ACI Multisite

V2.1

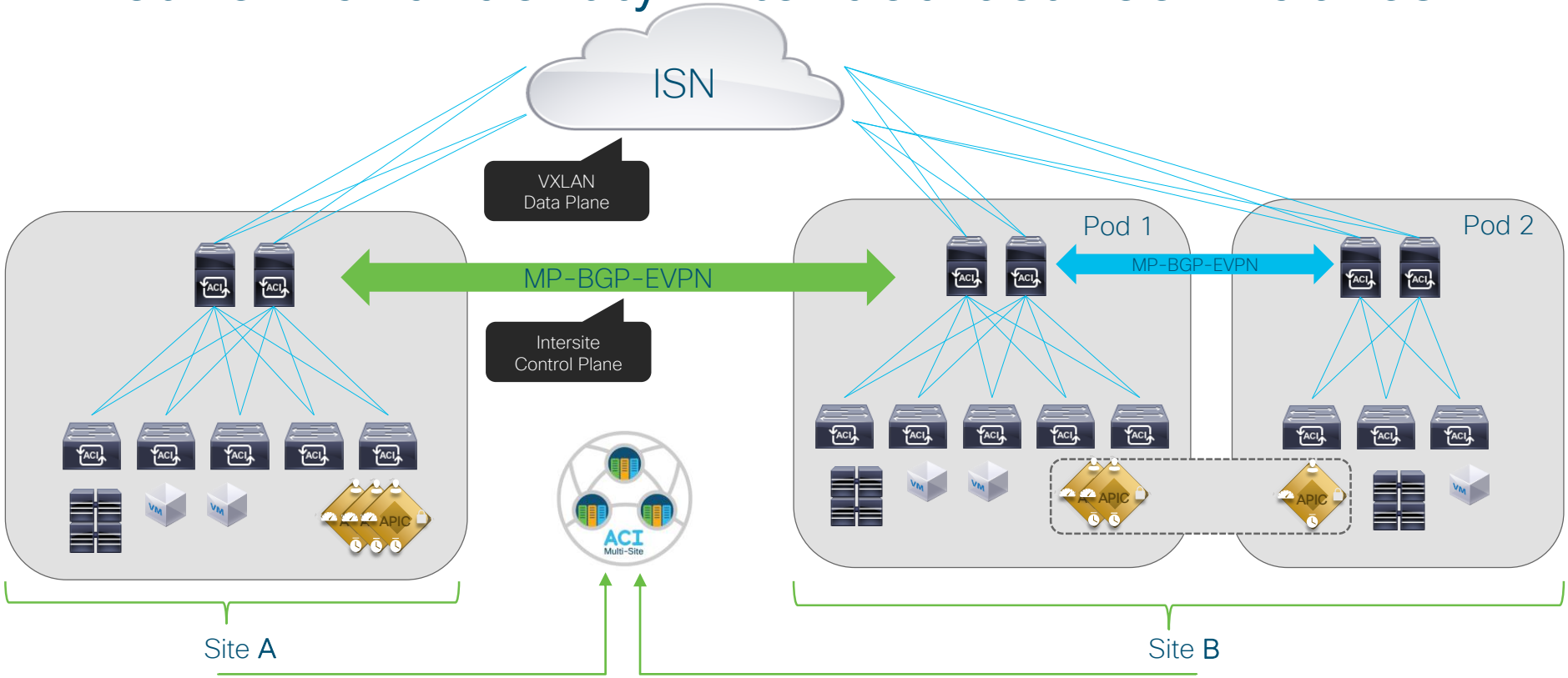
Roland Ducomble – CCIE 3745

EMEAR ACI solution Tac Team – Technical Leader

5th Dec 2019

Overview

Network and Identity Extended between Fabrics



Operational Consistency

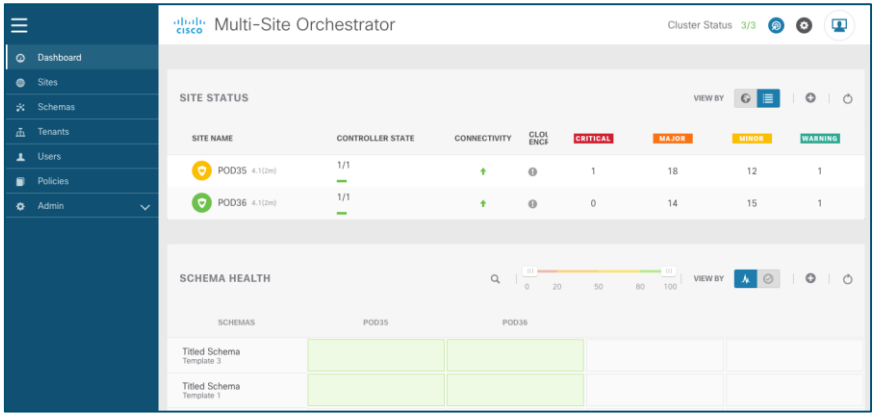
Policy Translation

Single Point Of Orchestration

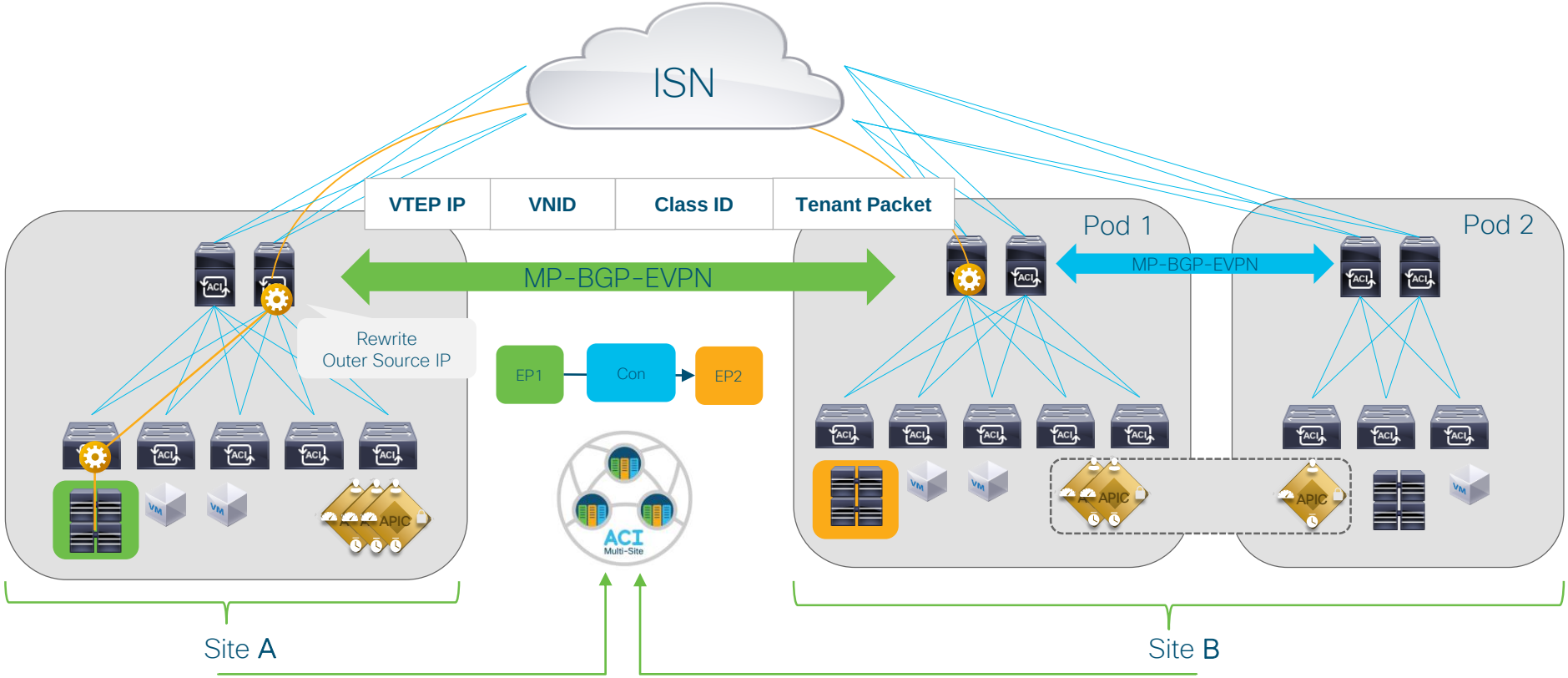
Independent Availability Zones

Multi-Site Orchestrator

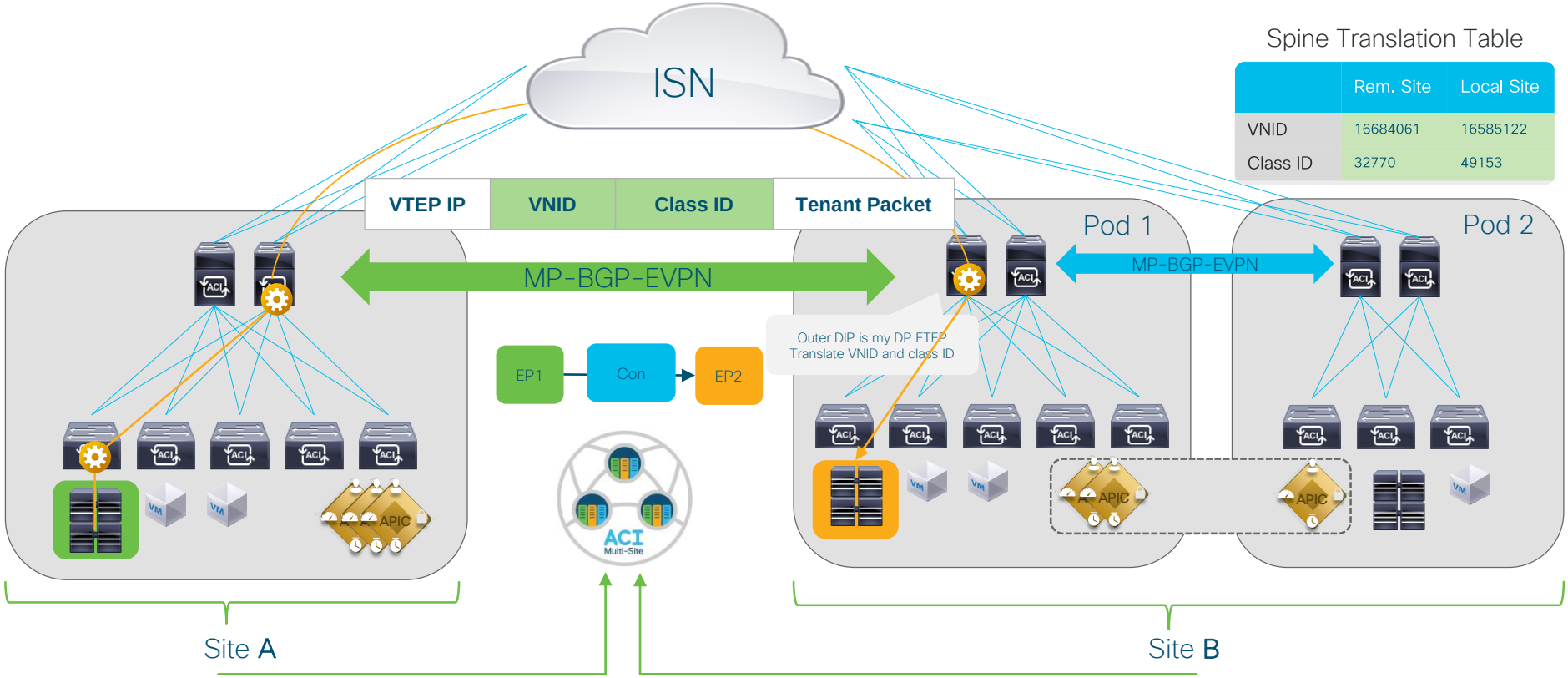
- Provision day-0 Infrastructure
- Create and deploy new tenants
- Define and publish policy Templates
- Add, Delete and Modify Sites
- Central health dashboard



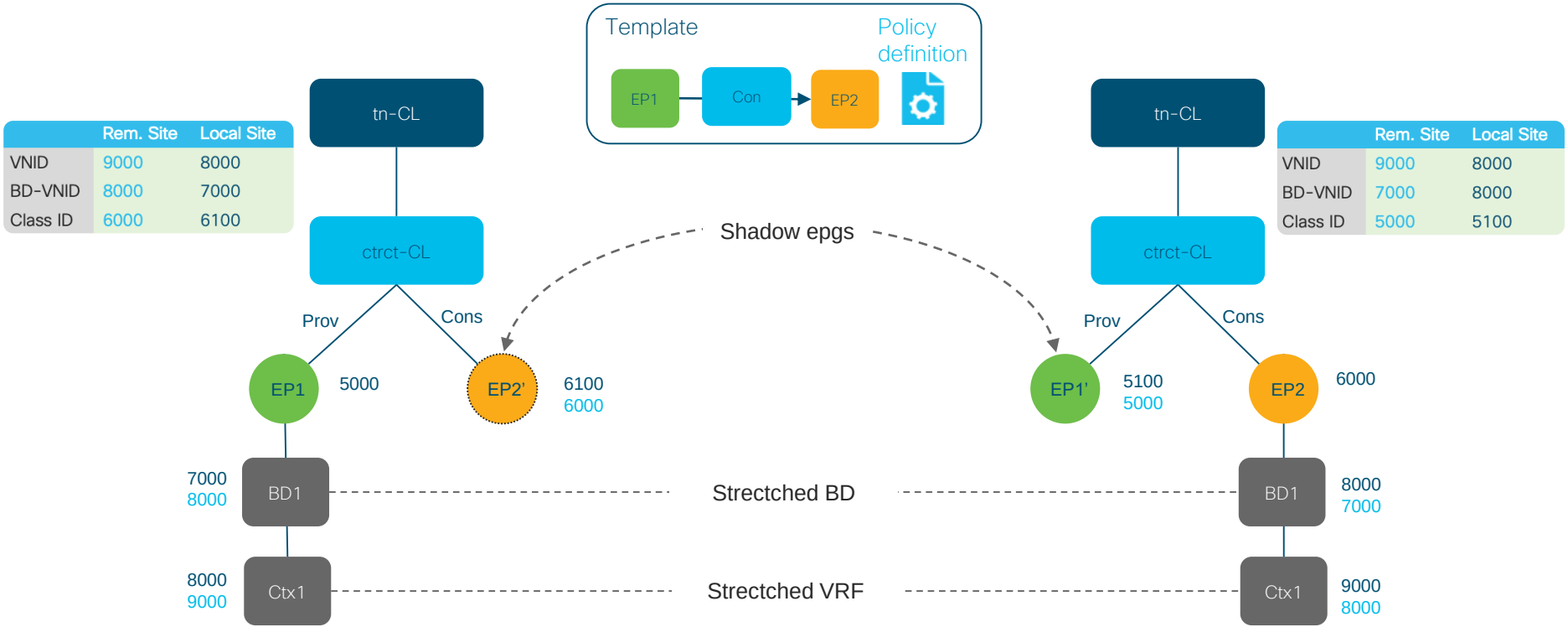
Overview



Overview



Multi-Site Orchestrator

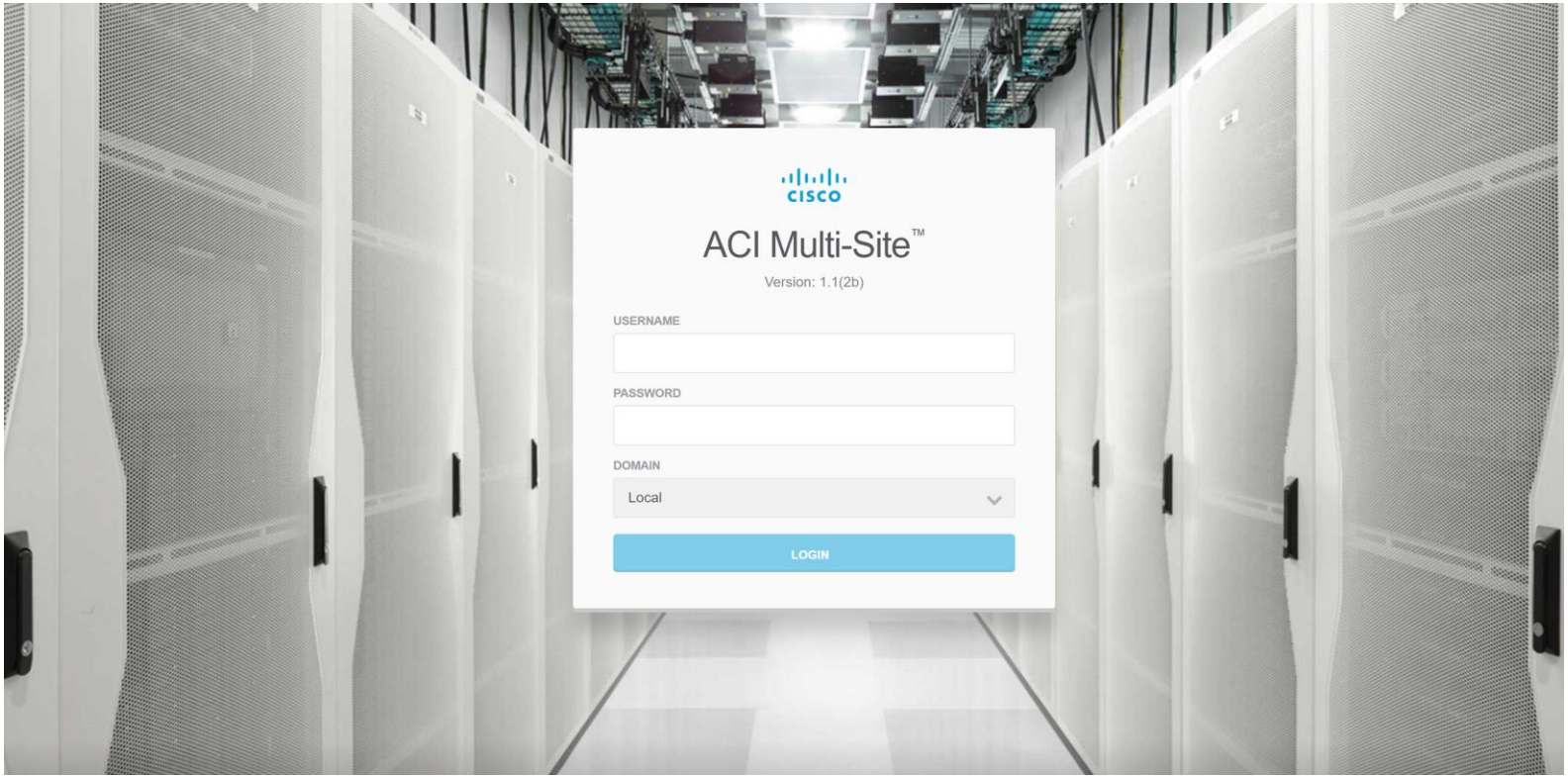


	Rem. Site	Local Site
VNID	9000	8000
BD-VNID	8000	7000
Class ID	6000	6100

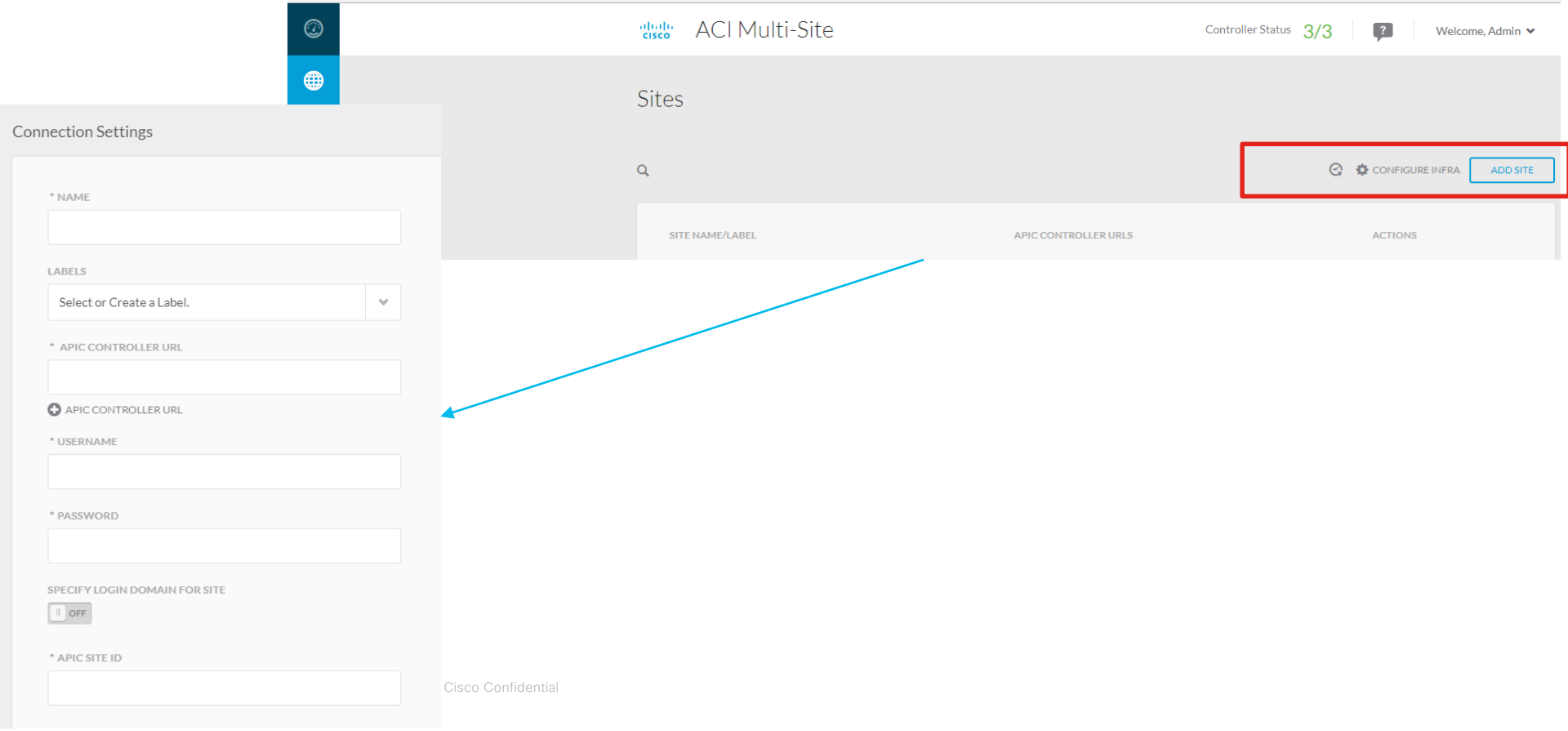
	Rem. Site	Local Site
VNID	9000	8000
BD-VNID	7000	8000
Class ID	5000	5100

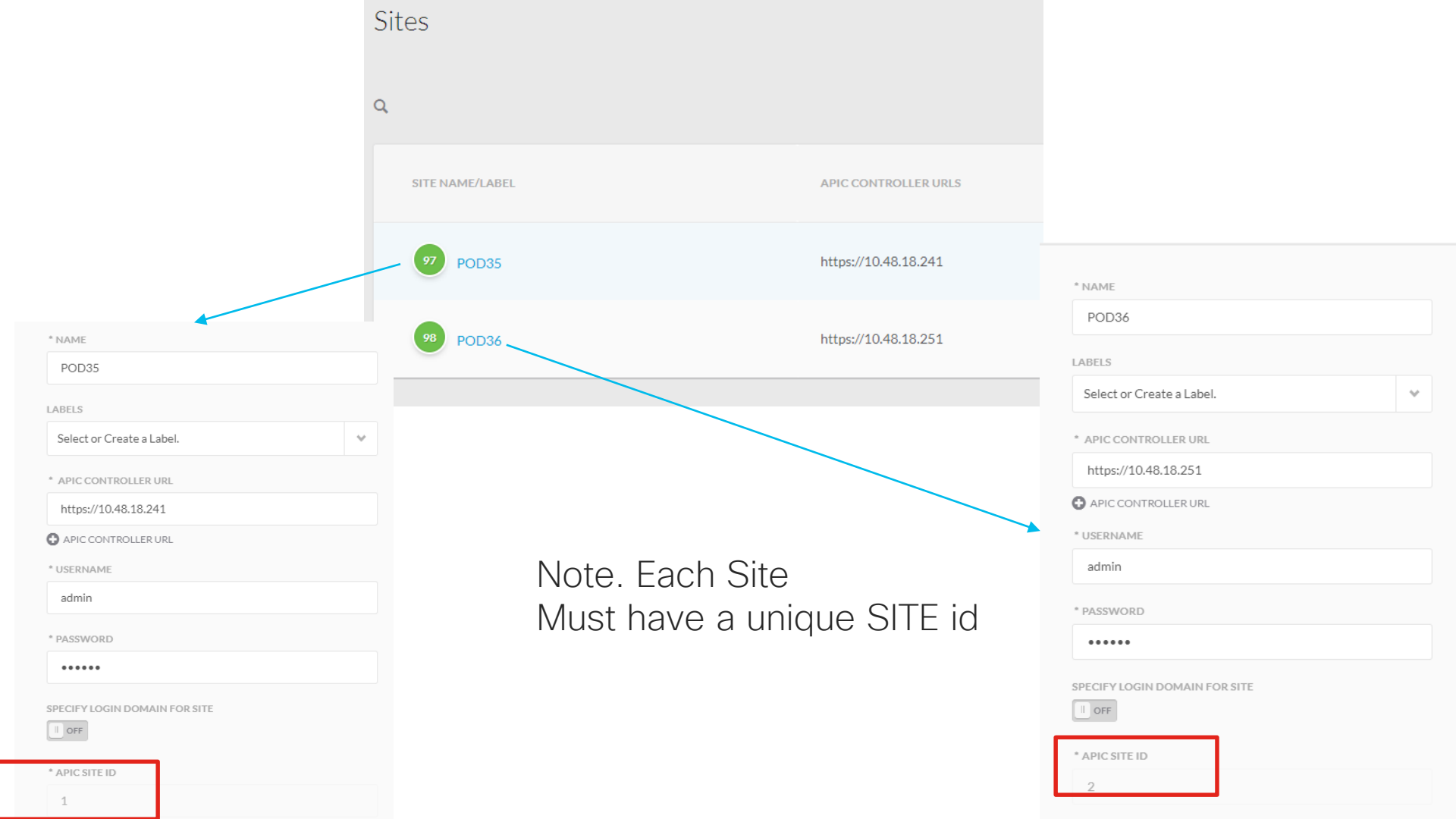
Multisite Day 0 Setup
Infra (overlay-1) config

Connect to MSC GUI



Step 1- In MSC – Add each site one by one



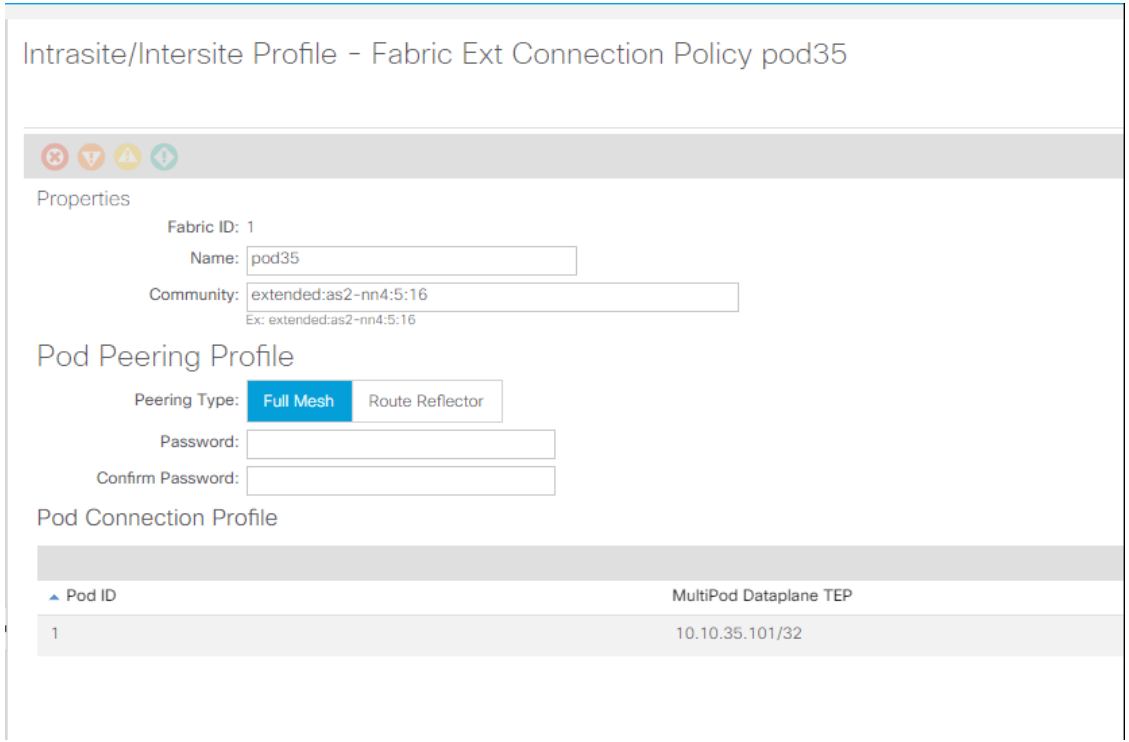

Sites
SITE NAME/LABEL
APIC CONTROLLER URLS
97
POD35
https://10.48.18.241
98
POD36
https://10.48.18.251
* NAME
POD35
LABELS
Select or Create a Label.
* APIC CONTROLLER URL
https://10.48.18.241
APIC CONTROLLER URL
* USERNAME
admin
* PASSWORD
SPECIFY LOGIN DOMAIN FOR SITE
OFF
* APIC SITE ID
1
Note. Each Site
Must have a unique SITE id
* NAME
POD36
LABELS
Select or Create a Label.
* APIC CONTROLLER URL
https://10.48.18.251
APIC CONTROLLER URL
* USERNAME
admin
* PASSWORD
SPECIFY LOGIN DOMAIN FOR SITE
OFF
* APIC SITE ID
2

Step 2 – in each APIC cluster

- On each site in the APIC you need to create the following :
  - iBGP AS and Route Reflector
  - Spine access policies for spine uplink to IPN with an AEP and phys domain
  - External Dataplane TEP per Site (unique Dataplane loopback per site)

- No Need to configure L3 out or ospf !!

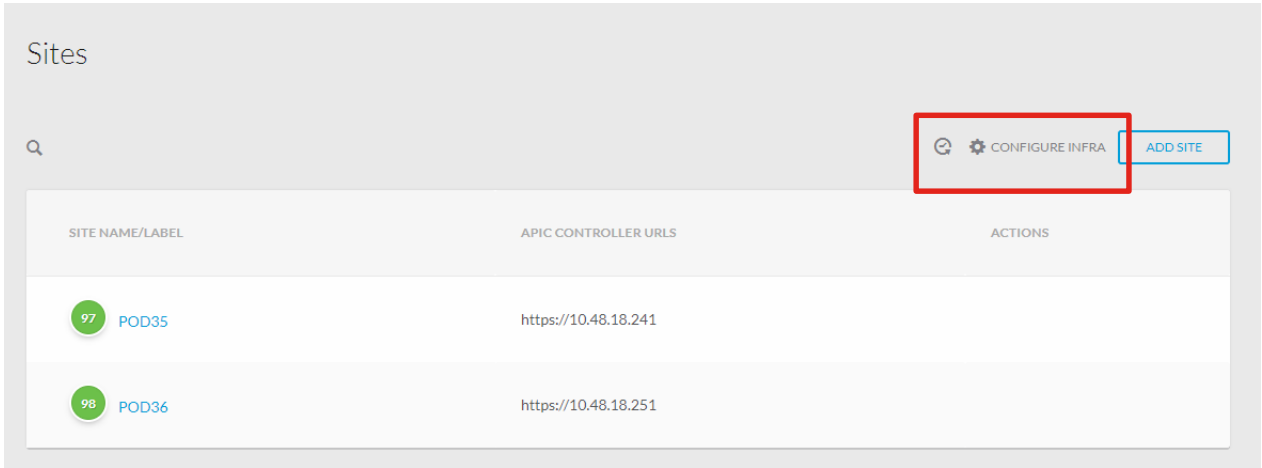
Configure Multipod External Dataplane TEP per site in Infra/Policies/Protocol/FabricExtConnProfiles



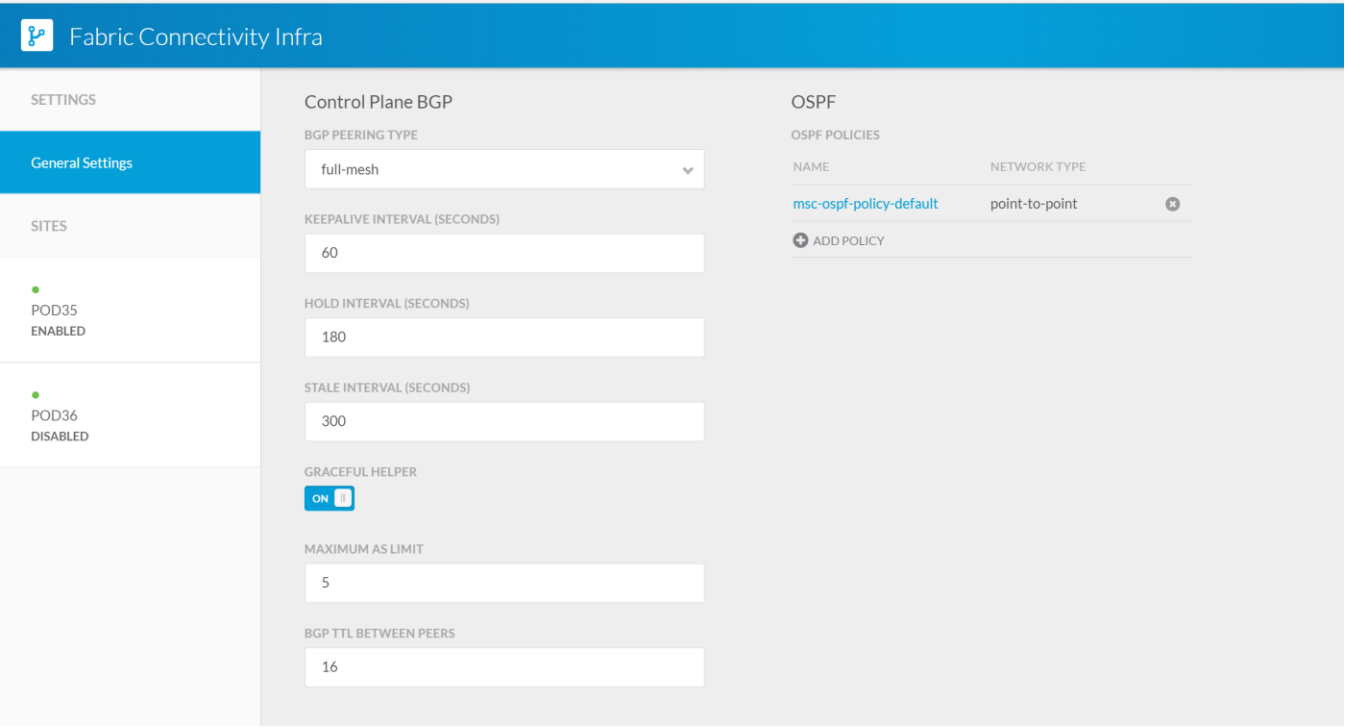
Specify Community for BGP and Mpod DP TEP

Step 3 – in MSC configure infra policies per Site

- (optional) modify BGP and OSPF policies
- Per site :
  - Specify Site policies (ospf area 0, multicast TEP, ..)
  - Pod policies (Dataplane TEP set in APIC)
  - Spine policies (vlan 4 subinterface addresses and bgp RID (Control plane TEP)



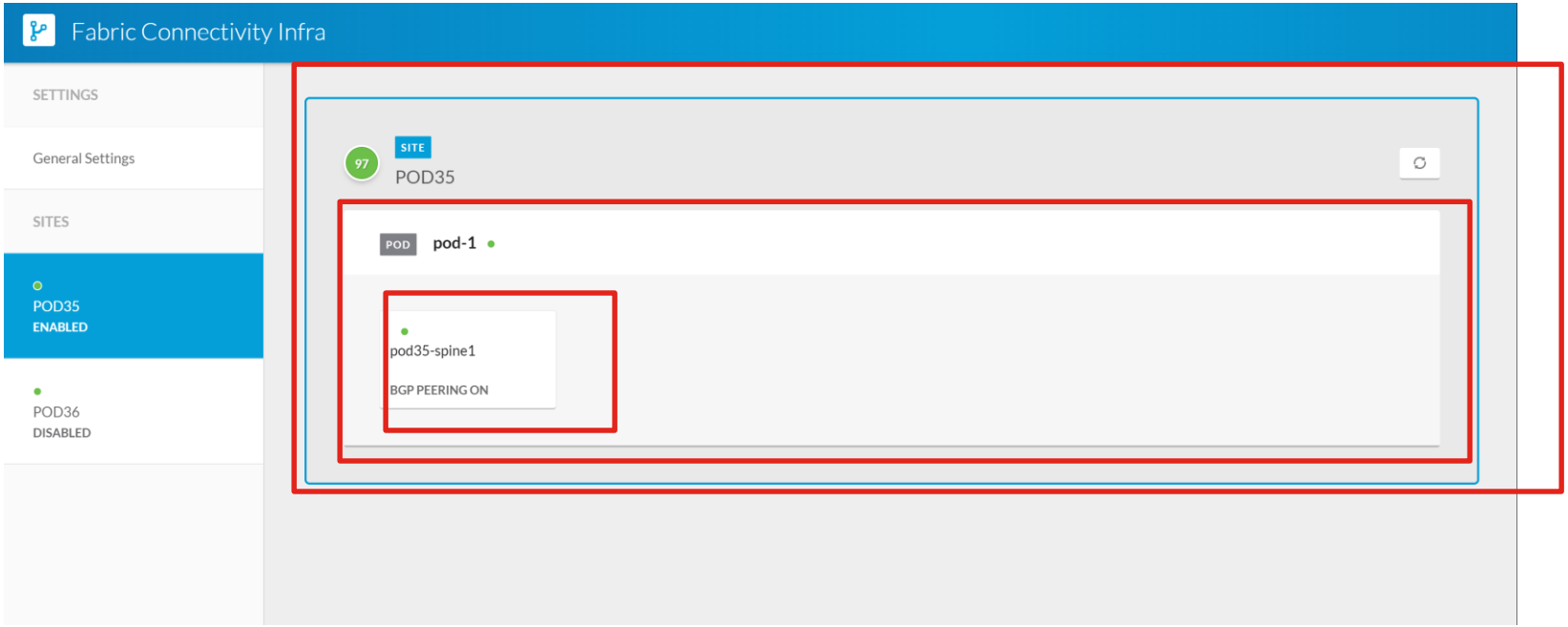
Fabric infra General settings



Allow to modify
Bgp and ospf policies

Once done,
Select one of the site
On the left

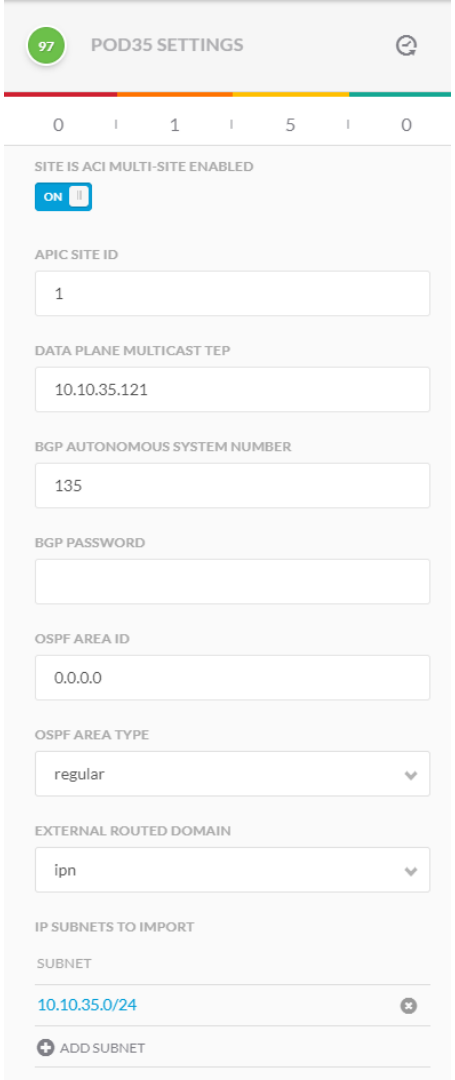
Site 1 (Pod35) infra settings on MSC



You can click on The Site , the Pod or the Spine,
It will allow different settings on the config pane (right pane)

