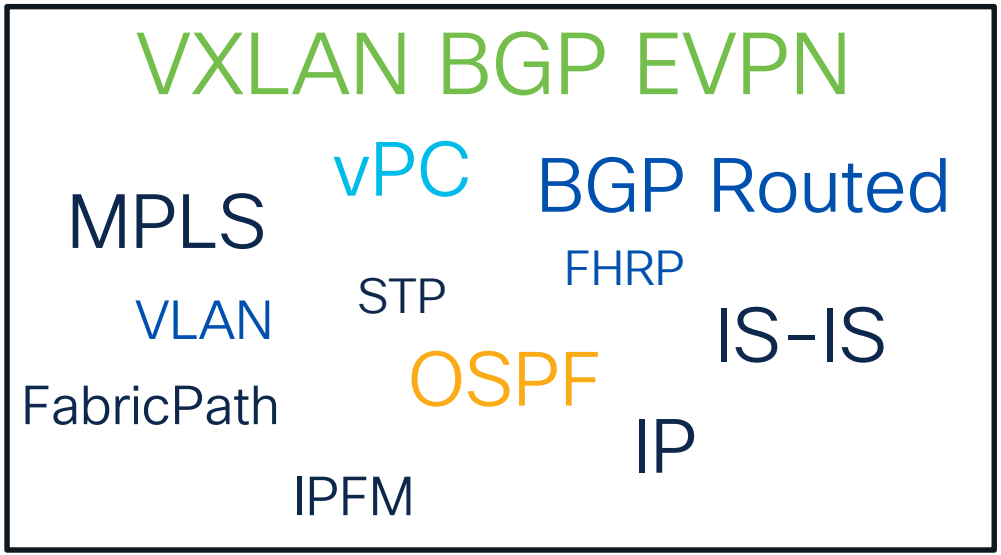




Why NDFC?

Multi-Topology, Multi-Protocol

Rich set of control plane and data plane possibilities available



Why NDFC?

Multi-Domain, Multi-Platform

NX-OS Nexus 9000 and 3000

IOS-XE Catalyst 9000

IOS-XR ASR 9000

NX-OS Nexus 7000

IOS-XE ASR 1000

TIP

Supported Hardware and Software might vary depending on NDFC version
Check compatibility matrix 12.1.3b

Why NDFC?

In a nutshell...

Step into SDN via VXLAN BGP EVPN

Config and Compliance across Cisco Products

Single Source of Truth

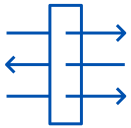
End to End Automation

Single Pane of Glass for Day-0/Day-1 Provisioning

Multi-OS management and support

Simplify Complex Network Operations

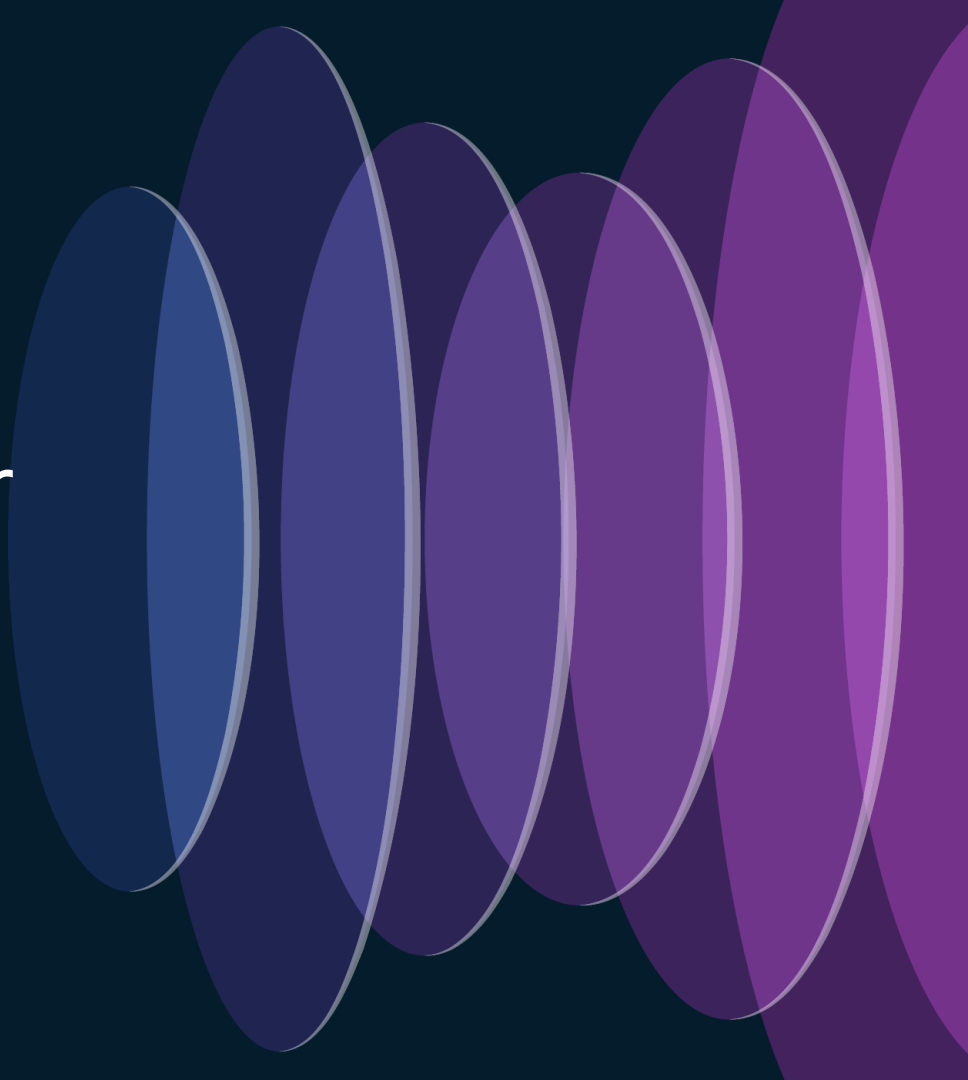
Automate, Manage, and Interconnect Multi-Fabric topologies



Layer-3 Boundary across Zones, L2/L3 across IOS-XE, NXOS, and Multicast Overlay

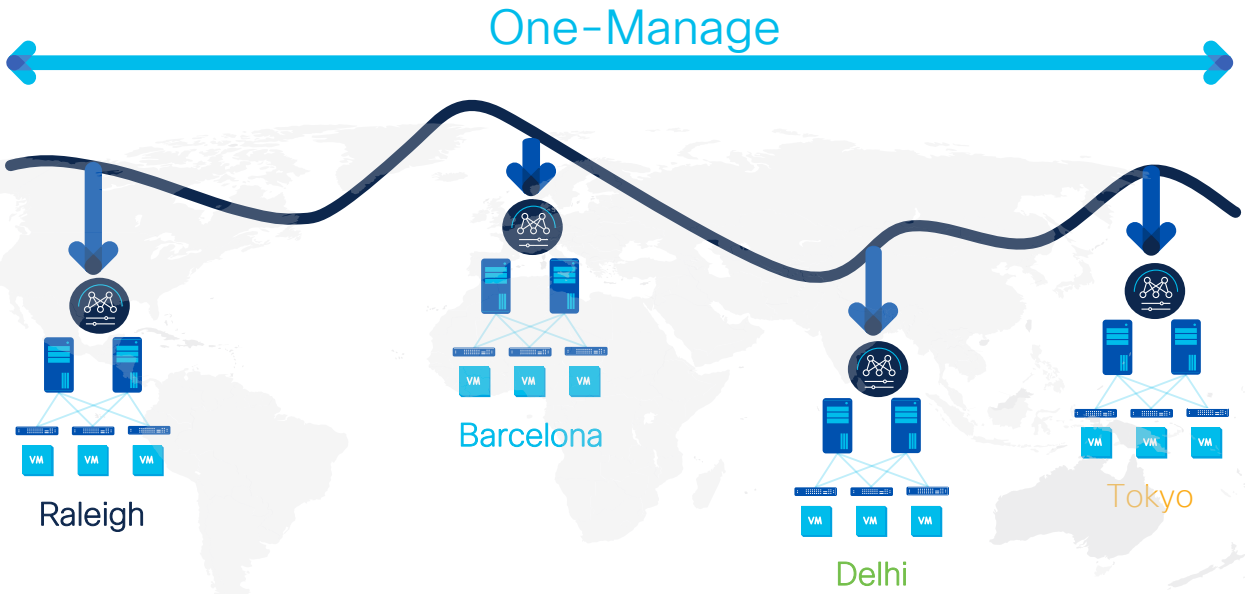
Programmability and Orchestration

NDFC Multi-Cluster
One-Manage



Multi-Cluster NDFC Deployments

Cisco NDFC 12.2(2)



- Single sign-on for clusters
  (With remote authentication)
- One view dashboard to see summary of fabrics, switch details etc.
- VXLAN EVPN Fabric
  - Unified view and automation workflows for Multi-Site deployments
  - Federation of MSDs
  - L2/L3 Network stretch
- No. of clusters in a federation: 4
  (12.2(2) scale)

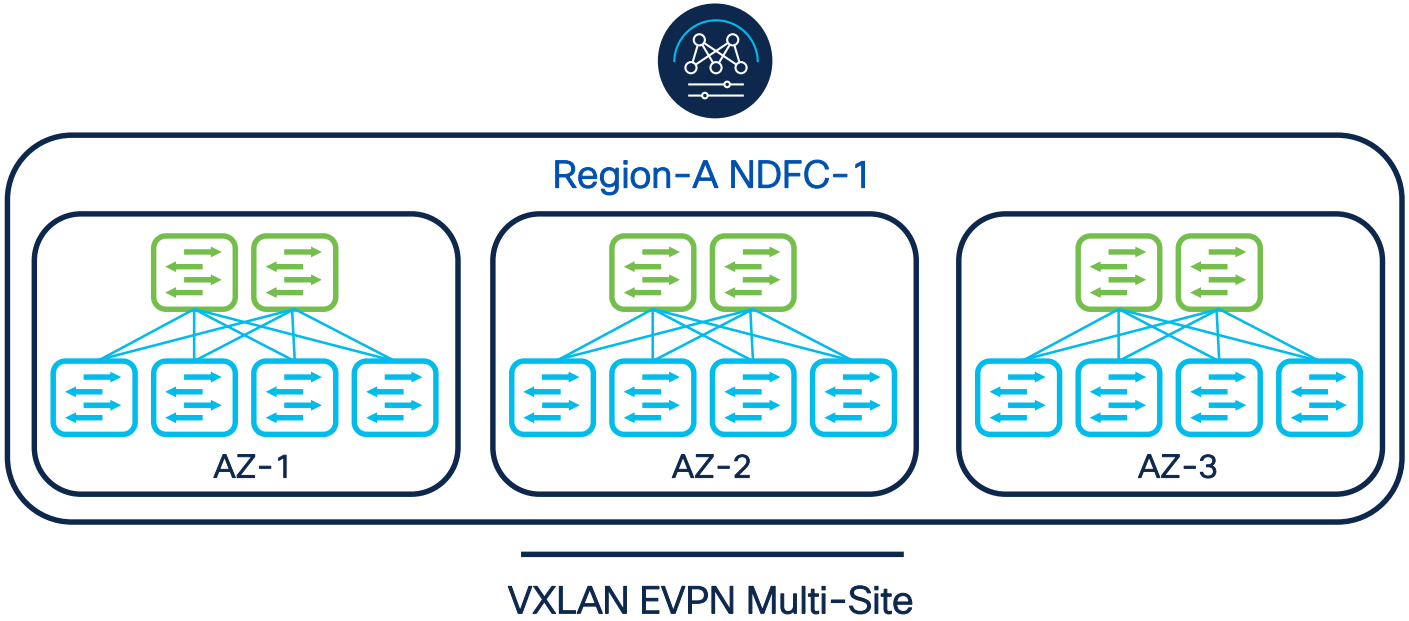
Benefits

End-to-End connectivity and policy between loosely coupled DC sites	Disaster Recovery, IP Mobility and Active-Active use cases	Network policy consistency	Change control domains separation	Flexible geo-redundancy for NDFC	Single point of Orchestration

NDFC-managed VXLAN Multi-Site

Currently Shipping

Use-case: Managing multiple DCs in a single instance of NDFC

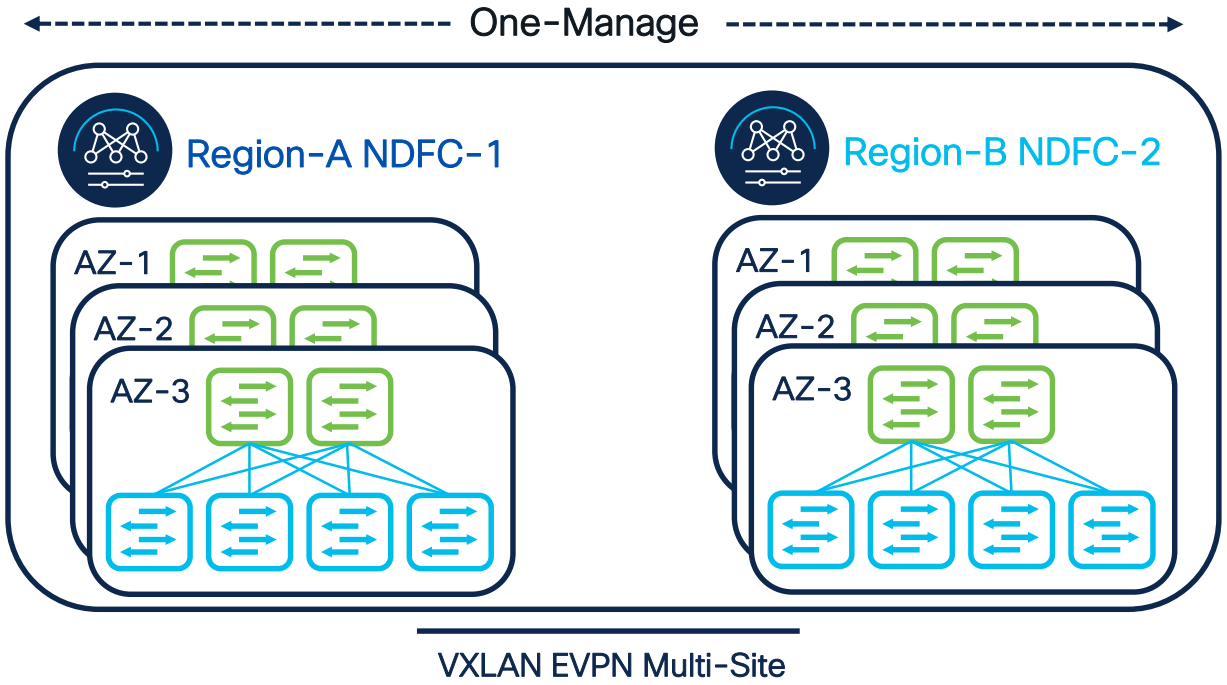


NDFC-managed VXLAN Multi-Site

One-Manage

Cisco NDFC 12.2(2)

Use-case 1: Managing multiple DCs with multiple instances of NDFC

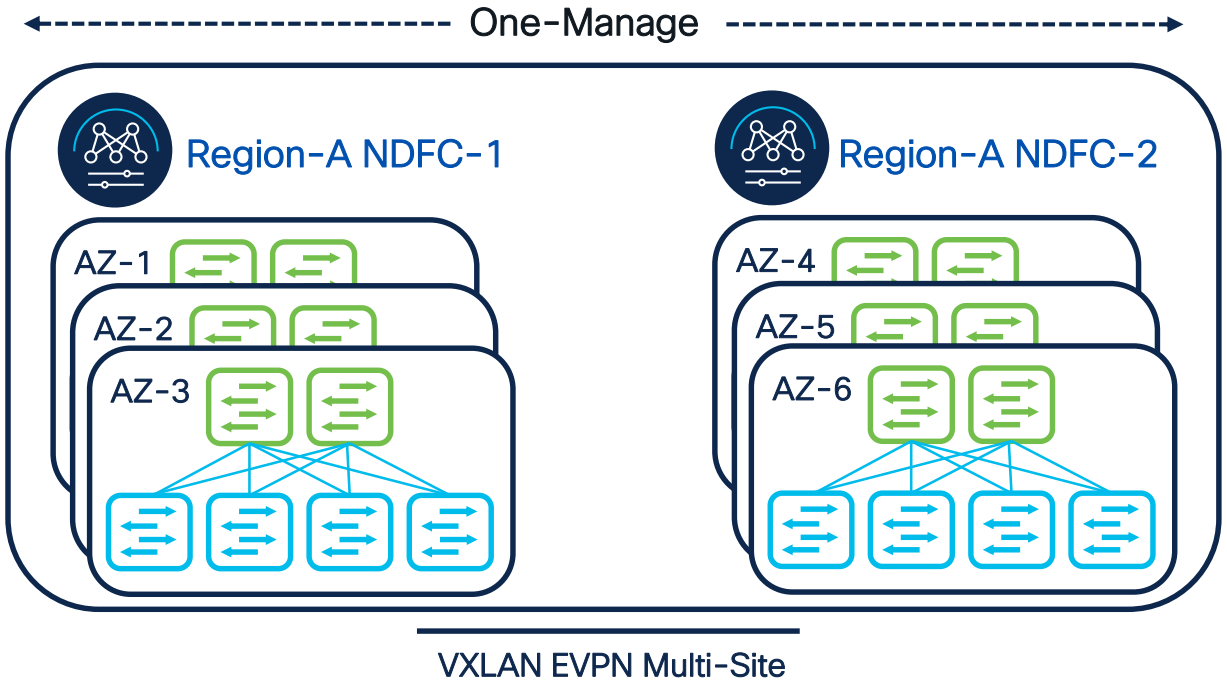


NDFC-managed VXLAN Multi-Site

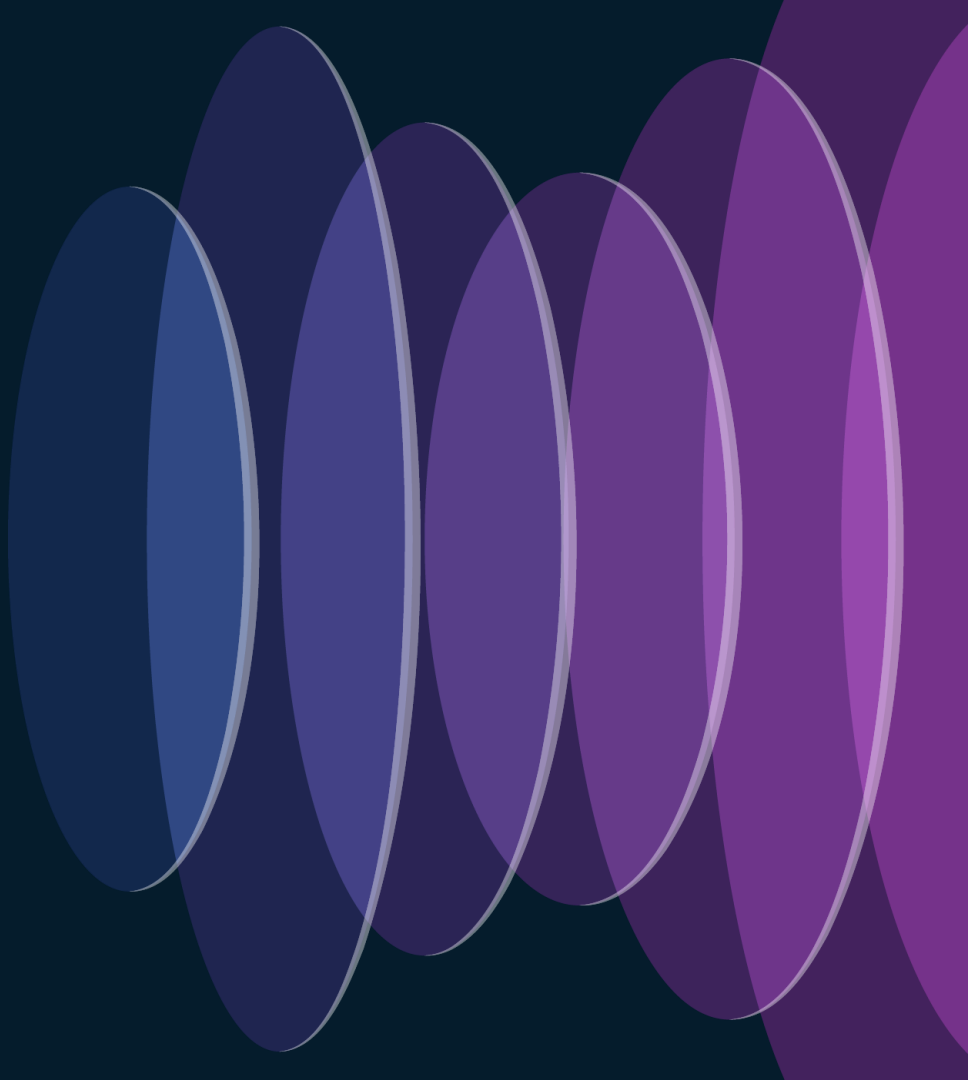
One-Manage

Cisco NDFC 12.2(2)

Use-case 2: Managing single large-scale DC with 500+ switches



Automate VXLAN Multi-Site with NDFC



cisco Live!

VXLAN BGP EVPN Greenfield

Not on VXLAN EVPN Today?

- NDFC Fabric Controller Mode
- Build VXLAN fabric in few minutes
- Templates already embed best practices
- IP addresses, overlay pool, routing profiles, replication attributes –all taken care by NDFC

Step 1
Create

Define fabric settings (Underlay, Overlay) – AS#, Replication Mode, IGP, IP Pools, etc.

Step 2
Discover

Import switches with POAP or Day-0 config

Define switch Roles (Border, Leaf, Spine, etc)

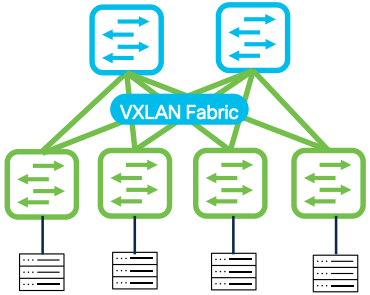
[Optional] Create vPC pairs

Step 3
Recalculate and Deploy

Generates config based on intent

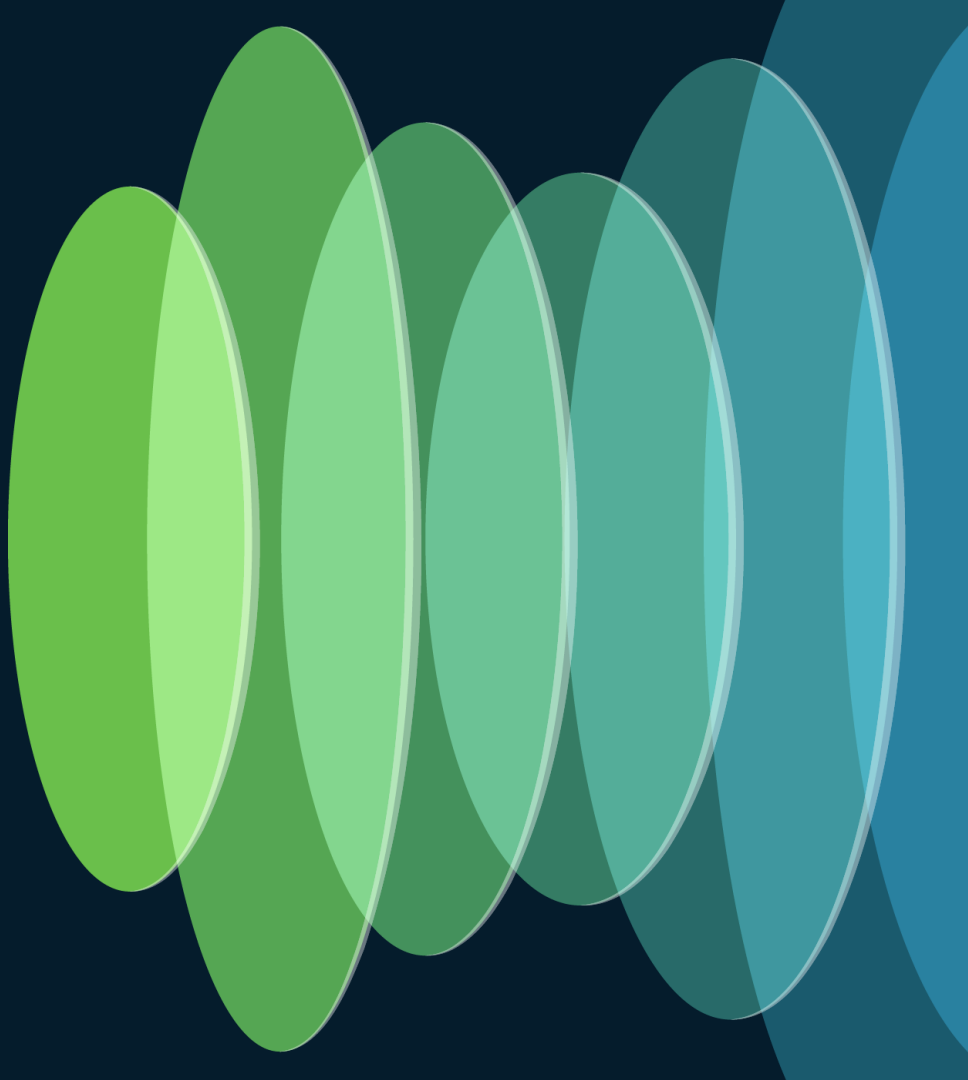
Preview side by side diffs

Cisco best practice implemented



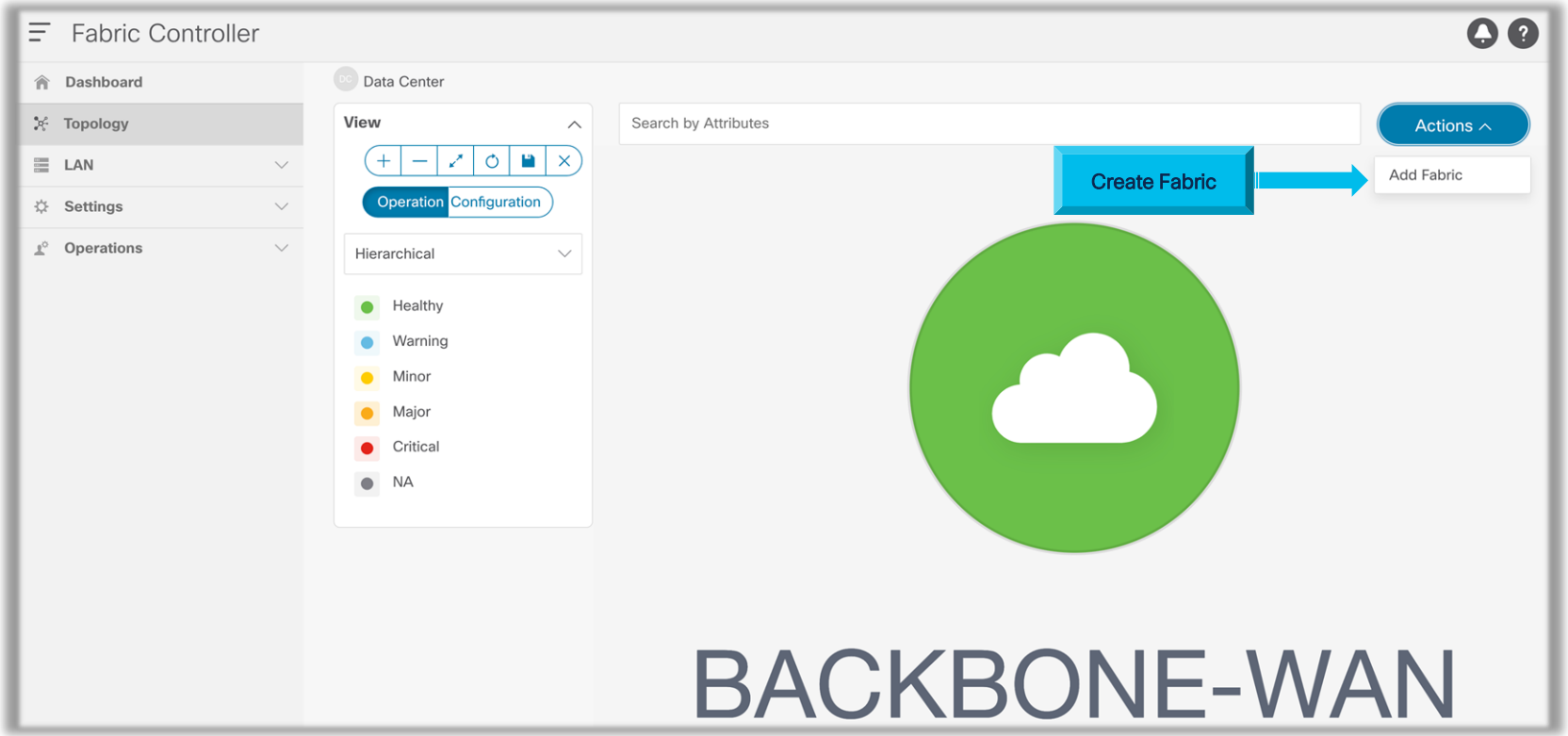
Fast, automated process

NDFC Day-0: VXLAN EVPN Underlay

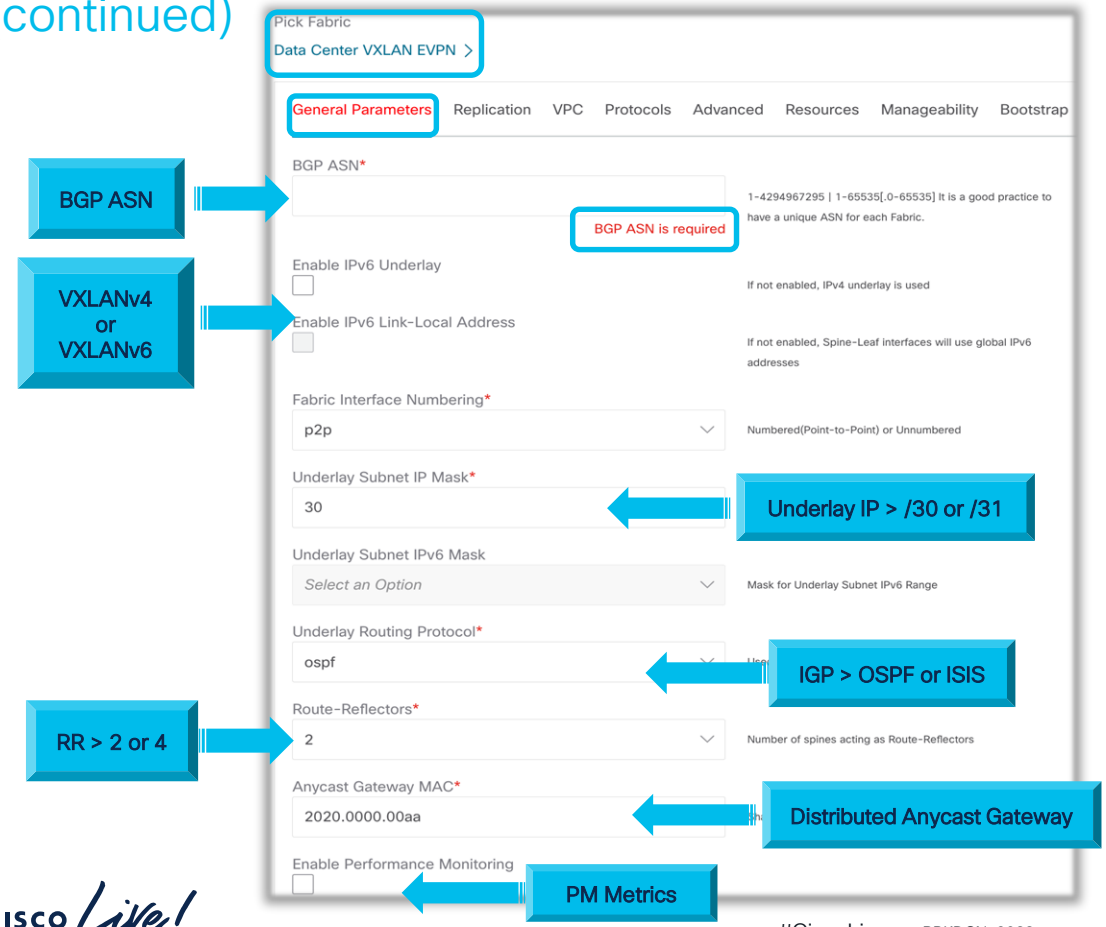


cisco Live!

Step1 -> Create a Fabric



Step1 -> Create a Fabric
(continued)

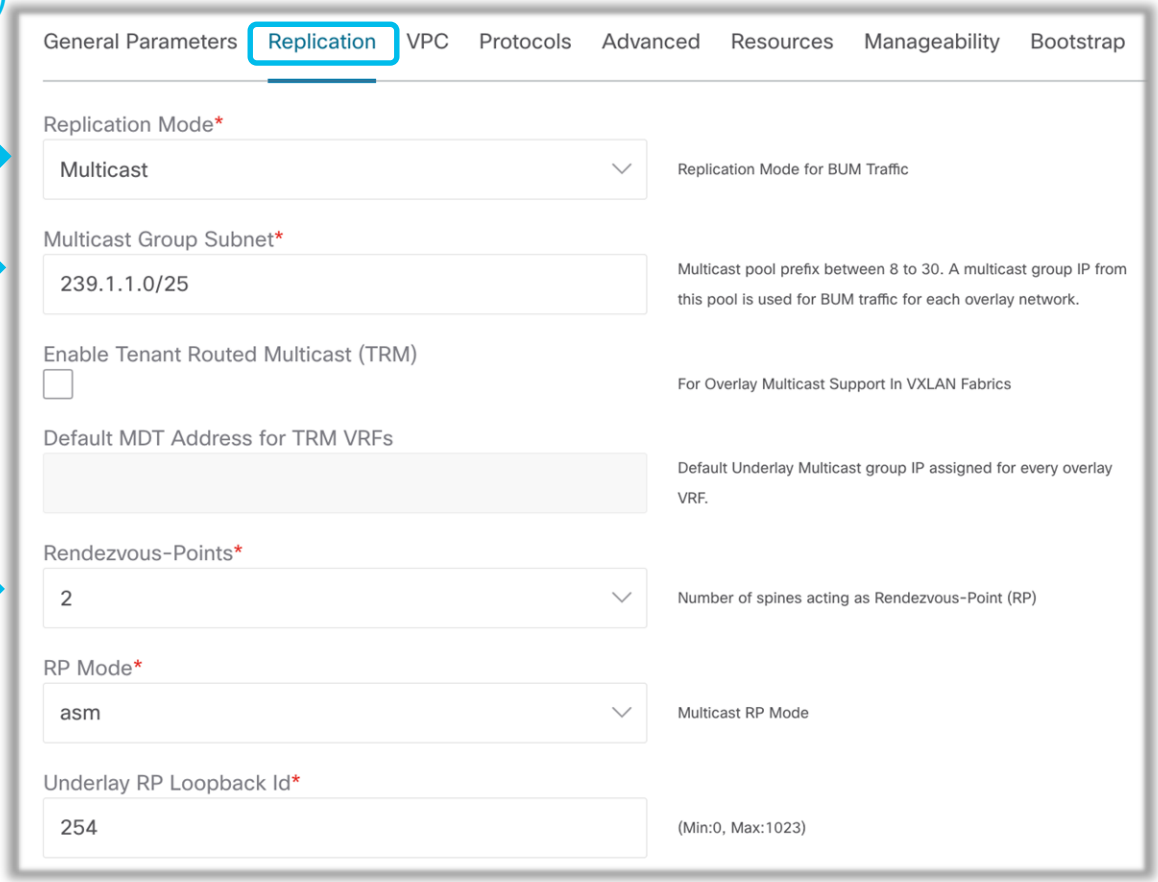


Step1 -> Create a Fabric
(continued)

BUM > Multicast or Ingress Replication

L2VNI Multicast Group

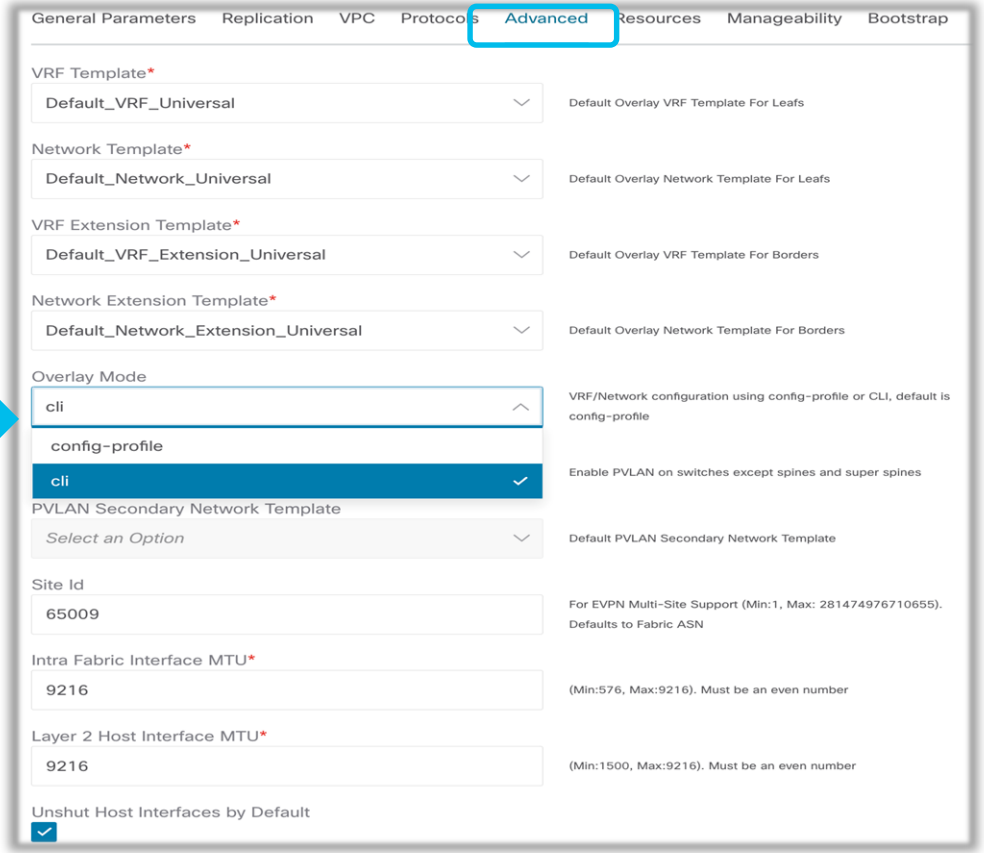
RP > 2 or 4



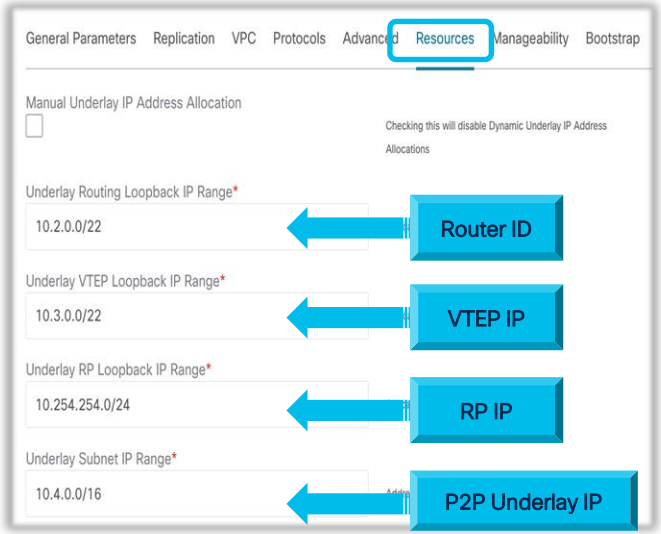
Step1 -> Create a Fabric
(continued)