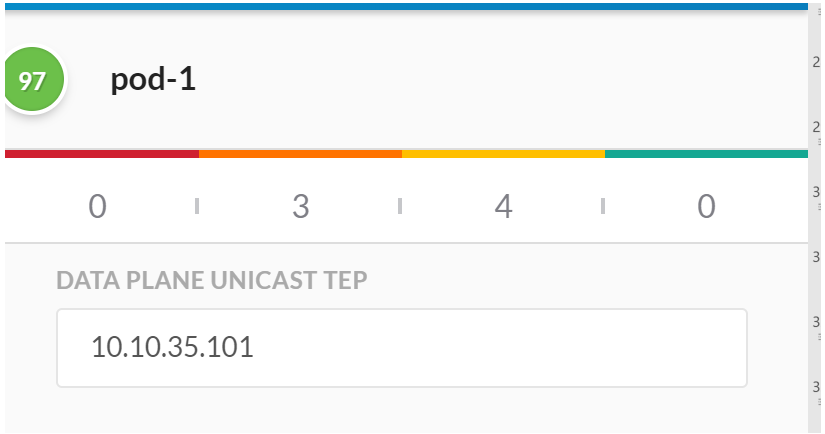
Site settings

- Specify Dataplane Multicast TEP (one lo per site) used for HREP
- BGP AS (matching AS from the site configured in apic)
- OSPF area type and area id
- Ext Routed Domain



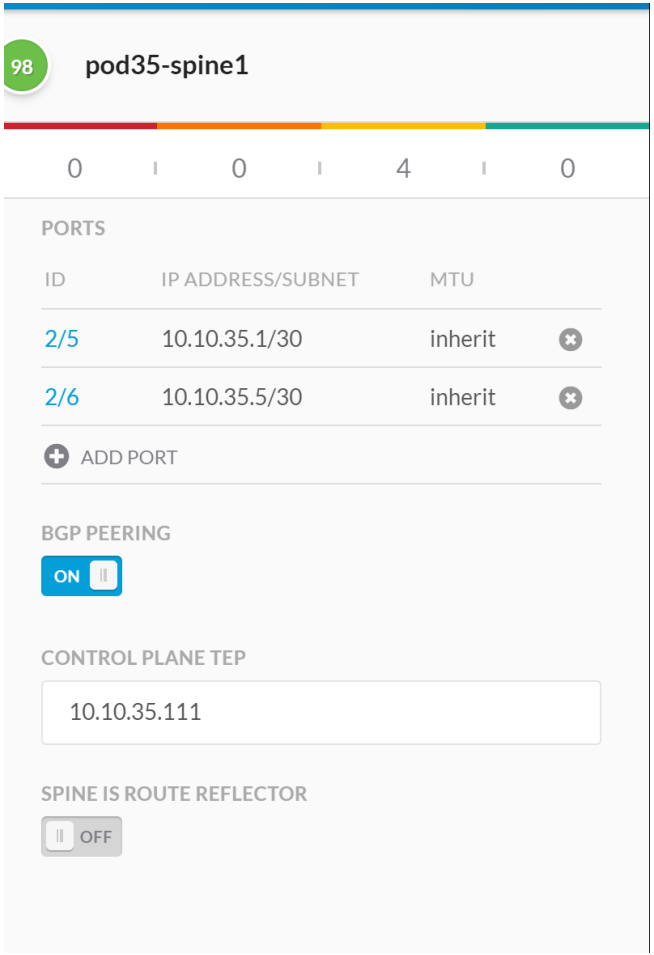
Pod infra settings

- Specify the Dataplane TEP set in APIC in step 2



Spine Infra settings

- For each interface from spine to IPN, set IP address and mask
- Enable BGP peering
- Configure Control plane TEP (Router ID)



Step 4 – Configure ISN (InterSite Network)with OSPF

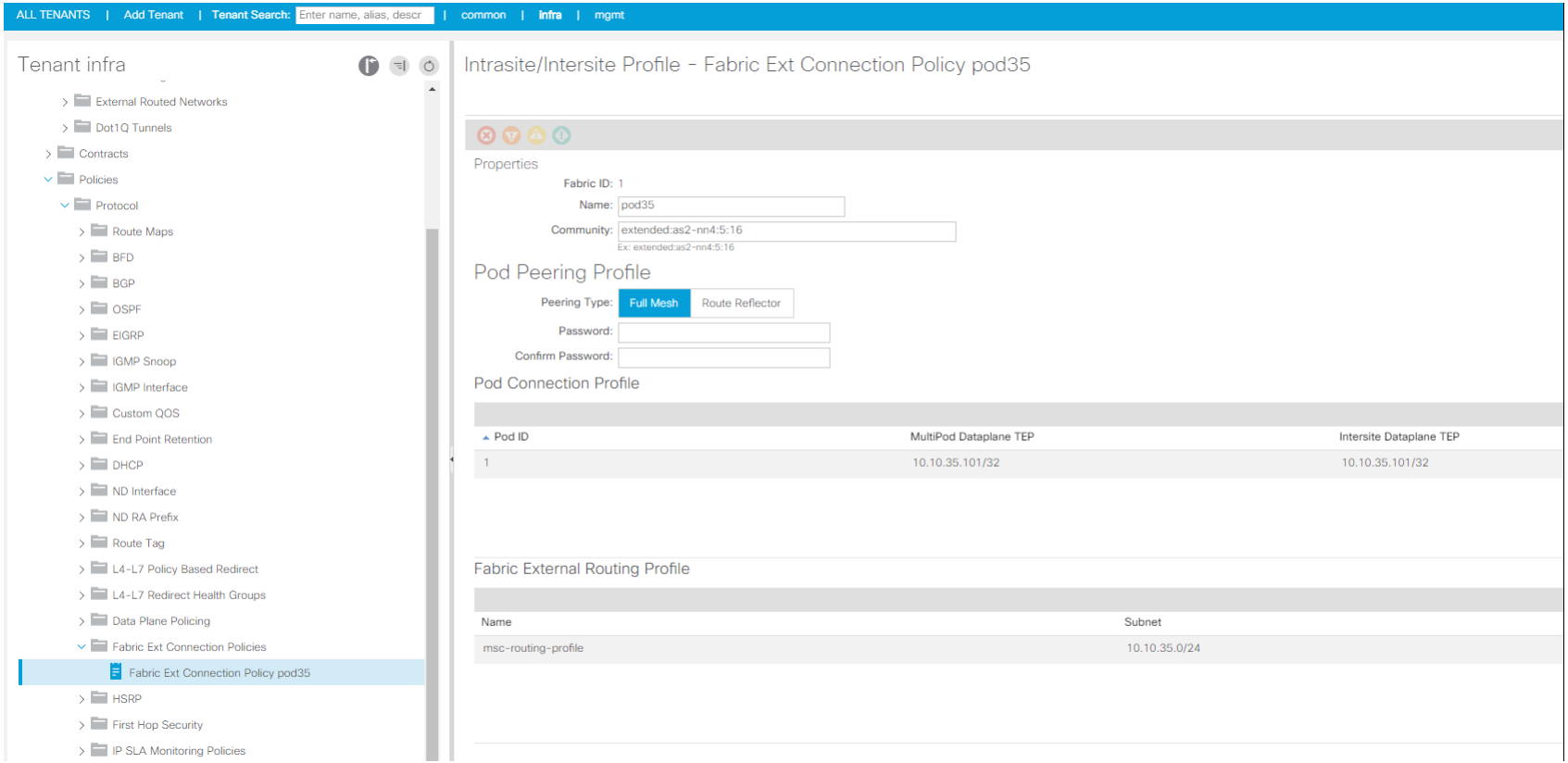
```
vrf context IPN
router ospf 1
  vrf IPN
    router-id 10.10.35.100



interface Ethernet1/49.4
  encapsulation dot1q 4
  vrf member IPN
  ip address 10.10.35.2/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  no shutdown

interface Ethernet1/50.4
  encapsulation dot1q 4
  vrf member IPN
  ip address 10.10.35.6/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  no shutdown
```

```
interface Ethernet1/51.4
  encapsulation dot1q 4
  vrf member IPN
  ip address 10.10.35.10/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  no shutdown


interface Ethernet1/52.4
  encapsulation dot1q 4
  vrf member IPN
  ip address 10.10.35.14/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  no shutdown

interface loopback1
  vrf member IPN
  ip address 10.10.35.100/32
```

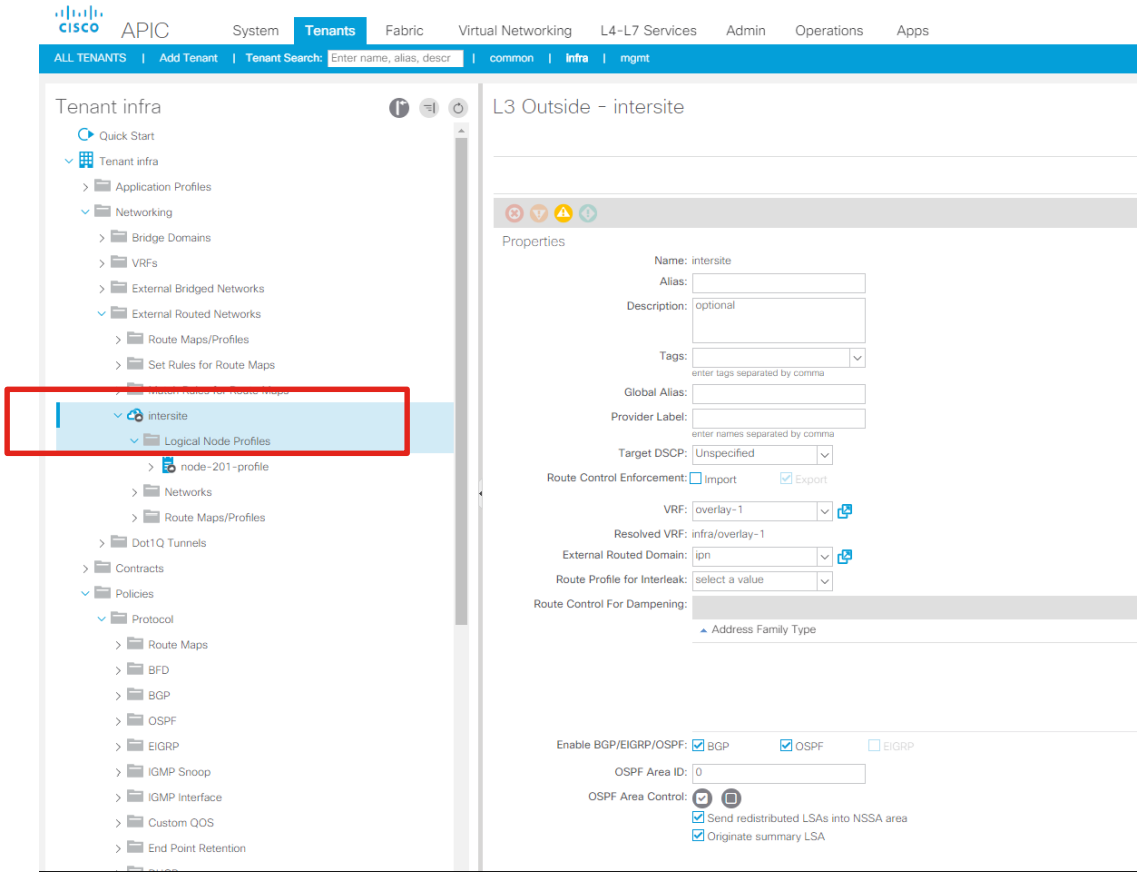
Step 4 – Verify config on each apic cluster

- Verify L3 out , ospf, bgp was set on each apic cluster
- Verify OSPF is up on spine and gets routes from IPN
- Verify BGP session is up to remote site

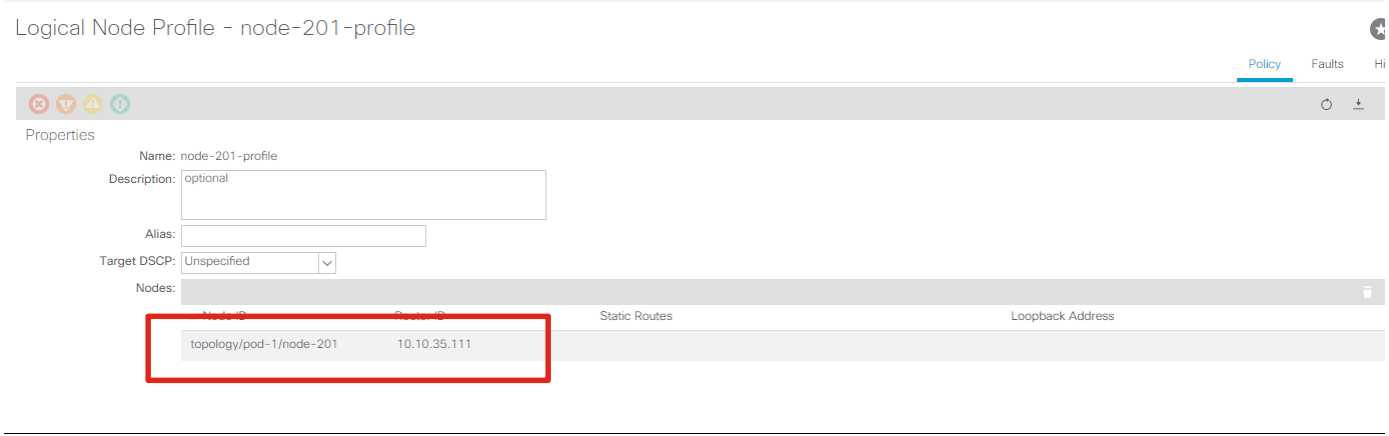
Resulting Intersite Profile when site is full configured in MSC

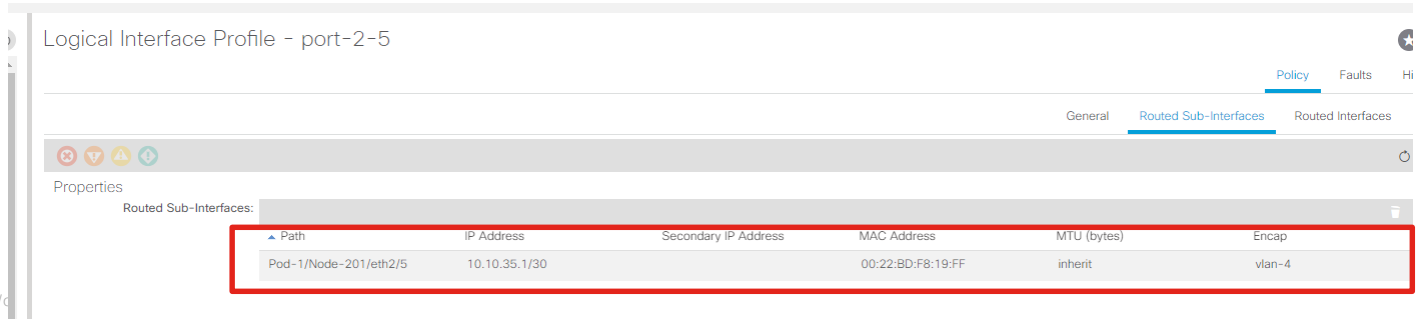


L3 out automatically create



L3 out logical node and l3 if in vlan 4





Overlay-1 check – Spine interface

```
loopback13, Interface status: protocol-up/link-up/admin-up, iod: 119, mode: etep, dci-ucast, vrf_vnid: 16777199
  IP address: 10.10.35.101, IP subnet: 10.10.35.101/32
  IP primary address route-preference: 1, tag: 0

loopback14, Interface status: protocol-up/link-up/admin-up, iod: 120, mode: mcast-hrep, vrf_vnid: 16777199
  IP address: 10.10.35.121, IP subnet: 10.10.35.121/32
  IP primary address route-preference: 1, tag: 0

loopback15, Interface status: protocol-up/link-up/admin-up, iod: 122, mode: mscp-etep, vrf_vnid: 16777199
  IP address: 10.10.35.111, IP subnet: 10.10.35.111/32
  IP primary address route-preference: 1, tag: 0

Ethernet2/5.37, Interface status: protocol-up/link-up/admin-up, iod: 121, mode: external, vrf_vnid: 16777199
  IP address: 10.10.35.1, IP subnet: 10.10.35.0/30
  IP primary address route-preference: 1, tag: 0
```

Dataplane unicast ETEP (anycast per site)

Dataplane multicast TEP

CP ETEP (Control plane - BGP router ID)

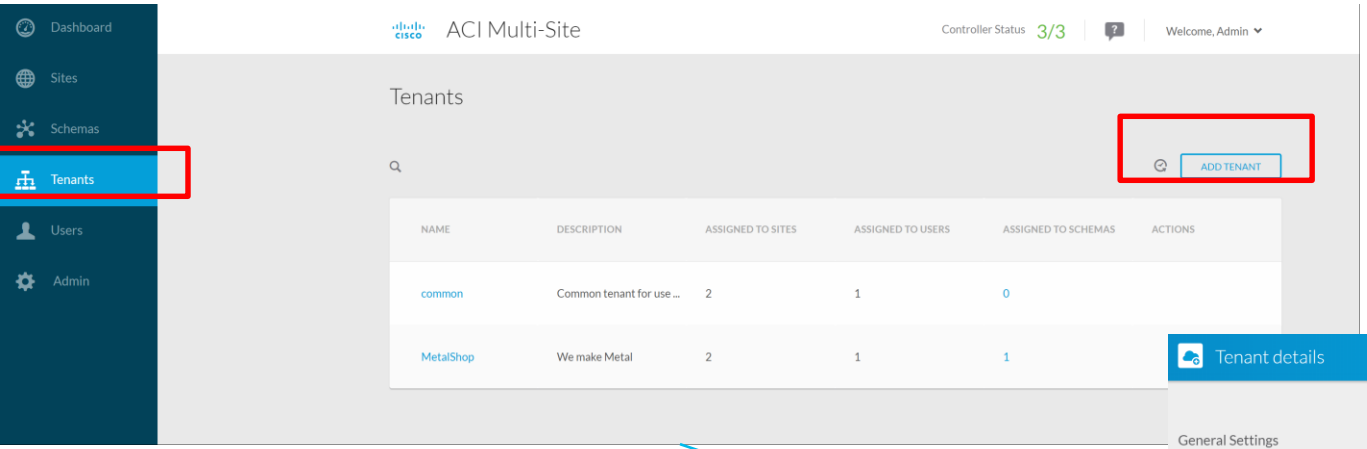
Verify BGP l2vpn evpn and OSPF is up on each spine

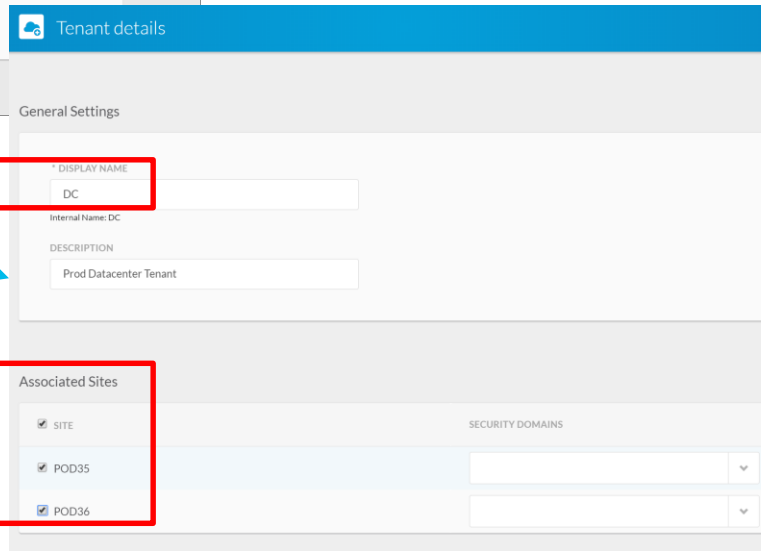
```
pod35-spine1#
pod35-spine1# show ip ospf  neighbors vrf overlay-1
 OSPF Process ID default VRF overlay-1
 Total number of neighbors: 2
 Neighbor ID     Pri State            Up Time  Address         Interface
 10.10.35.100      1 FULL/ -          02:06:51 10.10.35.2      Eth2/5.37
 10.10.35.100      1 FULL/ -          02:06:27 10.10.35.6      Eth2/6.38
pod35-spine1#
```

```
pod35-spine1# show bgp l2vpn evpn summary vrf overlay-1
BGP summary information for VRF overlay-1, address family L2VPN EVPN
BGP router identifier 10.10.35.111, local AS number 135
BGP table version is 26, L2VPN EVPN config peers 1, capable peers 1
13 network entries and 9 paths using 1864 bytes of memory
BGP attribute entries [4/576], BGP AS path entries [1/6]
BGP community entries [0/0], BGP clusterlist entries [0/0]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.10.35.112    4   136     126     124       26    0    0 01:54:15 3
```

Multisite Day 1 Tenant Creation

Create multisite Tenant in MSC

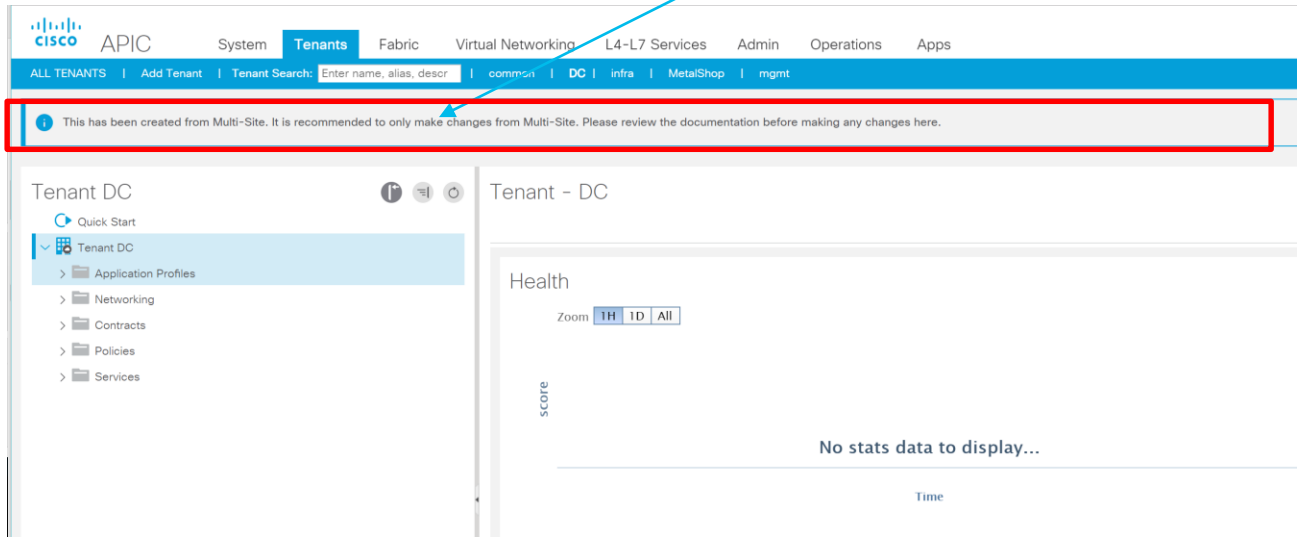




Add a tenant
Give tenant name
And associate it to one or more site

Tenant is created on each APIC cluster of site associated

See the warning that this tenant was created from Msite controller



Multisite Day 1 Schema Management

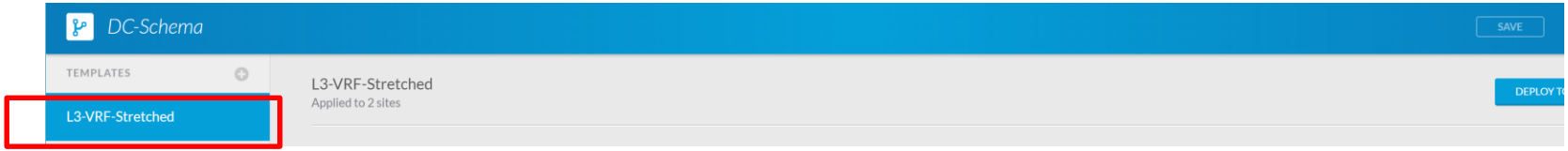
What are Schema ?

Create New schema
In Schemas- Add Schema

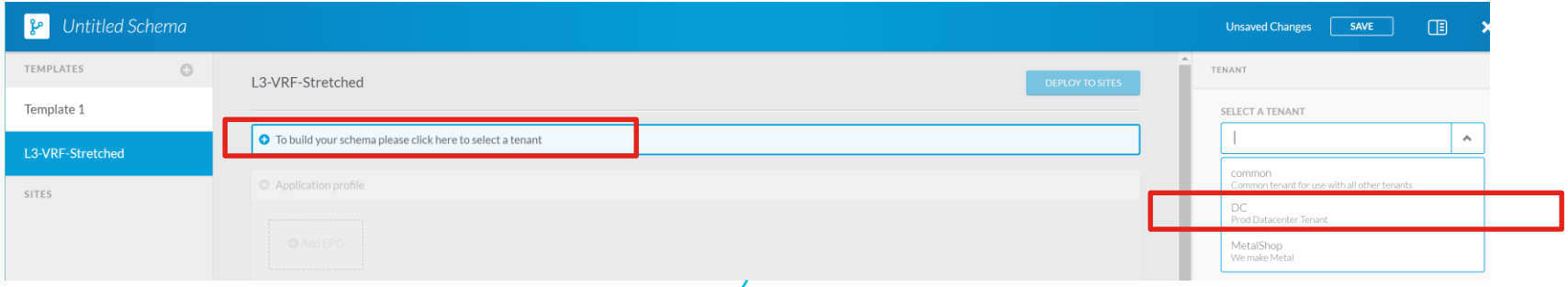
- Schema list displays tenants associated to a schema as well as templates in it.
- User can edit, delete and create a new schema from this page.
- Schema can be seen as a use-case skeleton which could spread over one or more tenant and one or more site
- Schema contains Templates

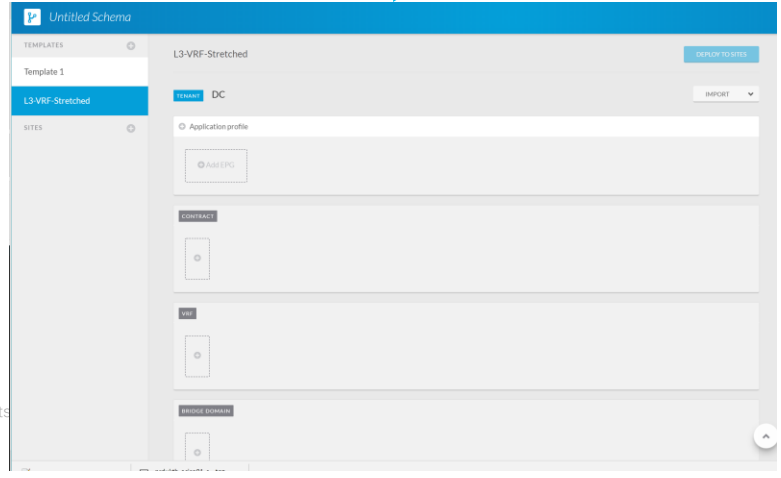
New schema Startup Screen – Templates

- You can name the Schema
- Schema starts with a default Template 1 that you can rename as well
- You can add extra templates in the scheme
- When building a templates, you must first assign it to a tenant



Select a tenant for the Template

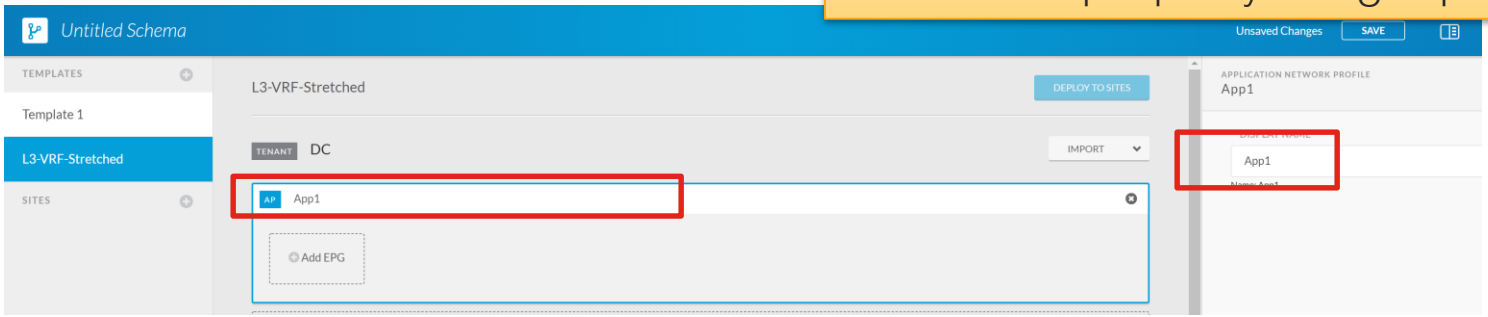




Once the tenant is Selected you can add App Prof/Epg VRF, BD and Contract

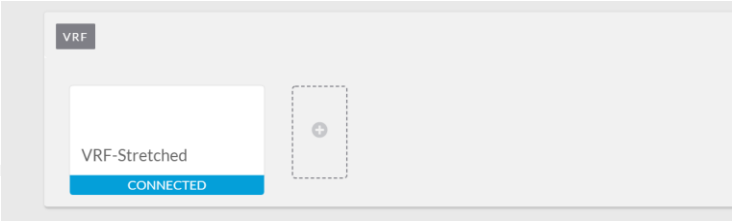
MSC GUI logic

Whenever you click somewhere
In middle panel you can modify
Name and property in right pane



We created ANP

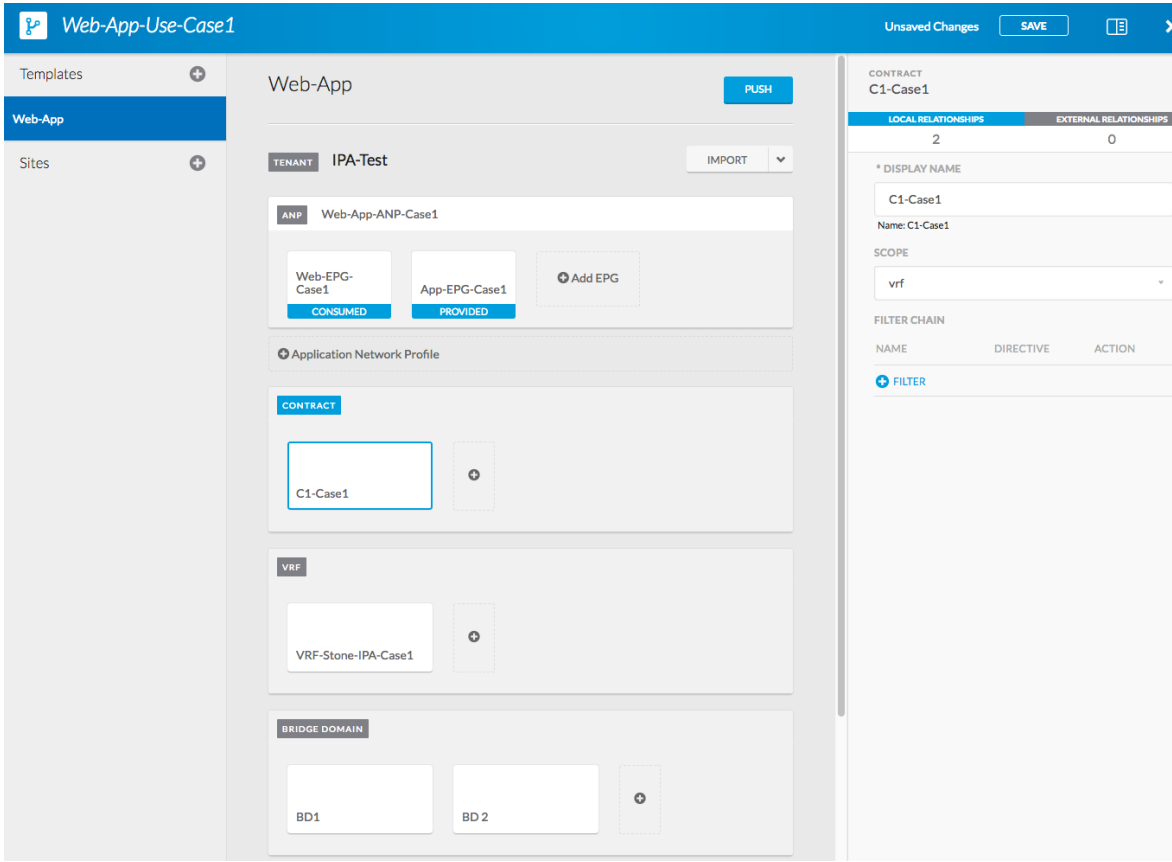
No we create a VRF in the Tenant (name VRF-stretched)



Schema builder: build templates and objects

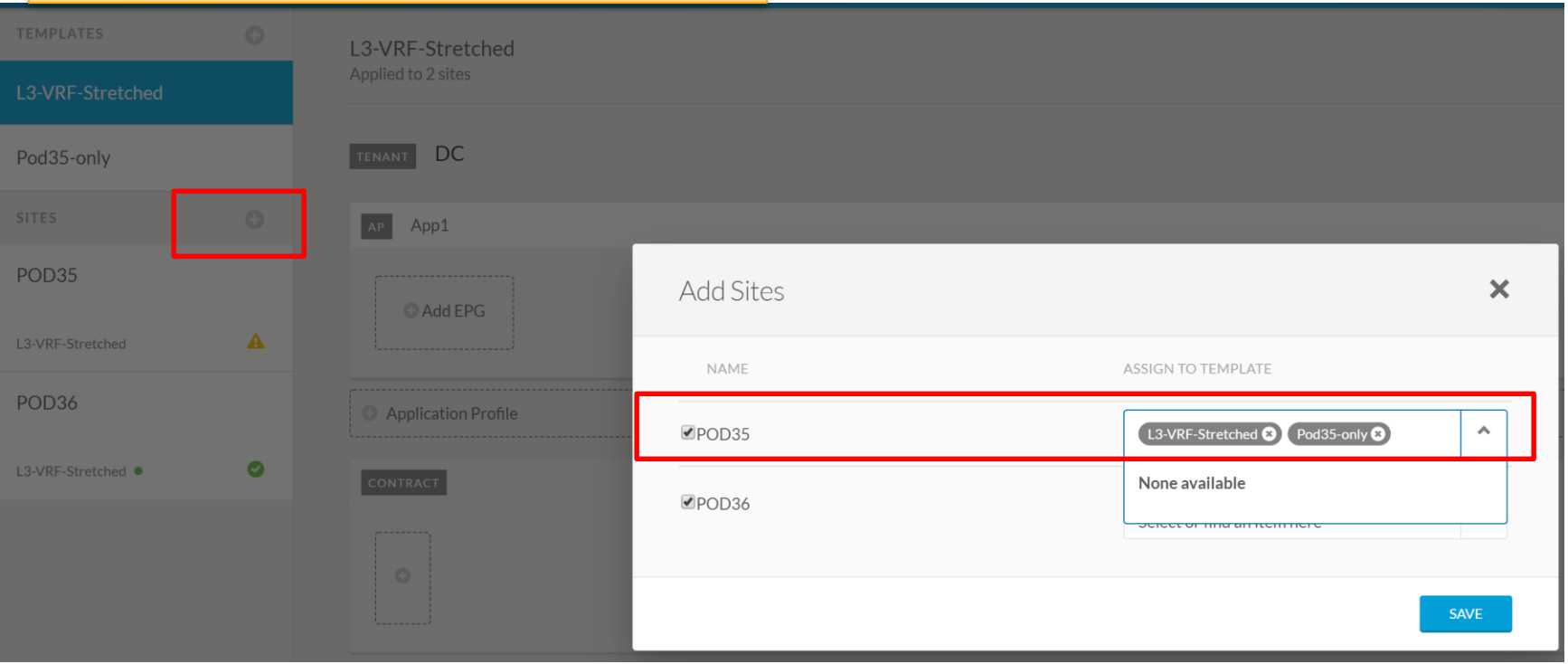
Templates allows to create objects:

- App Profile
- EPG
- VRF
- Contract
- BD
- Filter
- External EPG



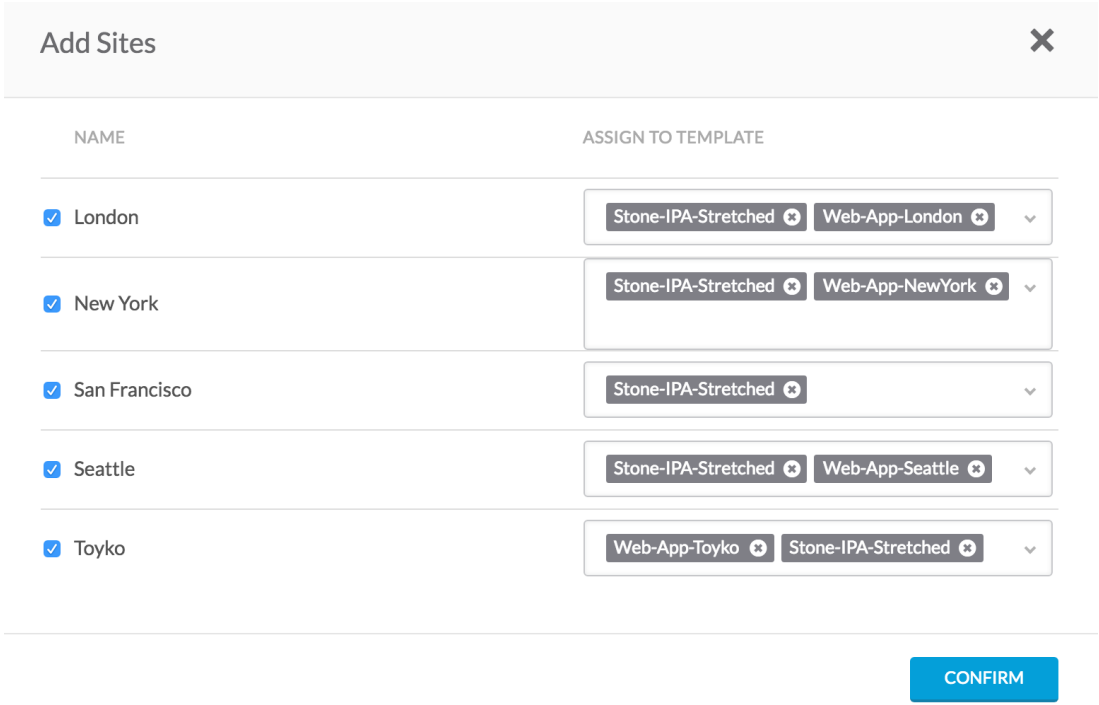
Schema builder: associate sites

Click the + sign next to Sites, there you can modify which templates applies to which sites.