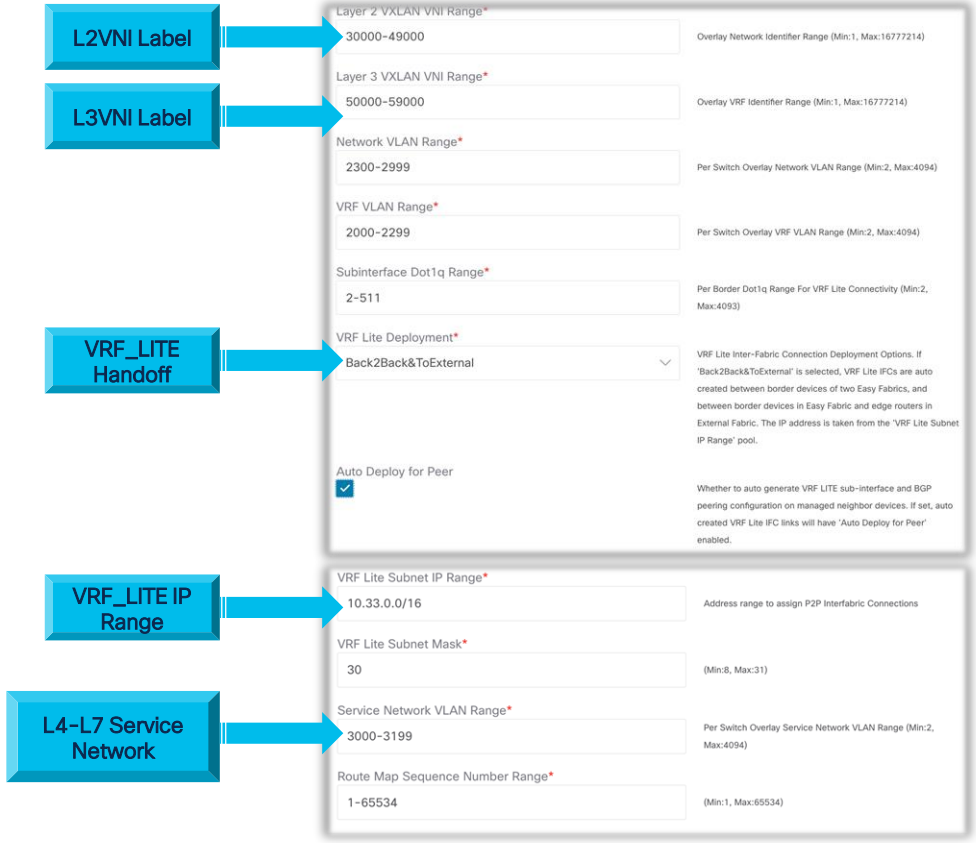
Cisco's Best Practice Configuration Templates

VXLAN Overlay Mode > CLI or Config-Profile



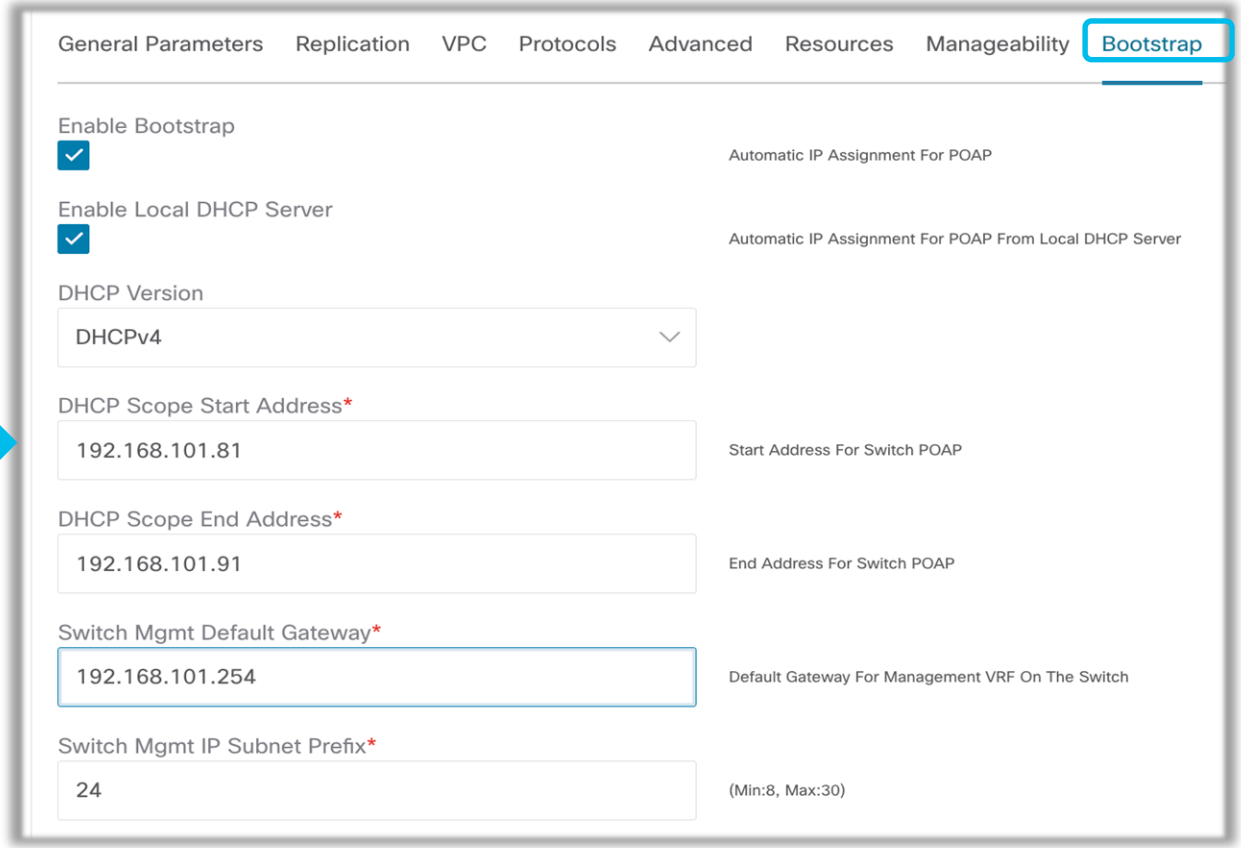
Step1 -> Create a Fabric
(continued)



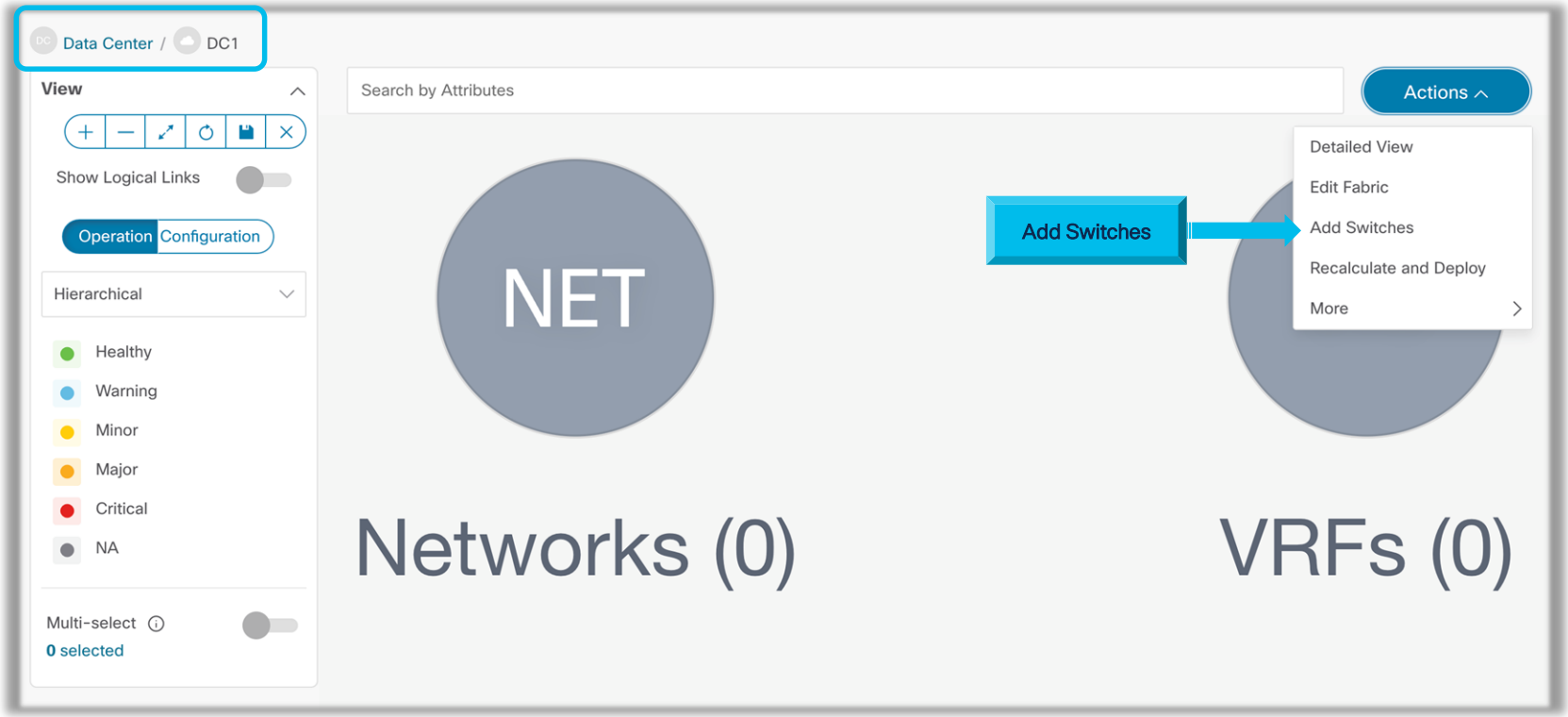


Step1 -> Create a Fabric
(continued)

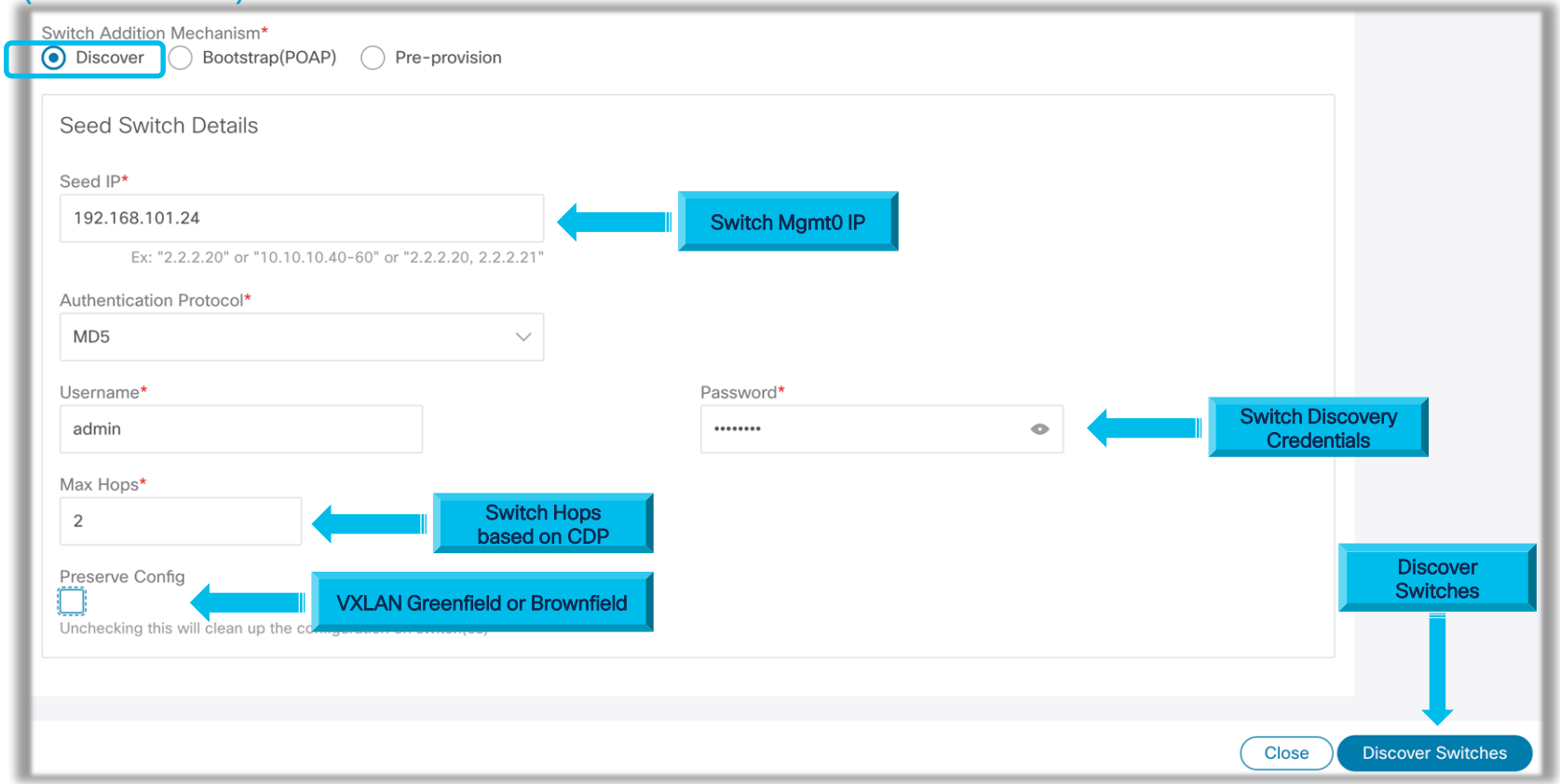
NDFC Built-In Bootstrap POAP Services. Supports OOB and In band POAP



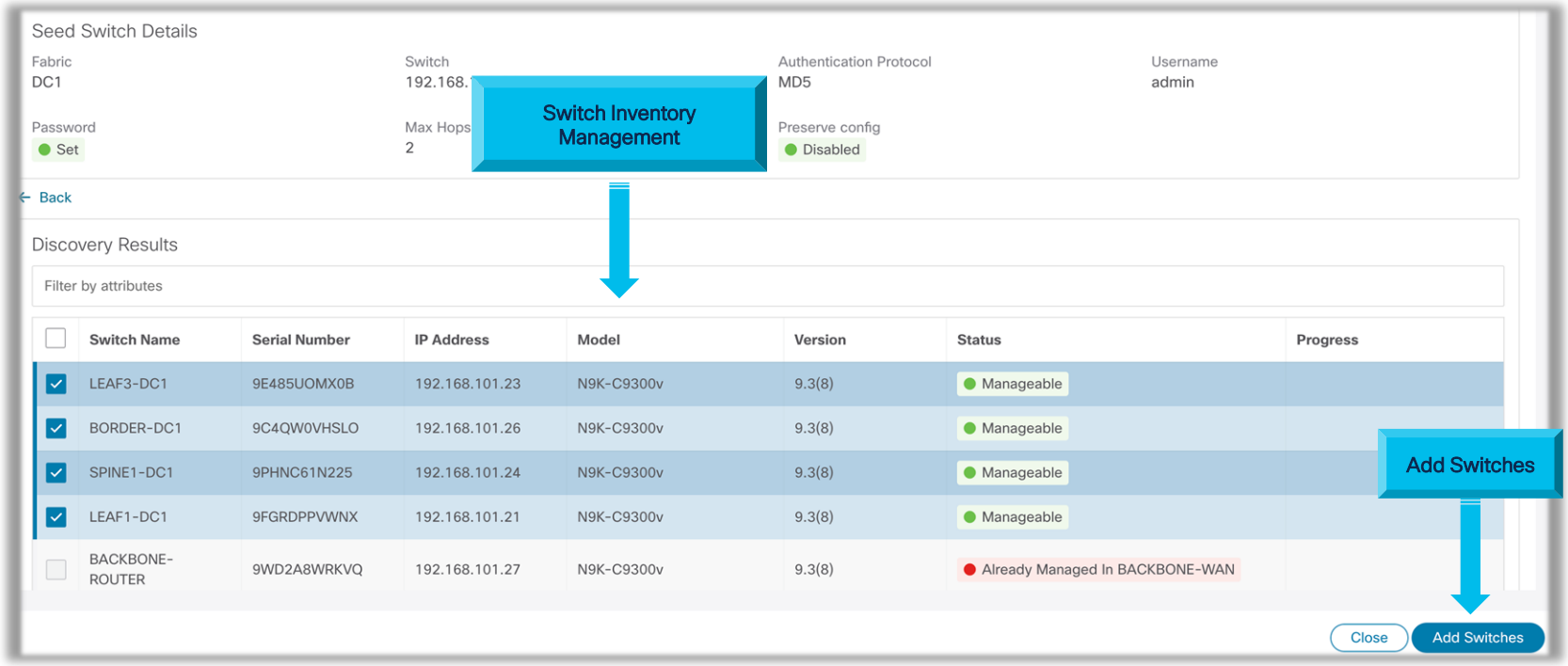
Step2 -> Add Switches



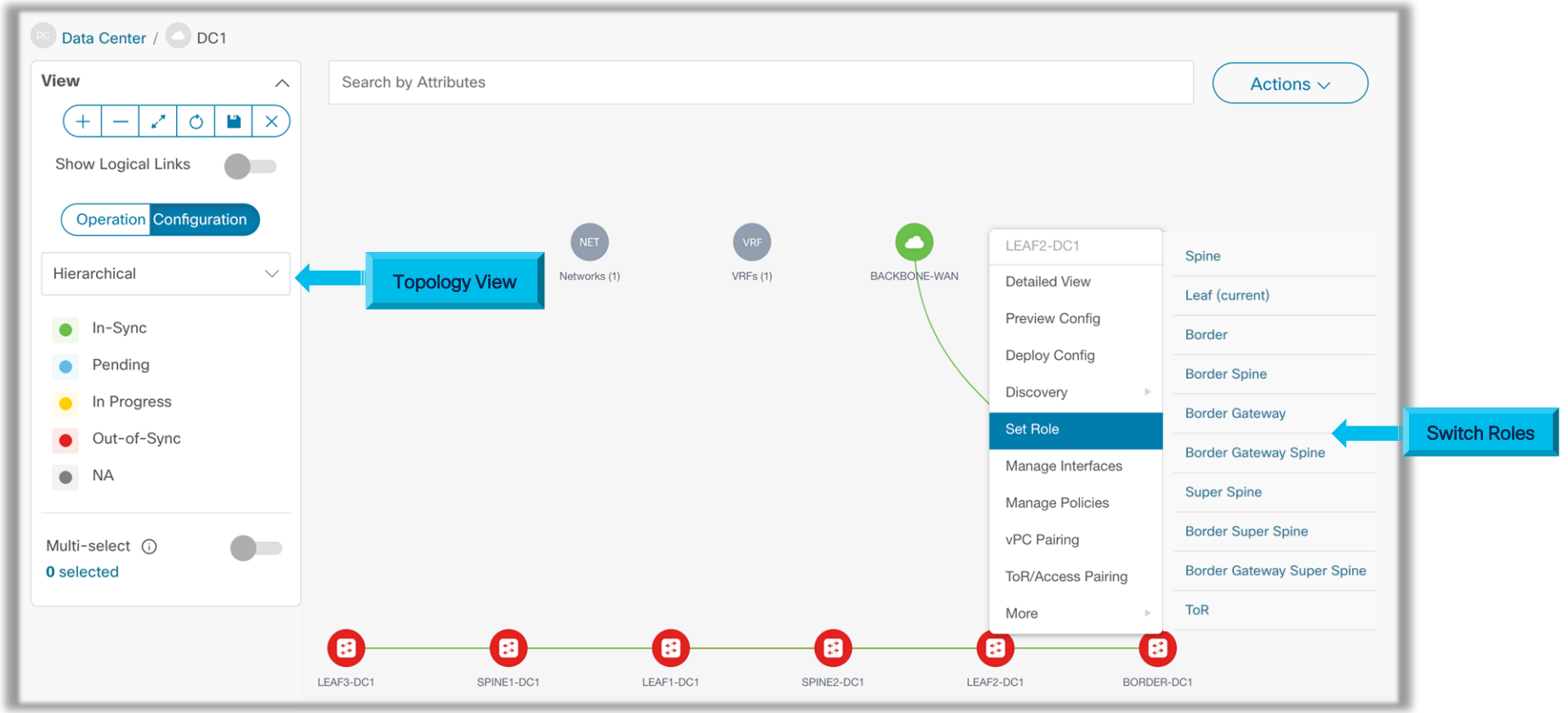
Step2 -> Add Switches
(continued)