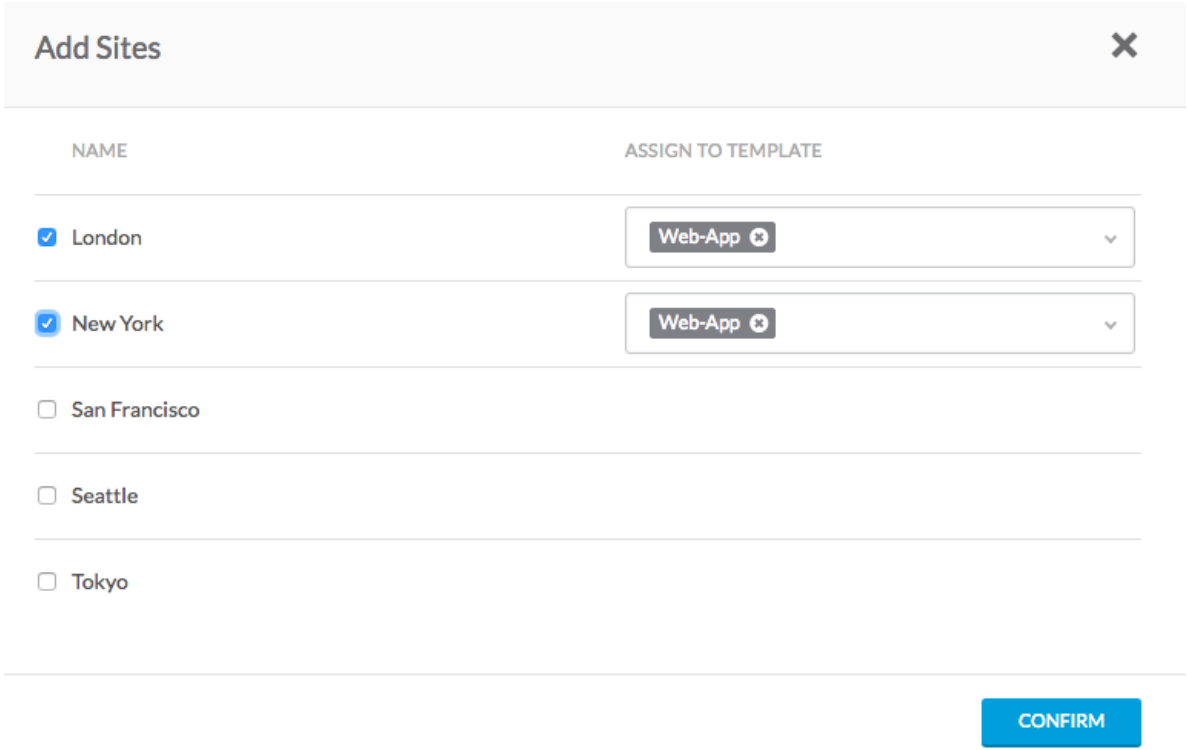
Schema builder: associate sites

- Templates containing policies have to be associated to all sites where they have to be stretched
- User can assign specific template to selected sites



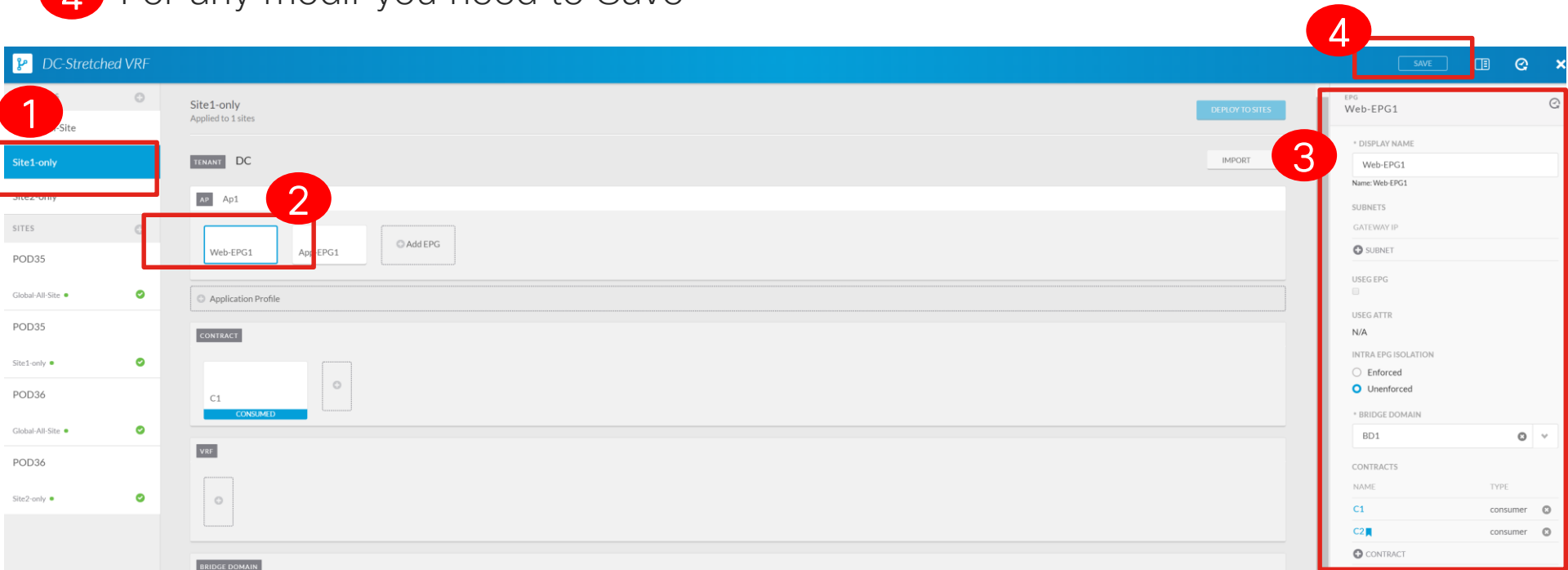
Schema builder: add sites

- Click on “+” next to sites to select which template should be deployed on which site



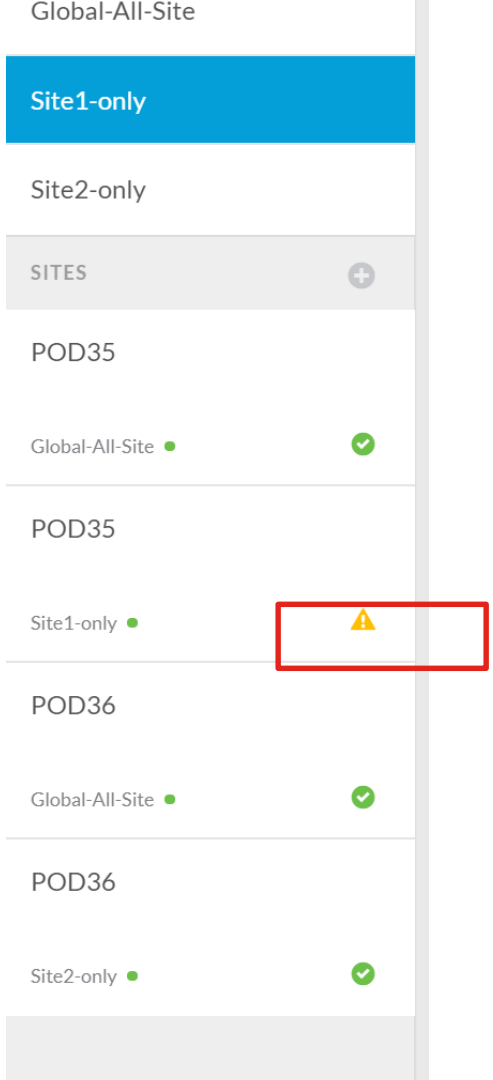
Example EPG screen in Template

1. Here Template Site1-only is selected
2. EPG : Web-EPG1 is selected in the template
3. We can modify EPG to BD assoc, contract cons or prov.
4. For any modif you need to Save



Changes Saved by not deployed yet

We see POD35/Site1-only
Template/site assoc which is not in sync
Click Deploy to site in Top right of UI to deploy

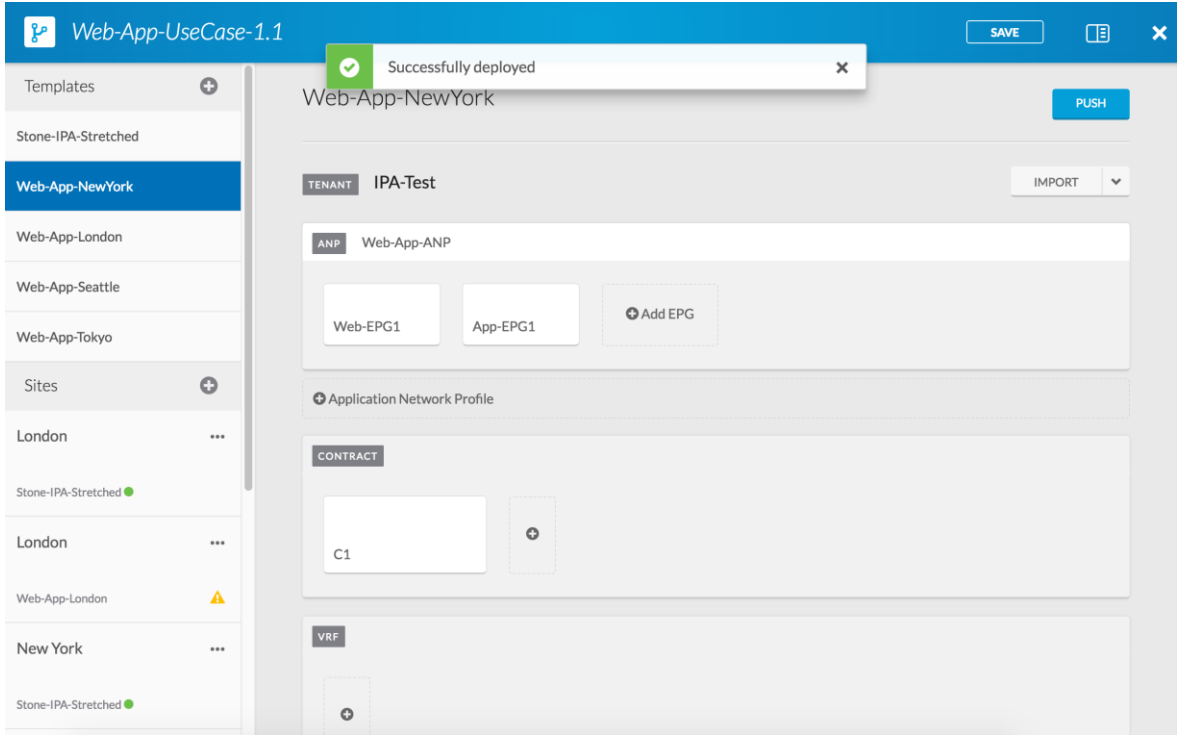


Modif in Template/Site association

1. Here we selected Site/Template combo (Pod35 – Site1only)
2. Same EPG
3. But we can edit/add more property (local to site) such as Domain or static path

Deploy schema

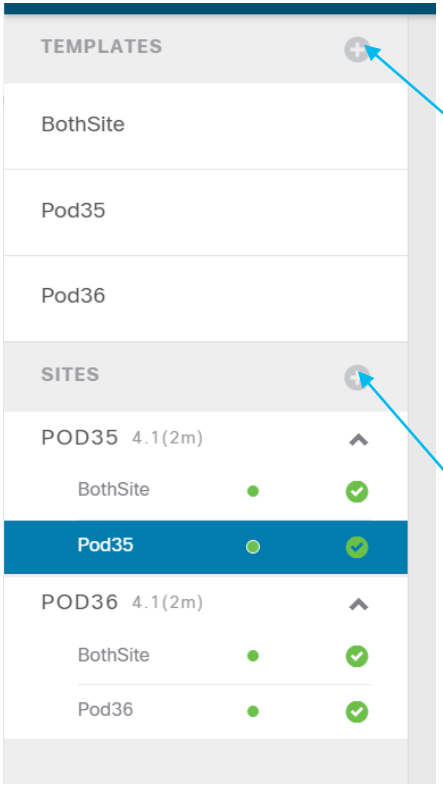
- We can deploy a template to associated sites.
  Push button provides this functionality.
- Note : Push button was replaced by “deploy to site” in latest MSC software
- **Push** saves entire schema and then deploys to associated sites.



Guidelines and Limitations

- If you modify an MO on APIC controlled via MSC, then MSC will not actively poll it. It'll overwrite it on the next push.
- If you modify any MO on APIC that is managed by MSO, APIC UI will warn you. If you ignore the warning and still go ahead and make your changes, your configuration on APIC will be different from the configuration on MSC.
- MSC does not actively detect configuration difference between APIC and MSC unless you trigger consistency check on MSO
- Pushing configuration back from MSC overwrites any local configuration changes made on APIC and brings it back to be the same as configured on MSC.

Templates and Sites differences



- You can have multiple template per schema
- In Template you define Object (VRF, BD, EPG, Contract, Filter, external Epg, L3 out, service graph)
- You can set global properties of those (cross site properties).
  - L3 multicast in VRF
  - Adding filter or service graph to contract
  - Adding contract to EPG
  - …

- Site Specific properties can only be set in sites sections of the templates :

  Static path or domain in EPG

  L3 out binding in BD

Why multiple Template per Schema

- If you do not want to stretch all resource you will need some template for every site containing what resource you want to stretch (VRF, BD, EPG, Contract,…)
- For resource you do not want to stretch they should be created in different template that will only be deployed in certain site
- Note if a resource only in site 1 needs to talk to a resource only deployed in site 2 (because you added a cross site contract), this is achieved by having the MSO pushing shadow resource on the peer site when cross site contract is pushed

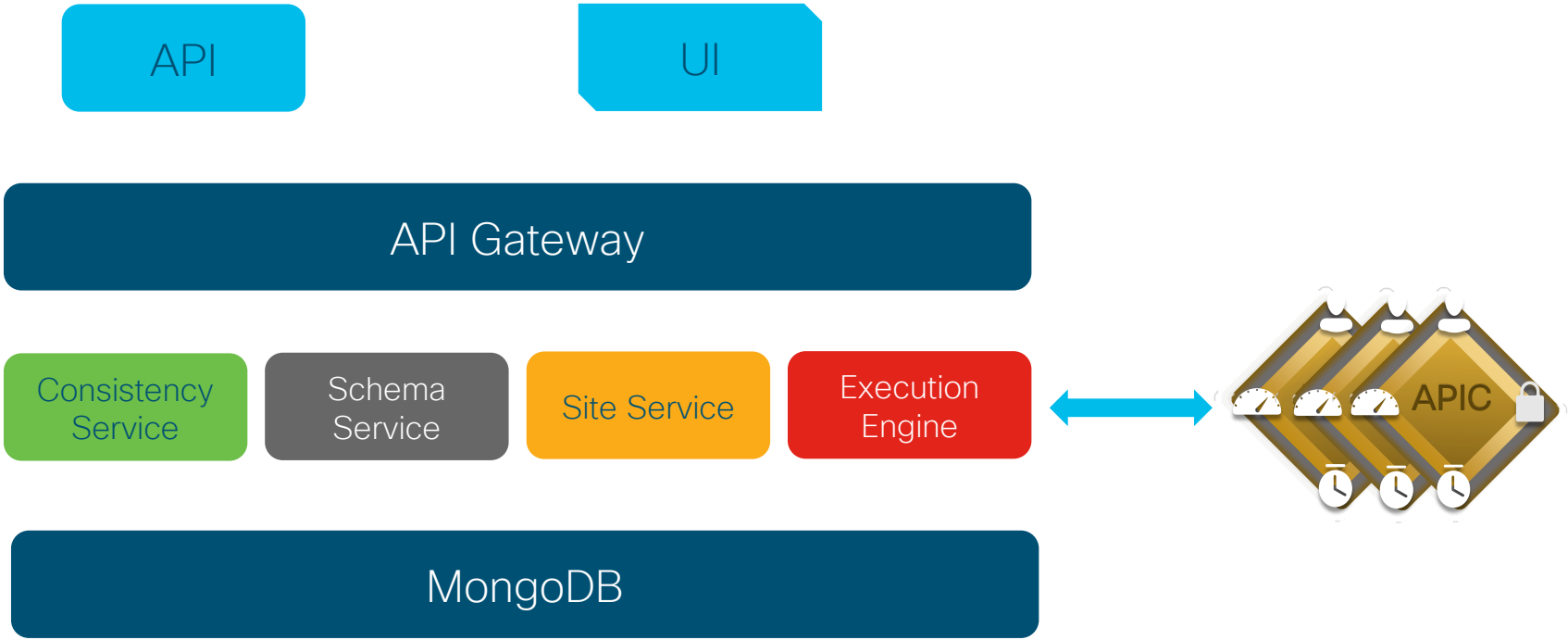
Example

- So if you want to stretch VRF but not BD/EPG you need different template
  - Template 1 : Stretched VRF pushed to both site
  - Template 2 : BD1 and EPG1 and subnet pushed to site 1 only
  - Template 3 : BD2 and EPG2 and subnet Pushed to site 2 only
- If you make a contract between EPG1 and EPG2 (cross site contract), we will push a shadow BD1/EPG1 to site2 and a shadow BD2/EPG2 to site1
- (See later section for more detail)

MSO Disaster and recovery

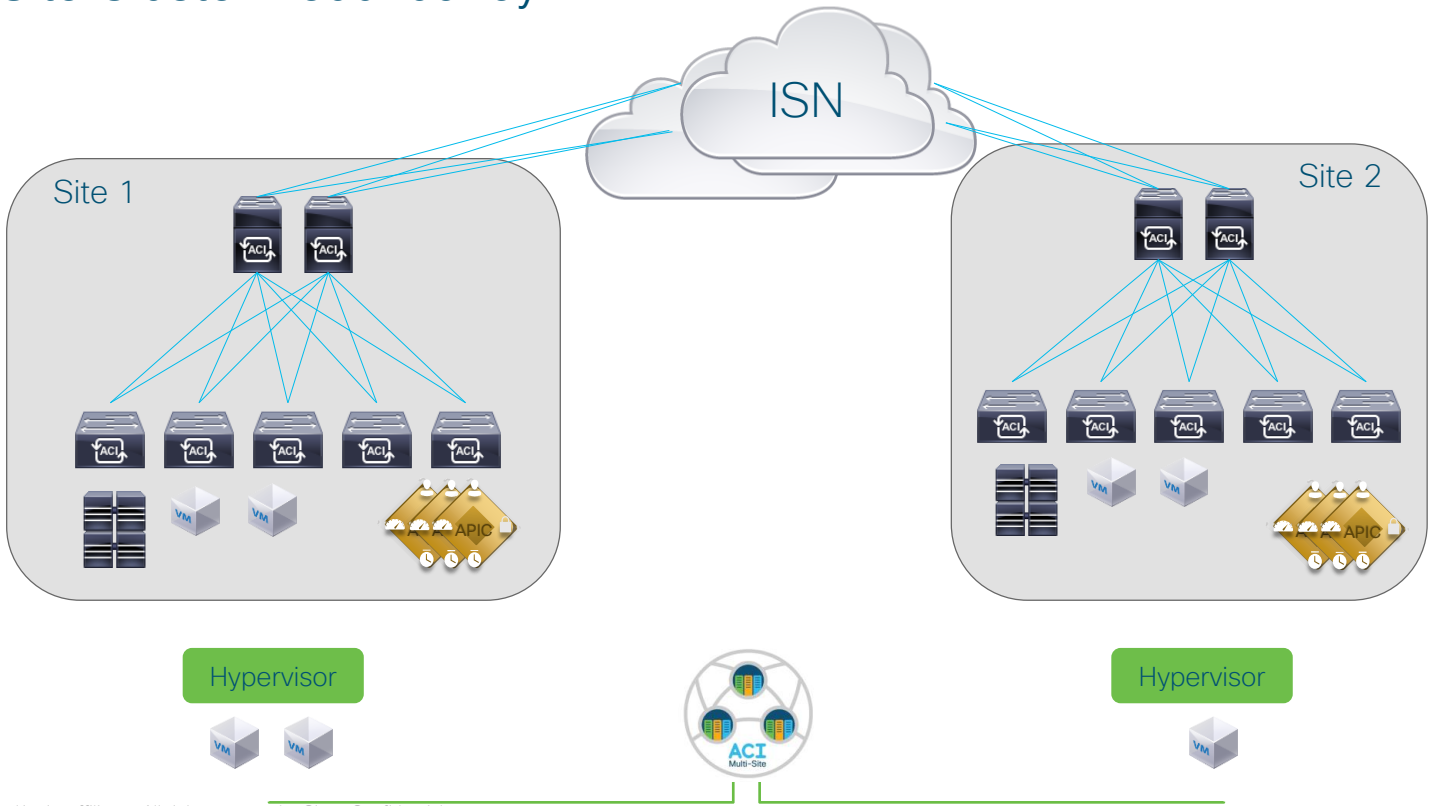
Disaster and Recovery

MSO Design – Microservice architecture



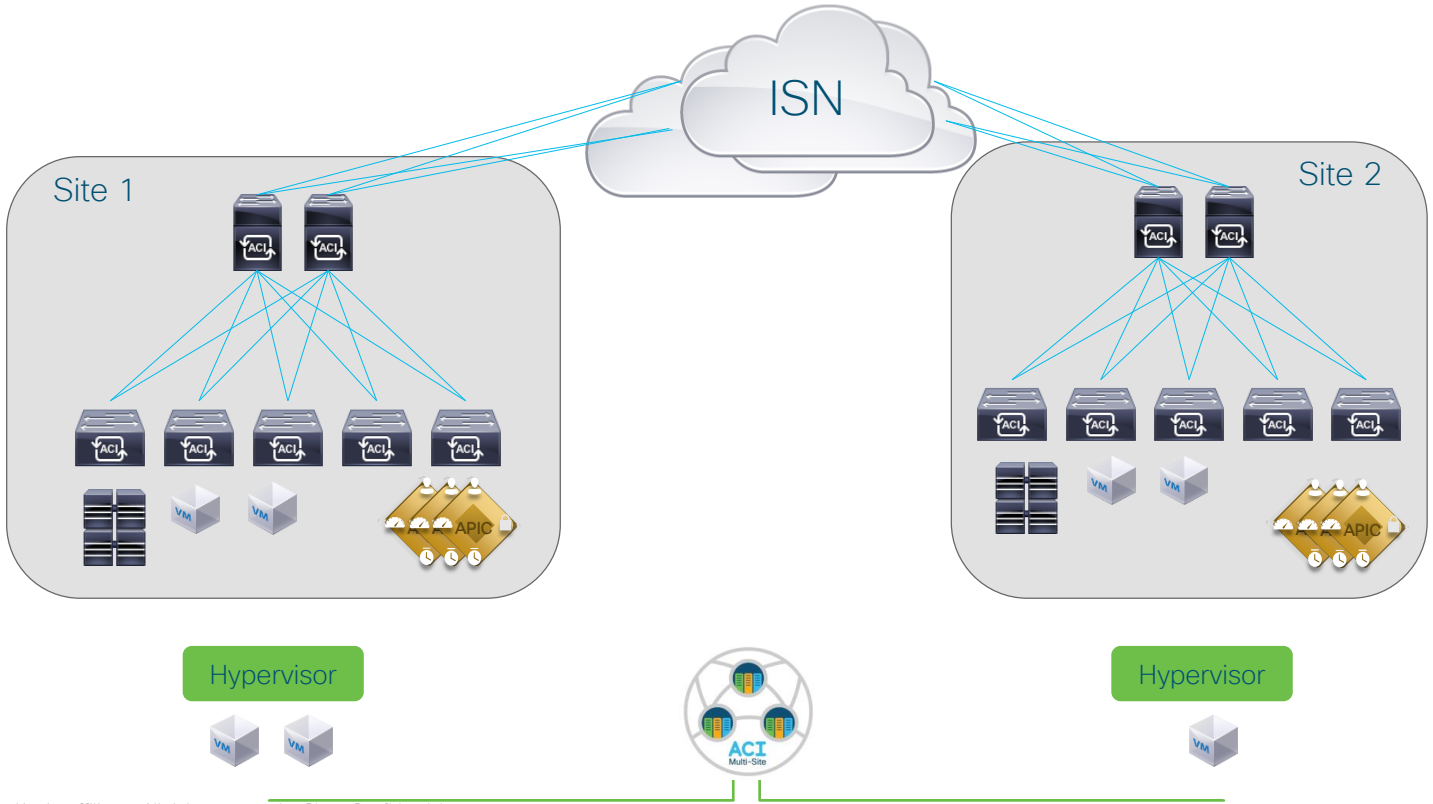
Disaster and Recovery

Multi-Site Cluster Redundancy



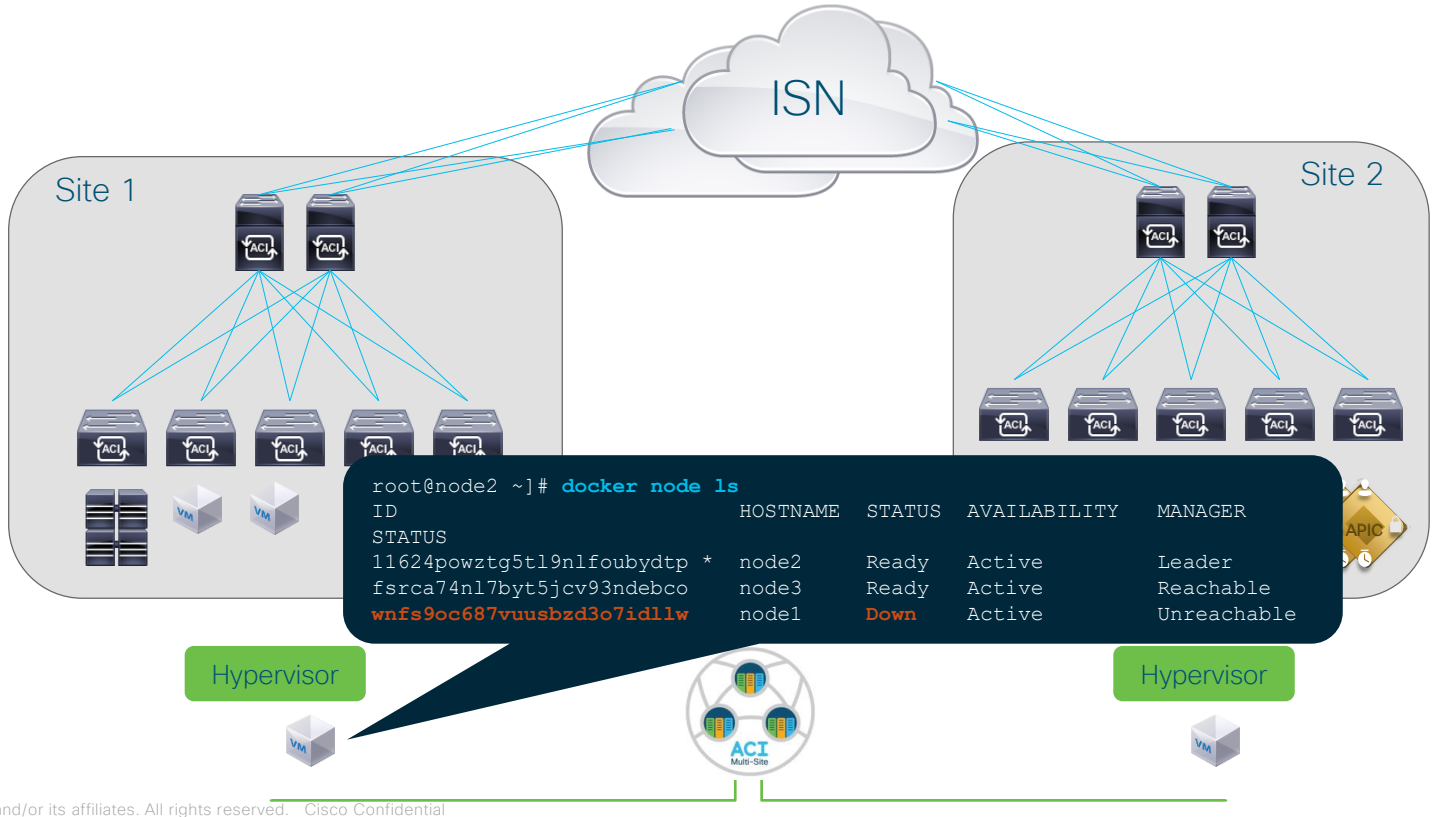
Disaster and Recovery

Single Node Failure



Disaster and Recovery

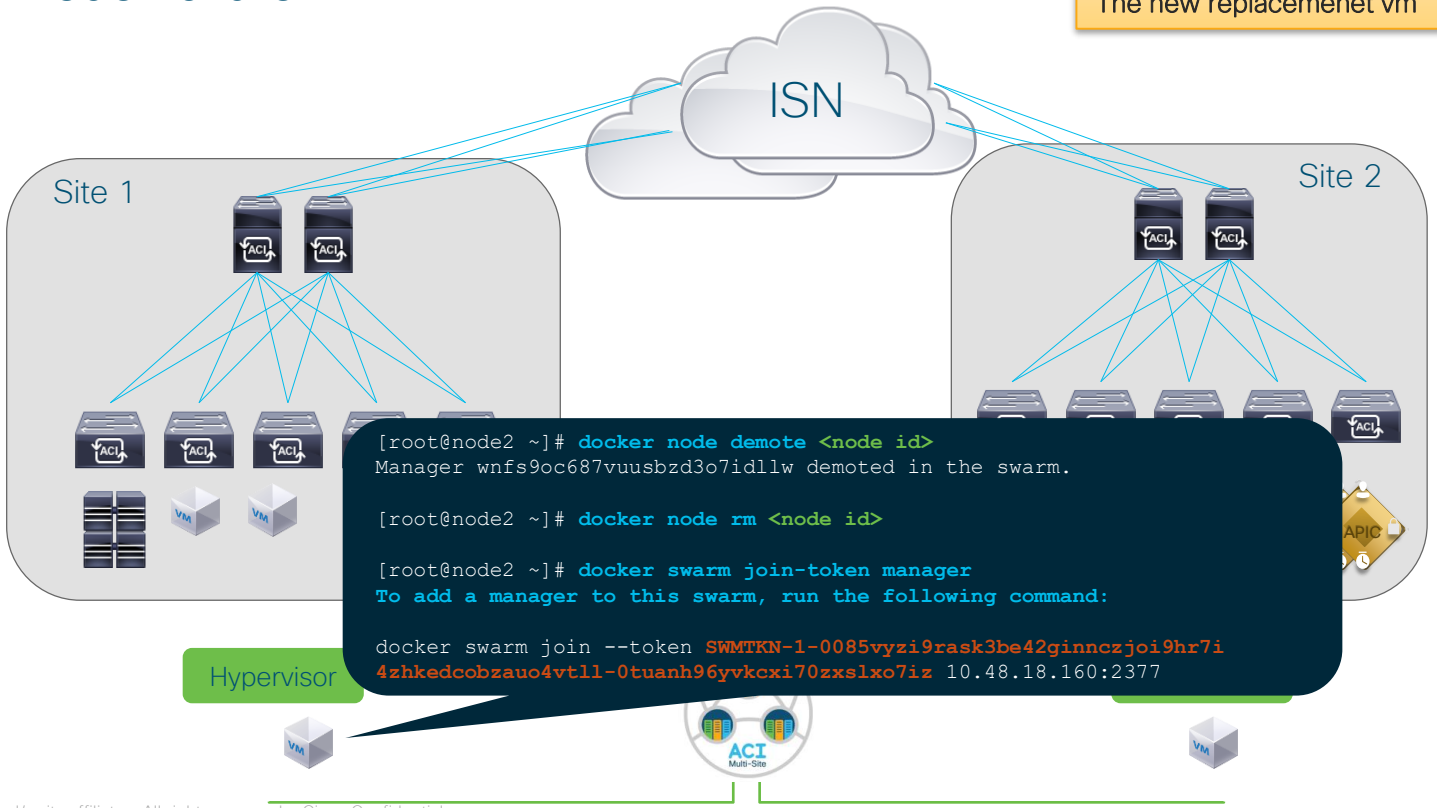
Single Node Failure



Disaster and Recovery

Single Node Failure

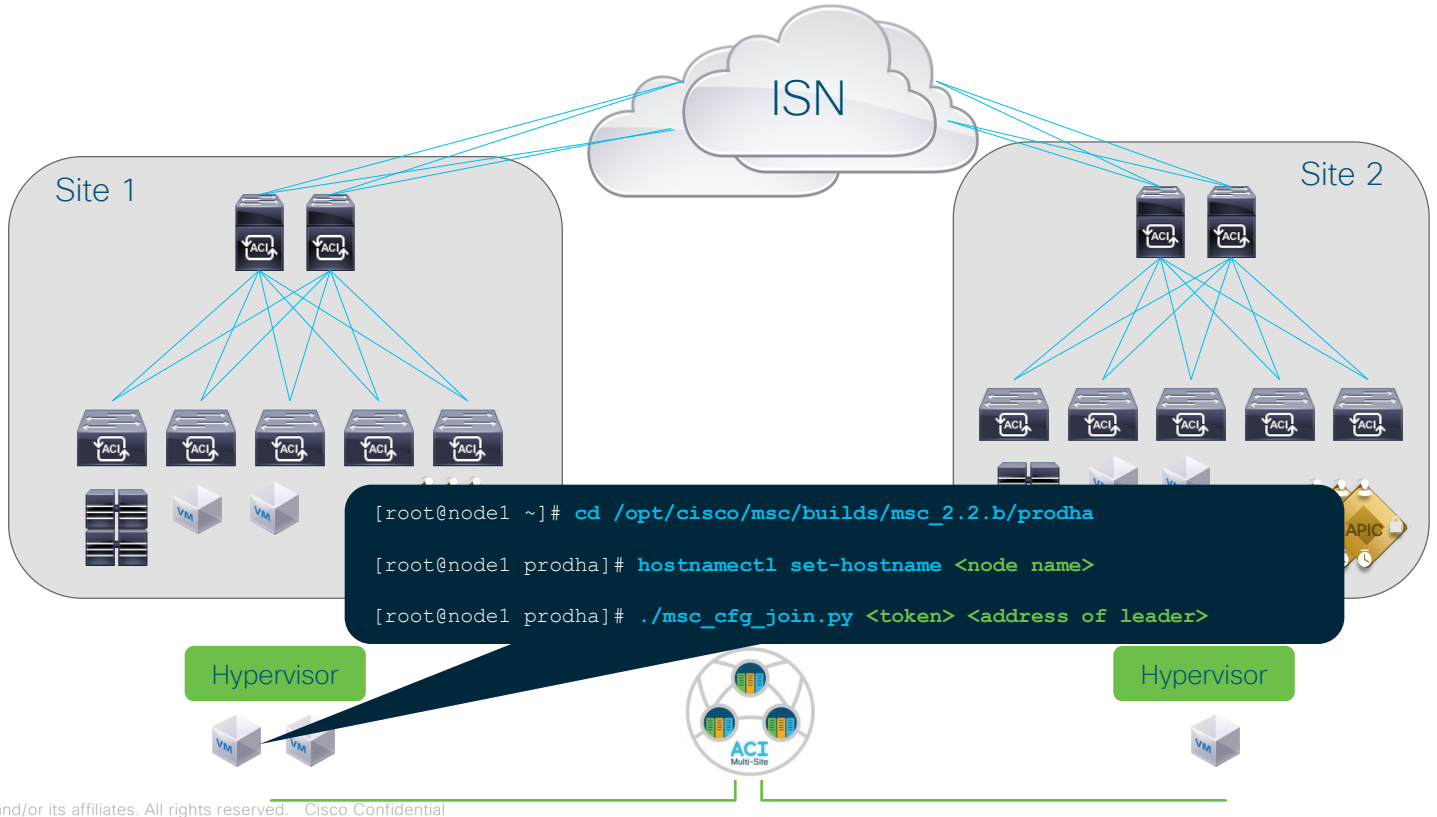
Node id to demote is here
Wnfs9oc687vuusbzd3o7idllw
Docker swarm join-token gen a
New token that we will use in
The new replacemenet vm



Disaster and Recovery

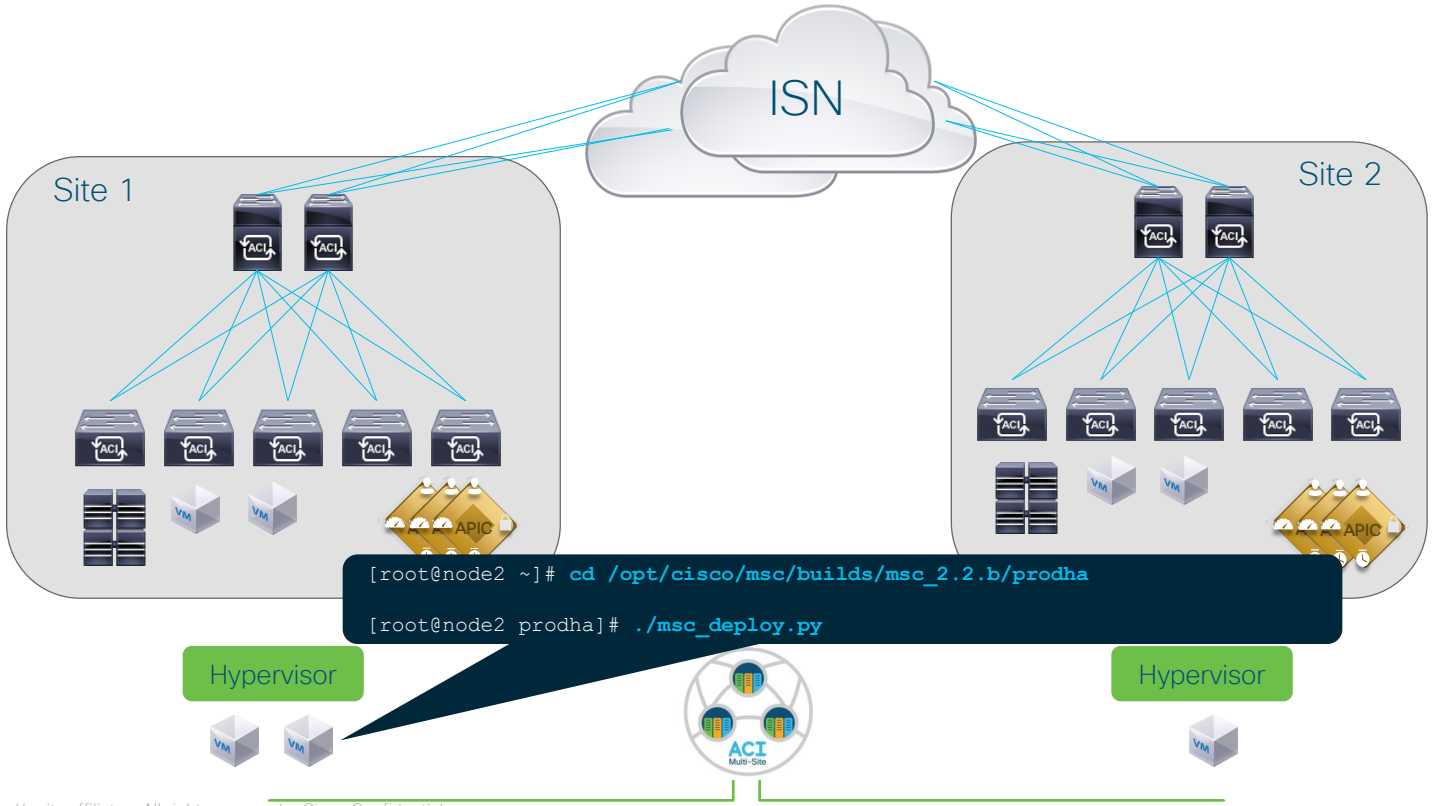
Single Node Failure

On new Node VM
Hostname must be node-X (here node-1)
Then run the join scrip



Disaster and Recovery

Single Node Failure



MSO Consistency Checker (3.2 and +)

Consistency Checker – Why is it Required?

- The deployment process in ACI Multi-Site is actually a two step process
  1. Provisioning policy objects on APICs
  2. Wait for APIC to generate vnids and pcTags, cross program these to other sites.

- Only the first step is deterministically complete when the user sees the “Successfully deployed” message on Multi-Site. The second step happens in the background after the first step has completed.
- Progress and success/failure status of the second step isn’t communicated to the user. This can lead to an odd state where the process failed and is the reason for a break in the flow of traffic, yet from a Multi-Site user’s perspective, it’s not entirely clear that something has gone wrong.

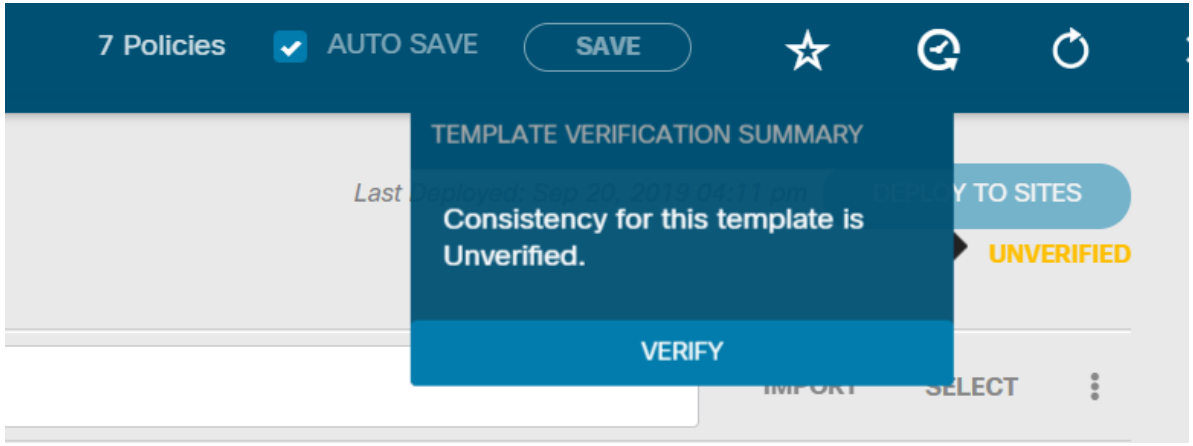
Consistency Checker – What it Achieves

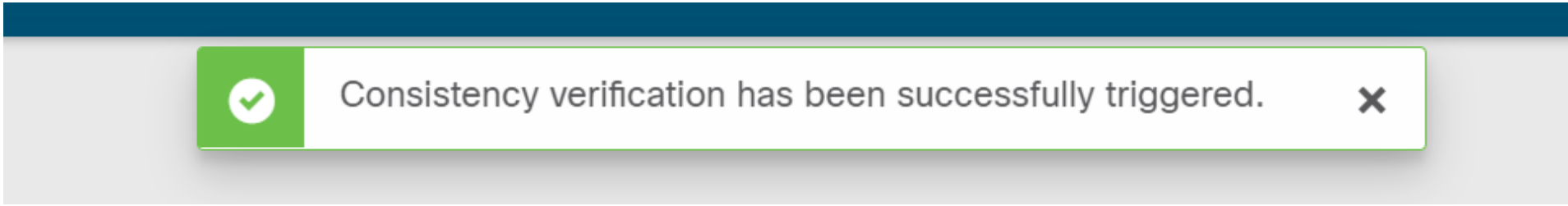
- Provides the user with a view into the success/failure state of the cross programming, alerting them when something goes wrong
- Provides Support people with a tool to look deeper into failures to help root cause them and provide solutions quicker
- Provides a per template-site view of cross programming status of EPGs, BDs, VRFs and External EPGs

- Consistency check can only be done for Template deployed to more than one site

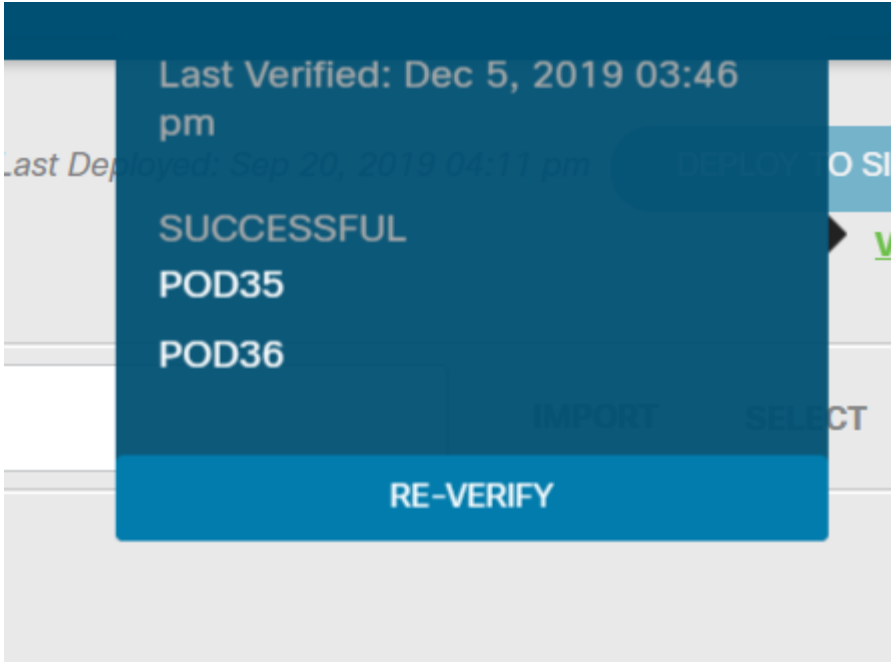
Trigger consistency checked

Here it was not verified since last template deploy
Hence we trigger verification

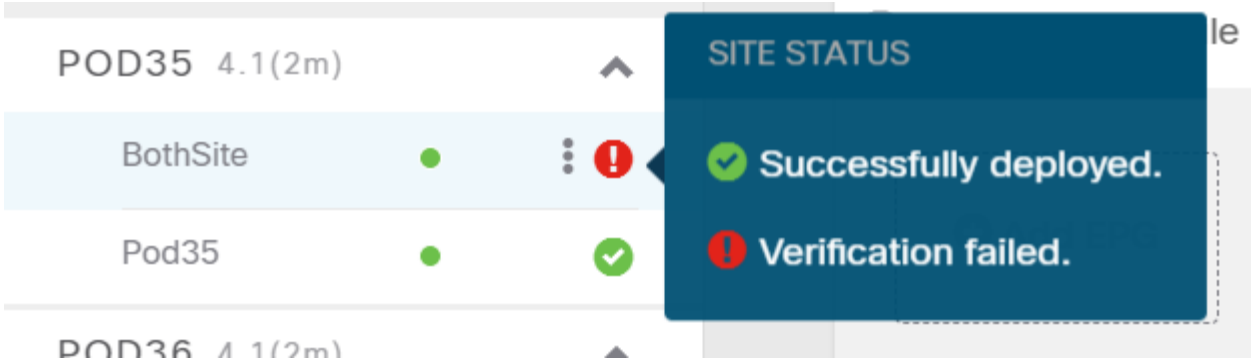




When consistency succeed ?

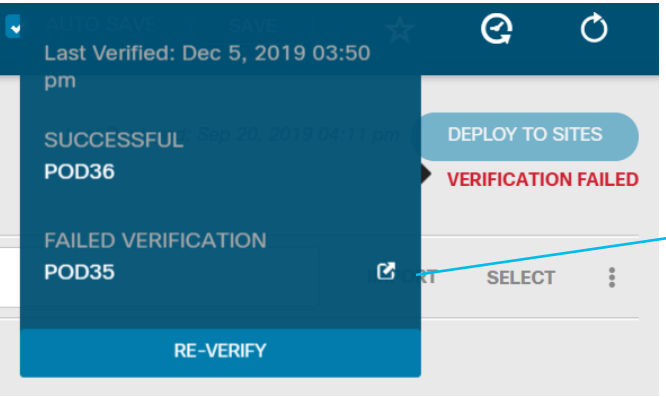


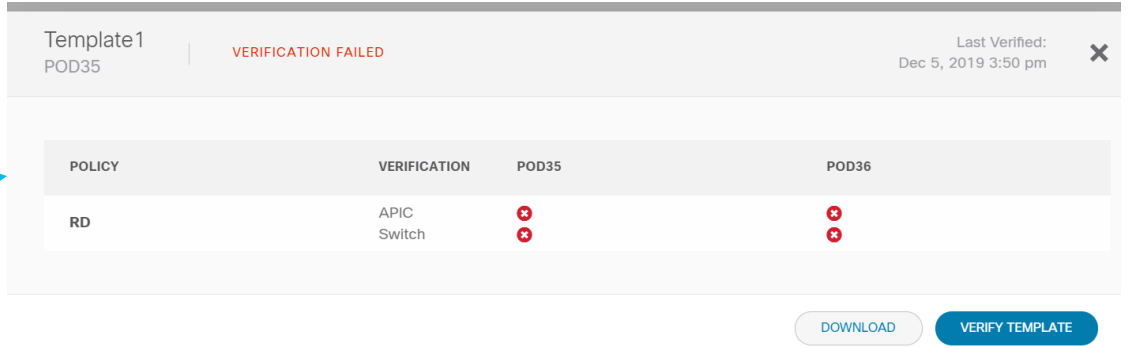
What if consistency Fails ?



What is Consistency fails ?

You can click to see name of not consistency policy
And download the report





```
"Final result" : "FAILED",
"Translation Report" : {
  "Comments" : [ {
    "Site Comments" : [ ],
    "Site Name" : "site-1"
  } ],
  "vrfs" : [ {
    "Overall Status" : "NOT OK",
    "Sites Report" : [ {
      "Local Site" : {
        "Site Name" : "site-1",
        "Spines" : [ {
          "bgpRtt" : {
            "status" : "Not OK"
          },
          "dciSubs" : {
            "encapOrPcTag" : "?",
            "pcTagStatus" : "Not OK"
          },
          "node" : "201"
        } ],
        "ctx Def" : {
          "encapOrPcTag" : "?",
          "pcTagStatus" : "Not OK"
        },
        "logical ctx" : {
          "encapOrPcTag" : "?",
          "pcTagStatus" : "Not OK"
        },
        "msc" : {
          "encapOrPcTag" : "2883584",
          "pcTagStatus" : "OK"
        }
      },
      "Remote Sites" : [ {
        "Site Name" : "site-2",
        "Spines" : [ {
          "bgpRtt" : {
            "status" : "Not OK"
          },
          "dciSubs" : {
            "encapOrPcTag" : "?",
            "pcTagStatus" : "Not OK"
          },
          "node" : "201"
        } ],
        "ctx Def" : {
          "encapOrPcTag" : "?",
          "pcTagStatus" : "Not OK"
        },
        "logical ctx" : {
          "encapOrPcTag" : "?",
          "pcTagStatus" : "Not OK"
        },
        "msc" : {
          "encapOrPcTag" : "2588673",
          "pcTagStatus" : "OK"
        }
```


Consistency checker report

We see the MO uni/tn-RD/ctx-RD is not OK
On site 1
(here it was deleted on APIC)


```
        }
      } ]
    } ],
    "common" : {
      "Context" : "?"
    },
    "dn" : "uni/tn-RD/ctx-RD",
    "mscRef" : "/schemas/5ce64ee30e000013048da8b8/templates/Template1/vrfs/RD"
  } ]
}
}
```

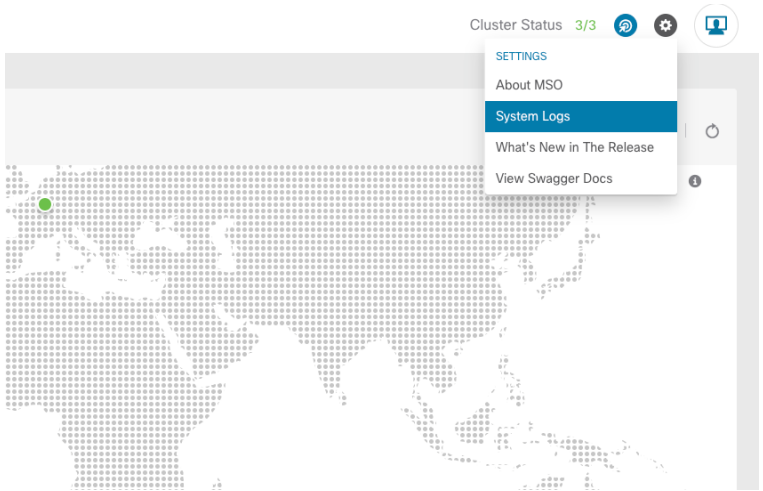
How to bring back consistency

- Redeploy template from MSO (with a little change)
- And run consistency check again

Multi-Site Orchestrator Troubleshooting

MSO Troubleshooting

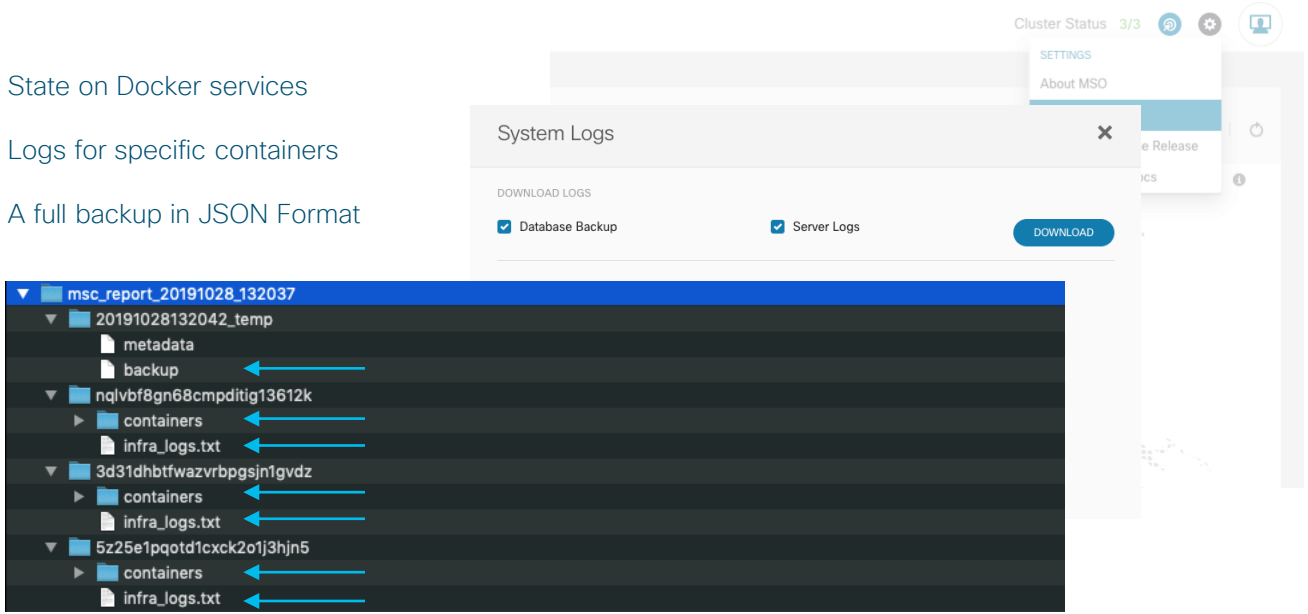
Generating Troubleshooting Report



MSO Troubleshooting

Generating Troubleshooting Report

- State on Docker services
- Logs for specific containers
- A full backup in JSON Format



MSO Troubleshooting

Inspect health of docker cluster

Execution engine is specially important as it push info to APIC

```
[root@node1 ~]# docker service ls

ID                  NAME                        MODE                REPLICAS            IMAGE                                          P
z41ny3pw9970        msc_cloudsecservice         replicated          1/1                 msc-cloudsecservice:2.2.2b
nvxo84bsnip4        msc_consistencyservice      replicated          1/1                 msc-consistencyservice:2.2.2b
mk3puj8qfm3b        msc_endpointservice         replicated          1/1                 msc-capic-sync:v4.2.34                         *
56inesks3e7p        msc_executionengine         replicated          1/1                 msc-executionengine:2.2.2b
hltytku6e5ip        msc_jobschedulerservice     replicated          1/1                 msc-jobschedulerservice:2.2.2b
ppgrncfh9ees        msc_kong                    global              3/3                 msc-kong:2.2.2b
ge5f2p6ehvn0        msc_kongdb                  replicated          1/1                 msc-postgres:9.4
4azb91e3qjca        msc_mongodb1                replicated          1/1                 msc-mongo:3.6
iwzdro1v1r71        msc_mongodb2                replicated          1/1                 msc-mongo:3.6
3ohpjlkq550h        msc_mongodb3                replicated          1/1                 msc-mongo:3.6
yhd0mddzwhzf        msc_platformservice         global              3/3                 msc-platformservice:2.2.2b
v5i60n9ou7rg        msc_policyservice           replicated          1/1                 msc-policyservice:2.2.2b                       *
si6nku7v0nk4        msc_schemaservice           global              3/3                 msc-schemaservice:2.2.2b
hkm339e6o5mv        msc_siteservice             global              3/3                 msc-siteservice:2.2.2b
l0rsw6jlmtf1        msc_syncengine              global              3/3                 msc-syncengine:2.2.2b
                                                                    3/3                 msc-ui:2.2.2b                                  *
                                                                    3/3                 msc-userservice:2.2.2b
```

The output is the expected health status.
All services have at least 1 container replicated.
If any of them is down, the system might not work correctly

Role Of execution engine

- Two kind of important log information
  - The Schema to push and the plan being generated
  - Websocket monitoring for cross vnid programming

- Look for
  - Log line with “error”
  - Stacktrace for exception

MSO Troubleshooting

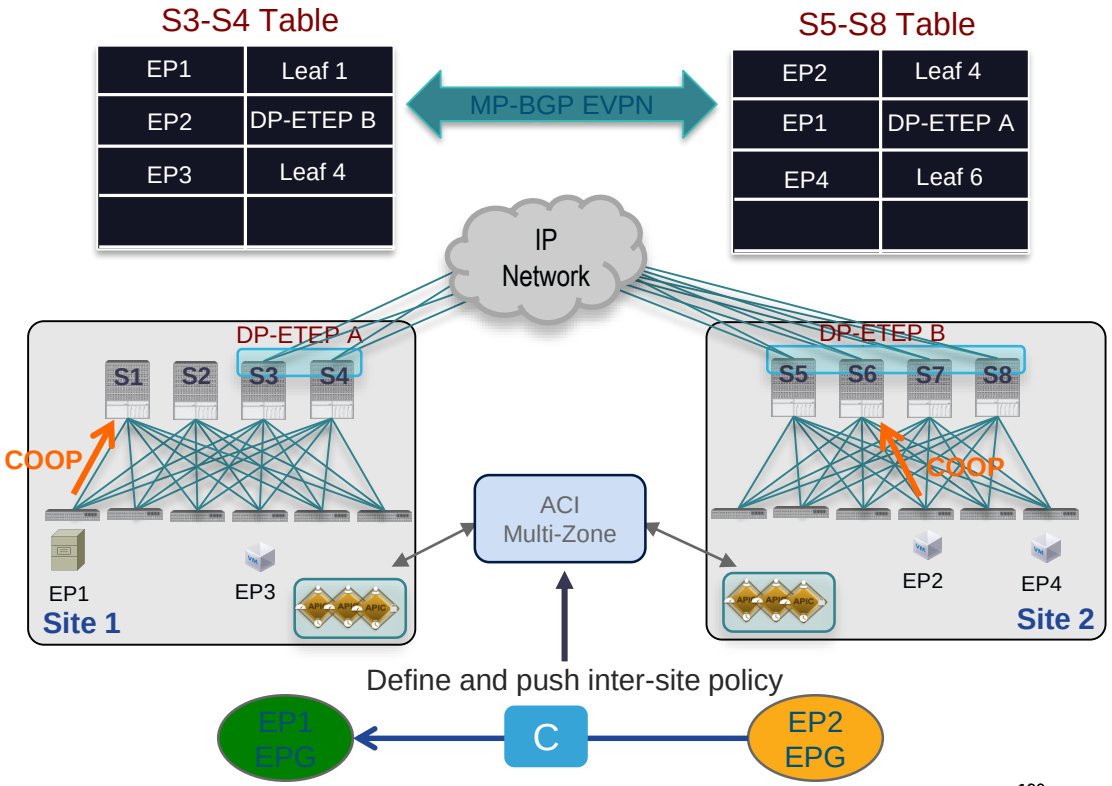
Getting container specific logs.

```
[root@node1 ~]# docker ps | egrep executionengine
8325bf8d9b70        msc-executionengine:2.2.2b      "bin/executionengine"    11 days ago         Up 11 days (healthy)

[root@node1 ~]# cd /var/lib/docker/containers
[root@node1 containers]# ls -al | egrep 8325bf8d9b70
drwx------.  4 root root 4096 Oct 28 06:50 8325bf8d9b7042448a260e75d4946df2edb508be391815c28524442098b74eff
[root@node1 ~]# cd 8325bf8d9b7042448a260e75d4946df2edb508be391815c28524442098b74eff

[root@node1 8325bf8d9b7042448a260e75d4946df2edb508be391815c28524442098b74eff]# ls -al
total 2060
drwx------.  4 root root    4096 Oct 28 06:50 .
drwx------. 38 root root    4096 Oct 17 04:10 ..
-rw-r-----.  1 root root 2067622 Oct 28 06:45 8325bf8d9b7042448a260e75d4946df2edb508be391815c28524442098b74eff-json.log
drwx------.  2 root root       6 Oct 17 04:06 checkpoints
-rw-------.  1 root root    6037 Oct 28 06:50 config.v2.json
-rw-r--r--.  1 root root    1468 Oct 28 06:50 hostconfig.json
-rw-r--r--.  1 root root      13 Oct 17 04:06 hostname
-rw-r--r--.  1 root root     173 Oct 17 04:06 hosts
drwx------.  3 root root      16 Oct 17 04:06 mounts
-rw-r--r--.  1 root root      38 Oct 17 04:06 resolv.conf
-rw-r--r--.  1 root root      71 Oct 17 04:06 resolv.conf.hash

[root@node1 ..]# less 8325bf8d9b7042448a260e75d4946df2edb508be391815c28524442098b74eff-json.log
```

Control Plane and Data plane High level

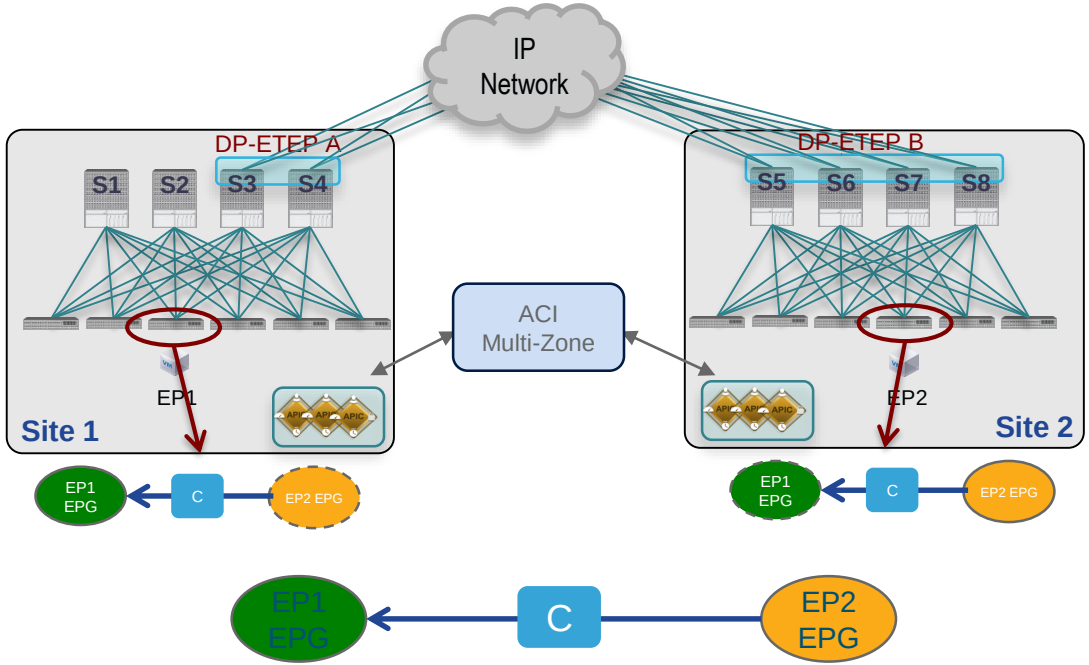
ACI Multi-Site
Inter-Site MP-BGP EVPN Control Plane

- MP-BGP EVPN used to communicate Endpoint (EP) information across Sites
  - MP-iBGP or MP-EBGP peering supported across sites
  - Remote host route entries (EVPN Type-2) are associated to the remote site Anycast DP-ETEP address
- Automatic filtering of endpoint information across Sites
  - Host routes are exchanged **only** if there is a cross-site contract requiring communication between endpoints

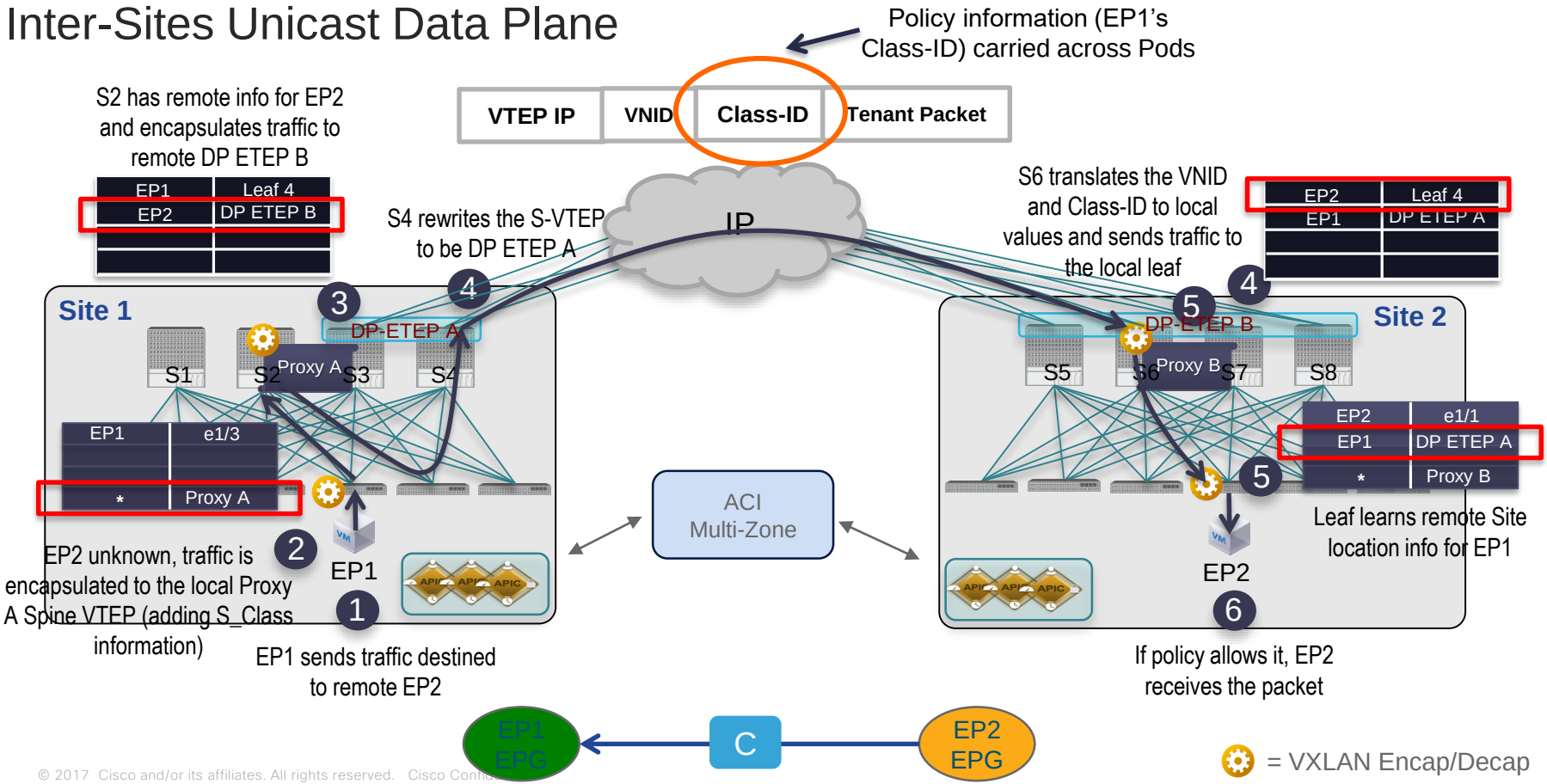


ACI Multi-Site
Inter-Site Policies

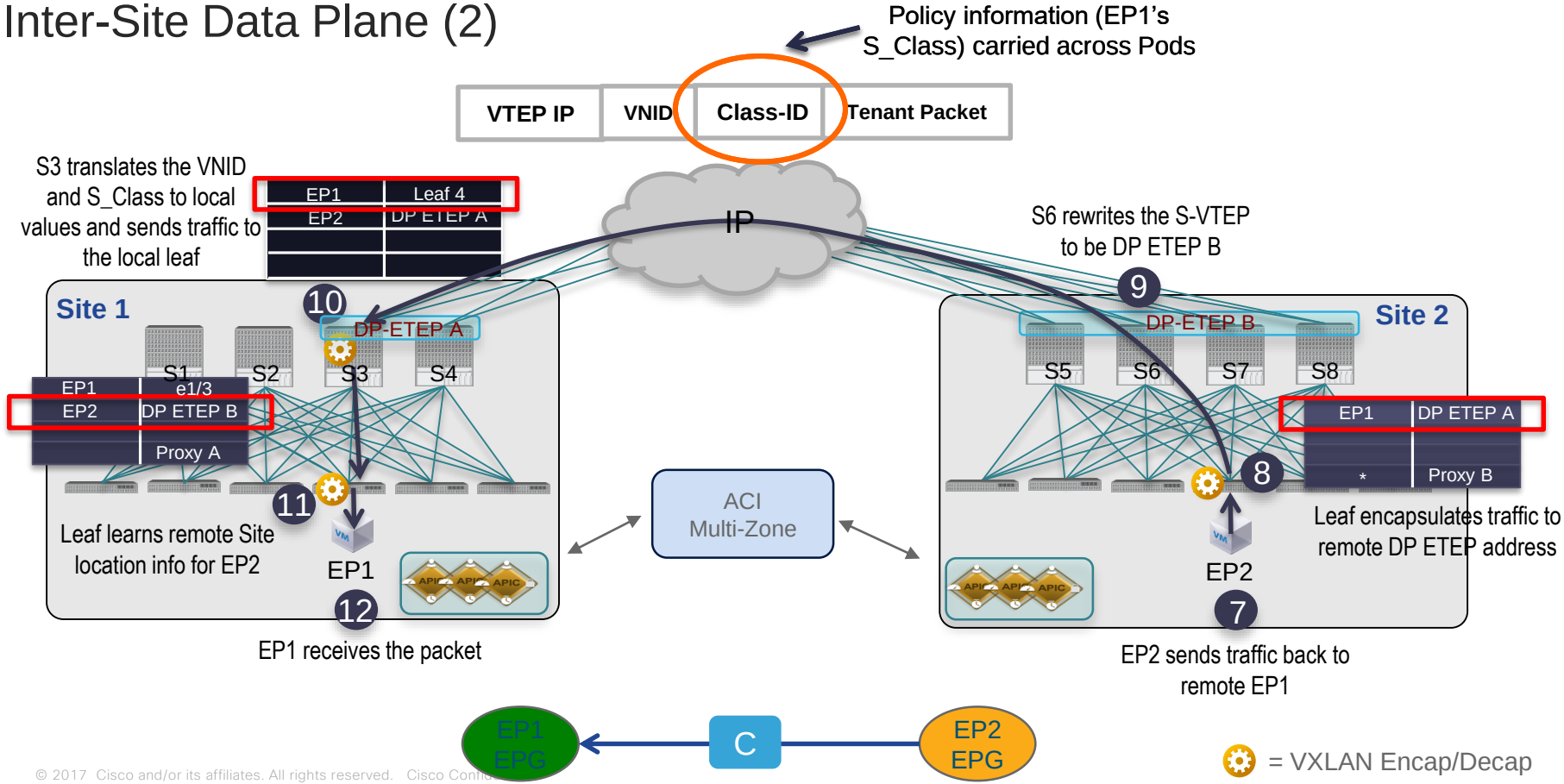
- Inter-Site policies defined on the ACI Multi-Zone are pushed to the respective APIC domains
- Policies are enforced at the ingress leaf node, once it has learned on the data plane info for remote endpoint



ACI Multi-Site
Inter-Sites Unicast Data Plane



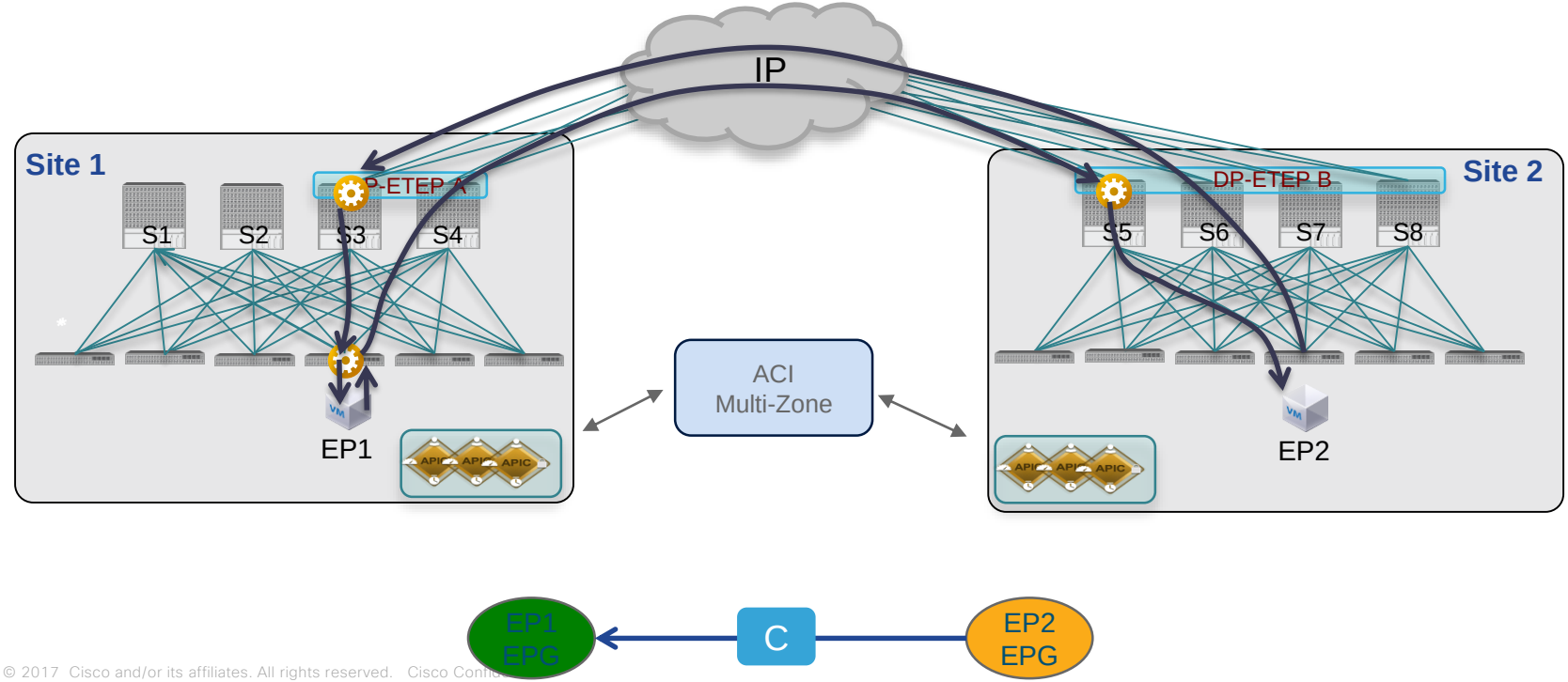
ACI Multi-Site
Inter-Site Data Plane (2)