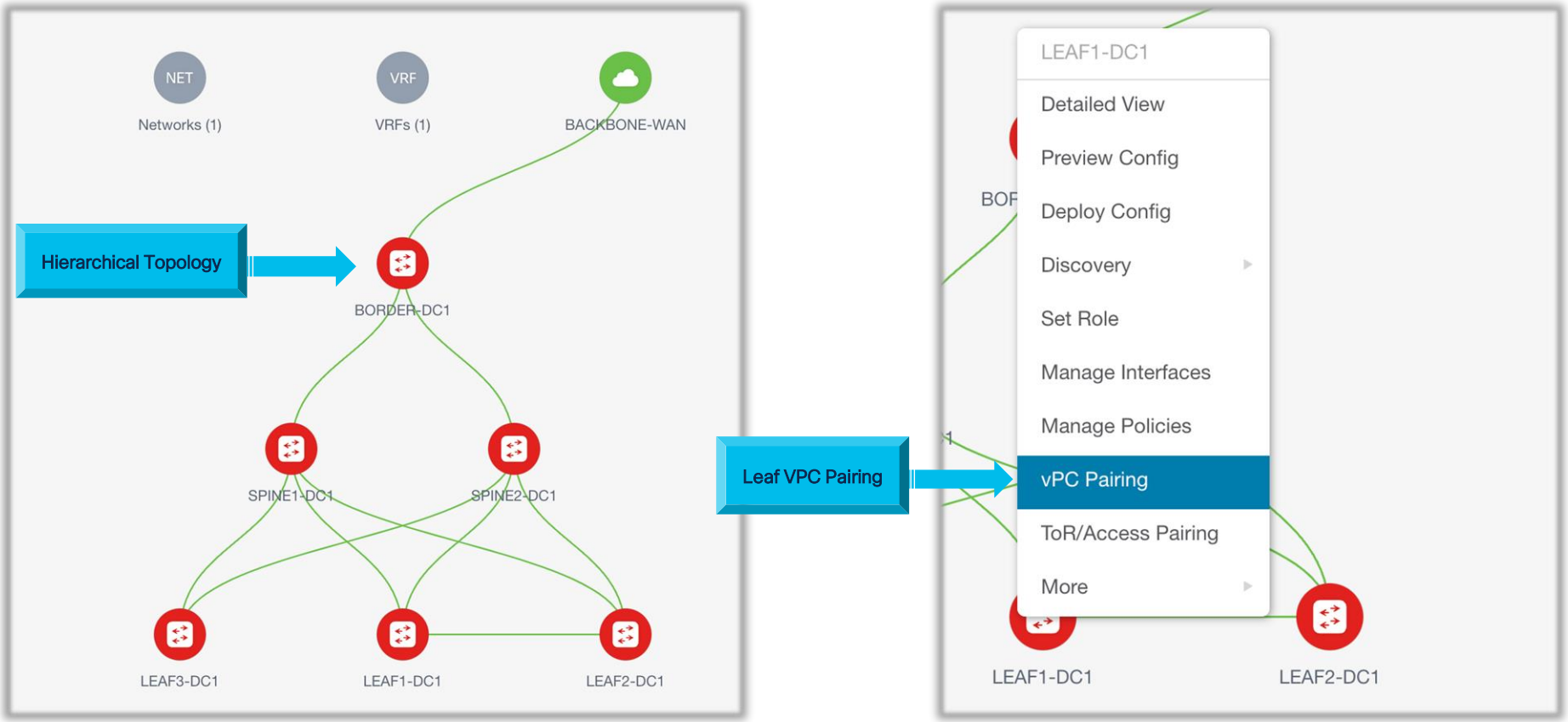
Step2 -> Add Switches
(continued)



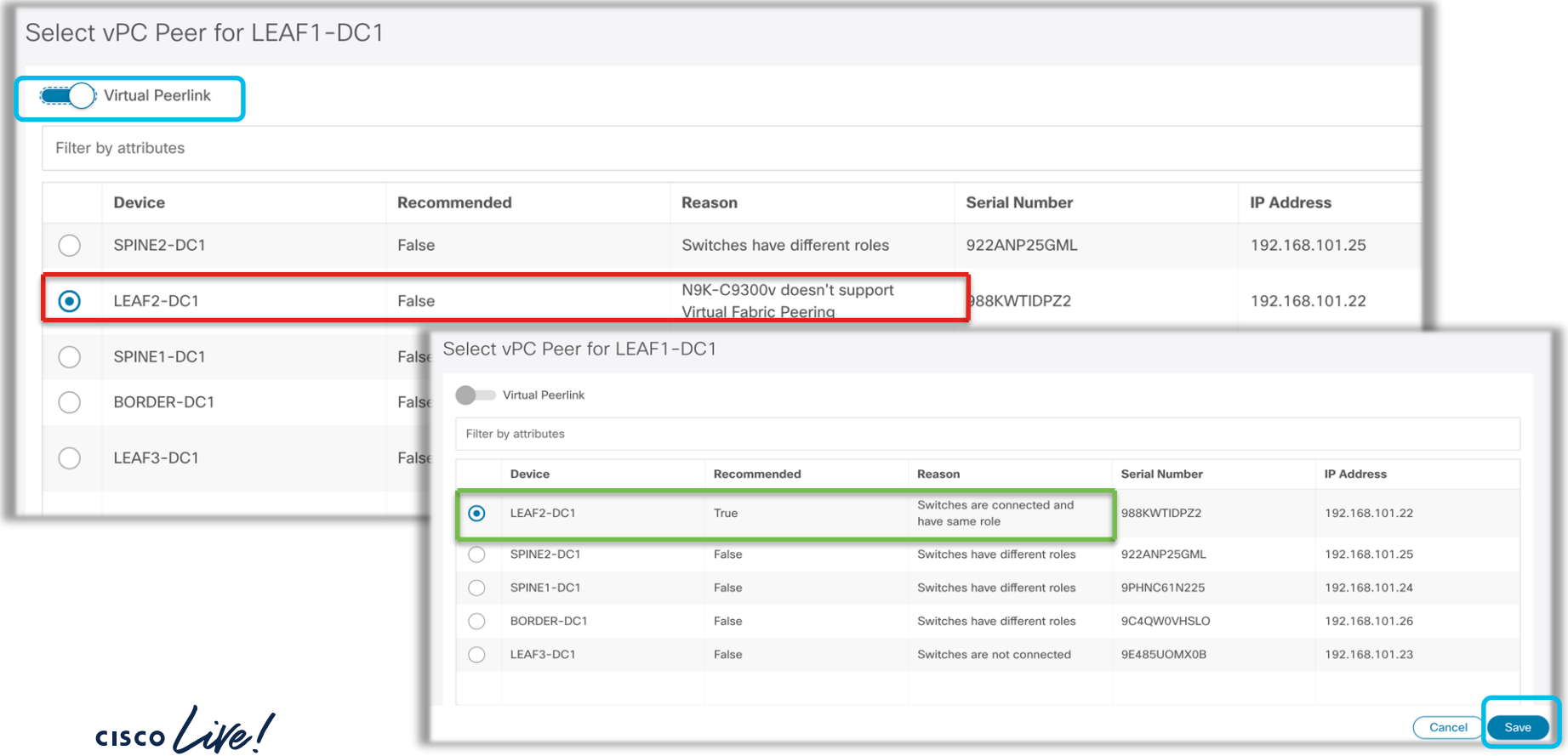
Step3 -> Set Role



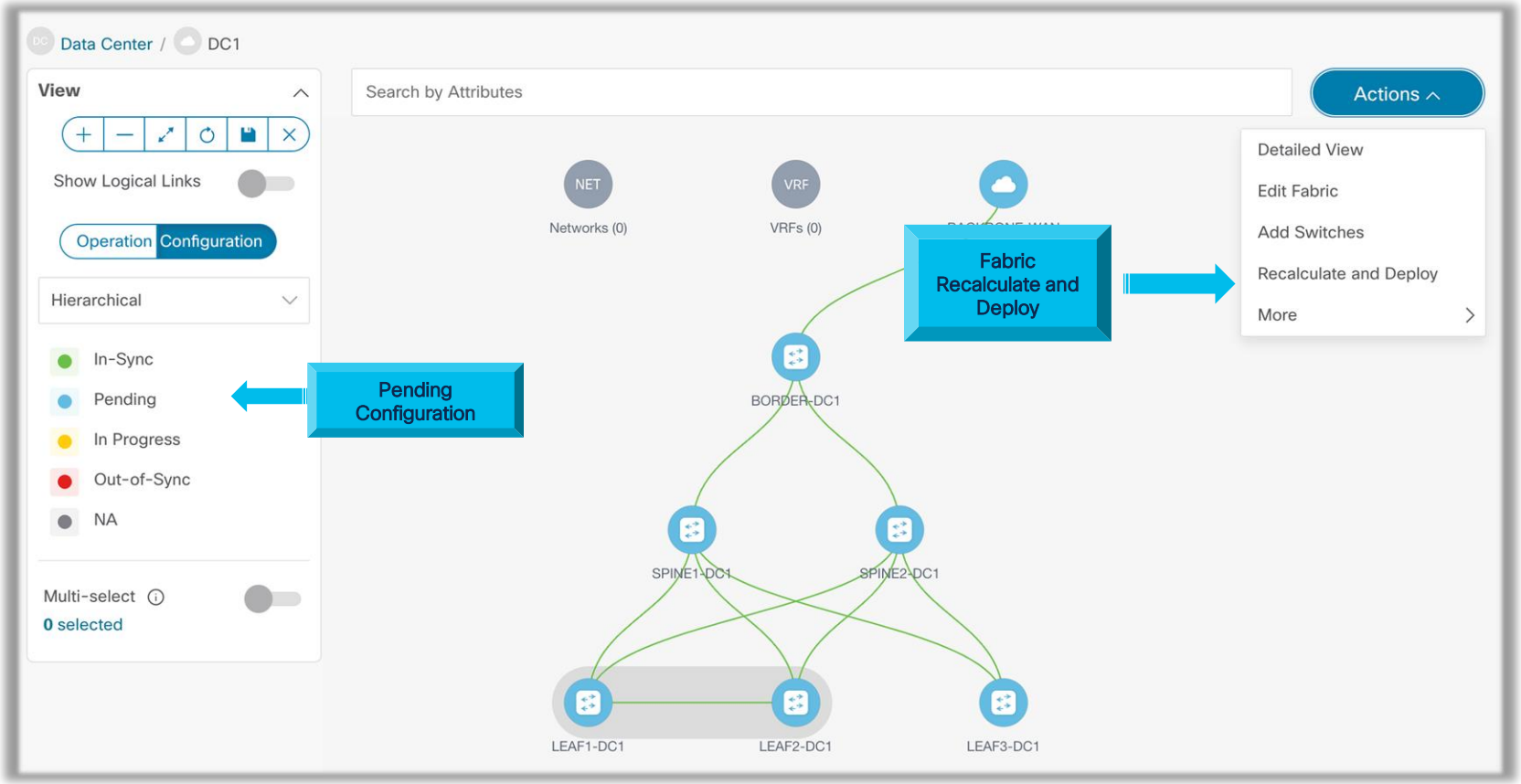
Step4 -> VPC Pairing
(optional)



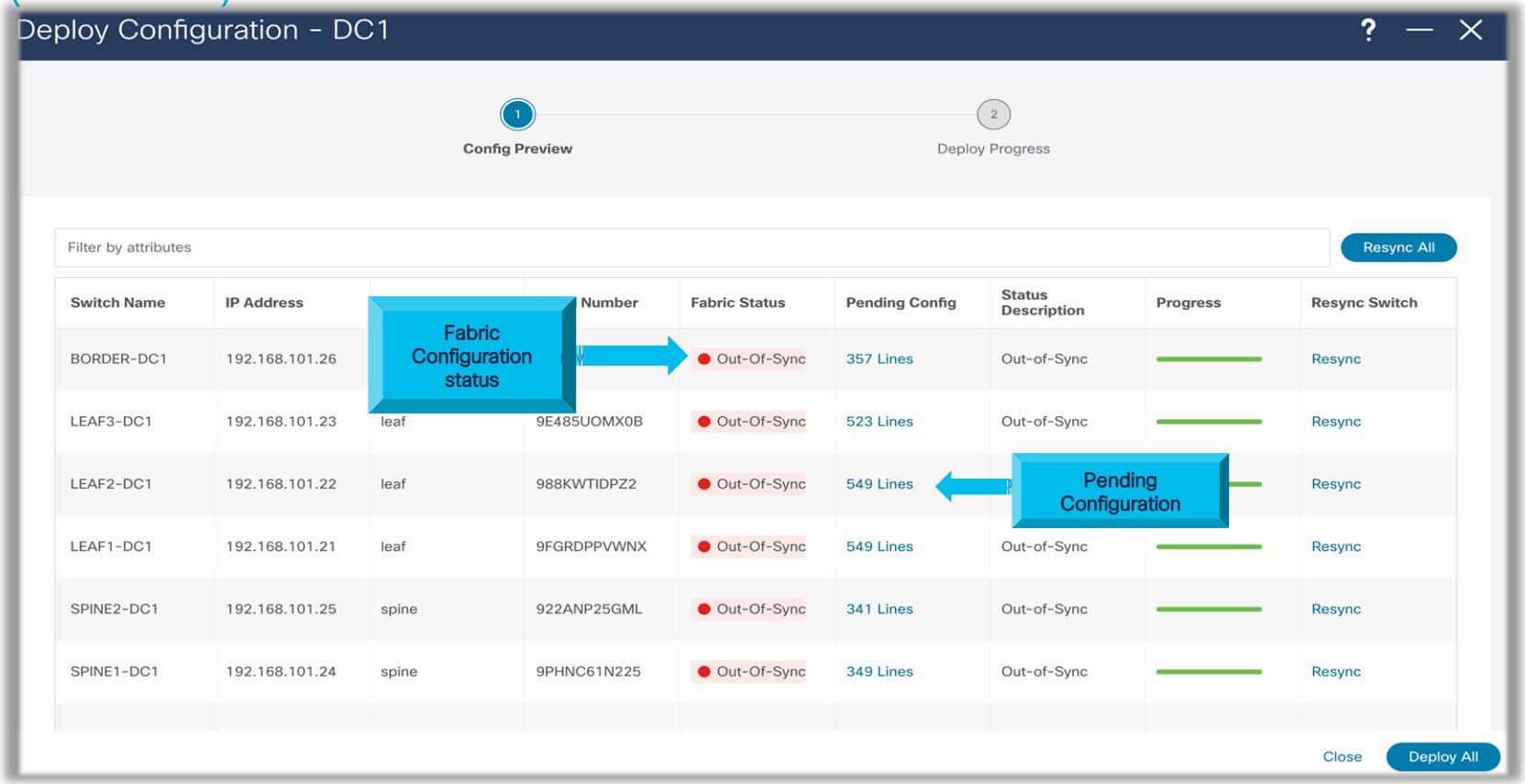
Step4 -> VPC Pairing
(optional)



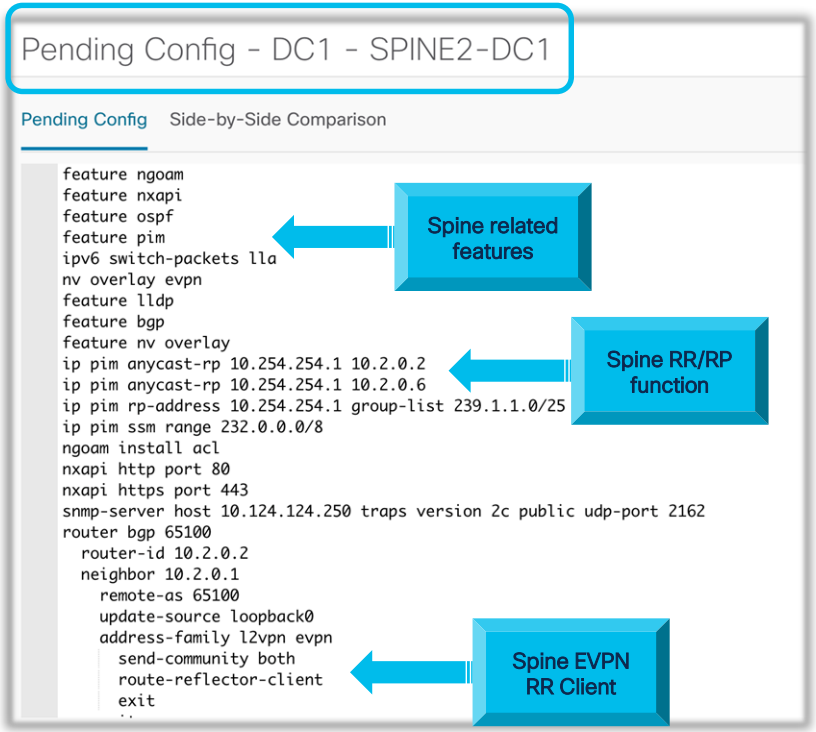
Step5 -> Recalculate and Deploy

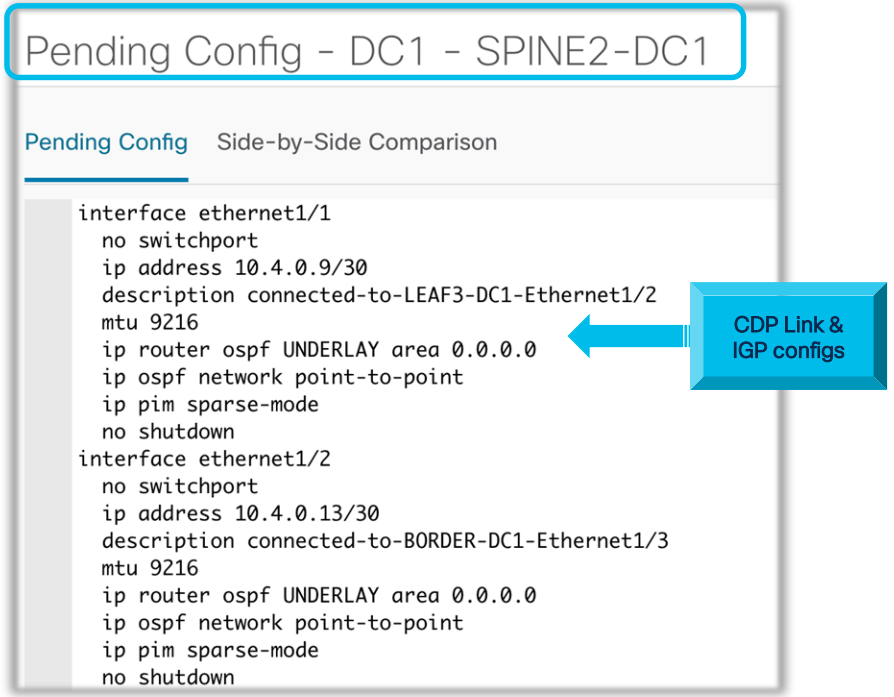


Step5 -> Recalculate and Deploy
(continued)

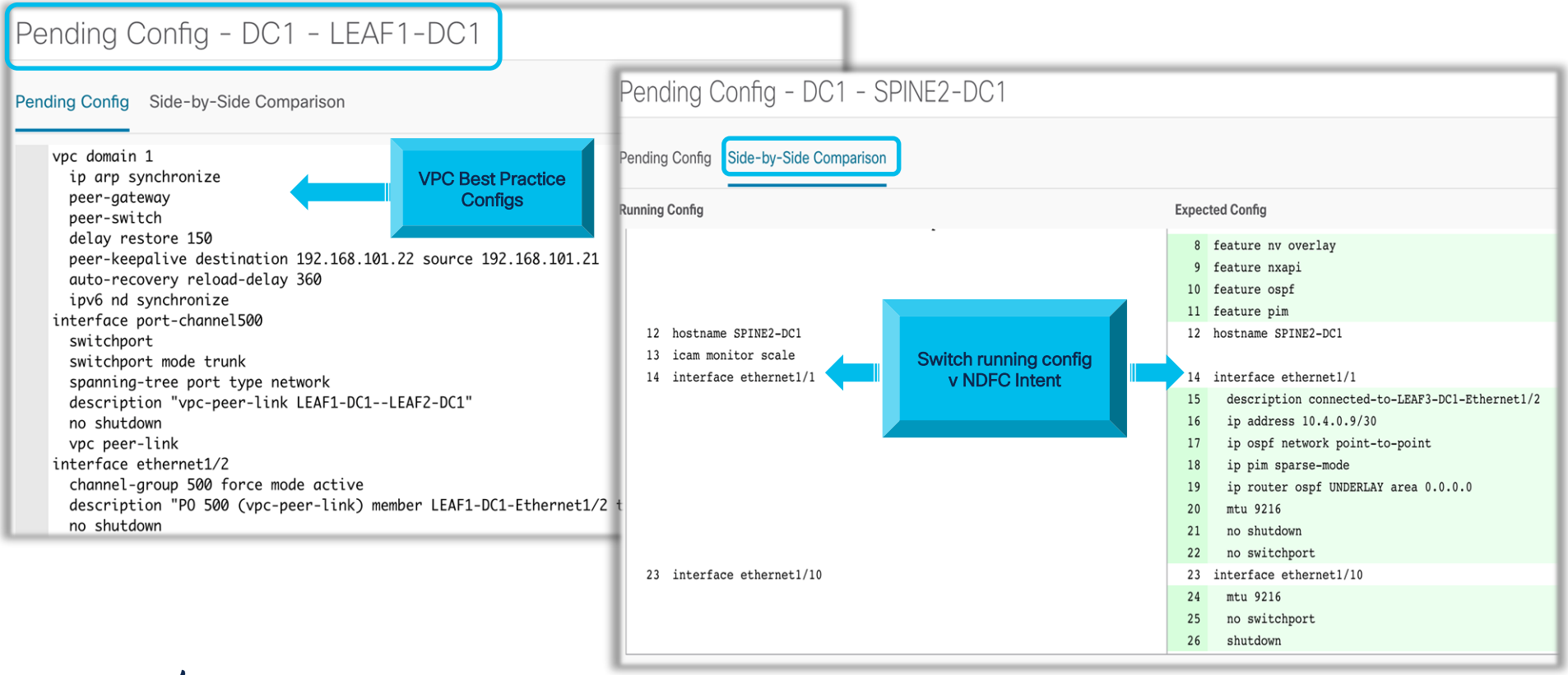


Step5 -> Recalculate and Deploy
(continued)