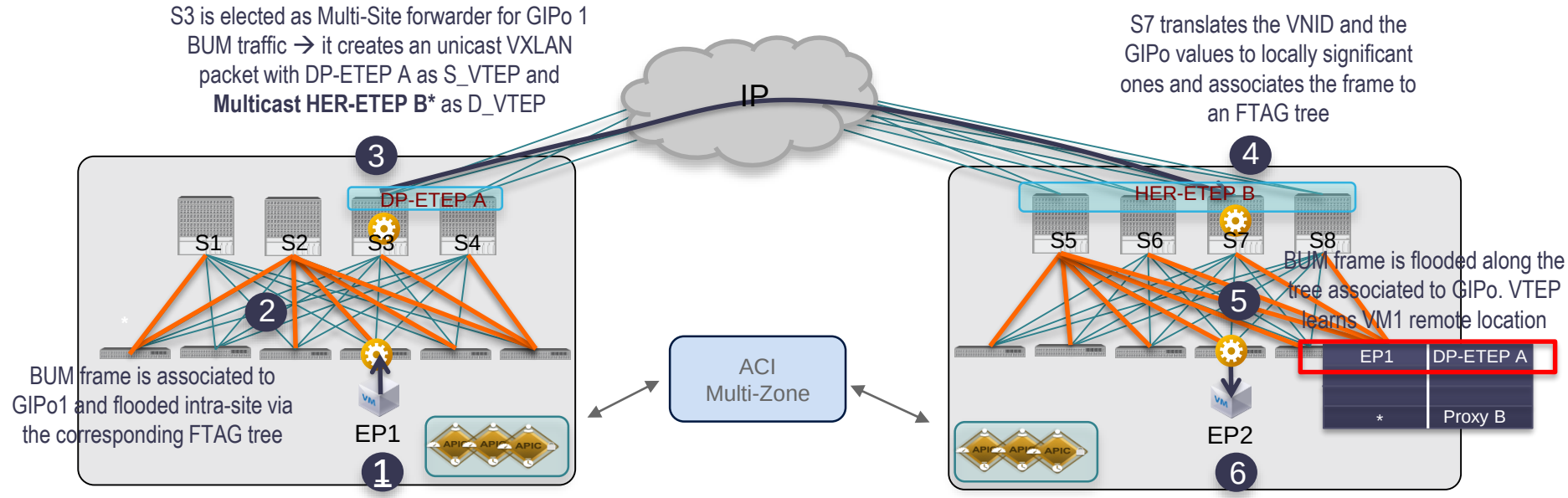
ACI Multi-Site
Inter-Site Data Plane (3)

= VXLAN Encap/Decap

From this point EP1 to EP2 communication is encapsulated Leaf to Remote Spine DP ETEPs in both directions

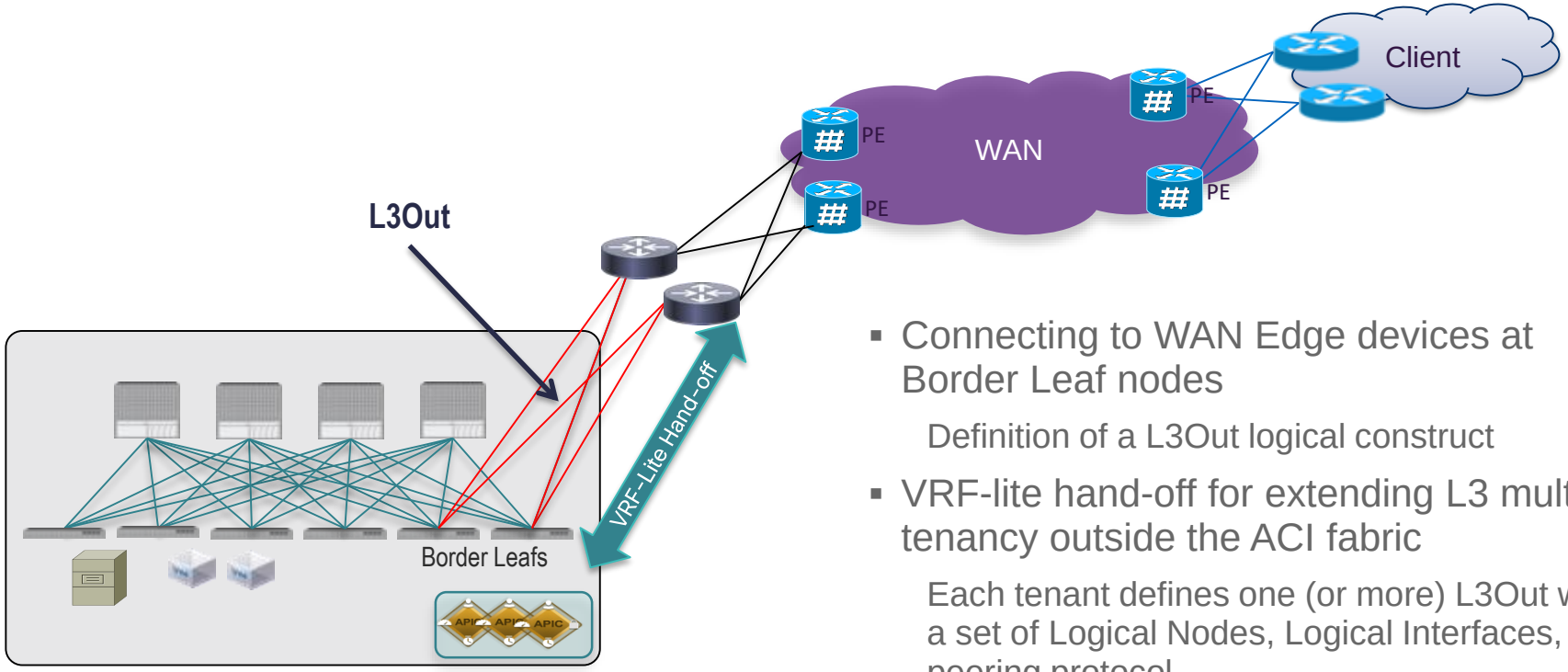


ACI Multi-Site
Layer 2 BUM Traffic Data Plane



*This is a different ETEP address than the one used for inter-site L3 unicast communication

Connecting ACI to Layer 3 Domain
‘Traditional’ L3Out on the BL Nodes



- Connecting to WAN Edge devices at Border Leaf nodes
  - Definition of a L3Out logical construct
- VRF-lite hand-off for extending L3 multi-tenancy outside the ACI fabric
  - Each tenant defines one (or more) L3Out with a set of Logical Nodes, Logical Interfaces, peering protocol

Translation troubleshooting Object Model

Multisite Translation Requirements

- In a typical APIC multisite environment, a particular site may expose certain services which other sites may desire to consume.
  - An example would be a shared DNS service provided by one of the sites which other sites want to use.
- The other common use case is to allow peer-to-peer communication between a pair of EPgs across sites.
  - This requires context stretching between the sites to give an illusion of same-context communication between the sites.

Translation Mechanism

- The policy mechanism we define to achieve this is *stretched EPgs, BDs, Contexts and InstPs*.
- MSO creates identical contracts on multiple sites with similar policy hierarchies and filters or make sure that the contract is replicated from one site to other sites.

Translation Mechanism

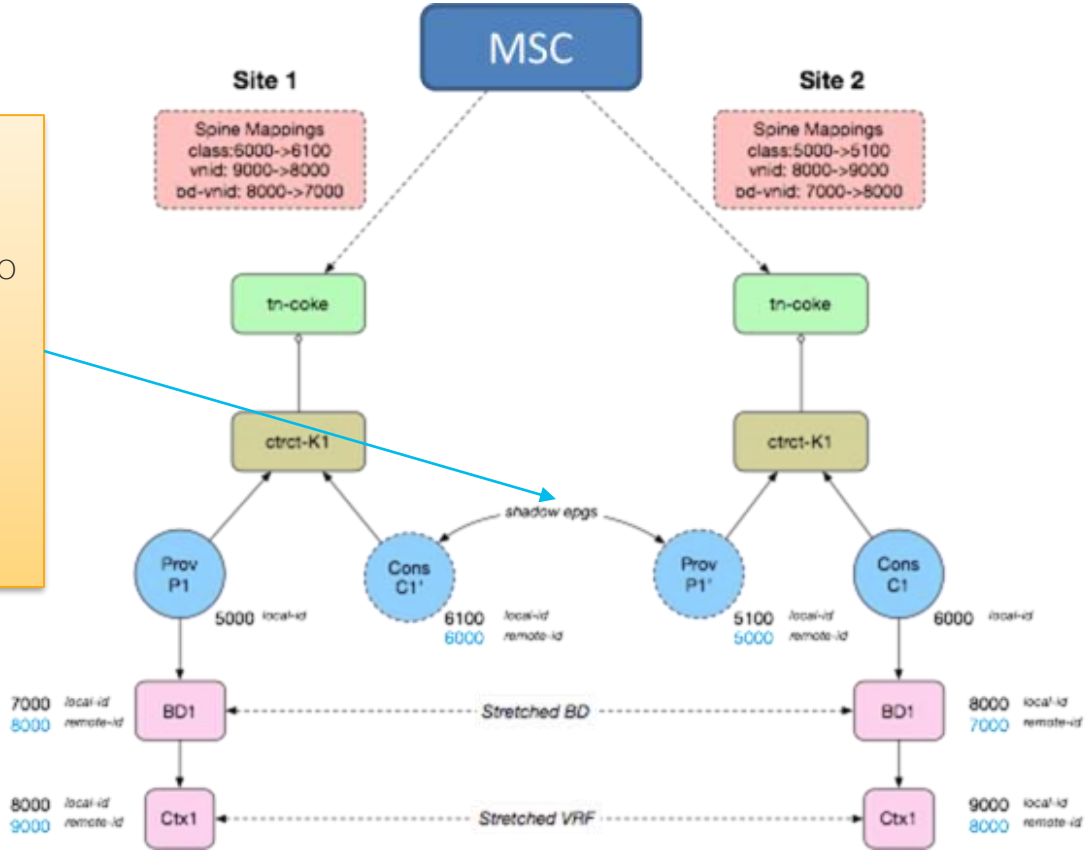
- The controller needs to be told which EPgs need to be stretched to other sites. It will then create identical EPgs on remote sites and associate them with the contract and configure the class ID mappings under it.
- Conceptually this can be viewed as *stretching* an EPg onto remote sites, though this EPg may not be physically deployed onto the TOR switches on remote sites.
- Note: Contracts should exist but EPGs specific actrlRules would exist on the ToR depending on Deployment Immediacy flag set to Immediate or Lazy.
- MSC has the flag present under static port deployment of EPGs.

BD Stretching
2 site local EPG in same BD

Here we have a stretched BD and EPG are not stretched
(local to each site)
EPG P1 is in site1
EPG C1 is in site2
Once there is contract between P1 and C1 (aka once in MSO you provide the same Contract on P1 as what you consume on C1) → this is what we call a cross site contract

This trigger creation of Shadow EPG on the other site
Shadow EPG P1' is created on site2
Shadow EPG C1' is created on site1

Shadow EPG have child object containing the needed translation of pcTag (see later slide)



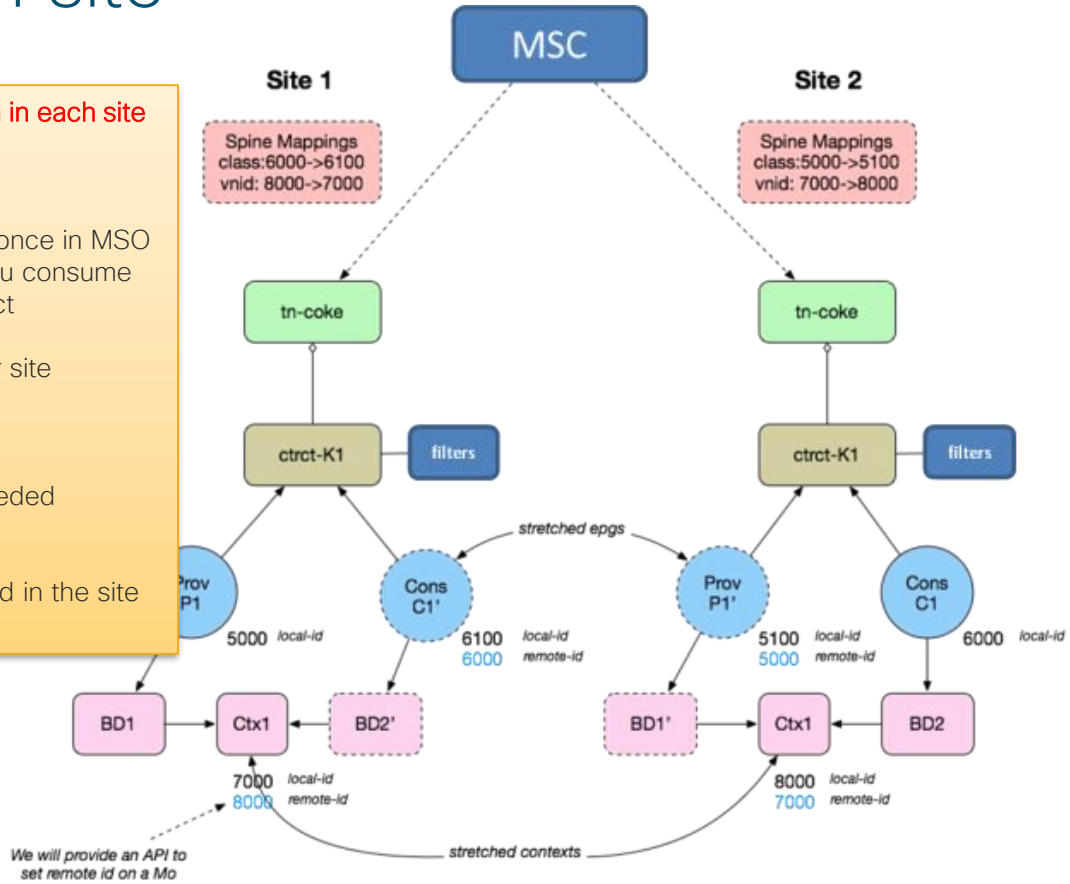
BD local to each site

Here we have a 2 BD one per site and one EPG in each site (local to each site)
EPG P1 in BD1 is in site1
EPG C1 in BD2 is in site2
Once there is contract between P1 and C1 (aka once in MSO you provide the same Contract on P1 as what you consume on C1) → this is what we call a cross site contract

This trigger creation of Shadow EPG on the other site
Shadow EPG P1' is created on site2
Shadow EPG C1' is created on site1

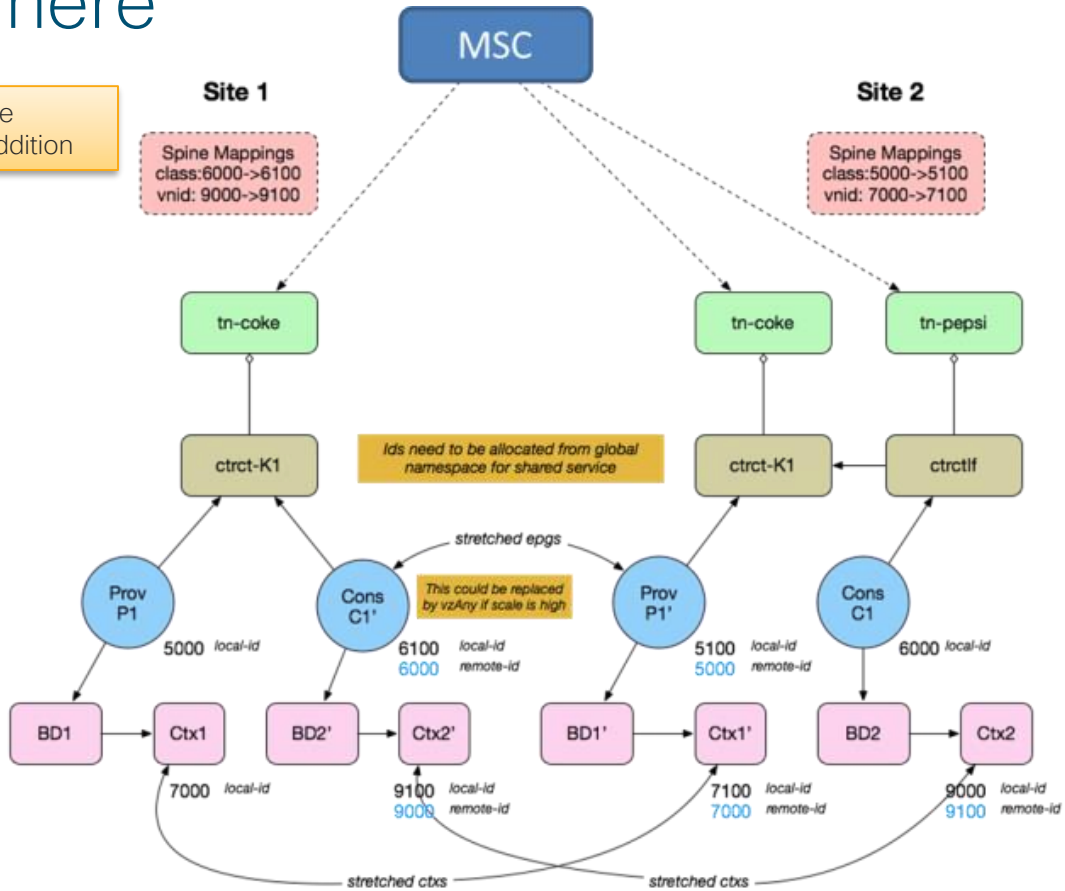
Shadow EPG have child object containing the needed translation of pcTag (see later slide)

Given the shadow EPG are in BD not yet deployed in the site we have shadow BD created as well

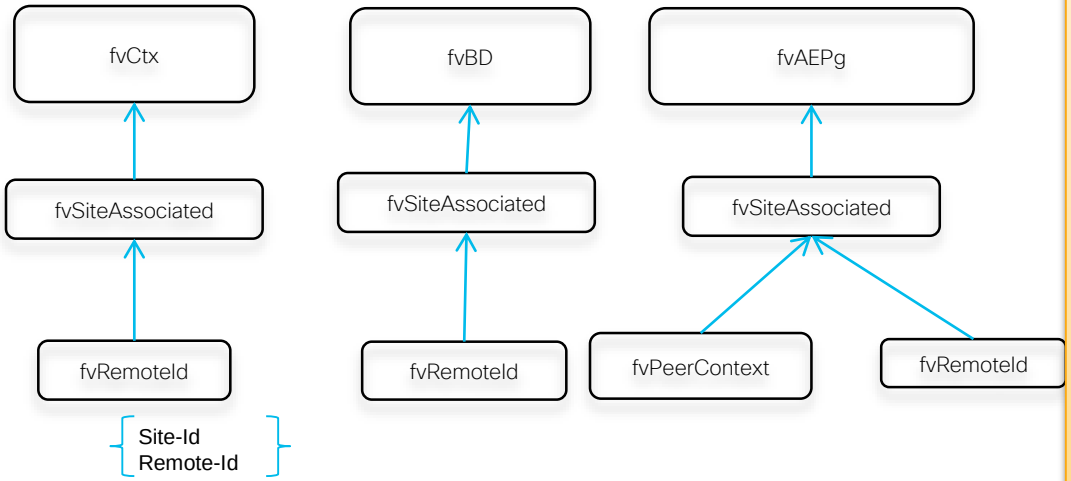


Shared Service EPG Stretching VRF is local to site here

Similar to previous scenario but we also need to create Shadow VRF to allow to have the translation child in addition



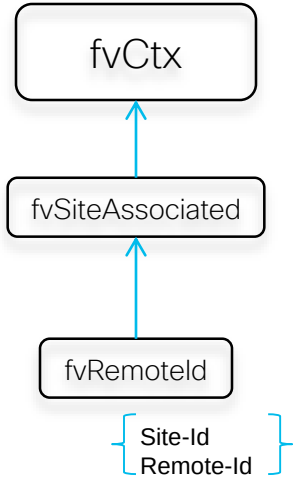
Logical Model



There are few reasons to extend Ressources (VRF, BD, EPG) in multisite

1. Either the resource exist locally in each site (EPG has EP in both side, or BD as EPG in both side even if EPG are not extended)
2. Either EPG or BD or VRF is fully local but it needs to communicate to a resource in the remote site (EPG/BD/VRF only in remote side) . This is usually happening when a cross site contract is set (or if preferred group is in use post 4.1)

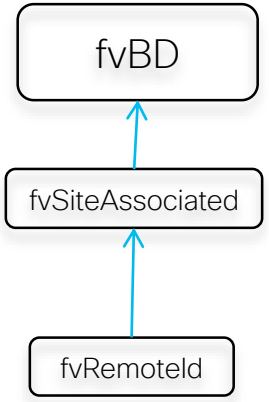
In both case we need to create translation of vnid and/or pcTag.
In Scenario 2, we need to create shadow resource as well to represent the remote resource that do not exist locally (shadow EPG or BD or VRF)




```
bdsol-aci35-apic1# moquery -d uni/tn-RD-L2/ctx-L2/stAsc
Total Objects shown: 1

# fv.SiteAssociated
childAction  :
descr        :
dn           : uni/tn-RD-L2/ctx-L2/stAsc
lcOwn        : local
modTs        : 2018-05-03T03:14:39.572+00:00
monPolDn     : uni/tn-common/monepg-default
name         : msc-local
nameAlias    :
ownerKey     :
ownerTag     :
rn           : stAsc
siteId       : 1
status       :
uid          : 15374
```

```
bdsol-aci35-apic1# moquery -d uni/tn-RD-L2/ctx-L2/stAsc/site-2
Total Objects shown: 1

# fv.RemoteId
siteId          : 2
childAction     :
descr           :
dn              : uni/tn-RD-L2/ctx-L2/stAsc/site-2
lcOwn           : local
modTs           : 2018-05-03T03:14:40.895+00:00
monPolDn        : uni/tn-common/monepg-default
name            :
nameAlias       :
ownerKey        :
ownerTag        :
remoteCtxPcTag  : 32770
remotePcTag     : 2162688
rn              : site-2
status          :
uid             : 15374
```

Logical BD – site 1



Site 2 BD

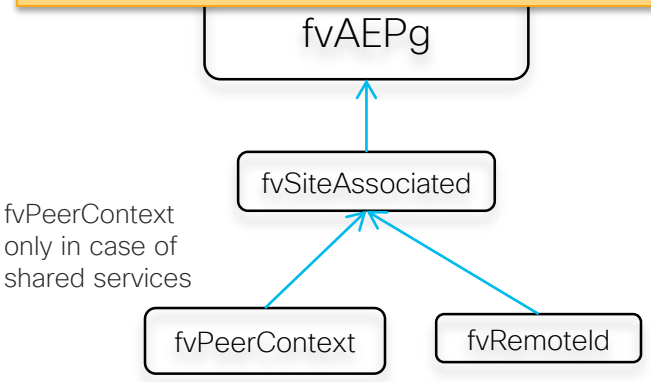
```
bdsol-aci36-apic1# moquery -c fvBD  -f 'fv.BD.seg
=="15073234"' | egrep "dn|scope|seg"
dn                          : uni/tn-RD-L2/BD-Web
scope                       : 2162688
seg                         : 15073234
```

```
bdsol-aci35-apic1# moquery -d  uni/tn-RD-L2/BD-Web/stAsc/
Total Objects shown: 1

# fv.SiteAssociated
childAction  :
descr        :
dn           : uni/tn-RD-L2/BD-Web/stAsc/stAsc
lcOwn        : local
modTs        : 2018-05-03T03:14:39.572+00:00
monPolDn     : uni/tn-common/monepg-default
name         : msc-local
nameAlias    :
ownerKey     :
ownerTag     :
rn           : stAsc
siteId       : 1
status       :
uid          : 15374
```

```
bdsol-aci35-apic1# moquery -d  uni/tn-RD-L2/BD-Web/stAsc/site-2
Total Objects shown: 1

# fv.RemoteId
siteId          : 2
childAction     :
descr           :
dn              : uni/tn-RD-L2/BD-Web/stAsc/site-2
lcOwn           : local
modTs           : 2018-05-03T03:14:40.895+00:00
monPolDn        : uni/tn-common/monepg-default
name            :
nameAlias       :
ownerKey        :
ownerTag        :
remoteCtxPcTag  : any
remotePcTag     : 15073234   Actually remote BD VNID
rn              : site-2
status          :
uid             : 15374
```

Logical shadow fvAEPg – site 1

EPG exist on site 2 →site 1 has a “shadow EPG” with same DN and as child as the fvSite Associated from the site where the epg is really deployed (here site2)



Site 2 REAL EPG

```
bdsol-aci36-apic1# moquery -c fvAEPg  -f 'fv.AEPg.pcTag
=="49155"'  | egrep "dn|scope|pcTag"
dn                  : uni/tn-RD-L2/ap-App/epg-Web
pcTag               : 49155
scope               : 2162688
```

Site 1 shadow EPG

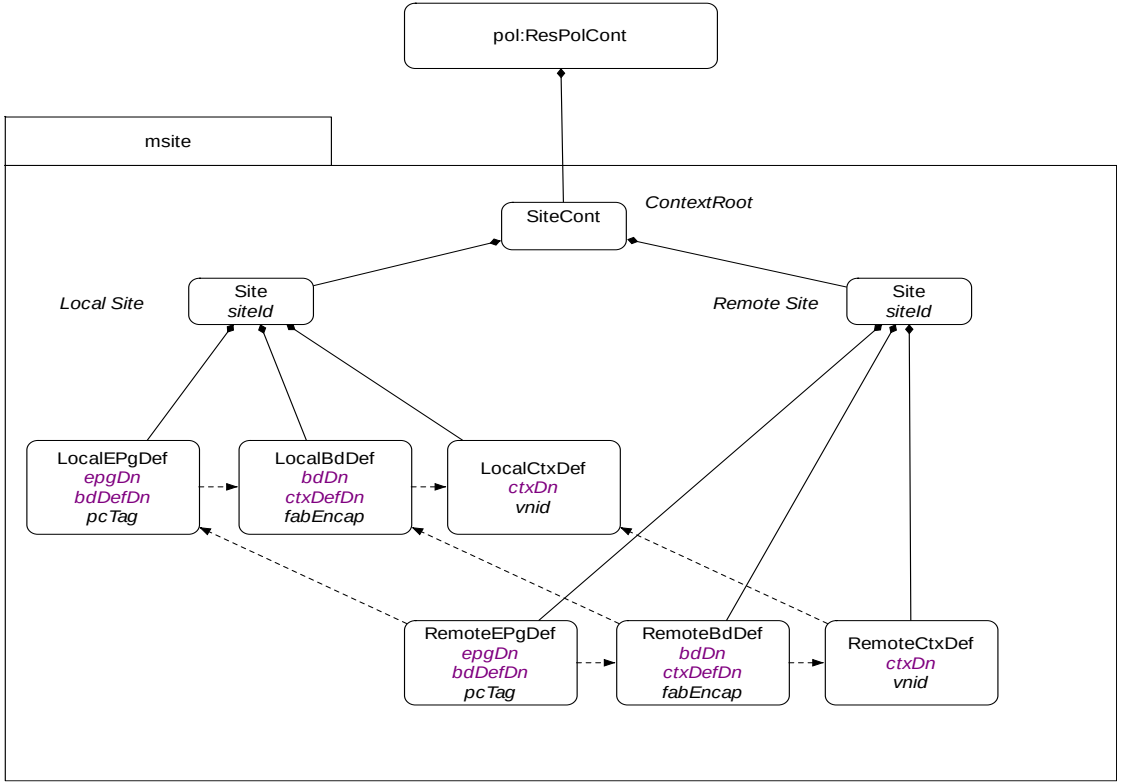
```
bdsol-aci35-apic1#  moquery -d uni/tn-RD/ap-App/epg-EPG36
Total Objects shown: 1
# fv.AEPg
dn                  : uni/tn-RD/ap-App/epg-Web
pcEnfPref           : unenforced
pcTag               : 16388
scope               : 2588673
```

```
bdsol-aci35-apic1# moquery -d uni/tn-RD-L2/ap-App/epg-Web/stAsc/
Total Objects shown: 1

# fv.SiteAssociated
childAction   :
descr         :
dn            : uni/tn-RD-L2/ap-App/epg-Web/stAsc/stAsc
lcOwn         : local
modTs         : 2018-05-03T03:14:39.572+00:00
monPolDn      : uni/tn-common/monepg-default
name          : msc-local
rn            : stAsc
siteId        : 1
```

Site 1 remote id for Xlate

```
d uni/tn-RD-L2/ap-App/epg-
Total Objects shown: 1

# fv.RemoteId
siteId          : 2
childAction     :
descr           :
dn              : uni/tn-RD-L2/ap-App/epg-Web/stAsc/site-2
ownerTag        :
remoteCtxPcTag  : any
remotePcTag     : 49155
rn              : site-2
status          :
uid             : 15374
```

Resolved Model



```
# fv.Site
siteId       : 1
childAction  :
dn           : resPolCont/sitecont/site-1/site-1
lcOwn        : local
modTs        : 2018-03-30T05:50:38.595+00:00
name         : msc-local
rn           : site-1
```

```
# fv.Site
siteId       : 2
childAction  :
dn           : resPolCont/sitecont/site-2/site-2
lcOwn        : local
modTs        : 2018-03-30T05:50:38.595+00:00
name         :
rn           : site-2
```

```
# fv.LocalEPgDef
moDn            : uni/tn-RD-L2/ap-App/epg-Web
LocalBdDefDn    : resPolCont/sitecont/site-1/localbddef-[uni/tn-RD-L2/BD-Web]
LocalCtxDefDn   : resPolCont/sitecont/site-1/localctxdef-[uni/tn-RD-L2/ctx-L2]
LocalDefDn      : resPolCont/sitecont/site-1/localepgdef-[uni/tn-RD-L2/ap-App/epg-Web]
childAction     :
dn              : resPolCont/sitecont/site-1/localepgdef-[uni/tn-RD-L2/ap-App/epg-Web]
lcOwn           : local
modTs           : 2018-05-03T03:14:40.903+00:00
pcTag           : 32771
rn              : localepgdef-[uni/tn-RD-L2/ap-App/epg-Web]
```

```
# fv.RemoteEPgDef
moDn            : uni/tn-RD-L2/ap-App/epg-Web
LocalDefDn      : resPolCont/sitecont/site-1/localepgdef-[uni/tn-RD-L2/ap-App/epg-Web]
RemoteBdDefDn   : resPolCont/sitecont/site-2/remotebddef-[uni/tn-RD-L2/BD-Web]
RemoteCtxDefDn  : resPolCont/sitecont/site-2/remotectxdef-[uni/tn-RD-L2/ctx-L2]
childAction     :
dn              : resPolCont/sitecont/site-2/remoteepgdef-[uni/tn-RD-L2/ap-App/epg-Web]
lcOwn           : local
modTs           : 2018-05-03T03:14:40.903+00:00
pcTag           : 49155
rn              : remoteepgdef-[uni/tn-RD-L2/ap-App/epg-Web]
```

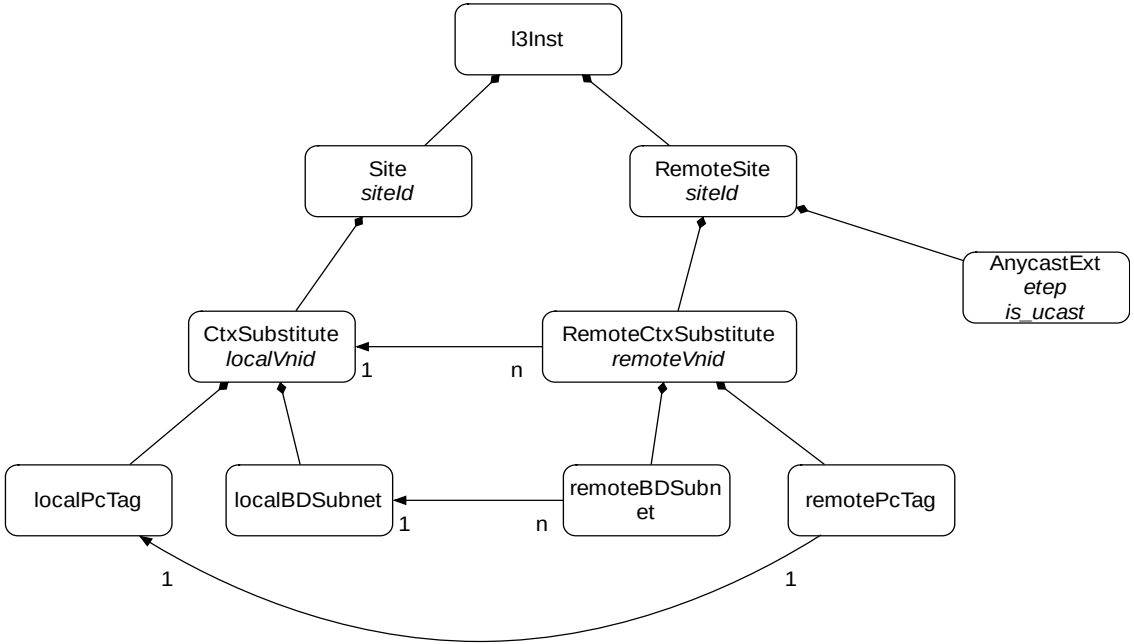
```
# fv.LocalBdDef
moDn                      : uni/tn-RD-L2/BD-Web
LocalCtxDefDn             : resPolCont/sitecont/site-1/localctxdef-[uni/tn-RD-L2/ctx-L2]
LocalDefDn                : resPolCont/sitecont/site-1/localbddef-[uni/tn-RD-L2/BD-Web]
bcastP                    : 225.0.216.80
childAction               :
dn                        : resPolCont/sitecont/site-1/localbddef-[uni/tn-RD-L2/BD-Web]
intersiteBumTrafficAllow  : yes
intersiteL2Stretch        : yes
lcOwn                     : local
modTs                     : 2018-05-03T03:14:40.903+00:00
pcTag                     : 15204288
rn                        : localbddef-[uni/tn-RD-L2/BD-Web]
```

```
# fv.RemoteBdDef
moDn            : uni/tn-RD-L2/BD-Web
LocalDefDn      : resPolCont/sitecont/site-1/localbddef-[uni/tn-RD-L2/BD-Web]
RemoteCtxDefDn  : resPolCont/sitecont/site-2/remotectxdef-[uni/tn-RD-L2/ctx-L2]
childAction     :
dn              : resPolCont/sitecont/site-2/remotebddef-[uni/tn-RD-L2/BD-Web]
lcOwn           : local
modTs           : 2018-05-03T03:14:40.903+00:00
pcTag           : 15073234
rn              : remotebddef-[uni/tn-RD-L2/BD-Web]
status          :
```

```
# fv.LocalCtxDef
moDn          : uni/tn-RD-L2/ctx-L2
LocalDefDn    : resPolCont/sitecont/site-1/localctxdef-[uni/tn-RD-L2/ctx-L2]
childAction   :
dn            : resPolCont/sitecont/site-1/localctxdef-[uni/tn-RD-L2/ctx-L2]
lcOwn         : local
modTs         : 2018-05-03T03:14:40.901+00:00
pcTag         : 2457600
rn            : localctxdef-[uni/tn-RD-L2/ctx-L2]
```

```
# fv.RemoteCtxDef
moDn          : uni/tn-RD-L2/ctx-L2
LocalDefDn    : resPolCont/sitecont/site-1/localctxdef-[uni/tn-RD-L2/ctx-L2]
childAction   :
dn            : resPolCont/sitecont/site-2/remotectxdef-[uni/tn-RD-L2/ctx-L2]
lcOwn         : local
modTs         : 2018-05-03T03:14:40.901+00:00
pcTag         : 2162688
rn            : remotectxdef-[uni/tn-RD-L2/ctx-L2]
status        :
```

Concrete Model



```
# dci.LocalSite
id            : 1
childAction   :
dn            : topology/pod-1/node-201/sys/inst-overlay-
1/localSite-1
lcOwn         : local
modTs         : 2018-03-30T05:50:38.558+00:00
name          : msc-local
rn            : localSite-1
```

```
# dci.RemoteSite
id            : 2
childAction   :
dn            : topology/pod-1/node-201/sys/inst-overlay-
1/remoteSite-2
lcOwn         : local
modTs         : 2018-03-30T05:50:38.558+00:00
name          :
rn            : remoteSite-2
```

```
# l3.LocalCtxSubstitute
FabEncap      : vxlan-2457600
DnName        : uni/tn-RD-L2/ctx-L2
childAction   :
dn            : topology/pod-1/node-201/sys/inst-overlay-1/localSite-1/localCtxSubstitute-
[vxlan-2457600]
lcOwn         : local
mcastEncap    : 0.0.0.0
modTs         : 2018-05-03T03:14:40.863+00:00
rn            : localCtxSubstitute-[vxlan-2457600]
status        :
```

```
# l3.RemoteCtxSubstitute
FabEncap      : vxlan-2162688
DnName        : uni/tn-RD-L2/ctx-L2
childAction   :
dn            : topology/pod-1/node-201/sys/inst-overlay-1/remoteSite-
2/remoteCtxSubstitute-[vxlan-2162688]
lcOwn         : local
modTs         : 2018-05-03T03:14:40.863+00:00
rn            : remoteCtxSubstitute-[vxlan-2162688]
status        :
```

```
# l3.RtToLocalCtxSubstitute
tDn           : topology/pod-1/node-201/sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-
[vxlan-2162688]
childAction   :
dn            : topology/pod-1/node-201/sys/inst-overlay-1/localSite-1/localCtxSubstitute-
[vxlan-2457600]/rttoLocalCtxSubstitute-[topology/pod-1/node-201/sys/inst-overlay-
1/remoteSite-2/remoteCtxSubstitute-[vxlan-2162688]]
rn            : rttoLocalCtxSubstitute-[topology/pod-1/node-201/sys/inst-overlay-
1/remoteSite-2/remoteCtxSubstitute-[vxlan-2162688]]
tCl           : l3RemoteCtxSubstitute
```

Child of local VRF Concrete.
tDn points to remote VRF concrete

```
# l2.LocalBdSubstitute
FabEncap      : vxlan-15204288
DnName        : uni/tn-RD-L2/BD-Web
childAction   :
ctrl          : bum-traffic
dn            : topology/pod-1/node-201/sys/inst-overlay-1/localSite-1/localCtxSubstitute-
[vxlan-2457600]/localBdSubstitute-[vxlan-15204288]
lcOwn         : local
mcastEncap    : 225.0.216.80
rn            : localBdSubstitute-[vxlan-15204288]
```

```
# l2.RemoteBdSubstitute
FabEncap      : vxlan-15073234
DnName        : uni/tn-RD-L2/BD-Web
childAction   :
dn            : topology/pod-1/node-201/sys/inst-overlay-1/remoteSite-
2/remoteCtxSubstitute-[vxlan-2162688]/remoteBdSubstitute-[vxlan-15073234]
lcOwn         : local
modTs         : 2018-05-03T03:14:40.868+00:00
rn            : remoteBdSubstitute-[vxlan-15073234]
```

```
# l2.LocalPcTagSubstitute
pcTag         : 32771
DnName        : uni/tn-RD-L2/ap-App/epg-Web
childAction   :
dn            : topology/pod-1/node-201/sys/inst-overlay-1/localSite-
1/localCtxSubstitute-[vxlan-2457600]/localPcTagSubstitute-32771
lcOwn         : local
modTs         : 2018-05-03T03:14:40.868+00:00
rn            : localPcTagSubstitute-32771
```

```
# l2.RemotePcTagSubstitute
pcTag         : 49155
DnName        : uni/tn-RD-L2/ap-App/epg-Web
childAction   :
dn            : topology/pod-1/node-201/sys/inst-overlay-1/remoteSite-
2/remoteCtxSubstitute-[vxlan-2162688]/remotePcTagSubstitute-49155
lcOwn         : local
modTs         : 2018-05-03T03:14:40.868+00:00
rn            : remotePcTagSubstitute-49155
```

Translation troubleshooting on switch (nxos) - dcimgr

Process involved for vnidmap/sclass/site-etep

Dcimgr/dcimc/sdkTraces for sclass/vnid map

- Dcimgr (on sup)
  - show dcimgr internal event-history events

And log file :

```
pod35-spine1# ls -al /var/sysmgr/tmp_logs/dcimgr.log
-rw-rw-rw- 1 root root 3162338 May  2 16:37 /var/sysmgr/tmp_logs/dcimgr.log
pod35-spine1#
```

- **HAL CLI :**

```
module-2# show platform internal  hal objects dci ?
  all             Dump All HAL objects
  remotesite      Remotesite or wan instance
  remotesiteetep  Unicast etep that belongs to this remotesite
  remotevrfvnid   Vrf for remotesite object
  sclassmap       Sclass mapping for remotesite vrf
  vnidmap         Vnid mapping for remotesite object
```

Dcimgr trace

Dcimgr consume concrete object to create translation

```
pod35-spine1# show dcimgr internal event-history events  | egrep -A 1 Event
1) Event:E_DEBUG, length:150, at 135850 usecs after Tue Apr 24 14:40:18 2018
    [1835623268] gr_objstore_bdvnid_map_mts_hdlr: [Create]: sys/inst-overlay-1/localSite-1/localCtxSubstitute-[vxlan-2490368]/localBdSubstitute-
[vxlan-15040468]
--
2) Event:E_DEBUG, length:278, at 135813 usecs after Tue Apr 24 14:40:18 2018
    [1835623268] gr_objstore_bdvnid_map_mts_hdlr: [Create]: sys/inst-overlay-1/localSite-1/localCtxSubstitute-[vxlan-2490368]/localBdSubstitute-
[vxlan-15040468]/rttoLocalBdSubstitute-[sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-[vxlan-2392064]/remoteBdSubstitute-[vxlan-15335344]]
--
3) Event:E_DEBUG, length:265, at 277412 usecs after Tue Apr 24 14:38:04 2018
    [1835623268] gr_objstore_sclass_map_mts_hdlr: [Create]: sys/inst-overlay-1/localSite-1/localCtxSubstitute-[vxlan-2490368]/localPcTagSubstitute-
16386/rttoLocalPcTagSubstitute-[sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-[vxlan-2392064]/remotePcTagSubstitute-16386]
--
4) Event:E_DEBUG, length:206, at 276933 usecs after Tue Apr 24 14:38:04 2018
    [1835623268] gr_objstore_vnid_map_mts_hdlr: [Create]: sys/inst-overlay-1/localSite-1/localCtxSubstitute-[vxlan-2490368]/rttoLocalCtxSubstitute-
[sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-[vxlan-2392064]]
--
5) Event:E_DEBUG, length:122, at 275994 usecs after Tue Apr 24 14:38:04 2018
    [1835623268] gr_objstore_remote_vrf_vnid_mts_hdlr: [Create]: sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-[vxlan-2392064]
--
6) Event:E_DEBUG, length:150, at 622469 usecs after Mon Apr 23 14:48:38 2018
    [1835623268] gr_objstore_bdvnid_map_mts_hdlr: [Create]: sys/inst-overlay-1/localSite-1/localCtxSubstitute-[vxlan-3014656]/localBdSubstitute-
[vxlan-16056262]
--
7) Event:E_DEBUG, length:278, at 622447 usecs after Mon Apr 23 14:48:38 2018
    [1835623268] gr_objstore_bdvnid_map_mts_hdlr: [Create]: sys/inst-overlay-1/localSite-1/localCtxSubstitute-[vxlan-3014656]/localBdSubstitute-
[vxlan-16056262]/rttoLocalBdSubstitute-[sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-[vxlan-2457600]/remoteBdSubstitute-[vxlan-15794151]]
--
```

DCI mgr – xlate with hex translation !

Vnid translate (vrf and bd)

```
pod36-spine1# show dcimgr repo vnid-maps detail
------------------------------------------------------------
     Remote             |          Local
site  Vrf       Bd      |    Vrf      Bd       Rel-state
------------------------------------------------------------
   1  2981888           |  2293760            [formed]
      0x2d8000          |  0x230000
------------------------------------------------------------
   1  2981888  16678778 |  2293760  16154554  [formed]
      0x2d8000 0xfe7f7a |  0x230000 0xf67fba
------------------------------------------------------------
   1  3014656           |  2457600            [formed]
      0x2e0000          |  0x258000
```

pcTag (sclass) translate

```
pod36-spine1# show dcimgr repo sclass-maps detail
------------------------------------------------------------
     Remote             |         Local
site  Vrf       PcTag   |  Vrf       PcTag     Rel-state
------------------------------------------------------------
  1  2981888   49153    | 2293760   49153    [formed]
     0x2d8000  0xc001   | 0x230000  0xc001
------------------------------------------------------------
  1  2981888   49154    | 2293760   49155    [formed]
     0x2d8000  0xc002   | 0x230000  0xc003
------------------------------------------------------------
  1  2981888   16387    | 2293760   16386    [formed]
     0x2d8000  0x4003   | 0x230000  0x4002
------------------------------------------------------------
  1  3014656   49153    | 2457600   49153    [formed]
     0x2e0000  0xc001   | 0x258000  0xc001
------------------------------------------------------------
  1  3014656   16387    | 2457600   32772    [formed]
     0x2e0000  0x4003   | 0x258000  0x8004
```

HAL – dci vnid Xlate

Tabular view

```
module-2# show platform internal hal dci vnidmap
Non-Sandbox Mode

Sandbox_ID: 0  Asic Bitmap: 0x0

Site    POD  isBD      Local      Remote  ----EPG table-----  -BDState Table
 ID     ID             vnid       vnid       idx   Localvnid   idx  isBD
----+------+------+--------+----------+--------+-----------+-----+------
   2      2      0    2457600    2162688   15372    2457600   15372      0
   2      2      1   16285610   16482195   15377   16285610   15377      1
   2      2      1   16056263   16121790   15364   16056263   15364      1
   2      2      1   15925206   16220082   15366   15925206   15366      1
   2      2      1   15040468   15335344   15371   15040468   15371      1
   2      2      0    2162689    2949121   15376    2162689   15376      0
```

Detail object

```
module-2# show platform internal hal objects dci vnidmap
## Get Objects for dci vnidmap for Asic 0

  OBJECT 1:
Handle                                                  : 339894
isbdvnid                                                : Enabled
localvnid                                               : 0xf87faa
localgipo                                               : 225.1.36.0/32
remotevnid                                              : 0xfb7f93
remotevrfvnid                                           : 0x2d0001
islocalbdctrl                                           : Enabled
siteid                                                  : 0x2
```

HAL – dci sclass translate

Tabular view

```
module-2# show platform internal hal dci sclassmap
Non-Sandbox Mode

Sandbox_ID: 0  Asic Bitmap: 0x0

            --- DCI Sclass table ---
 Site      Remote      Local      Remote      Remote      Local    Local
 ID        Vnid        Sclass     Sclass      Sclass      Sclass   Scope
------+----------+----------+---------+-----------+---------+------
   2     2293760      16387       16386       16386       16387       1
   2     2490368      32772       16389       16389       32772       6
   2     2162688      49154       16386       16386       49154       5
   2     2392064      16386       16386       16386       16386       4
   2     2293760      49154       49155       49155       49154       1
   2     2686976      32770       49153       49153       32770       3
```

Detail object

```
module-2# show platform internal hal objects dci sclassmap
## Get Objects for dci sclassmap for Asic 0

  OBJECT 0:
Handle                                                  : 38469
localsclass                                             : 0x4003
remotesclass                                            : 0x4002
remotevnid                                              : 0x230000
siteid                                                  : 0x2
```

BGP route exchange detail

BGP VNI

- Route Exchange issues can be seen either in the source or on the remote site.
  - Check if the BGP MOs are created for VNIs/RTs and RDs correctly. These MOs are created only on spines in every site. These MOs are created when the VRF/BD/EPGs are stretched or the contracts are created at EPG level
- Following shows mapping of BGP VNIDs and what routes are requested from COOP and why they are used:
  - Essentially we exchange all EP in BD VNID so bgp evi is the equivalent of (show bgp process used for vrf)

	VNID	Request Route Type from COOP	What routes are advertised	Why
1	VRF	IP Only with reserved MAC (0200.0000.0002)	Only Loop Back/SVI IPs	For inband management
2	BD	MAC-IP routes MAC routes only if L2 is Stretched	All EPs	For EP-EP communication

BGP MO for VNI – BD/VRF (site 1 and site 2)

BGP EVI contains the RD used to send Prefix for that BD /VRF

```
pod35-spine1# moquery -d sys/bgp/inst/encapgroupevi-
1/vni-bd-vrf-[vxlan-3014656]-bd-[vxlan-16351138]-epg-
[unknown]
Total Objects shown: 1

# bgp.Vni
type              : bd
vrfVnid           : vxlan-3014656
bdVnid            : vxlan-16351138
epgVnid           : unknown
bgpCfgFailedBmp   :
bgpCfgFailedTs    : 00:00:00:00.000
bgpCfgState       : 0
childAction       :
dn                : sys/bgp/inst/encapgroupevi-1/vni-bd-
vrf-[vxlan-3014656]-bd-[vxlan-16351138]-epg-[unknown]
l2Stretch         : enabled
lcOwn             : local
modTs             : 2018-04-11T04:28:21.600+00:00
name              :
rd                : rd:as2-nn4:1:33128354
rn                : vni-bd-vrf-[vxlan-3014656]-bd-
[vxlan-16351138]-epg-[unknown]
status            :
```

```
pod36-spine1# moquery -d  sys/bgp/inst/encapgroupevi-
1/vni-bd-vrf-[vxlan-2457600]-bd-[vxlan-16121791]-epg-
[unknown]
Total Objects shown: 1

# bgp.Vni
type              : bd
vrfVnid           : vxlan-2457600
bdVnid            : vxlan-16121791
epgVnid           : unknown
bgpCfgFailedBmp   :
bgpCfgFailedTs    : 00:00:00:00.000
bgpCfgState       : 0
childAction       :
dn                : sys/bgp/inst/encapgroupevi-1/vni-bd-
vrf-[vxlan-2457600]-bd-[vxlan-16121791]-epg-[unknown]
l2Stretch         : enabled
lcOwn             : local
modTs             : 2018-04-11T04:28:16.142+00:00
name              :
rd                : rd:as2-nn4:1:49676223
rn                : vni-bd-vrf-[vxlan-2457600]-bd-[vxlan-
16121791]-epg-[unknown]
status            :
```

Those are the local bgpVni of each site containing their own RD

BGP Route Target (same Stretched BD as previous slide) – part of subtree of BGP EVI

Pod 35 spine Import RT

```
# bgp.RttEntry
rtt           : route-target:as2-nn4:136:49676223
childAction   :
dn            : sys/bgp/inst/encapgroupevi-1/vni-bd-vrf-[vxlan-
3014656]-bd-[vxlan-16351138]-epg-[unknown]/rtp-import/ent-route-
target:as2-nn4:136:49676223
lcOwn         : local
modTs         : 2018-04-11T04:28:21.600+00:00
rn            : ent-route-target:as2-nn4:136:49676223
status        :
```

Pod 36 spine Import RT

```
# bgp.RttEntry
rtt           : route-target:as2-nn4:135:33128354
childAction   :
dn            : sys/bgp/inst/encapgroupevi-1/vni-bd-vrf-[vxlan-2457600]-
bd-[vxlan-16121791]-epg-[unknown]/rtp-import/ent-route-target:as2-
nn4:135:33128354
lcOwn         : local
modTs         : 2018-04-11T04:28:16.142+00:00
rn            : ent-route-target:as2-nn4:135:33128354
status        :
```

Pod 35 spine export RT

```
# bgp.RttEntry
rtt           : route-target:as2-nn4:135:33128354
childAction   :
dn            : sys/bgp/inst/encapgroupevi-1/vni-bd-vrf-[vxlan-
3014656]-bd-[vxlan-16351138]-epg-[unknown]/rtp-export/ent-route-
target:as2-nn4:135:33128354
lcOwn         : local
modTs         : 2018-04-11T04:28:21.600+00:00
rn            : ent-route-target:as2-nn4:135:33128354
status        :
```

Pod 36 spine export RT

```
# bgp.RttEntry
rtt           : route-target:as2-nn4:136:49676223
childAction   :
dn            : sys/bgp/inst/encapgroupevi-1/vni-bd-vrf-[vxlan-2457600]-
bd-[vxlan-16121791]-epg-[unknown]/rtp-export/ent-route-target:as2-
nn4:136:49676223
lcOwn         : local
modTs         : 2018-04-11T04:28:16.142+00:00
rn            : ent-route-target:as2-nn4:136:49676223
status        :
```

BGP EVI check (NXOS) – BD on site 1

Use **show bgp internal evi xx** to verify
RD and RT exp/import (where xx is BD VNID)
(kind of similar to show bgp process for GOLF)

```
pod35-spine1# show bgp internal evi 16351138

..
**************************************************
BGP L2VPN/EVPN RD Information for 1:33128354
  L2VNI ID                      : 16351138 (vni_16351138)
  #Prefixes Local/BRIB          : 2 / 2
  #Paths L3VPN->EVPN/EVPN->L3VPN : 0 / 0
**************************************************
==============================================
BGP Configured VNI Information:
  VNI ID (Index)                : 16351138 (0)
  RD                            : 1:33128354
  Export RTs                    : 1
    Export RT cfg list: 135:33128354(refcount:1
  Import RTs                    : 1
    Import RT cfg list: 136:49676223(refcount:1
  Topo Id                       : 16351138
  VTEP IP                       : 0.0.0.0
  VTEP VPC IP                   : 0.0.0.0
  Enabled                       : Yes
  Delete Pending                : No
  RD/Import RT/Export RT        : Yes/Yes/Yes
  Type                          : 3
  Usage                         : 2
  L2 stretch enabled            : 1
  VRF Vnid                      : 3014656
  Refcount                      : 00000003
  Encap                         : VxLAN

==============================================
++++++++++++++++++++++++++++++++++++
```

```
++++++++++++++++++++++++++++++++++++++++++++
BGP VNI Information for vni_16351138
  L2VNI ID                      : 16351138 (vni_16351138)
  RD                            : 1:33128354
  VRF Vnid                      : 3014656
  Prefixes (local/total)        : 2/2
  VNID registered with COOP     : Yes
  Enabled                       : Yes
  Delete pending                : 0
  Stale                         : No
  Import pending                : 0
  Import in progress            : 0
  Encap                         : VxLAN
  Topo Id                       : 16351138
  VTEP IP                       : 0.0.0.0
  VTEP VPC IP                   : 0.0.0.0
  Active Export RTs             : 1
  Active Export RT list         : 135:33128354
  Config Export RTs             : 1
    Export RT cfg list: 135:33128354(refcount:1
  Export RT chg/chg-pending     : 0/0
  Active Import RTs             : 1
  Active Import RT list         : 136:49676223
  Config Import RTs             : 1
    Import RT cfg list: 136:49676223(refcount:1
  Import RT chg/chg-pending     : 0/0
  IMET Reg/Unreg from L2RIB     : 1/0
  MAC Reg/Unreg from L2RIB      : 1/0
  MAC IP Reg/Unreg from L2RIB   : 1/0
  IP-only Reg/Unreg from L2RIB  : 0/0
  SMAD Reg/Unreg from L2RIB     : 1/0
  IMET Add/Del from L2RIB       : 0/0
  MAC Add/Del from L2RIB        : 3/2
  MAC IP Add/Del from L2RIB     : 3/2
  SMAD Add/Del from L2RIB       : 0/0
  IMET Dnld/Wdraw to L2RIB      : 0/0
  IMET Dnld/Wdraw to L2RIB failures : 0/0
  MAC Dnld/Wdraw to L2RIB       : 0/0
  MAC Dnld/Wdraw to L2RIB failures : 0/0
  SMAD Dnld/Wdraw to L2RIB      : 0/0
  SMAD Dnld/Wdraw to L2RIB failures : 0/0
```

Note the EVI number if the BD VNID we are looking for

BGP EVI check (NXOS) – BD on site 2

Use show bgp internal evi xx to verify
RD and RT exp/import (where xx is BD VNID)
(kind of similar to show bgp process for a VRF)

```
pod36-spine1# show bgp internal  evi  16121791

..
**********************************************
BGP L2VPN/EVPN RD Information for 1:49676223
  L2VNI ID                      : 16121791 (vni_16121791)
  #Prefixes Local/BRIB          : 0 / 2
  #Paths L3VPN->EVPN/EVPN->L3VPN : 0 / 0
**********************************************
==============================================
BGP Configured VNI Information:
  VNI ID (Index)                : 16121791 (0)
  RD                            : 1:49676223
  Export RTs                    : 1
    Export RT cfg list: 136:49676223(refcount:1
  Import RTs                    : 1
    Import RT cfg list: 135:33128354(refcount:1
  Topo Id                       : 16121791
  VTEP IP                       : 0.0.0.0
  VTEP VPC IP                   : 0.0.0.0
  Enabled                       : Yes
  Delete Pending                : No
  RD/Import RT/Export RT        : Yes/Yes/Yes
  Type                          : 3
  Usage                         : 2
  L2 stretch enabled            : 1
  VRF Vnid                      : 2457600
  Refcount                      : 00000003
  Encap                         : VxLAN
```

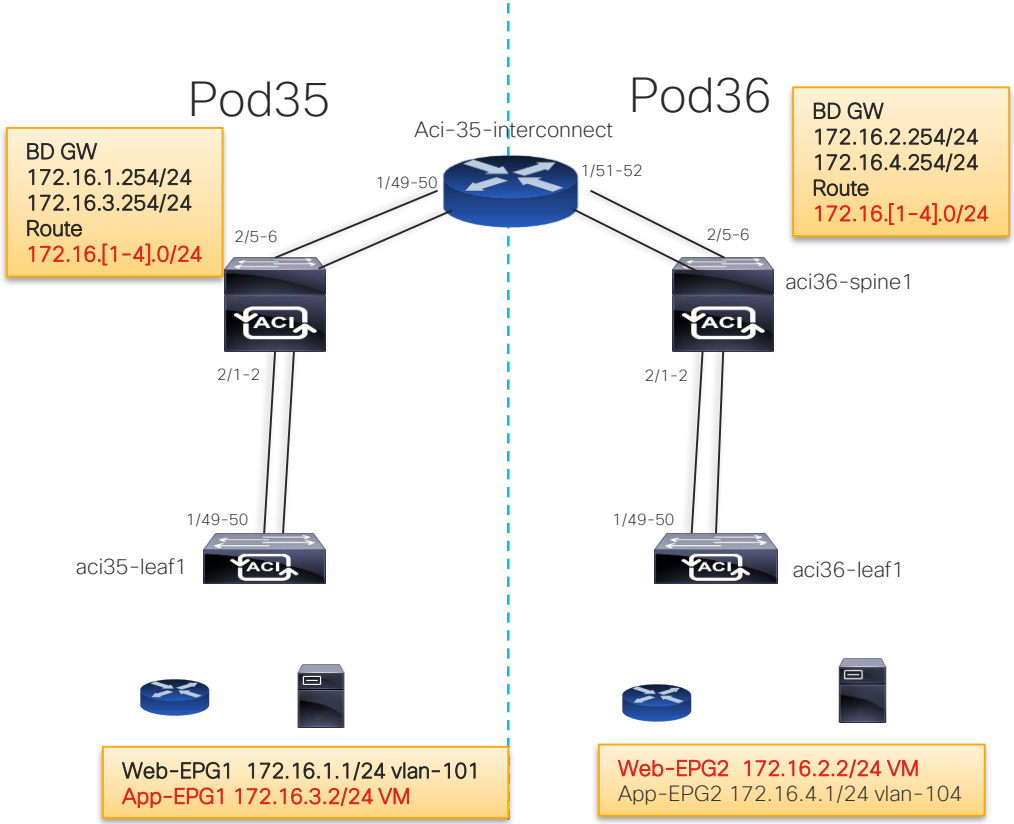
```
++++++++++++++++++++++++++++++++++++++++++++
BGP VNI Information for vni_16121791
  L2VNI ID                      : 16121791 (vni_16121791)
  RD                            : 1:49676223
  VRF Vnid                      : 2457600
  Prefixes (local/total)        : 0/2
  VNID registered with COOP     : Yes
  Enabled                       : Yes
  Delete pending                : 0
  Stale                         : No
  Import pending                : 0
  Import in progress            : 0
  Encap                         : VxLAN
  Topo Id                       : 16121791
  VTEP IP                       : 0.0.0.0
  VTEP VPC IP                   : 0.0.0.0
  Active Export RTs             : 1
  Active Export RT list         : 136:49676223
  Config Export RTs             : 1
    Export RT cfg list: 136:49676223(refcount:1
  Export RT chg/chg-pending     : 0/0
  Active Import RTs             : 1
  Active Import RT list         : 135:33128354
  Config Import RTs             : 1
    Import RT cfg list: 135:33128354(refcount:1
  Import RT chg/chg-pending     : 0/0
  IMET Reg/Unreg from L2RIB     : 1/0
  MAC Reg/Unreg from L2RIB      : 1/0
  MAC IP Reg/Unreg from L2RIB   : 1/0
  IP-only Reg/Unreg from L2RIB  : 0/0
  SMAD Reg/Unreg from L2RIB     : 1/0
  IMET Add/Del from L2RIB       : 0/0
  MAC Add/Del from L2RIB        : 0/0
  MAC IP Add/Del from L2RIB     : 0/0
  SMAD Add/Del from L2RIB       : 0/0
  IMET Dnld/Wdraw to L2RIB      : 0/0
  IMET Dnld/Wdraw to L2RIB failures : 0/0
  MAC Dnld/Wdraw to L2RIB       : 11/10
  MAC Dnld/Wdraw to L2RIB failures : 0/0
  SMAD Dnld/Wdraw to L2RIB      : 0/0
  SMAD Dnld/Wdraw to L2RIB failures : 0/0
  MAC-IP/SMAD Msite-RD routes   : 2
  MAC-IP WAN-RD routes          : 0
  MAC-IP network host routes    : 0
  Type                          : 3
```

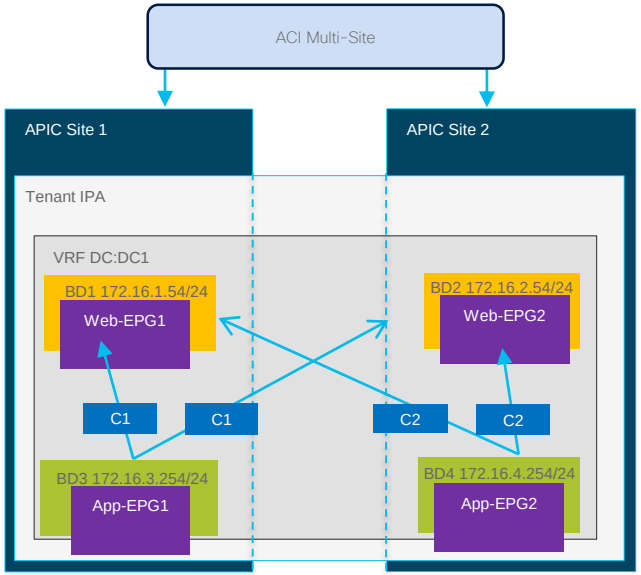
Unicast forwarding across site

Overview – spine behavior

- Unicast TX proxy/(local to remote site)
  - Leaf has not learned the remote site ep. Leaf sends the traffic to local spine proxy. Local spine looks up the route. The route for remote site Ep is programmed with next hop of remote site's ETEP. Dipo is re-written with remote site ETEP. Sipo is re-written with local site ETEP
- Unicast TX (local to remote site – no proxy)
  - Leaf has learned the remote site ep against remote site ETEP. Leaf sends the traffic to remote site ETEP. Local site spine will intercept this packet and re-write the sipo with Local site ETEP
- Unicast RX (remote to local site)
  - Incoming traffic destined to the local site's unicast ETEP goes through vnid and sclass translations. The receiving spine looks up the route for destination ep and sends the traffic to correct leaf.

Use case 1.1 – lab VRF DC:DC1



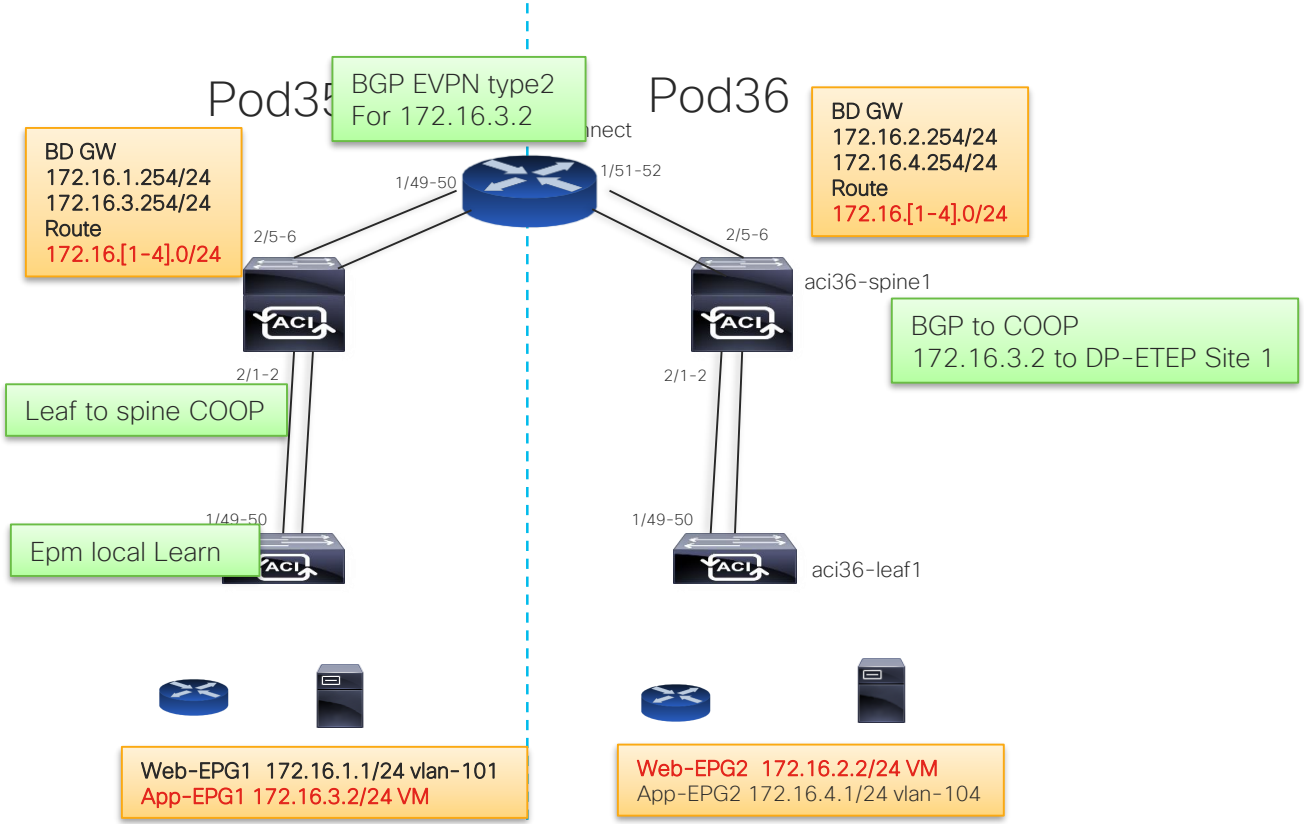


Test :
172.16.3.2 to 172.16.2.2

Src Site Leaf routing table

```
pod35-leaf1# show ip route vrf DC:DC1
IP Route Table for VRF "DC:DC1"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

172.16.1.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.88.66%overlay-1, [1/0], 3d00h, static, tag 4294967295
172.16.1.254/32, ubest/mbest: 1/0, attached, pervasive
    *via 172.16.1.254, vlan13, [1/0], 3d00h, local, local
172.16.2.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.88.66%overlay-1, [1/0], 00:33:15, static, tag 4294967295
172.16.3.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.88.66%overlay-1, [1/0], 01w08d, static, tag 4294967295
172.16.3.254/32, ubest/mbest: 1/0, attached, pervasive
    *via 172.16.3.254, vlan21, [1/0], 00:33:15, local, local
172.16.4.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.88.66%overlay-1, [1/0], 01w08d, static, tag 4294967295
```

Control Plane



Control plane EP in Pod1

Local EPM

```
pod35-leaf1# show system internal epm endpoint ip 172.16.3.2

MAC : 0050.56b1.4b52 ::: Num IPs : 1
IP# 0 : 172.16.3.2 ::: IP# 0 flags :
Vlan id : 22 ::: Vlan vnid : 8259 ::: VRF name : DC:DC1
BD vnid : 16351138 ::: VRF vnid : 3014656
Phy If : 0x1a001000 ::: Tunnel If : 0
Interface : Ethernet1/2
Flags : 0x80004c04 ::: sclass : 32770 ::: Ref count : 5
EP Create Timestamp : 04/19/2018 07:02:49.606635
EP Update Timestamp : 04/24/2018 05:08:19.642826
EP Flags : local|IP|MAC|sclass|timer|
::::
```

Local COOP site 1

```
pod35-spine1# show coop internal  info ip-db key 3014656 172.16.3.2


IP address : 172.16.3.2
Vrf : 3014656
Flags : 0
EP bd vnid : 16351138
EP mac :  00:50:56:B1:4B:52
Publisher Id : 10.0.112.64
Record timestamp : 04 24 2018 04:29:16 958852767
Publish timestamp : 04 24 2018 04:29:16 959066592
Seq No: 0
Remote publish timestamp: 01 01 1970 00:00:00 0
URIB Tunnel Info
Num tunnels : 1
        Tunnel address : 10.0.112.64
        Tunnel ref count : 1
```

Extract BGP table site 1

```
pod35-spine1# show bgp l2vpn  evpn vrf overlay-1 | egrep "Route Dis|172.16.3.2\]
Route Distinguisher: 1:33128354     (L2VNI 16351138)
*>l[2]:[0]:[16351138]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/272
Route Distinguisher: 10.10.35.101:135    (L2VNI 1)
*>l[2]:[0]:[16351138]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/272
```

Extract BGP table site 2

```
pod36-spine1#  show bgp l2vpn  evpn vrf overlay-1 | egrep "Route Dis|172.16.3.2\]"
Route Distinguisher: 1:33128354
*>e[2]:[0]:[0]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/272
Route Distinguisher: 1:49676223    (L2VNI 16121791)
*>e[2]:[0]:[16121791]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/272
Route Distinguisher: 10.10.35.102:136    (L2VNI 1)
*>e[2]:[0]:[16121791]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/272
```

Remote COOP entry site 2

```
pod36-spine1# show coop internal  info ip-db  | egrep -A
15 -B 1 "172.16.3.2$"
-------------------------------
IP address : 172.16.3.2
Vrf : 2457600
Flags : 0x4
EP bd vnid : 16121791
EP mac :  00:50:56:B1:4B:52
Publisher Id : 10.10.35.101
Record timestamp : 01 01 1970 00:00:00 0
Publish timestamp : 01 01 1970 00:00:00 0
Seq No: 0
Remote publish timestamp: 04 24 2018 05:05:25 371412024
URIB Tunnel Info
Num tunnels : 1
        Tunnel address : 10.10.35.101
        Tunnel ref count : 1
```

Control plane BGP path from site 1 to site 2 (detail on site 1)

Entry with BGP EVI 16351138
Which is BD vnid of the EP

```
pod35-spine1# show bgp internal evi  16351138 | egrep "RT list"
  Active Export RT list         : 135:33128354
  Active Import RT list         : 136:49676223
```

```
pod35-spine1# show bgp l2vpn  evpn 172.16.3.2  vrf overlay-1
Route Distinguisher: 1:33128354     (L2VNI 16351138)
BGP routing table entry for
[2]:[0]:[16351138]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/272,
version 79 dest ptr 0xa95a74d4
MSITE RD: 1:33128354     (L2VNI 16351138)
Local Route Distinguisher: 10.10.35.101:135     (L2VNI 1)
Paths: (1 available, best #1)
Flags: (0x00010a 00000000) on xmit-list, is not in rib/evpn
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: local 0x4000008c 0x0 ref 0, path is valid, is best
path
  AS-Path: NONE, path locally originated
    10.10.35.101 (metric 0) from 0.0.0.0 (10.10.35.111)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 16351138 3014656
      Extcommunity:
          RT:5:16

  Path-id 1 advertised to peers:
    10.10.35.112
```

```
Route Distinguisher: 10.10.35.101:135     (L2VNI 1)
BGP routing table entry for
[2]:[0]:[16351138]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/272,
version 79 dest ptr 0xa95a74d4
MSITE RD: 1:33128354     (L2VNI 16351138)
Local Route Distinguisher: 10.10.35.101:135     (L2VNI 1)
Paths: (1 available, best #1)
Flags: (0x00010a 00000000) on xmit-list, is not in rib/evpn
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: local 0x4000008c 0x0 ref 0, path is valid, is best path
  AS-Path: NONE, path locally originated
    10.10.35.101 (metric 0) from 0.0.0.0 (10.10.35.111)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 16351138 3014656
      Extcommunity:
          RT:5:16

  Path-id 1 advertised to peers:
    10.10.35.112
```

Control plane BGP path from site 1 to site 2 (detail on site 2)

Rx path from site 1

```
pod36-spine1#  show bgp l2vpn    evpn 172.16.3.2  vrf overlay-1
Route Distinguisher: 1:33128354
BGP routing table entry for
[2]:[0]:[0]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/272, version
828 dest ptr 0xa9603fa8
Paths: (1 available, best #1)
Flags: (0x000202 00000000) on xmit-list, is not in rib/evpn, is
locked
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: external 0x40000028 0x82040 ref 1, path is valid, is
best path, remote site path, remote nh not installed
  AS-Path: 135 , path sourced external to AS
    10.10.35.101 (metric 3) from 10.10.35.111 (10.10.35.111)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 16351138 3014656
      Received path-id 1
      Extcommunity:
          RT:135:33128354
          SOO:135:33554415
          COST:pre-bestpath:166:2684354560
          ENCAP:8
          Router MAC:0200.0a0a.2365

  Path-id 1 not advertised to any peer
```

```
pod36-spine1# show bgp internal evi 16121791 | egrep "RT list"
  Active Export RT list         : 136:49676223
  Active Import RT list         : 135:33128354
```

```
Route Distinguisher: 1:49676223    (L2VNI 16121791)
BGP routing table entry for
[2]:[0]:[16121791]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/272, version 841 dest
ptr 0xa9604646
MSITE RD: 1:49676223    (L2VNI 16121791)
Local Route Distinguisher: 10.10.35.102:136    (L2VNI 1)
Paths: (1 available, best #1)
Flags: (0x00021a 0x00000a) on xmit-list, is in rib/evpn, is in l2rib msite
shard, is in l2rib
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: external 0xc0000028 0xa0040 ref 56506, path is valid, is best path,
remote site path, remote nh not installed
          Imported from
1:33128354:[2]:[0]:[0]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/144
  AS-Path: 135 , path sourced external to AS
    10.10.35.101 (metric 3) from 10.10.35.111 (10.10.35.111)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 16351138 3014656
      Received path-id 1
      Extcommunity:
          RT:135:33128354
          SOO:135:33554415
          COST:pre-bestpath:166:2684354560
          ENCAP:8
          Router MAC:0200.0a0a.2365
Route Distinguisher: 10.10.35.102:136    (L2VNI 1)
BGP routing table entry for
[2]:[0]:[16121791]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/272, version 841 dest
ptr 0xa9604646
MSITE RD: 1:49676223    (L2VNI 16121791)
Local Route Distinguisher: 10.10.35.102:136    (L2VNI 1)
Paths: (1 available, best #1)
Flags: (0x00021a 0x00000a) on xmit-list, is in rib/evpn, is in l2rib msite
shard, is in l2rib
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: external 0xc0000028 0xa0040 ref 56506, path is valid, is best path,
remote site path, remote nh not installed
          Imported from
1:33128354:[2]:[0]:[0]:[48]:[0050.56b1.4b52]:[32]:[172.16.3.2]/144
  AS-Path: 135 , path sourced external to AS
    10.10.35.101 (metric 3) from 10.10.35.111 (10.10.35.111)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 16351138 3014656
      Received path-id 1
      Extcommunity:
          RT:135:33128354
```

In rib/evpn and l2rib msite shard – means It is in coop and this spine is shard ownerFor coop

Imported path
To local RD

Coop site 2

- Checking COOP in local VRF VNID in site 2 we see it is indeed in coop DB
- 2457600 is the local vnid for the VRF

```
pod36-spine1# show coop internal  info  ip-db key
2457600  172.16.3.2


IP address : 172.16.3.2
Vrf : 2457600
Flags : 0x4
EP bd vnid : 16121791
EP mac :  00:50:56:B1:4B:52
Publisher Id : 10.10.35.101
Record timestamp : 01 01 1970 00:00:00 0
Publish timestamp : 01 01 1970 00:00:00 0
Seq No: 0
Remote publish timestamp: 05 02 2018 05:11:56
694567370
URIB Tunnel Info
Num tunnels : 1
        Tunnel address : 10.10.35.101
        Tunnel ref count : 1
```

DCI Mgr on spine pod 35 (site 1)

Remote Site
DP-ETEP and
Mcast ETEP(dcimgr and Object model)

```
pod35-spine1# show dcimgr repo eteps

Remote site=2 :
Rem Etep=10.10.35.102/32, is_ucast=yes
Rem Etep=10.10.35.122/32, is_ucast=no
```

```
pod35-spine1# moquery -c  dci.AnycastExtn
Total Objects shown: 2

# dci.AnycastExtn
etep          : 10.10.35.102/32
childAction   :
dn            : sys/inst-overlay-1/remoteSite-2/anycastExtn-
[10.10.35.102/32]
is_ucast      : yes
lcOwn         : local
modTs         : 2018-03-30T05:50:38.558+00:00
rn            : anycastExtn-[10.10.35.102/32]
status        :

# dci.AnycastExtn
etep          : 10.10.35.122/32
childAction   :
dn            : sys/inst-overlay-1/remoteSite-2/anycastExtn-
[10.10.35.122/32]
is_ucast      : no
lcOwn         : local
modTs         : 2018-03-30T05:50:38.558+00:00
rn            : anycastExtn-[10.10.35.122/32]
status        :
```