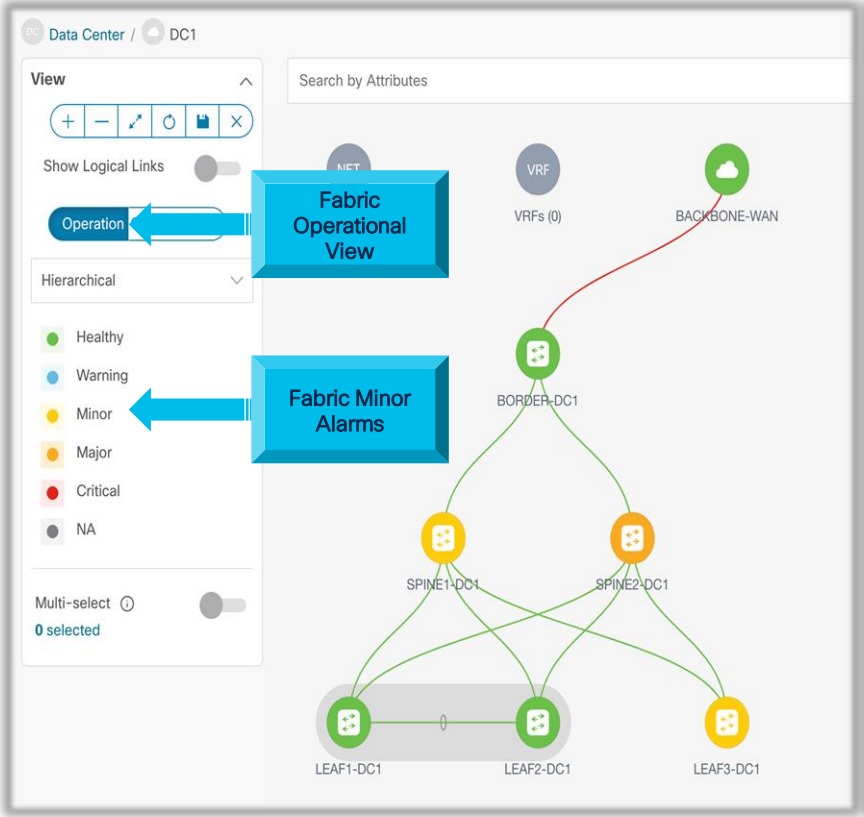


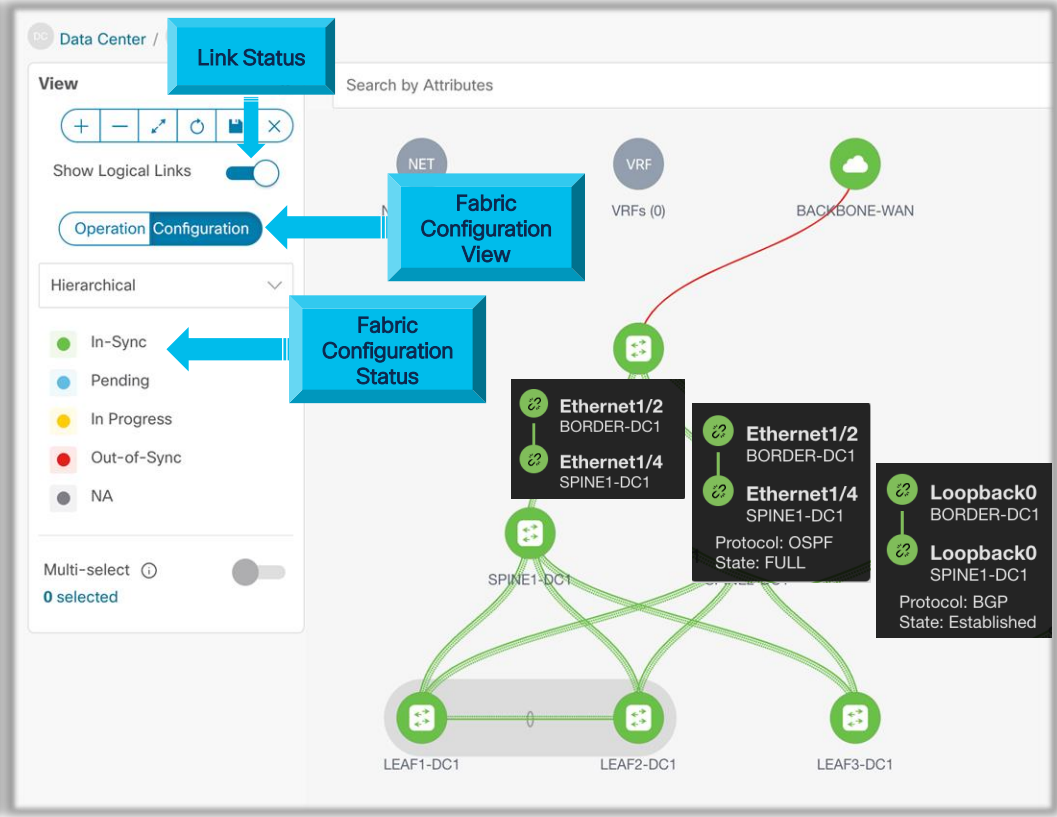


Step5 -> Recalculate and Deploy
(continued)

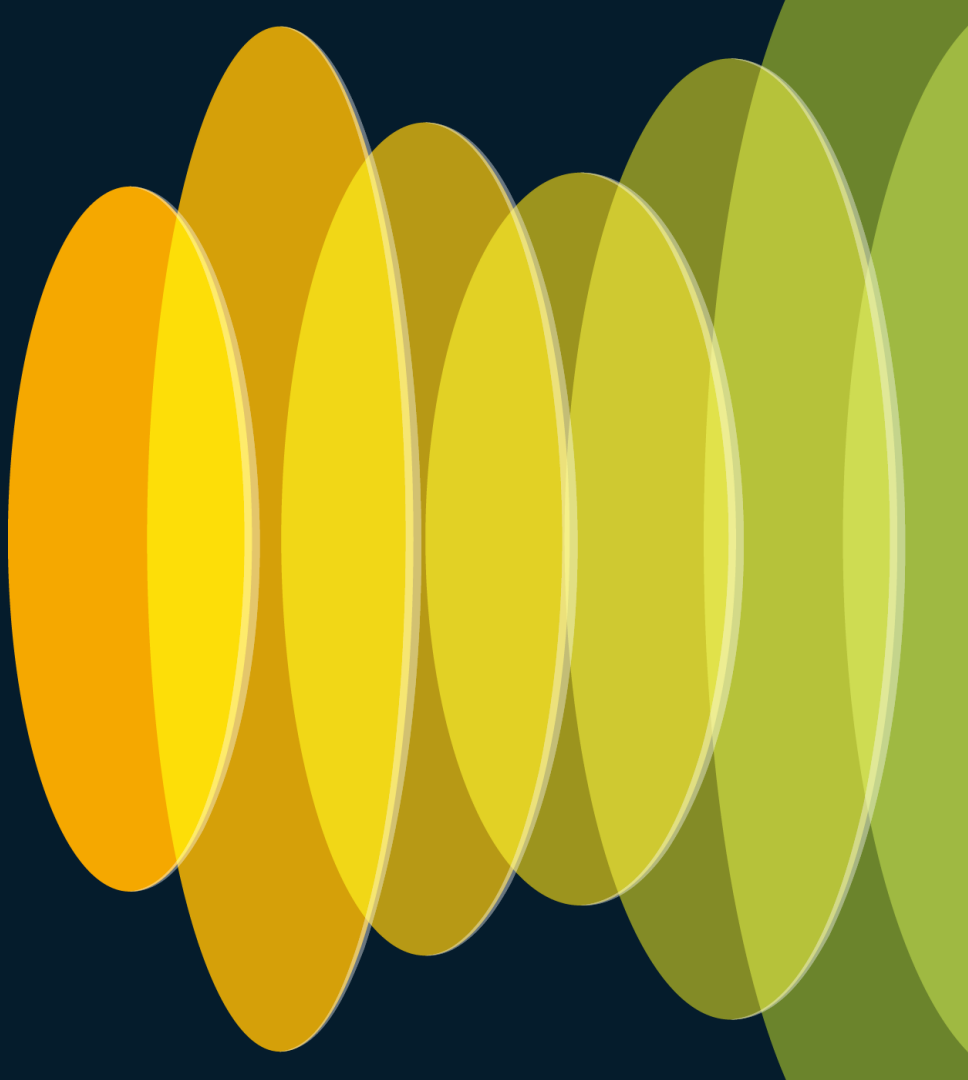


NDFC VXLAN EVPN Topology View





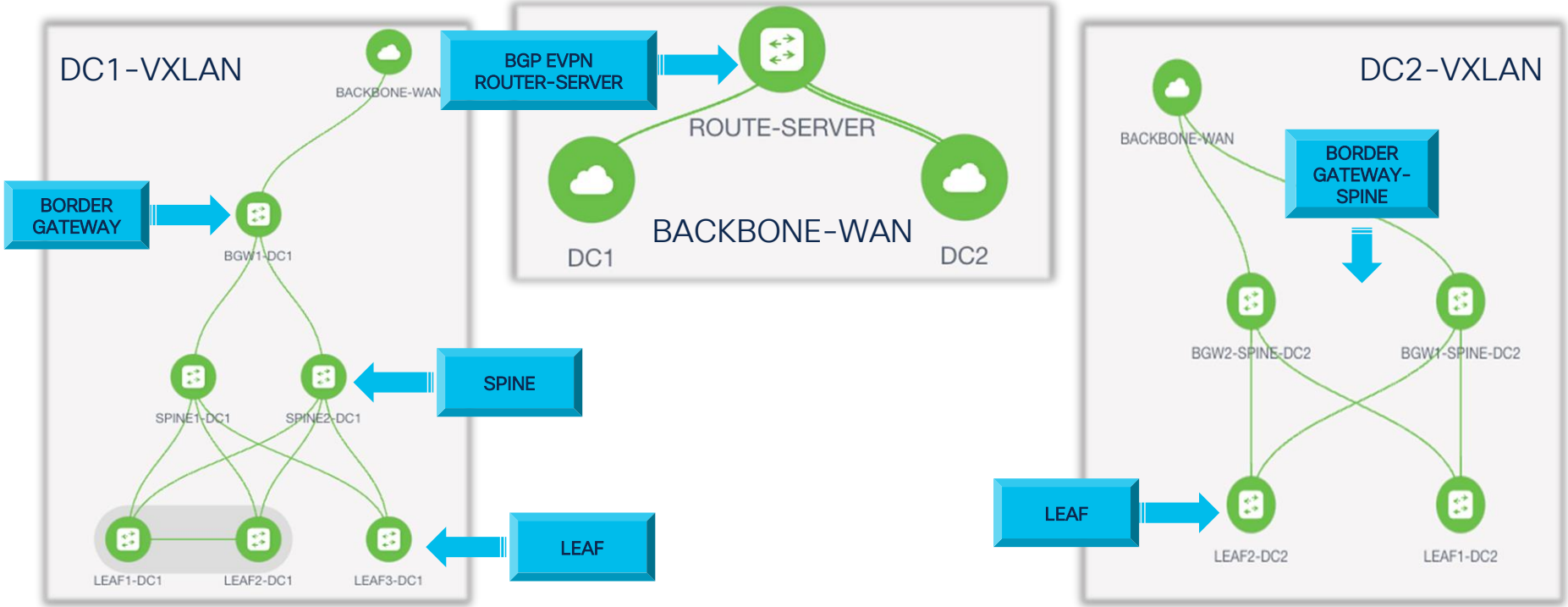
NDFC Day-0:
Multi-Site DCI



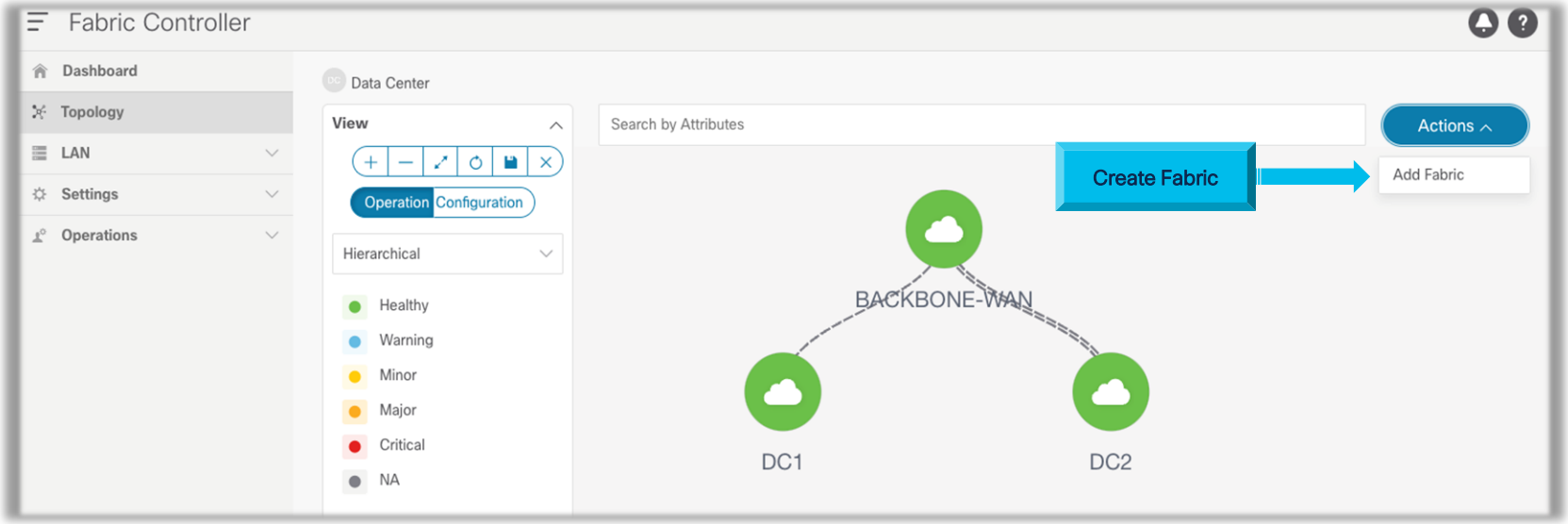
cisco Live!

Step 1 & 2 -> Create Individual Fabrics

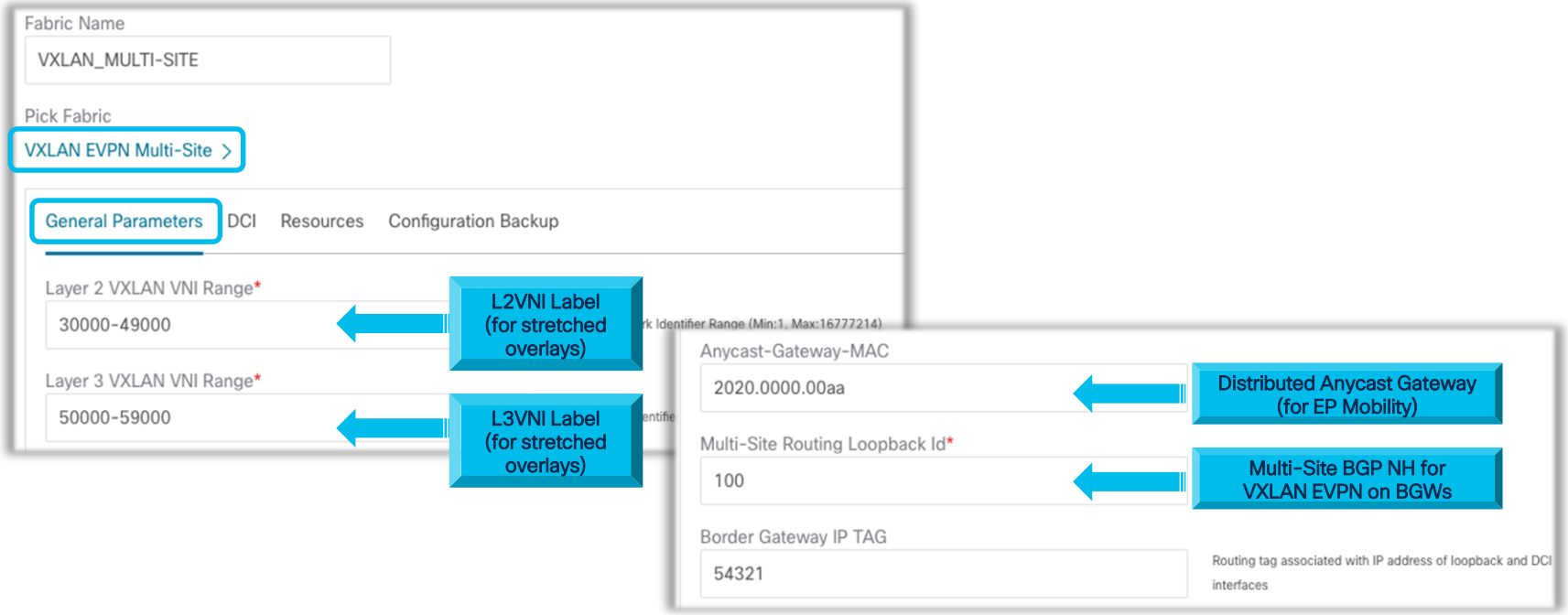
Set appropriate roles



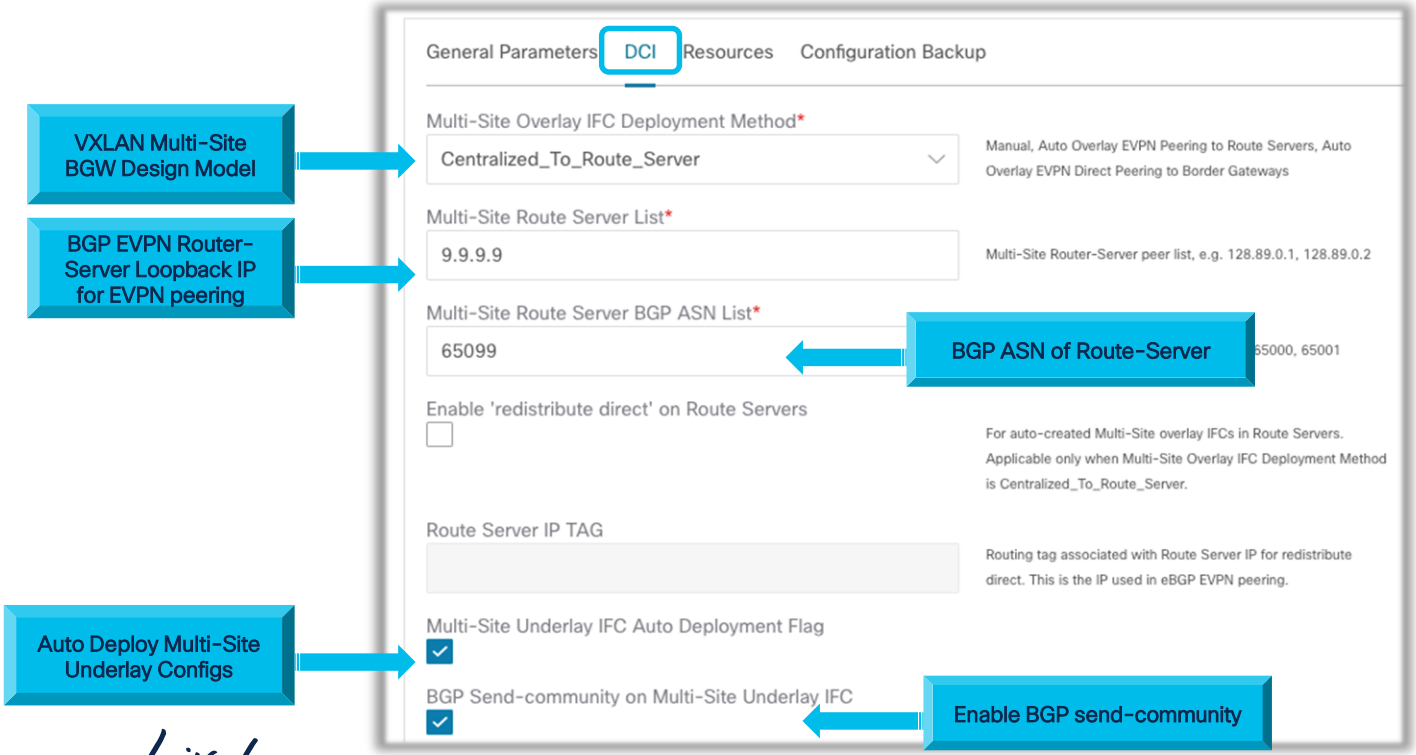
Step 3 -> Create VXLAN EVPN Multi-Site Fabric