DCI Mgr on spine pod 35 (site 1) – VNID MAP

DCI mgr vnid map

```
pod35-spine1# show dcimgr repo vnid-maps
------------------------------------------------------------
     Remote              |          Local
site  Vrf       Bd       |    Vrf      Bd       Rel-state
------------------------------------------------------------
  2   2457600             |   3014656             [formed]
  2   2457600   16121790  |   3014656  16056263   [formed]
  2   2457600   16121791  |   3014656  16351138   [formed]
  2   2457600   16220082  |   3014656  15925206   [formed]
  2   2457600   15794151  |   3014656  16056262   [formed]
```

Vrf vnid map in Obj Model

```
pod35-spine1# moquery -d sys/inst-overlay-1/remoteSite-
2/remoteCtxSubstitute-[vxlan-2457600]
Total Objects shown: 1

# l3.RemoteCtxSubstitute
FabEncap      : vxlan-2457600
DnName        : uni/tn-DC/ctx-DC1
childAction   :
dn            : sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-[vxlan-
2457600]
lcOwn         : local
modTs         : 2018-04-11T04:06:17.167+00:00
rn            : remoteCtxSubstitute-[vxlan-2457600]
status        :
```

BD vnid map in Obj Model

```
pod35-spine1# moquery -d sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-[vxlan-2457600]/remoteBdSubstitute-
[vxlan-16121790]
Total Objects shown: 1

# l2.RemoteBdSubstitute
FabEncap      : vxlan-16121790
DnName        : uni/tn-DC/BD-BD4
childAction   :
dn            : sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-[vxlan-2457600]/remoteBdSubstitute-[vxlan-
16121790]
lcOwn         : local
modTs         : 2018-04-11T04:27:30.868+00:00
rn            : remoteBdSubstitute-[vxlan-16121790]
status        :
```

DCI Mgr on spine pod 35 (site 1) – SCLASS MAP

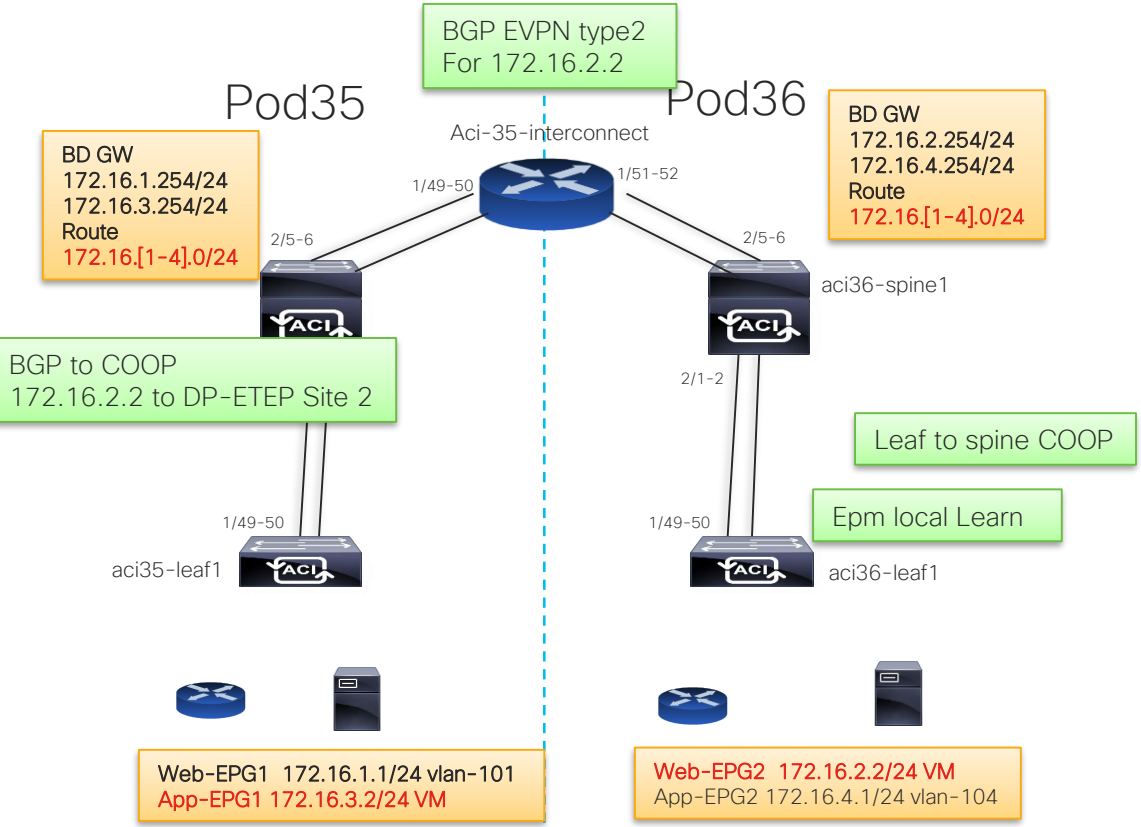
DCI mgr vnid map

```
pod35-spine1# show dcimgr repo sclass-maps
------------------------------------------------------------
     Remote             |        Local
site  Vrf       PcTag   |  Vrf      PcTag     Rel-state
------------------------------------------------------------
  2   2457600   49153   |  3014656   49153   [formed]
  2   2457600   16387   |  3014656   16388   [formed]
  2   2457600   32772   |  3014656   16387   [formed]
  2   2457600   16390   |  3014656   32770   [formed]
  2   2457600   32771   |  3014656   32772   [formed]
```

Sclass translate in object model

```
pod35-spine1# moquery -d sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-[vxlan-
2457600]/remotePcTagSubstitute-16387
Total Objects shown: 1

# 12.RemotePcTagSubstitute
pcTag        : 16387
DnName       : uni/tn-DC/ap-App2/epg-App-EPG2
childAction  :
dn           : sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-[vxlan-2457600]/remotePcTagSubstitute-16387
lcOwn        : local
modTs        : 2018-04-11T04:27:30.868+00:00
rn           : remotePcTagSubstitute-16387
status       :
```

Control Plane EP – in site 2 to COOP in site 1



Control plane EP in Site 2

Local COOP site 2

Local EPM

```
pod36-leaf1# show system internal  epm endpoint ip 172.16.2.2

MAC : 0050.56b1.4403 ::: Num IPs : 1
IP# 0 : 172.16.2.2 ::: IP# 0 flags :
Vlan id : 21 ::: Vlan vnid : 8194 ::: VRF name : DC:DC1
BD vnid : 16220082 ::: VRF vnid : 2457600
Phy If : 0x1a001000 ::: Tunnel If : 0
Interface : Ethernet1/2
Flags : 0x80004c04 ::: sclass : 32771 ::: Ref count : 5
EP Create Timestamp : 04/19/2018 07:03:23.999543
EP Update Timestamp : 05/02/2018 02:33:29.507208
EP Flags : local|IP|MAC|sclass|timer|
::::
```

```
pod36-spine1# show coop internal   info  ip-db key 2457600 172.16.2.2

IP address : 172.16.2.2
Vrf : 2457600
Flags : 0
EP bd vnid : 16220082
EP mac :  00:50:56:B1:44:03
Publisher Id : 10.1.48.64
Record timestamp : 05 02 2018 02:29:12 339899902
Publish timestamp : 05 02 2018 02:29:12 340145880
Seq No: 0
Remote publish timestamp: 01 01 1970 00:00:00 0
URIB Tunnel Info
Num tunnels : 1
        Tunnel address : 10.1.48.64
        Tunnel ref count : 1
```

Extract BGP table site 2

```
pod36-spine1# show bgp l2vpn  evpn vrf overlay-1 | egrep "Route Dis|172.16.2.2\]"
Route Distinguisher: 1:49774514     (L2VNI 16220082)
*>l[2]:[0]:[16220082]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/272
Route Distinguisher: 10.10.35.102:136    (L2VNI 1)
*>l[2]:[0]:[16220082]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/272
```

Extract BGP table site 1

```
pod35-spine1#  show bgp l2vpn  evpn vrf overlay-1 | egrep "Route Dis|172.16.2.2\]"
Route Distinguisher: 1:49774514
*>e[2]:[0]:[0]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/272
Route Distinguisher: 1:32702422    (L2VNI 15925206)
*>e[2]:[0]:[15925206]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/272
Route Distinguisher: 10.10.35.101:135    (L2VNI 1)
*>e[2]:[0]:[15925206]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/272
```

Remote COOP entry site 1

```
pod35-spine1# show coop internal  info ip-db  | egrep -A
15 -B 1 "172.16.2.2$"
-------------------------------
IP address : 172.16.2.2
Vrf : 3014656
Flags : 0x4
EP bd vnid : 15925206
EP mac :  00:50:56:B1:44:03
Publisher Id : 10.10.35.102
Record timestamp : 01 01 1970 00:00:00 0
Publish timestamp : 01 01 1970 00:00:00 0
Seq No: 0
Remote publish timestamp: 04 24 2018 05:05:34 611613733
URIB Tunnel Info
Num tunnels : 1
        Tunnel address : 10.10.35.102
        Tunnel ref count : 1
```

Control plane BGP path from site 2 to site 1 (detail on site 2)

```
pod36-spine1# show bgp l2vpn  evpn 172.16.2.2  vrf overlay-1
Route Distinguisher: 1:49774514    (L2VNI 16220082)
BGP routing table entry for
[2]:[0]:[16220082]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/272,
version 69 dest ptr 0xa9603608
MSITE RD: 1:49774514    (L2VNI 16220082)
Local Route Distinguisher: 10.10.35.102:136    (L2VNI 1)
Paths: (1 available, best #1)
Flags: (0x00010a 00000000) on xmit-list, is not in rib/evpn
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: local 0x4000008c 0x0 ref 0, path is valid, is best
path
  AS-Path: NONE, path locally originated
    10.10.35.102 (metric 0) from 0.0.0.0 (10.10.35.112)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 16220082 2457600
      Extcommunity:
          RT:5:16

  Path-id 1 advertised to peers:
    10.10.35.111
```

```
Route Distinguisher: 10.10.35.102:136    (L2VNI 1)
BGP routing table entry for
[2]:[0]:[16220082]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/272,
version 69 dest ptr 0xa9603608
MSITE RD: 1:49774514    (L2VNI 16220082)
Local Route Distinguisher: 10.10.35.102:136    (L2VNI 1)
Paths: (1 available, best #1)
Flags: (0x00010a 00000000) on xmit-list, is not in rib/evpn
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: local 0x4000008c 0x0 ref 0, path is valid, is best path
  AS-Path: NONE, path locally originated
    10.10.35.102 (metric 0) from 0.0.0.0 (10.10.35.112)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 16220082 2457600
      Extcommunity:
          RT:5:16

  Path-id 1 advertised to peers:
    10.10.35.111
```

Control plane BGP path from site 2 to site 1 (detail on site 1)

Rx path from site 2

```
pod35-spine1#  show bgp l2vpn    evpn 172.16.2.2  vrf overlay-1
Route Distinguisher: 1:49774514
BGP routing table entry for
[2]:[0]:[0]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/272, version
529 dest ptr 0xa95a7a3e
Paths: (1 available, best #1)
Flags: (0x000202 00000000) on xmit-list, is not in rib/evpn, is
locked
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: external 0x40000028 0x82040 ref 1, path is valid, is
best path, remote site path, remote nh not installed
  AS-Path: 136 , path sourced external to AS
    10.10.35.102 (metric 3) from 10.10.35.112 (10.10.35.112)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 16220082 2457600
      Received path-id 1
      Extcommunity:
          RT:136:49774514
          SOO:136:50331631
          COST:pre-bestpath:166:2684354560
          ENCAP:8
          Router MAC:0200.0a0a.2366

  Path-id 1 not advertised to any peer
```


Imported path
To local RD

```
Route Distinguisher: 1:32702422    (L2VNI 15925206)
BGP routing table entry for
[2]:[0]:[15925206]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/272, version 542 dest
ptr 0xa95a8646
MSITE RD: 1:32702422    (L2VNI 15925206)
Local Route Distinguisher: 10.10.35.101:135    (L2VNI 1)
Paths: (1 available, best #1)
Flags: (0x00021a 0x00000a) on xmit-list, is in rib/evpn, is in l2rib msite
shard, is in l2rib
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: external 0xc0000028 0xa0040 ref 56506, path is valid, is best path,
remote site path, remote nh not installed
            Imported from
1:49774514:[2]:[0]:[0]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/144
  AS-Path: 136 , path sourced external to AS
    10.10.35.102 (metric 3) from 10.10.35.112 (10.10.35.112)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 16220082 2457600
      Received path-id 1
      Extcommunity:
          RT:136:49774514
          SOO:136:50331631
          COST:pre-bestpath:166:2684354560
          ENCAP:8
          Router MAC:0200.0a0a.2366
Route Distinguisher: 10.10.35.101:135    (L2VNI 1)
BGP routing table entry for
[2]:[0]:[15925206]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/272, version 542 dest
ptr 0xa95a8646
MSITE RD: 1:32702422    (L2VNI 15925206)
Local Route Distinguisher: 10.10.35.101:135    (L2VNI 1)
Paths: (1 available, best #1)
Flags: (0x00021a 0x00000a) on xmit-list, is in rib/evpn, is in l2rib msite
shard, is in l2rib
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: external 0xc0000028 0xa0040 ref 56506, path is valid, is best path,
remote site path, remote nh not installed
            Imported from
1:49774514:[2]:[0]:[0]:[48]:[0050.56b1.4403]:[32]:[172.16.2.2]/144
  AS-Path: 136 , path sourced external to AS
    10.10.35.102 (metric 3) from 10.10.35.112 (10.10.35.112)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 16220082 2457600
      Received path-id 1
      Extcommunity:
          RT:136:49774514
```

DCI Mgr on spine pod 36 (site 2)

Remote Site
DP-ETEP and
Mcast ETEP(dcimgr and Object model)

```
pod36-spine1# show dcimgr repo eteps

Remote site=1 :
Rem Etep=10.10.35.101/32, is_ucast=yes
Rem Etep=10.10.35.121/32, is_ucast=no
pod36-spine1#
```

```
pod36-spine1# moquery -c dciAnycastExtn
Total Objects shown: 2

# dci.AnycastExtn
etep          : 10.10.35.101/32
childAction   :
dn            : sys/inst-overlay-1/remoteSite-1/anycastExtn-
[10.10.35.101/32]
is_ucast      : yes
lcOwn         : local
modTs         : 2018-03-30T05:50:34.562+00:00
rn            : anycastExtn-[10.10.35.101/32]
status        :

# dci.AnycastExtn
etep          : 10.10.35.121/32
childAction   :
dn            : sys/inst-overlay-1/remoteSite-1/anycastExtn-
[10.10.35.121/32]
is_ucast      : no
lcOwn         : local
modTs         : 2018-03-30T05:50:34.562+00:00
rn            : anycastExtn-[10.10.35.121/32]
status        :
```

DCI Mgr on spine pod 36 (site 2) – VNID MAP

DCI mgr vnid map

```
pod35-spine1# show dcimgr repo vnid-maps
------------------------------------------------------------
     Remote              |          Local
site  Vrf       Bd       |    Vrf      Bd       Rel-state
------------------------------------------------------------
  1   3014656             |   2457600             [formed]
  1   3014656  16056263   |   2457600  16121790   [formed]
  1   3014656  16351138   |   2457600  16121791   [formed]
  1   3014656  15925206   |   2457600  16220082   [formed]
  1   3014656  16056262   |   2457600  15794151   [formed]
```

Vrf vnid map in Obj Model

```
pod36-spine1# moquery -c l3RemoteCtxSubstitute

# l3.RemoteCtxSubstitute
FabEncap      : vxlan-3014656
DnName        : uni/tn-DC/ctx-DC1
childAction   :
dn            : sys/inst-overlay-1/remoteSite-1/remoteCtxSubstitute-[vxlan-
3014656]
lcOwn         : local
modTs         : 2018-04-11T04:06:11.695+00:00
rn            : remoteCtxSubstitute-[vxlan-3014656]
status        :
```

BD vnid map in Obj Model

```
pod36-spine1# moquery -d sys/inst-overlay-1/remoteSite-1/remoteCtxSubstitute-[vxlan-3014656]/remoteBdSubstitute-
[vxlan-15925206]
Total Objects shown: 1

# l2.RemoteBdSubstitute
FabEncap      : vxlan-15925206
DnName        : uni/tn-DC/BD-BD2
childAction   :
dn            : sys/inst-overlay-1/remoteSite-1/remoteCtxSubstitute-[vxlan-3014656]/remoteBdSubstitute-[vxlan-
15925206]
lcOwn         : local
modTs         : 2018-04-11T04:28:16.146+00:00
rn            : remoteBdSubstitute-[vxlan-15925206]
status        :
```

DCI Mgr on spine pod 36 (site 2) – SCLASS MAP

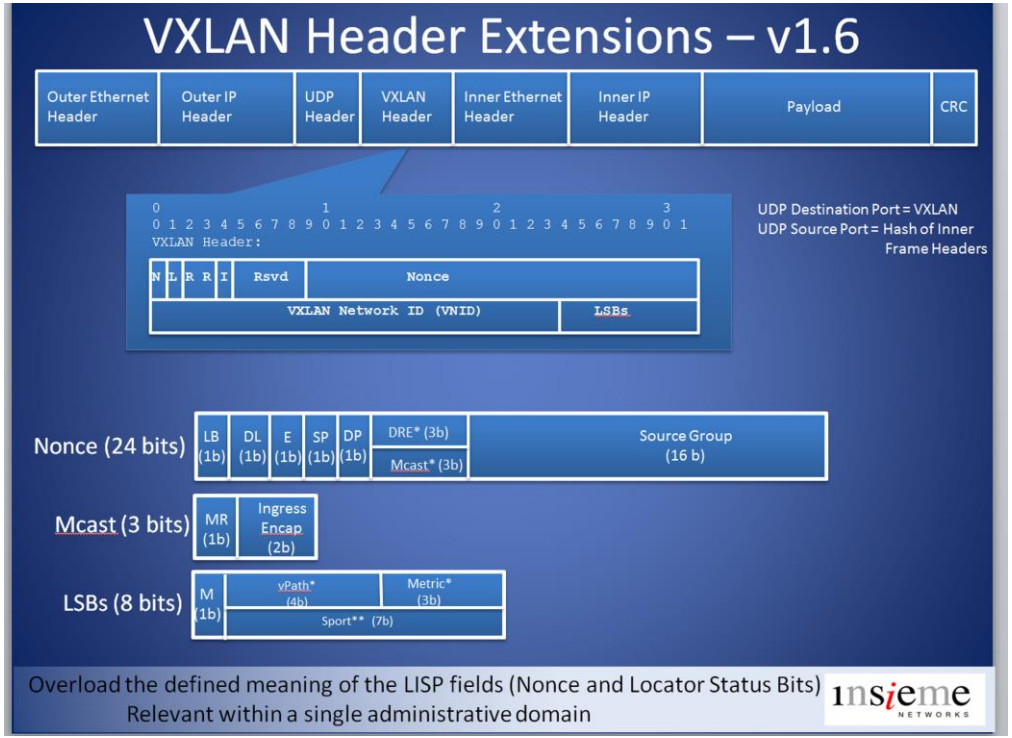
DCI mgr vnid map

```
pod36-spine1# show dcimgr repo sclass-maps
-------------------------------------------------------------
     Remote             |        Local
site  Vrf       PcTag   |  Vrf       PcTag     Rel-state
-------------------------------------------------------------
  1   3014656   49153   |  2457600   49153    [formed]
  1   3014656   16387   |  2457600   32772    [formed]
  1   3014656   16388   |  2457600   16387    [formed]
  1   3014656   32770   |  2457600   16390    [formed]
  1   3014656   32772   |  2457600   32771    [formed]
```

Sclass translate in object model

```
pod36-spine1# moquery -c l2RemotePcTagSubstitute
Total Objects shown: 12

# l2.RemotePcTagSubstitute
pcTag         : 16388
DnName        : uni/tn-DC/ap-App2/epg-App-EPG2
childAction   :
dn            : sys/inst-overlay-1/remoteSite-1/remoteCtxSubstitute-[vxlan-3014656]/remotePcTagSubstitute-16388
lcOwn         : local
modTs         : 2018-04-11T04:27:26.433+00:00
rn            : remotePcTagSubstitute-16388
status        :
```

ivxlan header review

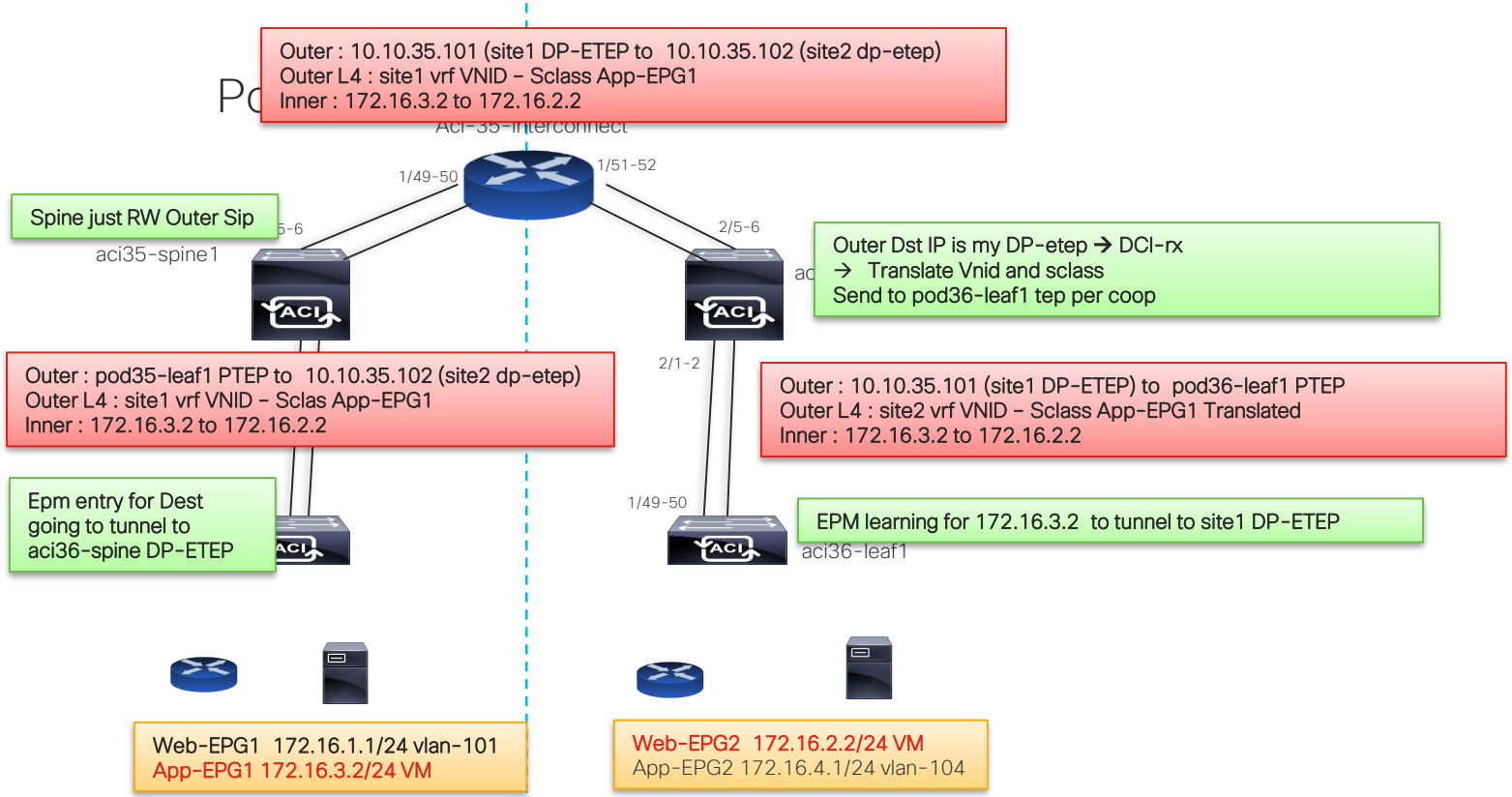


Note in Outer L4 header you can Get :
VNID (BD or VRF)
Sclass (src sclass) as part of Nounce field (last 4 nibble):

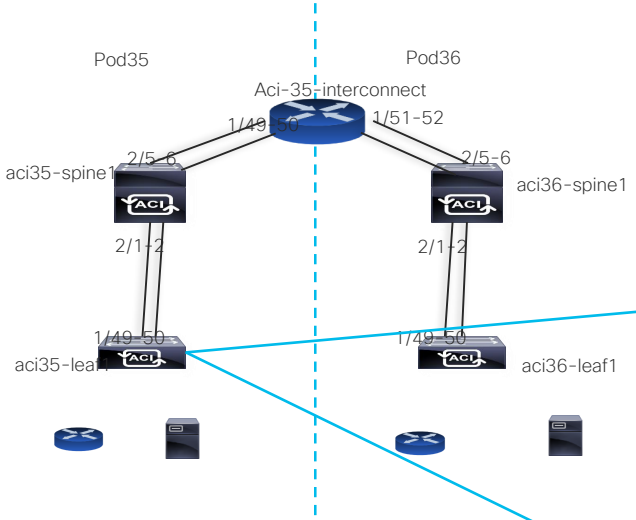
Ex :

```
hom_elam_in_l4v_tn.tn_nonce_info: 0x188002

Sclass of Rx frame is 0x8002
```

Data path – known EP Site 1 to Site 2 (Known unicast on ingress leaf)



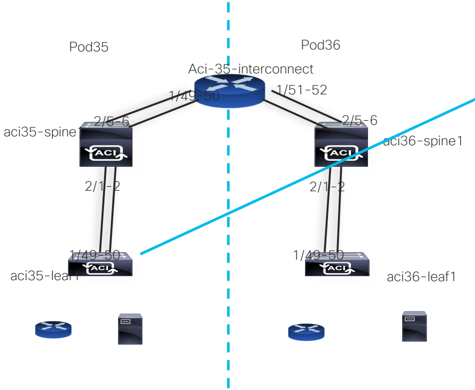
Ingress Leaf Known EP




```
pod35-leaf1# show system internal  epm  endpoint  ip 172.16.2.2

MAC : 0000.0000.0000 ::: Num IPs : 1
IP# 0 : 172.16.2.2 ::: IP# 0 flags :
Vlan id : 0 ::: Vlan vnid : 0 ::: VRF name : DC:DC1
BD vnid : 0 ::: VRF vnid : 3014656
Phy If : 0 ::: Tunnel If : 0x18010007
Interface : Tunnel7
Flags : 0x80004400 ::: sclass : 32772 ::: Ref count : 3
EP Create Timestamp : 04/24/2018 05:05:32.831665
EP Update Timestamp : 04/25/2018 04:58:50.374323
EP Flags : IP|sclass|timer|
::::
```

```
pod35-leaf1# show interface tunnel  7
Tunnel7 is up
    MTU 9000 bytes, BW 0 Kbit
    Transport protocol is in VRF "overlay-1"
    Tunnel protocol/transport is ivxlan
    Tunnel source 10.0.112.64/32 (lo0)
    Tunnel destination 10.10.35.102/32
    Last clearing of "show interface" counters never
    Tx
    0 packets output, 1 minute output rate 0 packets/sec
    Rx
    0 packets input, 1 minute input rate 0 packets/sec
```

ELAM Ingress leaf Site 1 – EP known

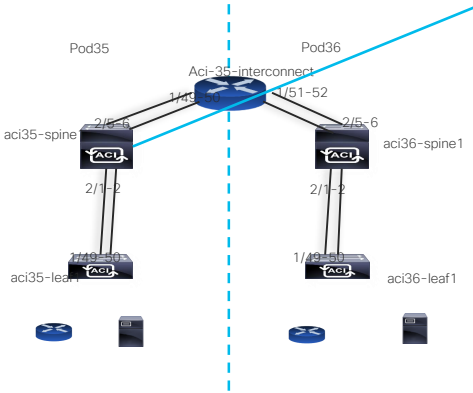



```
module-1# debug platform internal roc elam asic 0
module-1(DBG-elam)# trigger init in-select 6 out-select  1
module-1(DBG-elam-insel6)# set outer ipv4 src_ip 172.16.3.2 dst_ip 172.16.2.2
```

```
module-1(DBG-elam-insel6)# report | egrep "SRCID|ovector|encap"
   SRCID: 22
  hom_lurw_vec.encap_l2_idx: 0x2
  hom_lurw_vec.encap_pcid: 0x0
  hom_lurw_vec.encap_idx: 0x1
  hom_lurw_vec.encap_vld: 0x1
  hom_elam_out_sidebnd_no_spare_vec.ovector_idx: 0x78
```

```
module-1(DBG-elam-insel6)# show platform internal hal tunnel rtep apd
=====================================================================================================
ifId     IP              HwVrfId BDXlate SrcTepIdx DstInfoIdx RwEncapIdx ECMPIdx  ECMPMbrIdx Num     L2Index  RwDmacIdx
=====================================================================================================
18010007 10.10.35.102    4       1       3aa3      2800       1          0        0          2       1a030000 2
                                                                                                     1a031000 3
module-1(DBG-elam-insel6)# show platform internal hal l2  port gpd
================================================================================================================
                     Uc   Uc                                |        Reprogram      |                                         | Rep |
                   I PC   Pc                              L |   R I R D     R  U U X | L Xla Ovx N  NI Vif     RwV    Ing  Egr  | V R | PROF H
IfId     Ifname    P Cfg  MbrID As AP Sl Sp Ss Ovec S | P P P S P Sp Sp C M L | 3 Idx Idx L3 L3 Tid     Tid    Lbl  Lbl  | S V | ID   I
================================================================================================================
1a001000 Eth1/2    0 77   58    0  12 0  11 22 22   1   0 0 0 0 0 0  0  0 0 0   0 0   0   0  0  D-116   -      0    0      0 0   24   0
1a030000 Eth1/49   0 1    42    0  3d 0  3c 78 78   1   0 0 0 0 0 0  0  0 0 0   1 6   4   2  2  D-16a   -      400  0      0 0   19   0
```

ELAM Ingress LC Spine Site 1 – EP known

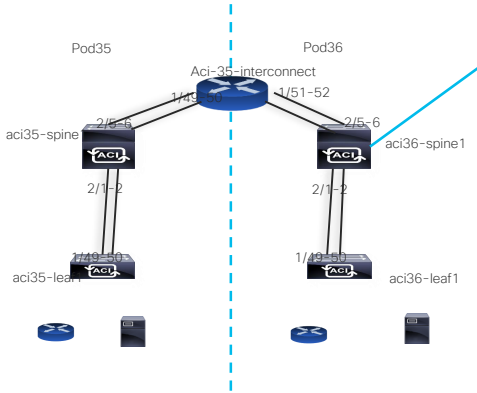



```
module-2# debug platform internal roc elam asic 0
module-2(DBG-elam)# trigger init in-select 14 out-select 1
module-2(DBG-elam-insel15)# set inner ipv4 src_ip 172.16.3.2 dst_ip 172.16.2.2
```

```
##########################HOMEWOOD ELAM REPORT START##########################
Dumping report for asic type 8 inst 0 slice 0 a_to_d 1 insel 15 outsel 1
LUA captured data with :
   SRCID: 20
*** Parsed Outer l2 vector
  hom_elam_in_l2v_da_sa_qtag0.qtag0_vlan: 0x2
*** Parsed Outer l3 vector
  hom_elam_in_l3v_ipv4.da: 0xA0A2366     - 10.0.35.102 (Dp-ETEP site2)
  hom_elam_in_l3v_ipv4.sa: 0xA007040     - 10.0.112.64  (leaf1 pod35 PTEP)
*** Parsed Outer l4 vector
  hom_elam_in_l4v_tn.tn_seg_id: 0x2E0000                        - 3014656
  hom_elam_in_l4v_tn.tn_nonce_info: 0x8002

hom_lua_latch_results_vec.lua4_1.lux_ispine_dci_rx: 0x0
 hom_lua_latch_results_vec.lua4_1.lux_ispine_dci_tx: 0x0
```

```
module-2(DBG-elam-insel15)# show platform internal  hal l2 port gpd  | egrep "Eth2/1|==|IfId|Uc|Xla"
=============================================================================================================================================
                      Uc   Uc                          |       Reprogram       |                                                  | Rep |
                     I PC   Pc                        L |   R I R D    R  U U X | L Xla Ovx N  NI Vif    RwV    Ing  Egr  | V R | PROF H
IfId     Ifname      P Cfg  MbrID As AP Sl Sp Ss Ovec S | P P P S P Sp Sp C M L | 3 Idx Idx L3 L3 Tid    Tid    Lbl  Lbl  | S V | ID   I
=============================================================================================================================================
1a080000 Eth2/1      0 9a   28    0  11 0  10 20 20   1   0 0 0 0 0 0  0  0 0 0   1 1   1   1  1  D-f3   D-61   100  0      0 0   4    0

pod35-spine1# show lldp neighbors  | egrep "Eth2/1"
pod35-leaf1          Eth2/1          120        BR          Eth1/49
pod35-spine1#
```

ELAM Ingress LC Spine Site 2 – Proxy

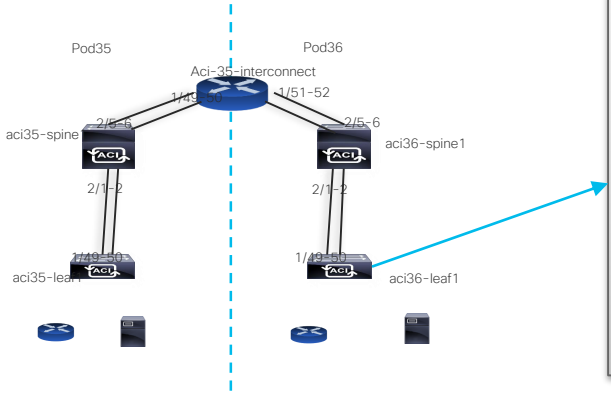



```
module-2# debug platform internal roc elam asic 0
module-2(DBG-elam)# trigger reset
module-2(DBG-elam)# trigger init in-select 15 out-select 1
module-2(DBG-elam-insel15)# set inner ipv4 src_ip 172.16.3.2 dst_ip 172.16.2.2
```

```
##########################HOMEWOOD ELAM REPORT START##########################
Dumping report for asic type 8 inst 0 slice 0 a_to_d 1 insel 15 outsel 1
LUA captured data with :
 SRCID: 0
*** Parsed Outer 12 vector
 hom_elam_in_l2v_da_sa_qtag0.qtag0_vlan: 0x4
*** Parsed Outer 13 vector
  hom_elam_in_l3v_ipv6_da_only.da: 0x00000000000000000A0A2366 - 10.10.35.102 (site2 - DP-ETEP)
  hom_elam_in_l3v_ipv6_da_only.sa: 0xA0A2365                  - 10.10.35.101 (site1 - DP-ETEP)
*** Parsed Outer 14 vector
  hom_elam_in_l4v_tn.tn_nonce_info: 0x188002                  - Rx sclass is 0x8002 = 16387
  hom_elam_in_l4v_tn.tn_seg_id: 0x2E0000                      - 3014656 (vnid before rewrite)


 hom_elam_out_sidebnd_no_spare_vec.ovector_idx: 0x78     (useless internal port to FC)
 hom_lua_latch_results_vec.lua4_1.lux_ispine_dci_rx: 0x1
 hom_lua_latch_results_vec.lua4_1.lux_ispine_dci_tx: 0x0

 hom_lurw_vec.info.ifabric_spine.vnid: 0x258000     - Vnid after rewrite = 2457600
 hom_lurw_vec.info.ifabric_spine.sclass: 0x4006      - rewritten Sclass is 16390

======== lux_fwd_mode = 0x09516040
LUX_FWD_MODE: ISPINE_LC      bit is set     ingress LC
LUX_FWD_MODE: ISPINE_DCI     bit is set
..
```

```
pod36-spine1# show dcimgr repo sclass-maps |  egrep "3014656.*16387"
  1    3014656    16387   |  2457600    32772    [formed]
pod36-spine1# show dcimgr repo vnid-maps  | egrep 3014656
  1    3014656            |   2457600            [formed]
```

ELAM egress leaf Site 2 – EP known

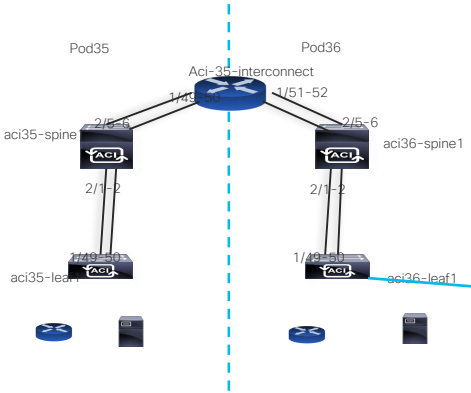



```
module-1# debug platform internal roc elam asic 0
module-1(DBG-elam)# trigger reset
module-1(DBG-elam)# trigger init in-select 15 out-select 1
module-1(DBG-elam-insel15)# set inner ipv4 src_ip 172.16.3.2 dst_ip 172.16.2.2
```

```
##########################HOMEWOOD ELAM REPORT START############################
Dumping report for asic type 8 inst 0 slice 0 a_to_d 1 insel 15 outsel 1
LUA captured data with :
 SRCID: 80
*** Parsed Outer l2 vector
 hom_elam_in_l2v_da_sa_qtag0.qtag0_vlan: 0x2
*** Parsed Outer l3 vector
  hom_elam_in_l3v_ipv6_da_only.da: 0x00000000000000000A013040 - 10.1.48.64 (pod36-leaf1 PTEP)
  hom_elam_in_l3v_ipv6_da_only.sa: 0xA0A2365                  - 10.10.35.101 (site 1 - DP-ETEP)
*** Parsed Outer l4 vector
 hom_elam_in_l4v_tn.tn_nonce_info: 0x384006                   - sclass in packet is 0x4006
 hom_elam_in_l4v_tn.tn_seg_id: 0x258000                       - vnid is the rewritten vnid


 hom_lurw_vec.info.ifabric_leaf.dclass: 0x8003
 hom_lurw_vec.info.ifabric_leaf.sclass: 0x4006
```

```
module-2(DBG-elam-insel15)# show platform internal  hal l2 port gpd  | egrep "Eth2/1|==|IfId|Uc|Xla"
=============================================================================================================================================
                      Uc   Uc                          |       Reprogram       |                                                  | Rep |
                     I PC   Pc                        L |   R I R D    R  U U X | L Xla Ovx N  NI Vif    RwV     Ing  Egr   | V R | PROF H
IfId     Ifname      P Cfg  MbrID As AP Sl Sp Ss Ovec S | P P P S P Sp Sp C M L | 3 Idx Idx L3 L3 Tid    Tid     Lbl  Lbl   | S V | ID   I
=============================================================================================================================================
1a084000 Eth2/5      0 9e   24    0  1  0  0  0  0    1   0 0 0 0 0 0  0  0 0 0   1 b   b   1  1  D-19a  D-2ee   300  0       1 0   2    0

pod35-spine1# show lldp neighbors  | egrep "Eth2/1"
pod35-leaf1          Eth2/1          120        BR          Eth1/49
pod35-spine1#
```