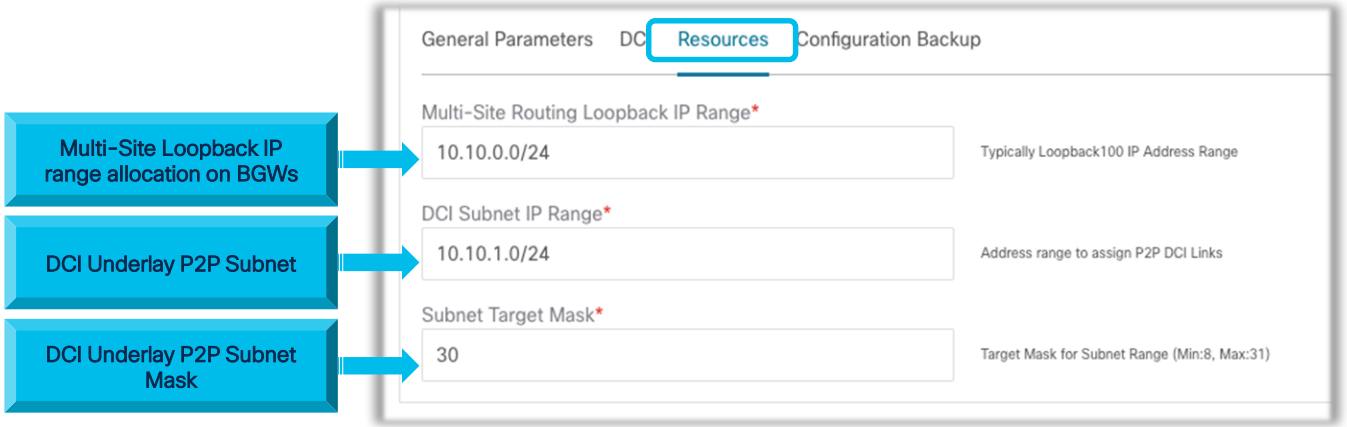
Step 3 -> Create VXLAN EVPN Multi-Site Fabric (continued)



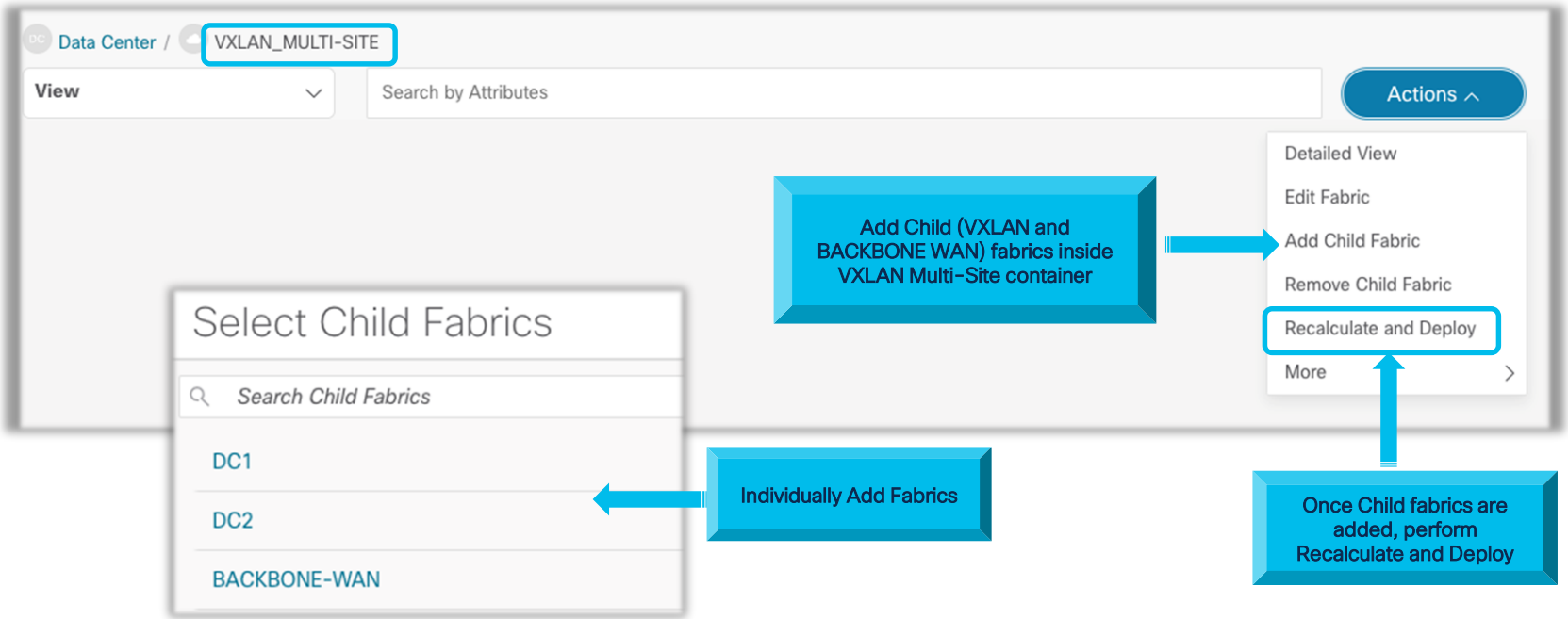
Step 3 -> Create VXLAN EVPN Multi-Site Fabric (continued)



Step 3 -> Create VXLAN EVPN Multi-Site Fabric
(continued)

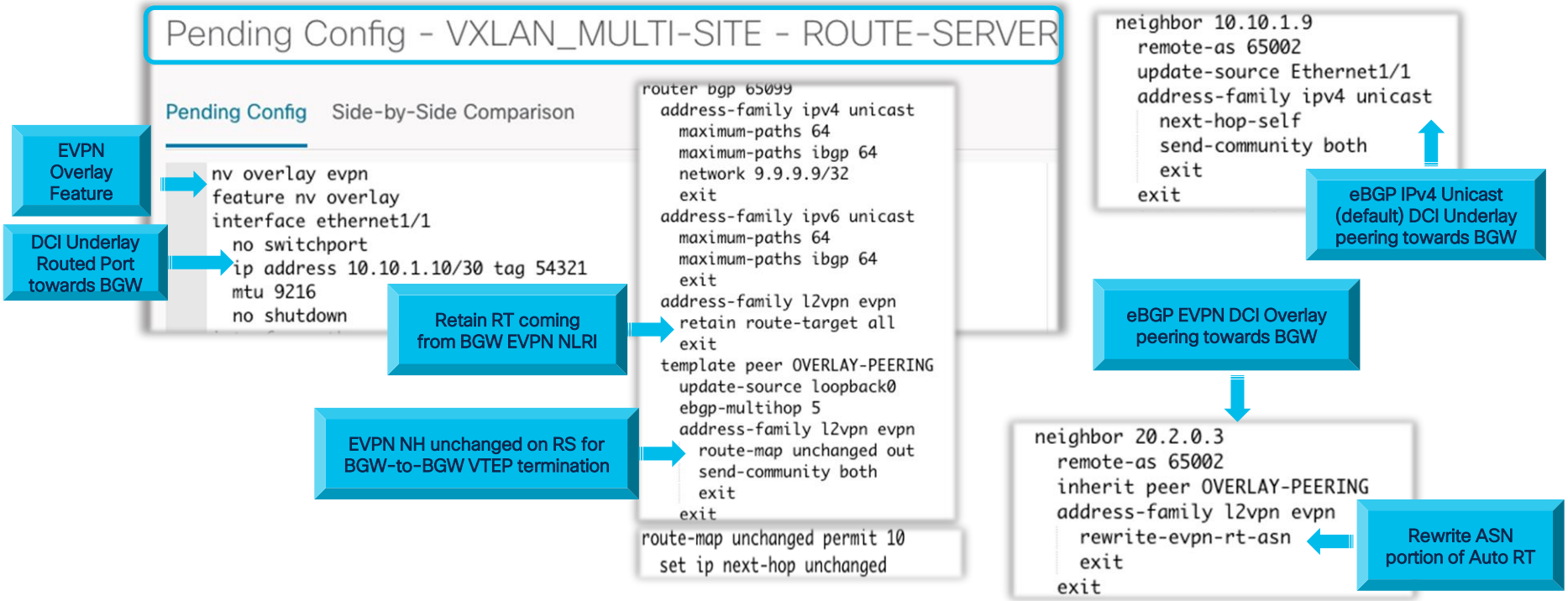


Step 3 -> Create VXLAN EVPN Multi-Site Fabric
(continued)



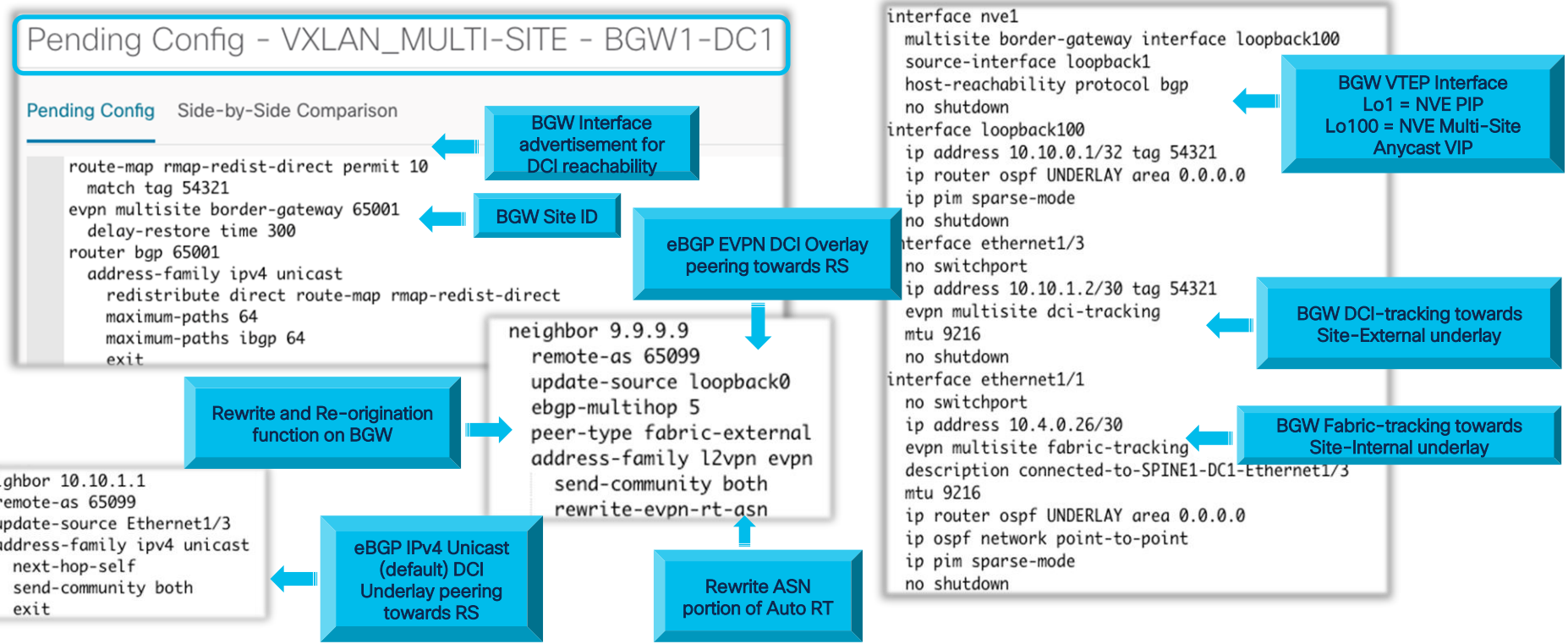
Step 4 -> Recalculate & Deploy

In VXLAN EVPN Multi-Site Fabric



Step 4 -> Recalculate & Deploy

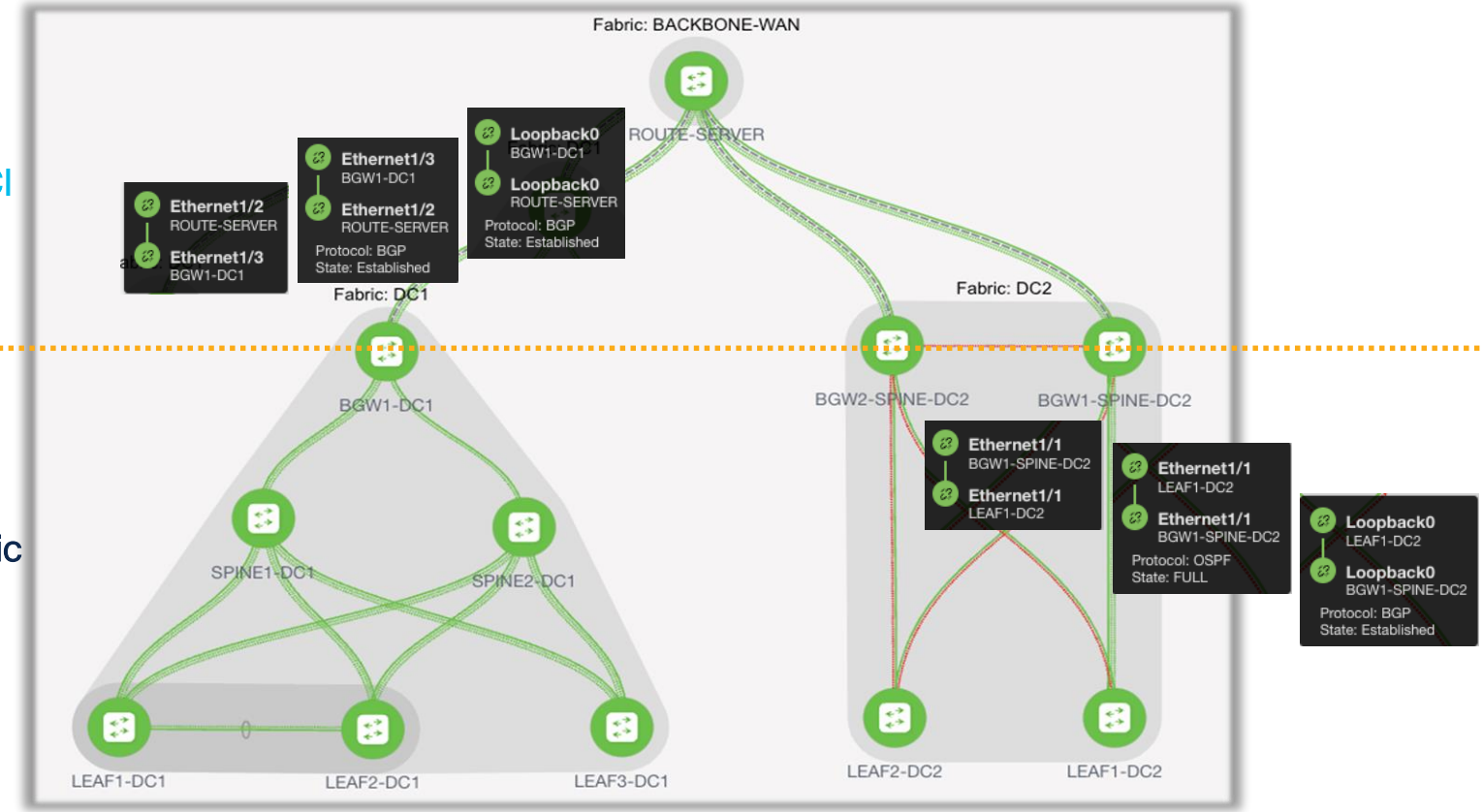
In VXLAN EVPN Multi-Site Fabric



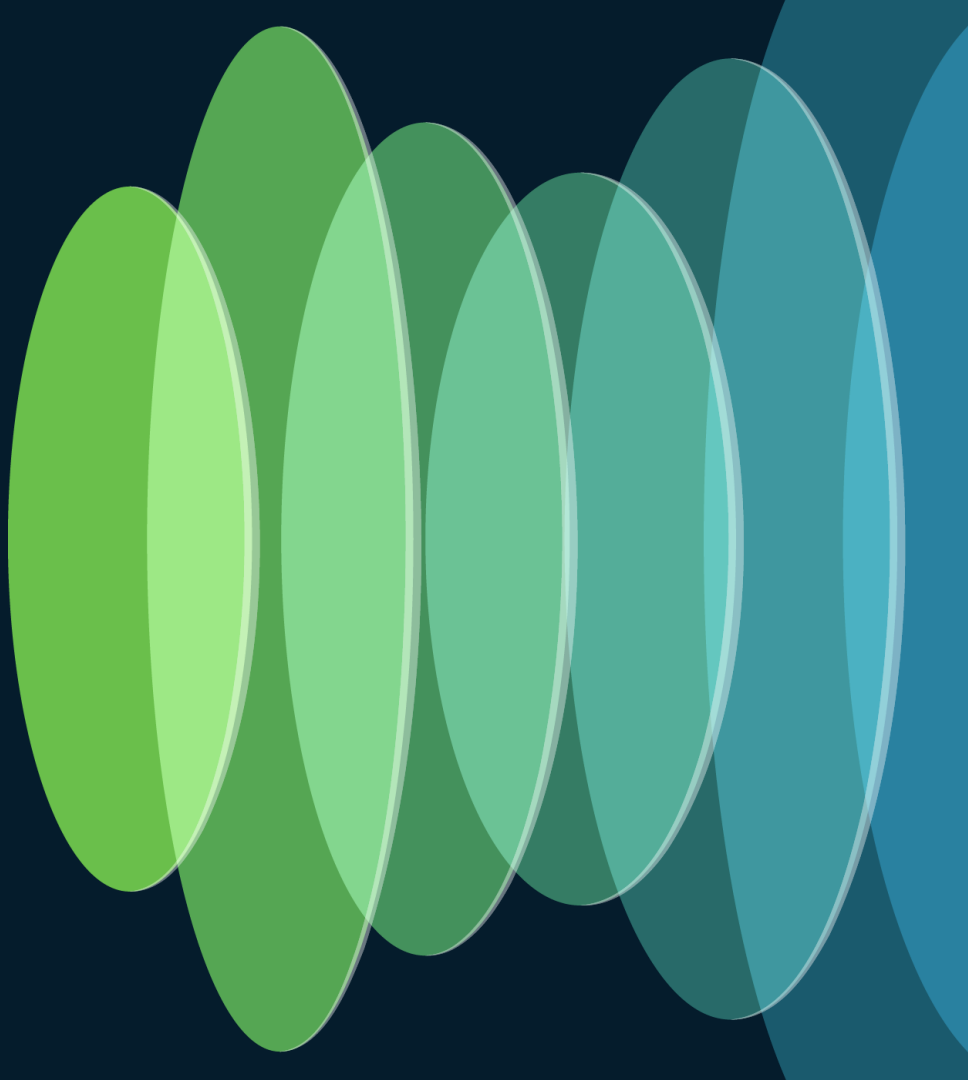
NDFC VXLAN EVPN Multi-Site Topology View

Site-External DCI

Site-Internal Fabric



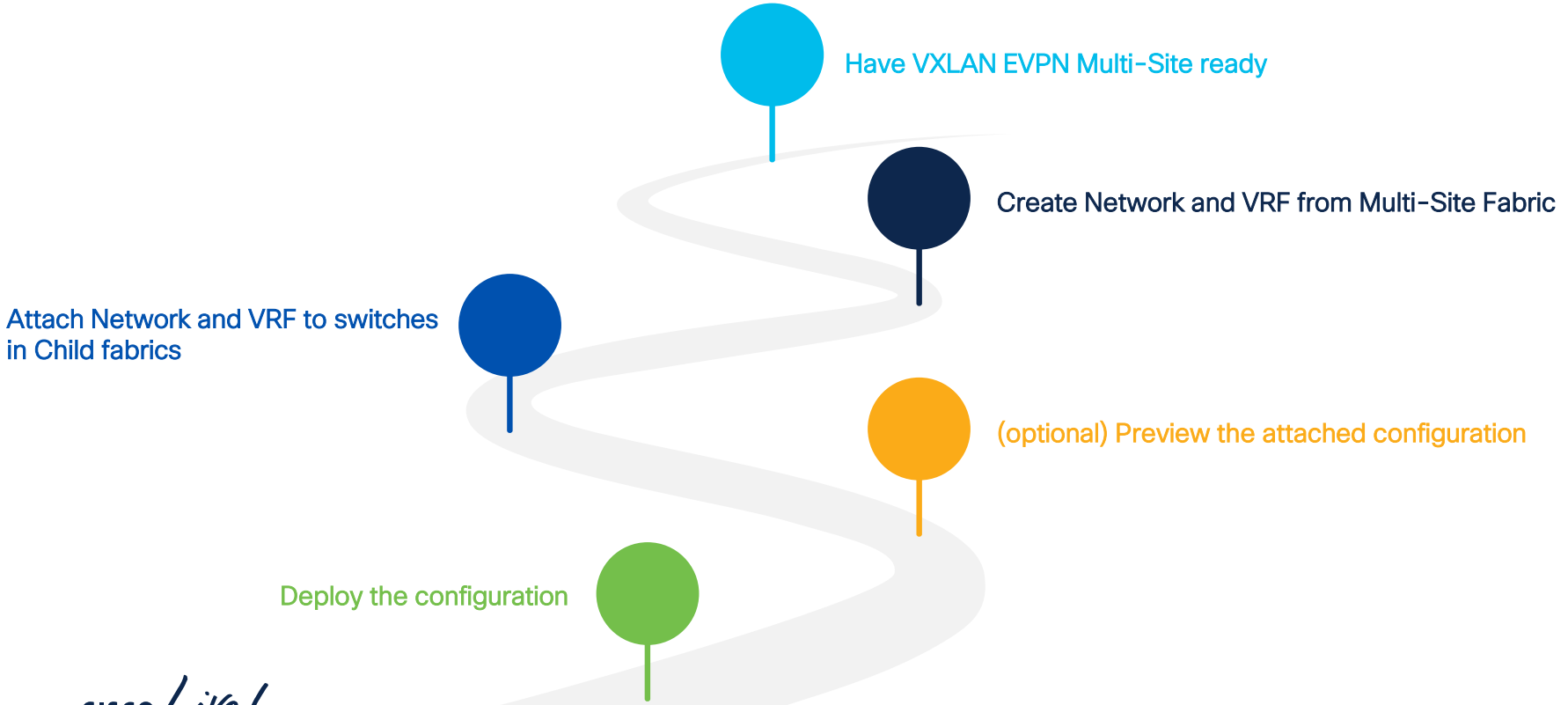
NDFC Day-1: VXLAN EVPN Overlay



cisco Live!

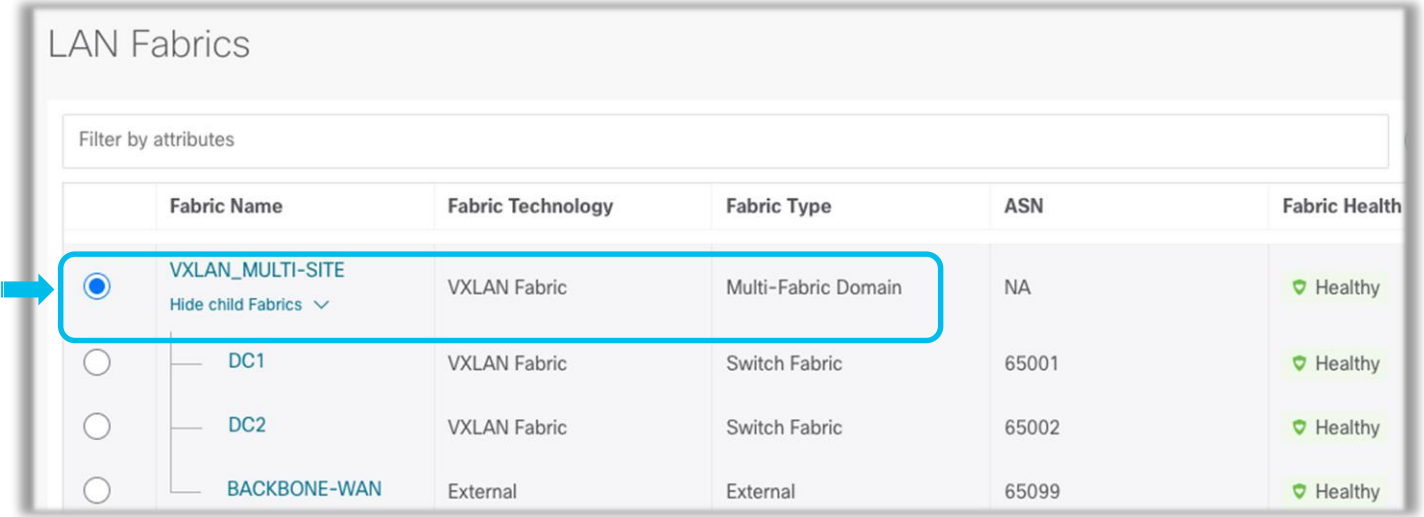
Deploy VXLAN EVPN Overlays

in Multi-Site Architecture



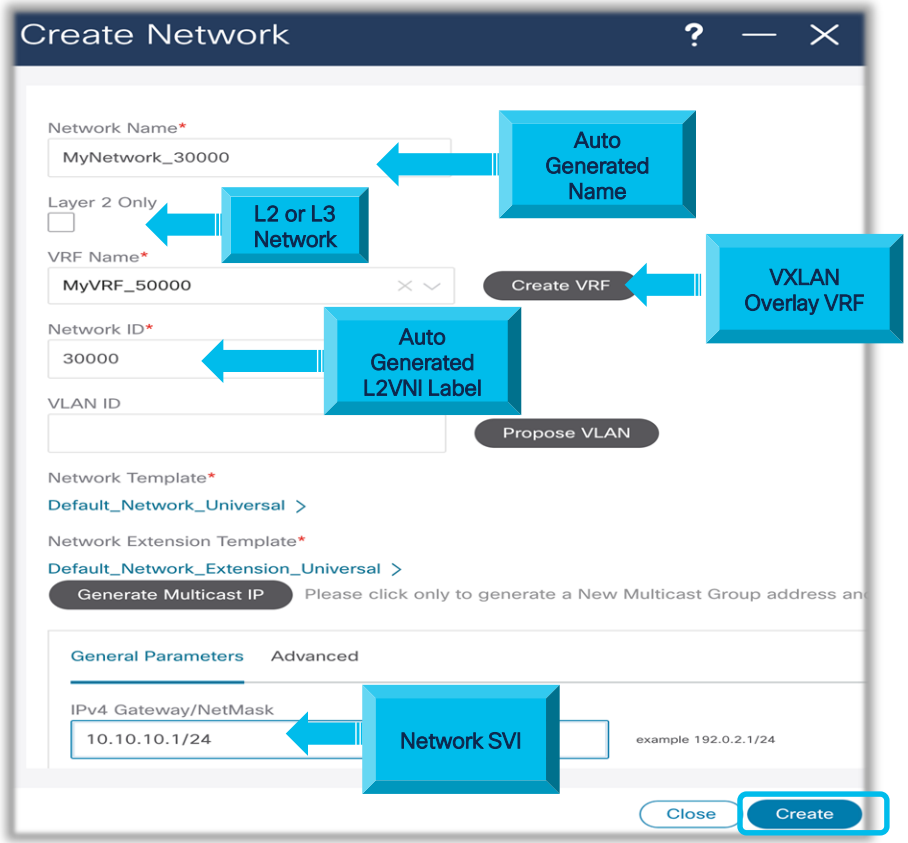
Step 1 -> Navigate to VXLAN Multi-Site Fabric

Select Multi-Site Fabric

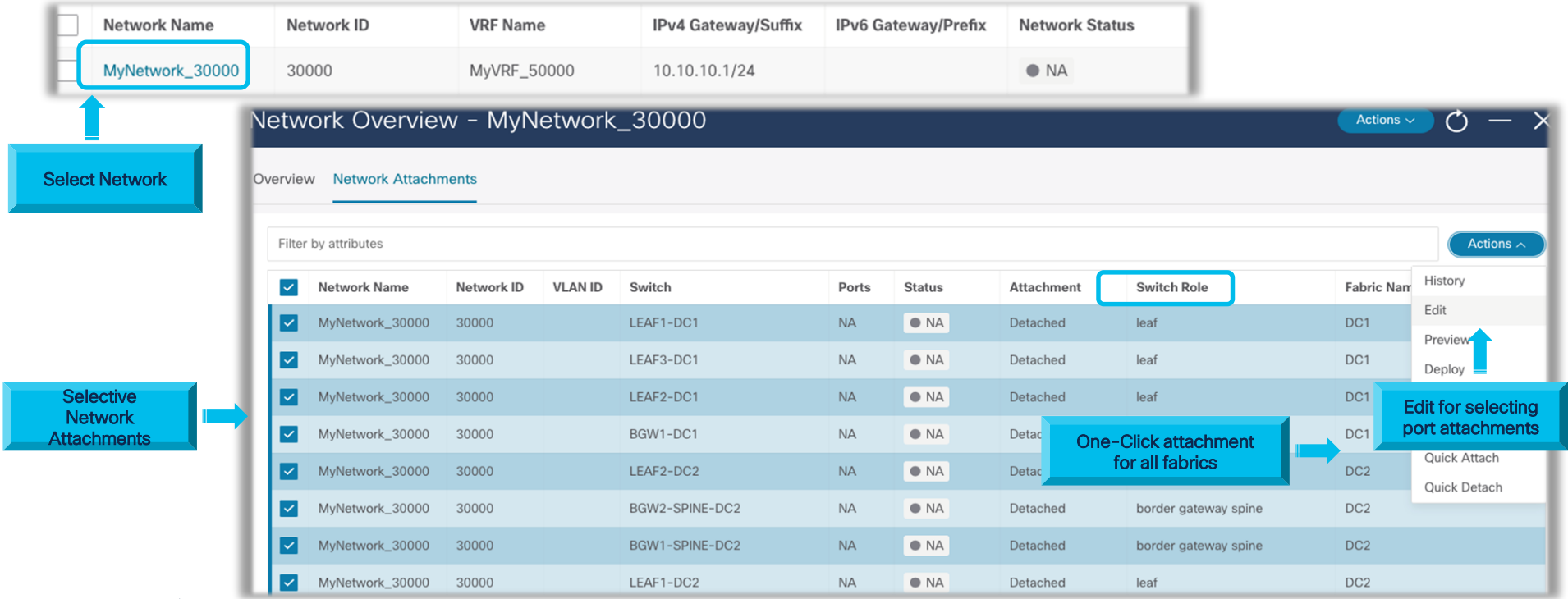


*UI Navigation: LAN Fabrics > Select Fabric > Fabric Overview > Networks > Create

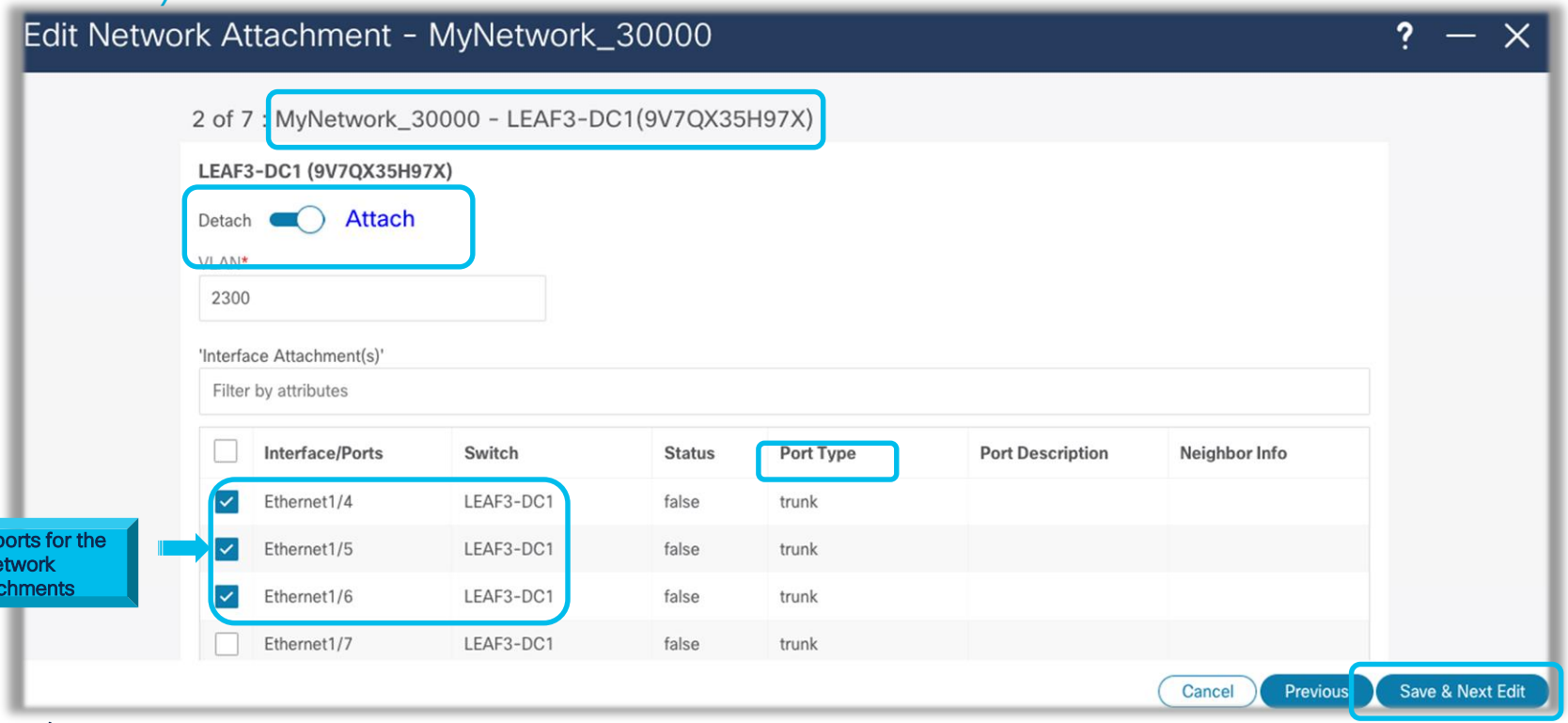
Step 2 -> Create VRF and Network



Step 3 -> Attach VRF and Network

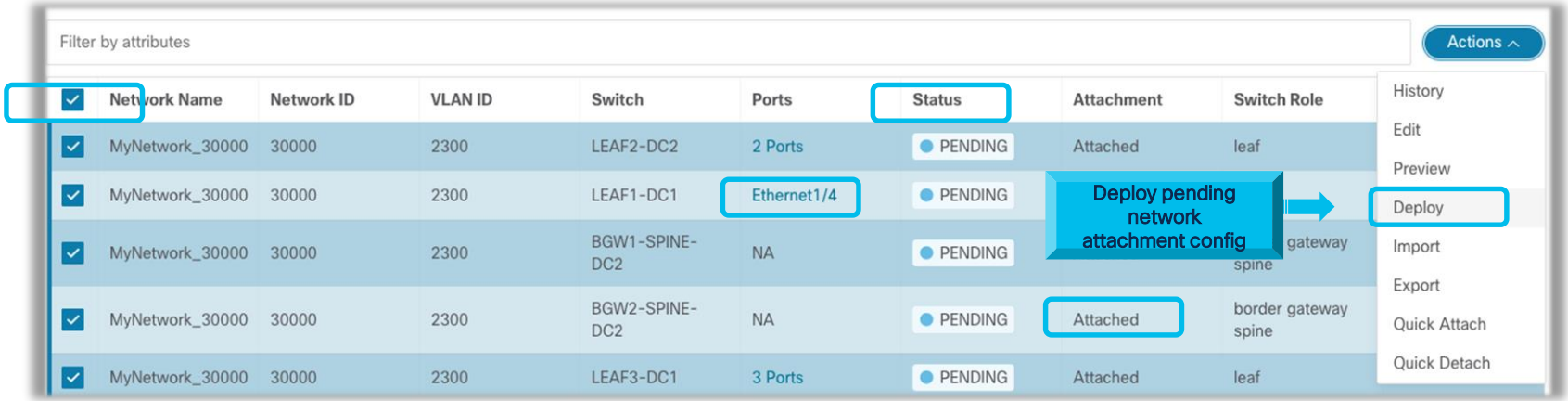


Step 3 -> Attach VRF and Network
(continued)



Select ports for the network attachments

Step 4 -> Preview and Deploy VRF and Network



VRF and Network Configs

Pending Config - VXLAN_MULTI-SITE - BGW1-DC1

L3VNI VRF Configs

```
vlan 2000
  vn-segment 50000
vrf context myvrf_50000
  vni 50000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn
exit
interface Vlan2000
  vrf member myvrf_50000
  ip forward
  ipv6 address use-link-local-only
  no ip redirects
  no ipv6 redirects
  mtu 9216
  no shutdown
```

VXLAN Tenant

```
router bgp 65001
  vrf myvrf_50000
    address-family ipv4 unicast
      advertise l2vpn evpn
      redistribute direct route-map fabric-rmap-redist-subnet
      maximum-paths ibgp 2
      exit
    address-family ipv6 unicast
      advertise l2vpn evpn
      redistribute direct route-map fabric-rmap-redist-subnet
      maximum-paths ibgp 2
configure terminal
interface nve1
  member vni 30000
    multisite ingress-replication
    mcast-group 239.1.1.1
  member vni 50000 associate-vrf
```

NVE Tunnel Configs

L2VNI Network Configs

```
vlan 2300
  vn-segment 30000
evpn
  vni 30000 l2
    rd auto
    route-target import auto
    route-target export auto
```

VRF and Network Configs

Pending Config - VXLAN_MULTI-SITE - LEAF3-DC1

```
vlan 2000
  vn-segment 50000
vrf context myvrf_50000
  vni 50000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn
exit
interface Vlan2000
  vrf member myvrf_50000
  ip forward
  ipv6 address use-link-local-only
  no ip redirects
  no ipv6 redirects
  mtu 9216
  no shutdown
```

L3VNI VRF Configs

```
router bgp 65001
  vrf myvrf_50000
    address-family ipv4 unicast
      advertise l2vpn evpn
      redistribute direct route-map fabric-rmap-redist-subnet
      maximum-paths ibgp 2
      exit
    address-family ipv6 unicast
      advertise l2vpn evpn
      redistribute direct route-map fabric-rmap-redist-subnet
      maximum-paths ibgp 2
configure terminal
interface nve1
  member vni 30000
    mcast-group 239.1.1.1
  member vni 50000 associate-vrf
```

VXLAN Tenant

NVE Tunnel Configs

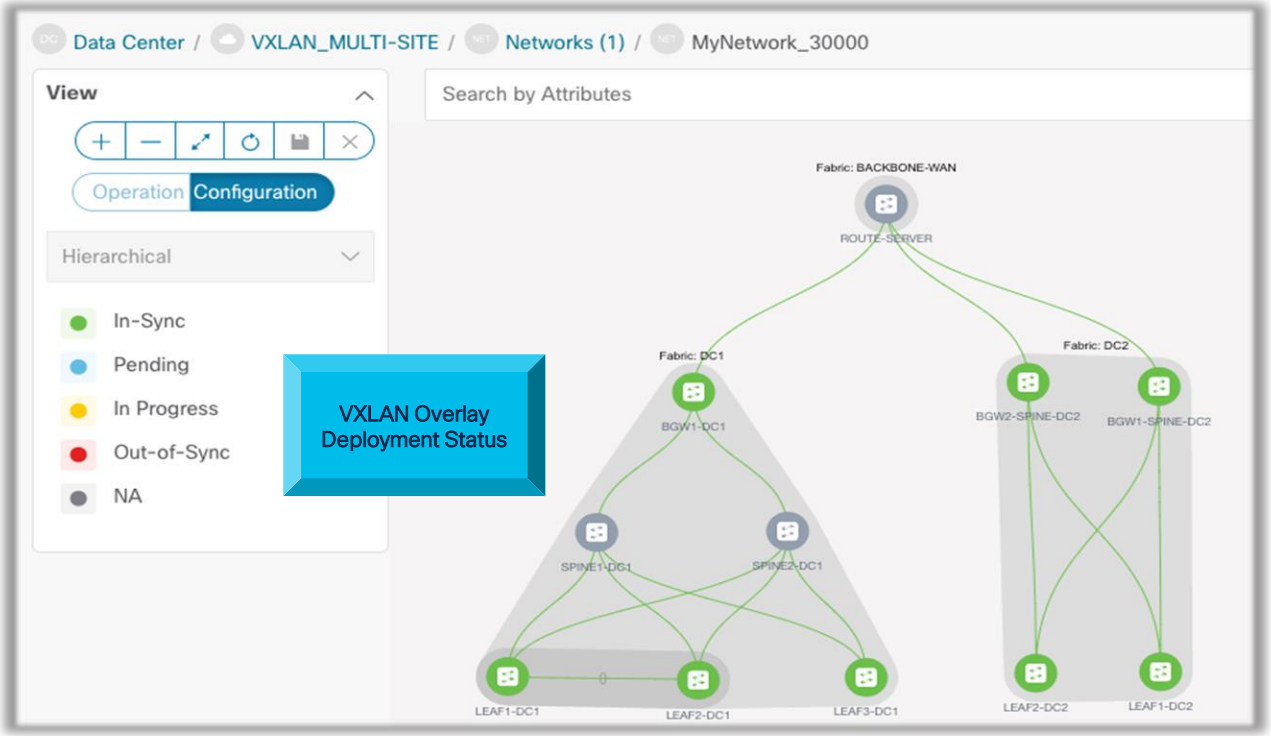
```
vlan 2300
  vn-segment 30000
interface Vlan2300
  vrf member myvrf_50000
  ip address 10.10.10.1/24 tag 12345
  fabric forwarding mode anycast-gateway
  no shutdown
exit
evpn
  vni 30000 l2
    rd auto
    route-target import auto
    route-target export auto
configure terminal
interface ethernet1/4
  switchport trunk allowed vlan add 2300
interface ethernet1/5
  switchport trunk allowed vlan add 2300
interface ethernet1/6
  switchport trunk allowed vlan add 2300
```

L3 SVI Configs

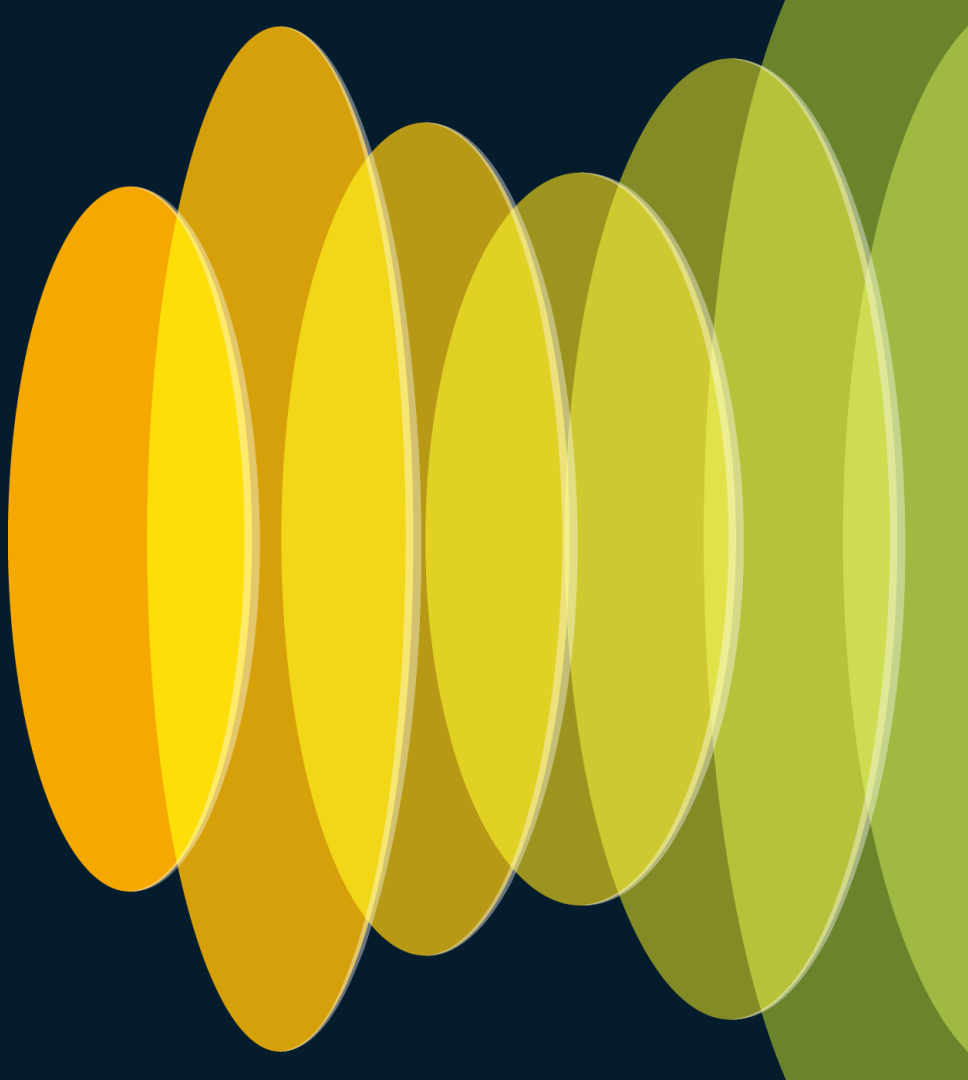
L2VNI Network Configs

VRF and Network Deployment Status

In VXLAN EVPN Multi-Site Fabric



Verification and Validation with NDFC



cisco Live!

Verification through NDFC

Keeping you away from CLI

Step 1

Verify Network and VRF attachments

Configuration Execution Status:

Network Status **Deployed**

VRF Status **Deployed**

Step 2

Deployment History

Configuration Execution Status:

Verify Deployment History Status **Success** for Underlay, Overlay, Interfaces, and more

Step 3

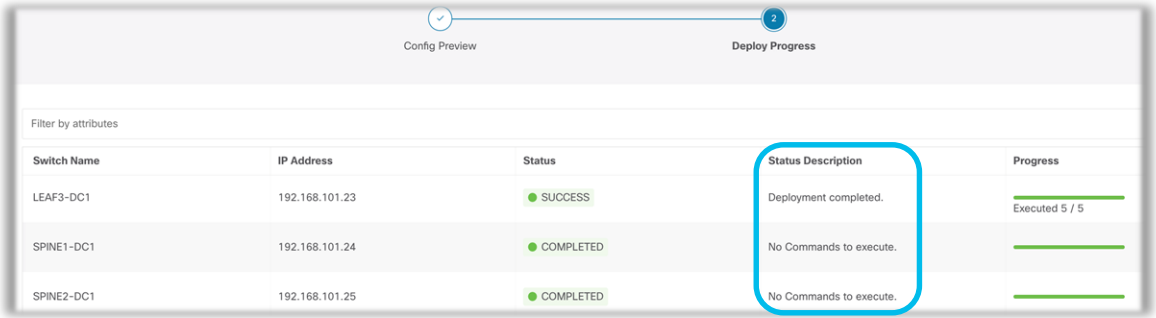
Show commands

Service / features status

(CLI through NDFC)

Attachment deployment status

Job execution perspective



Success or Failure deployment details

Multi-Stage Preview and Deployment

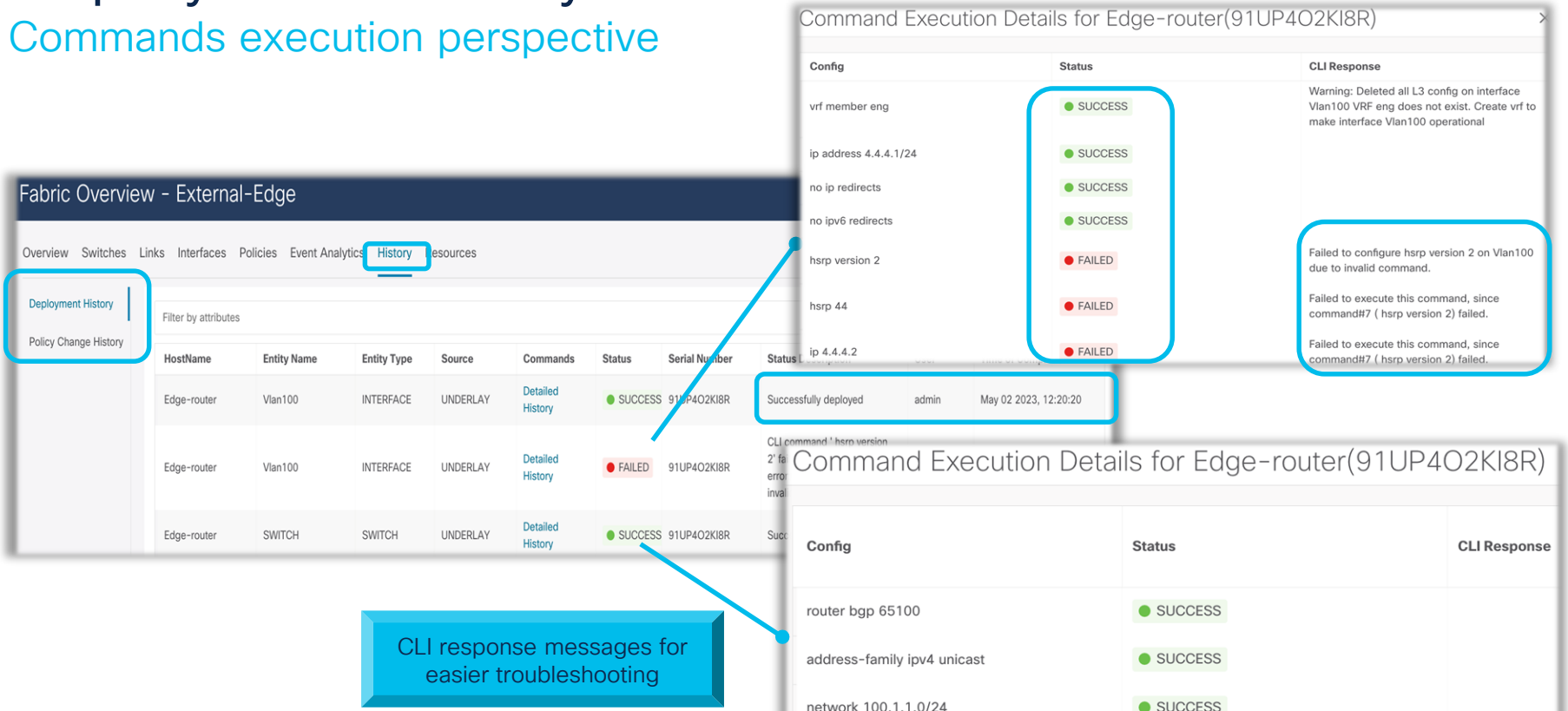
Switch Name	IP Address	Status	Status Description	Progress
Edge-router	192.168.101.99	FAILED	Deployment failed. Check deployment history for more information.	Executed 1 / 12
Edge-Catalyst	192.168.101.101	COMPLETED	No Commands to execute.	

Role	VRF Status	Status Description	Progress
border gateway	Deployment In-Pr	Adding diff to deployment queue	

Role	VRF Status	Status Description	Progress
border gateway	In-Sync	Config compliance sync completed	

Deployment History Tool

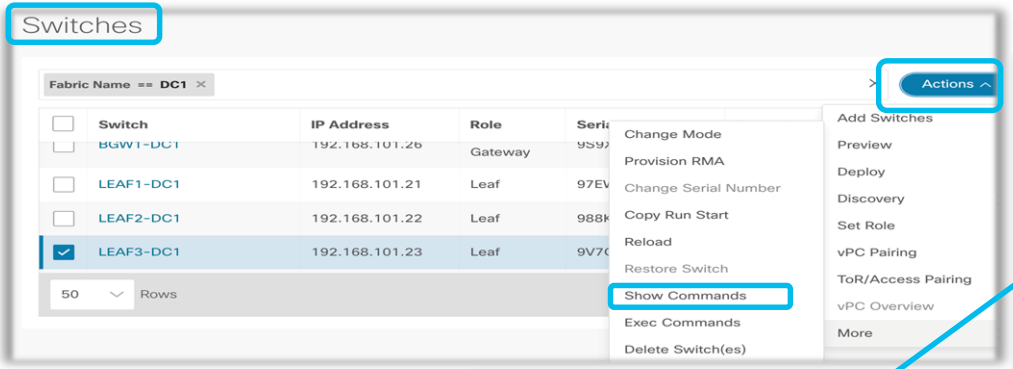
Commands execution perspective

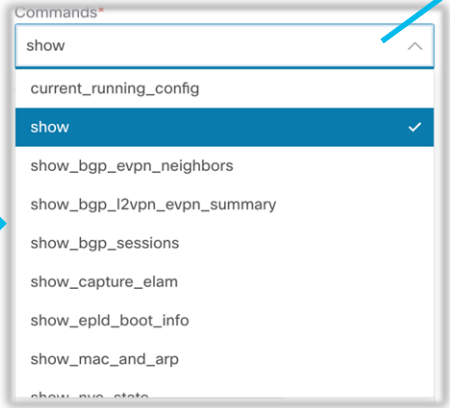


CLI response messages for easier troubleshooting

Show Commands Tool

Switch Config perspective





NDFC pre-built commands or user commands

```
#show ip ospf nei
OSPF Process ID UNDERLAY VRF default
 Total number of neighbors: 2
 Neighbor ID     Pri State            Up Time  Address         Interface
 10.2.0.6          1 FULL/ -          2d22h    10.4.0.29       Eth1/2
 10.2.0.1          1 FULL/ -          2d22h    10.4.0.21       Eth1/3

#show bgp l2vpn evpn su
BGP summary information for VRF default, address family L2VPN EVPN
BGP router identifier 10.2.0.3, local AS number 65001
BGP table version is 29, L2VPN EVPN config peers 2, capable peers 2
9 network entries and 16 paths using 2436 bytes of memory
BGP attribute entries [16/2752], BGP AS path entries [1/10]
BGP community entries [0/0], BGP clusterlist entries [6/24]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.2.0.1        4 65001    4252    4240       29    0    0    2d22h 5
10.2.0.6        4 65001    4252    4239       29    0    0    2d22h 5

#show int nve1
nve1 is up
admin state is up,  Hardware: NVE
  MTU 9216 bytes
  Encapsulation VXLAN
```

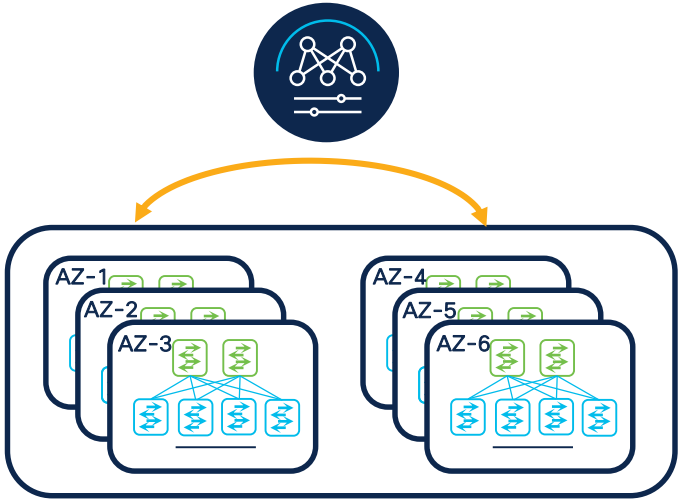
Appendix

White Paper and Document References

- VXLAN BGP EVPN
  - https://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/white-paper-listing.html
- NDFC
  - https://www.cisco.com/c/en/us/products/cloud-systems-management/prime-data-center-network-manager/white-paper-listing.html

Conclusion

Key points to remember



- VXLAN EVPN Multi-Site maintains clear change and fault domain separation to deploy large-scale and highly available DC architectures.
- NDFC provides simplified mechanism to extend and provide end-to-end network and policy consistency across regions, all with a single point of orchestration.
- NDFC provides a single plane of glass solution to automate and manage Nexus and Non-Nexus devices

Complete Your Session Evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to **win 1 of 5 full conference passes** to Cisco Live 2025.

Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn **exclusive prizes!**

Complete your surveys in the **Cisco Live mobile app.**

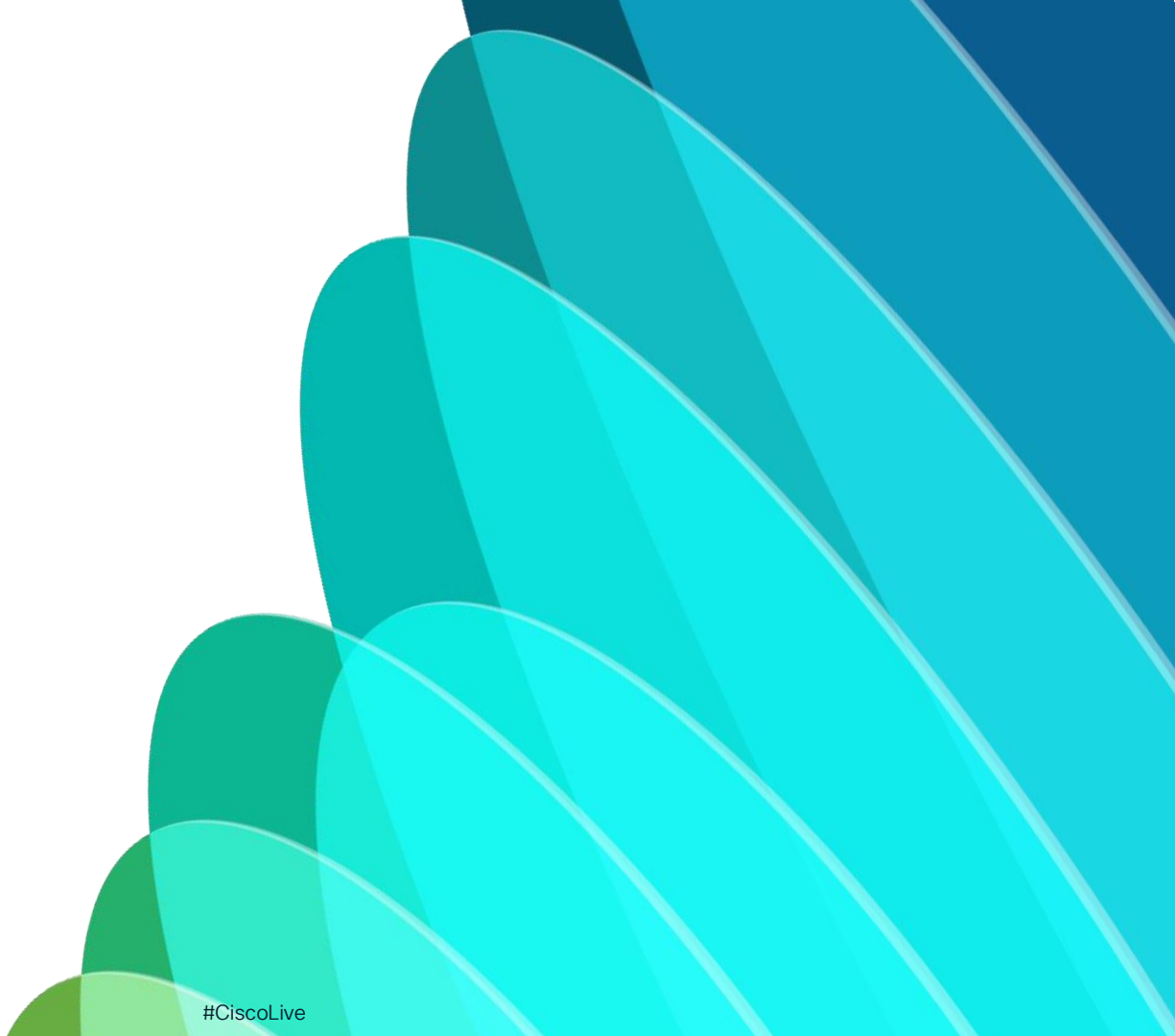
Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer 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

Contact me at: Cisco Webex

The bridge to possible

Thank you



cisco Live!