EPM dest leaf in site 2

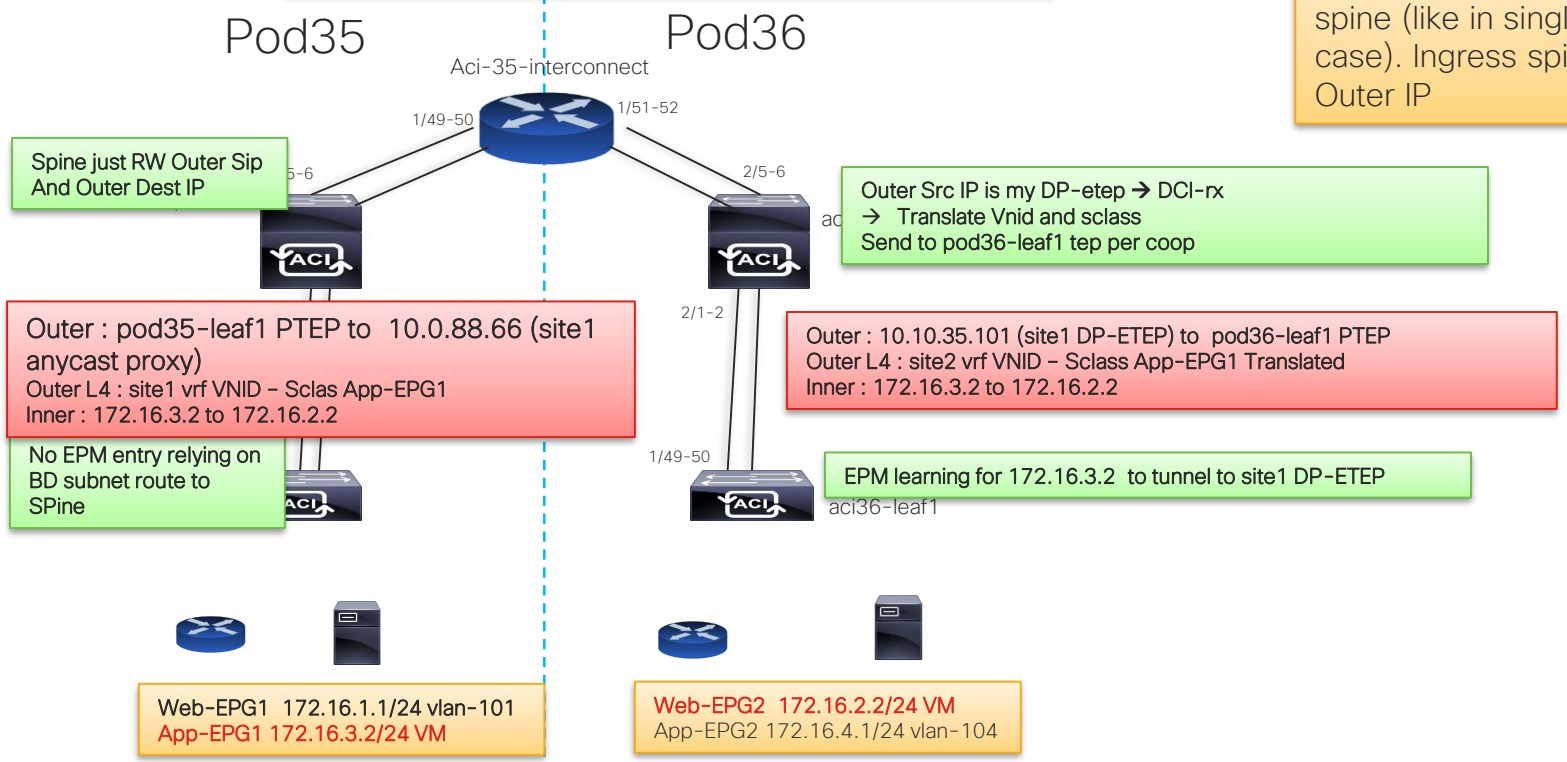



```
pod36-leaf1# show system internal  epm  endpoint  ip 172.16.3.2

MAC : 0000.0000.0000 ::: Num IPs : 1
IP# 0 : 172.16.3.2 ::: IP# 0 flags :
Vlan id : 0 ::: Vlan vnid : 0 ::: VRF name : DC:DC1
BD vnid : 0 ::: VRF vnid : 2457600
Phy If : 0 ::: Tunnel If : 0x18010007
Interface : Tunnel7
Flags : 0x80004400 ::: sclass : 16390 ::: Ref count : 3
EP Create Timestamp : 04/24/2018 05:05:25.720716
EP Update Timestamp : 04/25/2018 05:22:30.240899
EP Flags : IP|sclass|timer|
::::


pod36-leaf1# show interface tunnel  7
Tunnel7 is up
    MTU 9000 bytes, BW 0 Kbit
    Transport protocol is in VRF "overlay-1"
    Tunnel protocol/transport is ivxlan
    Tunnel source 10.1.48.64/32 (lo0)
    Tunnel destination 10.10.35.101/32
    Last clearing of "show interface" counters never
    Tx
    0 packets output, 1 minute output rate 0 packets/sec
    Rx
    0 packets input, 1 minute input rate 0 packets/sec
```

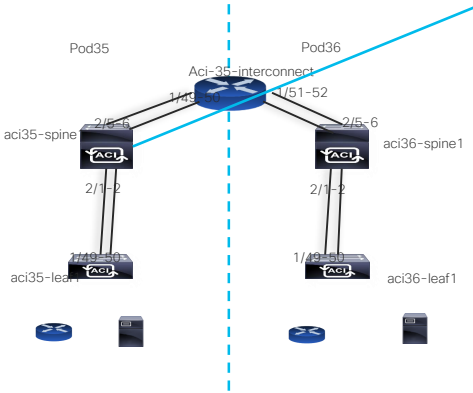
Data path – unknown EP on leaf Site 1 to Site 2 - Proxy

Outer : 10.10.35.101 (site1 DP-ETEP to 10.10.35.102 (site2 dp-etep)
Outer L4 : site1 vrf VNID – Sclass App-EPG1
Inner : 172.16.3.2 to 172.16.2.2

Only differences is that ingress leaf does send to Local Proxy spine (like in single site pod case). Ingress spine does Rw Outer IP



ELAM Ingress LC Spine Site 1 – Proxy



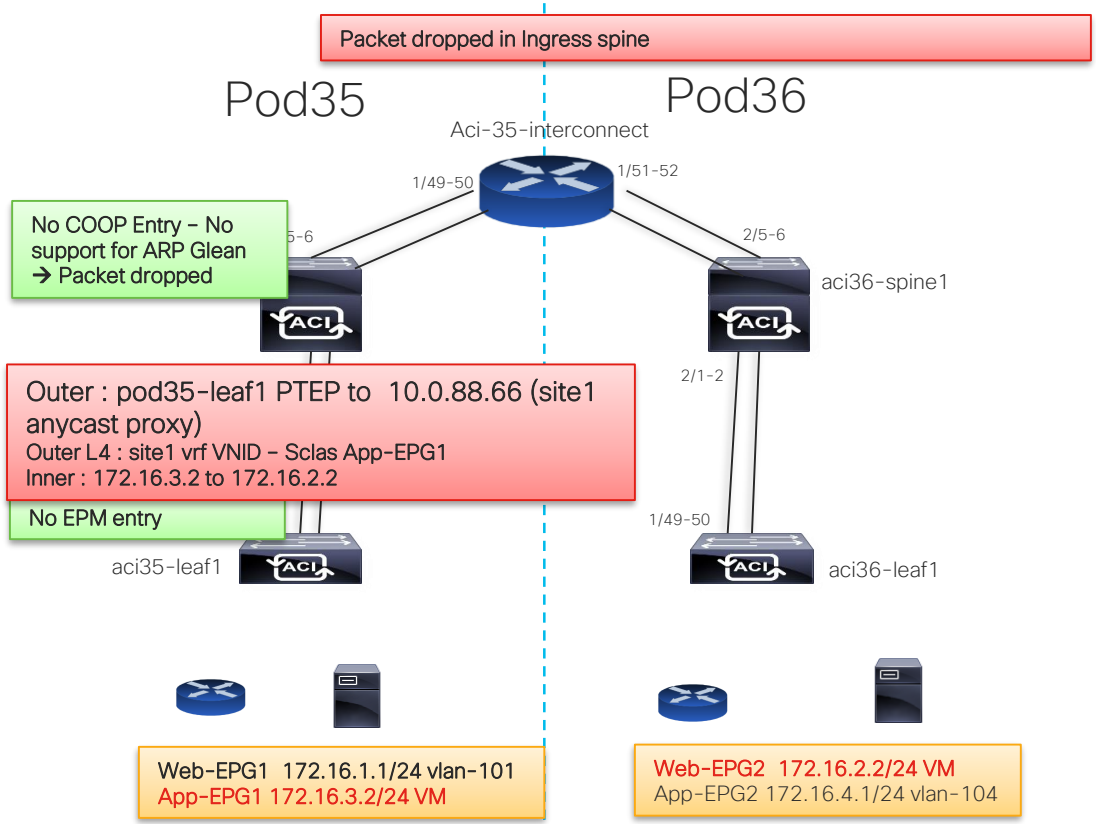

```
module-2# debug platform internal roc elam asic 0
module-2(DBG-elam)# trigger init in-select 14 out-select 1
module-2(DBG-elam-insel15)# set inner ipv4 src_ip 172.16.3.2 dst_ip 172.16.2.2
```

```
##########################HOMEWOOD ELAM REPORT START##########################
Dumping report for asic type 8 inst 0 slice 0 a_to_d 1 insel 15 outsel 1
LUA captured data with :
   SRCID: 20
*** Parsed Outer l2 vector
  hom_elam_in_l2v_da_sa_qtag0.qtag0_vlan: 0x2
*** Parsed Outer l3 vector
  hom_elam_in_l3v_ipv4.da: 0xA005842     - 10.0.88.66 (spine proxy site 1)
  hom_elam_in_l3v_ipv4.sa: 0xA007040     - 10.0.112.64  (leaf1 pod35 PTEP)
*** Parsed Outer l4 vector
  hom_elam_in_l4v_tn.tn_seg_id: 0x2E0000                        - 3014656
  hom_elam_in_l4v_tn.tn_nonce_info: 0x8002

hom_lua_latch_results_vec.lua4_1.lux_ispine_dci_rx: 0x0
 hom_lua_latch_results_vec.lua4_1.lux_ispine_dci_tx: 0x0
```

```
module-2(DBG-elam-insel15)# show platform internal  hal l2 port gpd  | egrep "Eth2/1|==|IfId|Uc|Xla"
=============================================================================================================================================
                        Uc   Uc                          |       Reprogram       |                                                  | Rep |
                      I PC   Pc                        L |   R I R D    R  U U X | L Xla Ovx N  NI Vif    RwV     Ing  Egr  | V R | PROF H
IfId     Ifname       P Cfg  MbrID As AP Sl Sp Ss Ovec S | P P P S P Sp Sp C M L | 3 Idx Idx L3 L3 Tid    Tid     Lbl  Lbl  | S V | ID   I
=============================================================================================================================================
1a080000 Eth2/1       0 9a   28    0  11 0  10 20 20   1   0 0 0 0 0 0  0  0 0 0   1 1   1   1  1  D-f3   D-61    100  0      0 0   4    0

pod35-spine1# show lldp neighbors  | egrep "Eth2/1"
pod35-leaf1          Eth2/1          120        BR          Eth1/49
pod35-spine1#
```

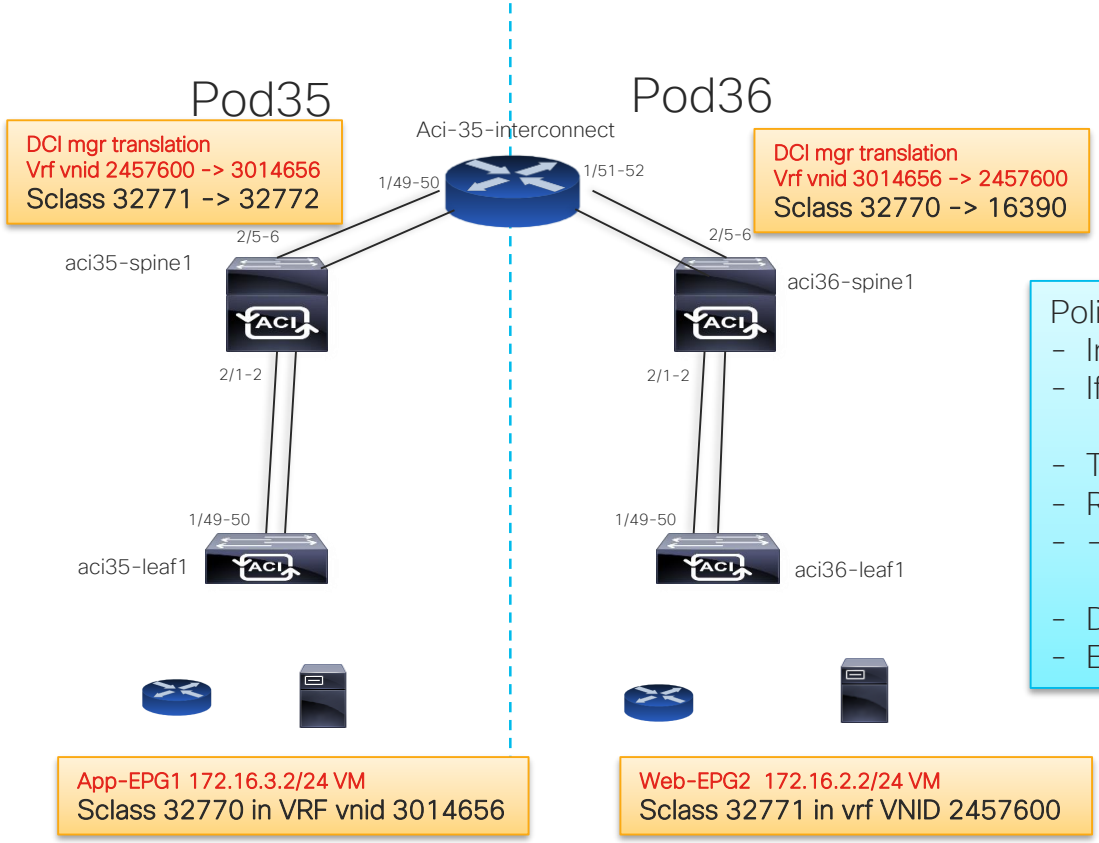
Data path – unknown EP Site 1 to Site 2 – 3.0 or 3.1

3.0. and 3.1 do not support ARP glean across msite



Policy enforcement

Sclass Translationg



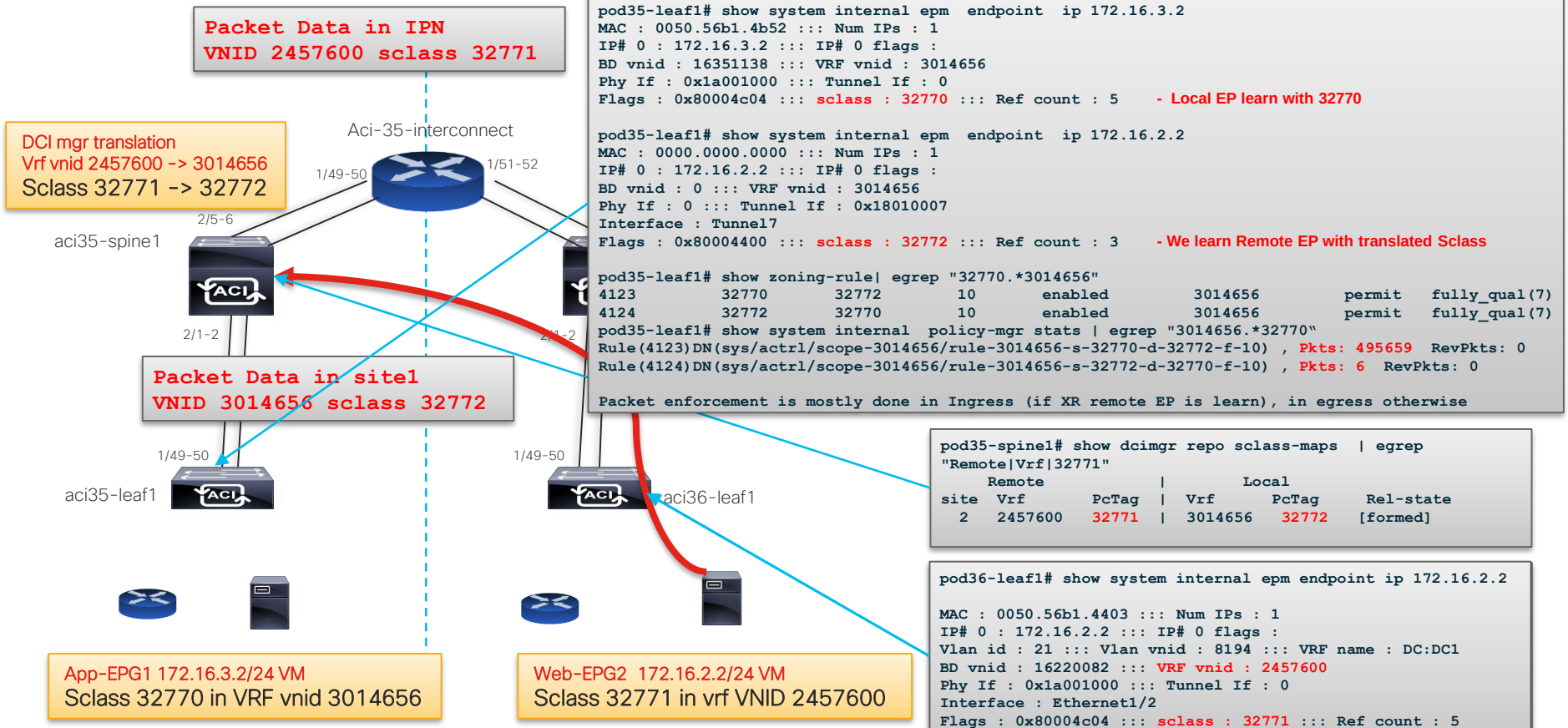
Policy Enforcement

- Ingress leaf derives sclass and vnid based on local EPM
- If Remote EPM is populated – Enforce Policy (as usual)

- Transmit to Remote Spine Site
- Remote spine site translate sclass and VNId
- - sent it to Dest leaf

- Dest leaf learn remote EP entry in translated sclass
- Enforce policy if not done on ingress

Policy from Site 2 (172.16.2.2) to Site 1 (172.16.3.2)

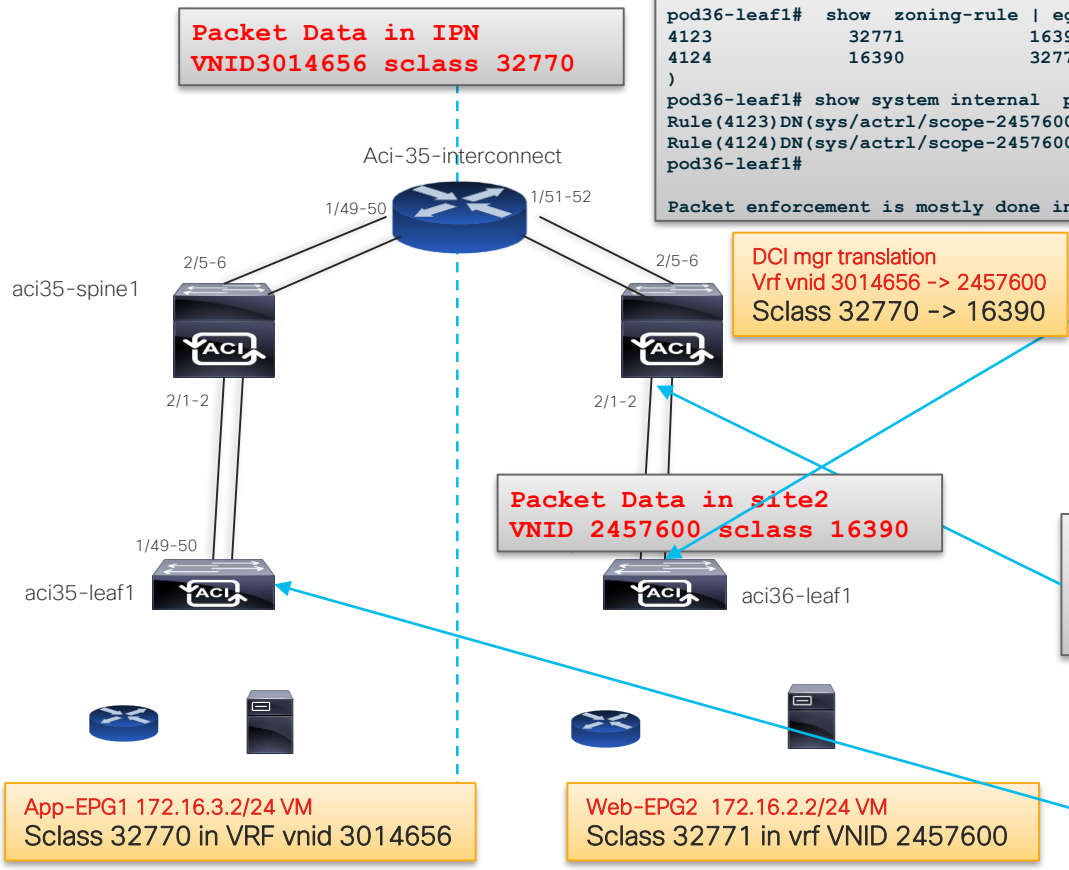



```
pod35-leaf1# show system internal epm  endpoint  ip 172.16.3.2
MAC : 0050.56b1.4b52 ::: Num IPs : 1
IP# 0 : 172.16.3.2 ::: IP# 0 flags :
BD vnid : 16351138 ::: VRF vnid : 3014656
Phy If : 0x1a001000 ::: Tunnel If : 0
Flags : 0x80004c04 ::: sclass : 32770 ::: Ref count : 5     - Local EP learn with 32770

pod35-leaf1# show system internal epm  endpoint  ip 172.16.2.2
MAC : 0000.0000.0000 ::: Num IPs : 1
IP# 0 : 172.16.2.2 ::: IP# 0 flags :
BD vnid : 0 ::: VRF vnid : 3014656
Phy If : 0 ::: Tunnel If : 0x18010007
Interface : Tunnel7
Flags : 0x80004400 ::: sclass : 32772 ::: Ref count : 3     - We learn Remote EP with translated Sclass

pod35-leaf1# show zoning-rule| egrep "32770.*3014656"
4123          32770        32772        10       enabled          3014656          permit   fully_qual(7)
4124          32772        32770        10       enabled          3014656          permit   fully_qual(7)
pod35-leaf1# show system internal  policy-mgr stats | egrep "3014656.*32770"
Rule(4123)DN(sys/actrl/scope-3014656/rule-3014656-s-32770-d-32772-f-10) , Pkts: 495659  RevPkts: 0
Rule(4124)DN(sys/actrl/scope-3014656/rule-3014656-s-32772-d-32770-f-10) , Pkts: 6  RevPkts: 0

Packet enforcement is mostly done in Ingress (if XR remote EP is learn), in egress otherwise
```

```
pod35-spine1# show dcimgr repo sclass-maps  | egrep
"Remote|Vrf|32771"
     Remote                |        Local
site  Vrf        PcTag     |  Vrf       PcTag      Rel-state
  2   2457600    32771     |  3014656   32772      [formed]
```

```
pod36-leaf1# show system internal epm endpoint ip 172.16.2.2

MAC : 0050.56b1.4403 ::: Num IPs : 1
IP# 0 : 172.16.2.2 ::: IP# 0 flags :
Vlan id : 21 ::: Vlan vnid : 8194 ::: VRF name : DC:DC1
BD vnid : 16220082 ::: VRF vnid : 2457600
Phy If : 0x1a001000 ::: Tunnel If : 0
Interface : Ethernet1/2
Flags : 0x80004c04 ::: sclass : 32771 ::: Ref count : 5
```

Policy from Site 1 (172.16.3.2) to Site 2 (172.16.3.2)




```
pod36-leaf1#  show  zoning-rule | egrep  "16390.*2457600"
4123            32771           16390     11      enabled  2457600  permit           fully_qual(7)
4124            16390           32771     11      enabled  2457600  permit           fully_qual(7)
)
pod36-leaf1# show system internal  policy-mgr stats | egrep 16390
Rule(4123)DN(sys/actrl/scope-2457600/rule-2457600-s-32771-d-16390-f-11) Pkts: 505314  RevPkts: 0
Rule(4124)DN(sys/actrl/scope-2457600/rule-2457600-s-16390-d-32771-f-11) Pkts: 4  RevPkts: 0
pod36-leaf1#

Packet enforcement is mostly done in Ingress (if XR remote EP is learn), in egress otherwise
```

DCI mgr translation
Vrf vnid 3014656 -> 2457600
Sclass 32770 -> 16390

```
pod36-spine1# show dcimgr repo sclass-maps  | egrep
"Remote|Vrf|32771"
     Remote           |        Local
site  Vrf       PcTag  |  Vrf       PcTag     Rel-state
 1   3014656   32770  |  2457600   16390    [formed]
```

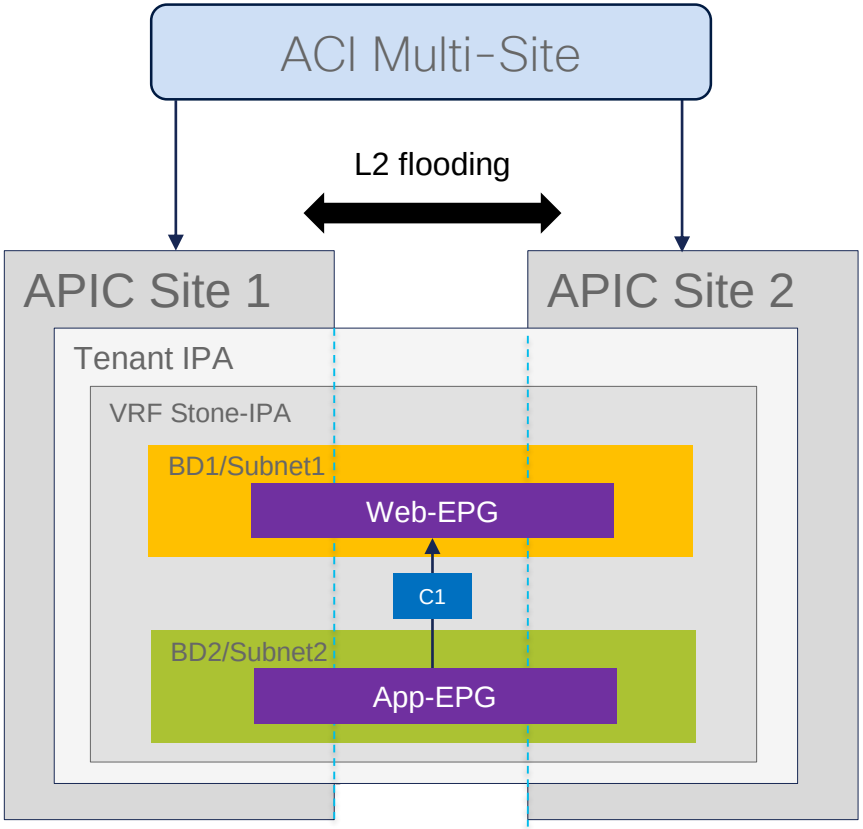
```
pod35-leaf1# show  system internal epm  endpoint  ip
172.16.3.2

MAC : 0050.56b1.4b52 ::: Num IPs : 1
IP# 0 : 172.16.3.2 ::: IP# 0 flags :
Vlan id : 22 ::: Vlan vnid : 8259 ::: VRF name : DC:DC1
BD vnid : 16351138 ::: VRF vnid : 3014656
Phy If : 0x1a001000 ::: Tunnel If : 0
Interface : Ethernet1/2
Flags : 0x80004c04 ::: sclass : 32770 ::: Ref count : 5
```

Multicast Multisite

Overview – Layer 2 BUM traffic across Sites

- TX (local to remote site)
  - GIPo (BUM) traffic sourced from the local site is Head-end replicated (HREP) to each remote site from the Spine. DIPo is rewritten to a unicast address called as Multicast HREP TEP IP (also called Multicast DP-TEP IP) of the remote site. SIPo is rewritten with the Unicast ETEP IP
- RX (remote to local site)
  - Incoming traffic destined to the local site's Multicast HREP TEP IP gets translated, derives the local site's BD-GIPo, and follows the regular GIPo lookup path from there

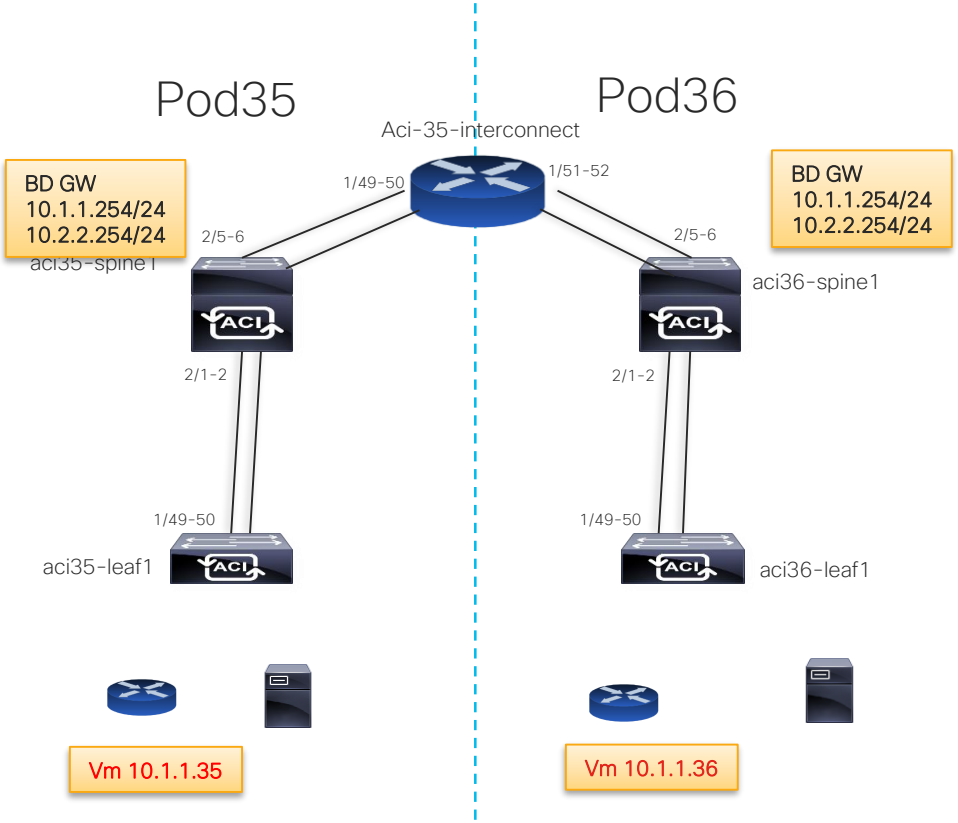
Multi-Site
Stretched BD with L2 Broadcast Extension

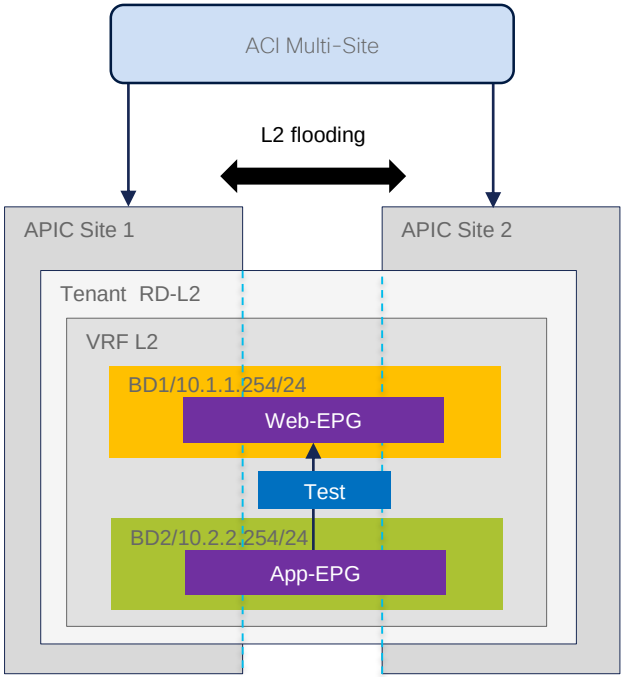


Use Case Properties

- Active/Active deployment with inter-site Layer 2 extension
- Objects stretched across sites:
  - Tenant ID
  - VRF context
  - BD/Subnet
  - Provider and Consumer EPGs
  - Policy between EPGs
- L2 flooding enabled at the BD level
  - L2 BUM traffic forwarded over head-end replicated VXLAN tunnels
- L2 application clustering and 'live' VM migration

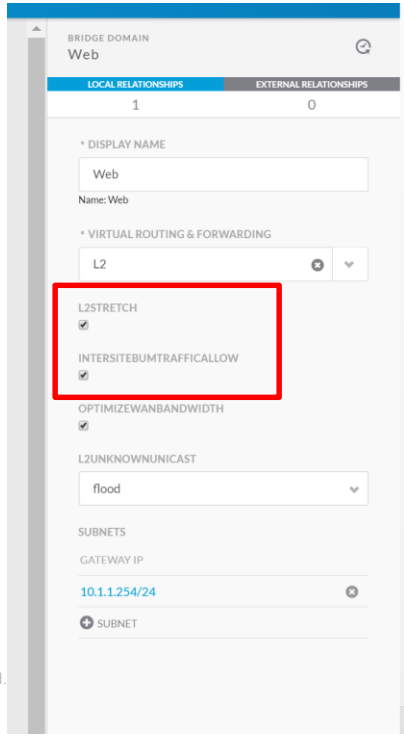
Use case – lab VRF RD-L2:L2





Config Check

- BD must be set with intersite BUM allow flag




```
admin@bdsol-aci35-apic1:~> moquery -d uni/tn-RD-L2/BD-Web
Total Objects shown: 1

# fv.BD
name                          : Web
OptimizeWanBandwidth          : yes
arpFlood                      : yes
bcastP                        : 225.0.216.80
childAction                   :
configIssues                  :
descr                         :
dn                            : uni/tn-RD-L2/BD-Web
epClear                       : no
epMoveDetectMode              :
extMngdBy                     : msc
intersiteBumTrafficAllow      : yes
intersiteL2Stretch            : yes
ipLearning                    : yes
lcOwn                         : local
limitIpLearnToSubnets         : yes
llAddr                        : ::
mac                           : 00:22:BD:F8:19:FF
mcastAllow                    : no
modTs                         : 2018-05-03T03:14:39.650+00:00
monPolDn                      : uni/tn-common/monepg-default
mtu                           : inherit
multiDstPktAct                : bd-flood
nameAlias                     :
ownerKey                      :
ownerTag                      :
pcTag                         : 32770
rn                            : BD-Web
scope                         : 2457600
seg                           : 15204288
status                        :
type                          : regular
uid                           : 15374
unicastRoute                  : yes
unkMacUcastAct                : flood
unkMcastAct                   : flood
vmac                          : not-applicable
```

Config check

- Multicast HREP TEP IP per Site
- Tunnel to each Remote site's Multicast HREP TEP

```
pod35-spine1# show ip interface vrf overlay-1 | egrep -A 1 mcast-hrep
loopback14, Interface status: protocol-up/link-up/admin-up, iod: 120, mode: mcast-hrep, vrf_vnid: 16777199
  IP address: 10.10.35.121, IP subnet: 10.10.35.121/32

pod35-spine1# show interface tunnel 5
Tunnel5 is up
    MTU 9000 bytes, BW 9 Kbit
    Transport protocol is in VRF "overlay-1"
    Tunnel protocol/transport ivxlan
    Tunnel source 10.0.112.65, destination 10.10.35.122
```

Control Plane interaction

- ISIS
  - For the stretched BDs (with intersiteBUMTrafficAllow), based on HREP-TEP configuration, ISIS adds the Remote site's HREP Tunnel If to the BD-GIPO of the Stretched BD.
  - BD-GIPOs are striped across the Multisite-capable Spines – meaning HREP Tunnel If is added to BD-GIPo only on one of the Multi-site capable Spines in a site
  - Unlike Multi-pod, no IGMP joins are sent out towards IPN, since native multicast is not used for forwarding BUM traffic across the sites

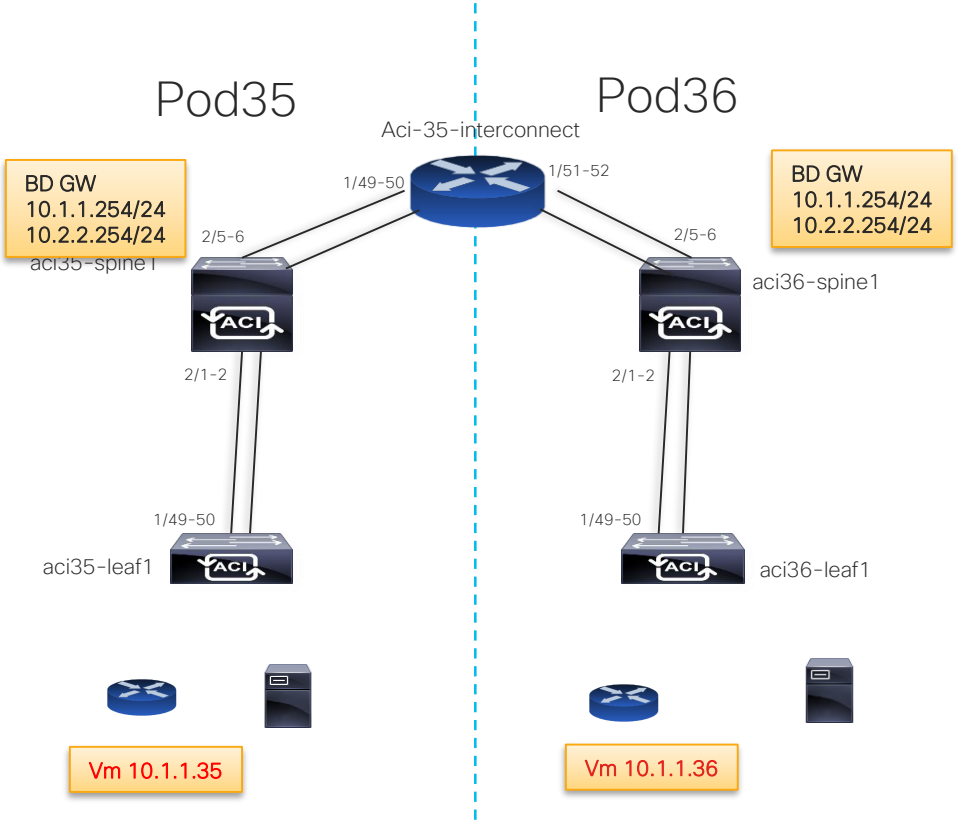
```
pod35-spine1# show isis internal mcast routes gipo  | egrep -A 6 "225.0.216.80"
 GIPo: 225.0.216.80 [LOCAL]
    OIF List:
      Ethernet2/1.35
      Ethernet2/2.36
      Tunnel5
```

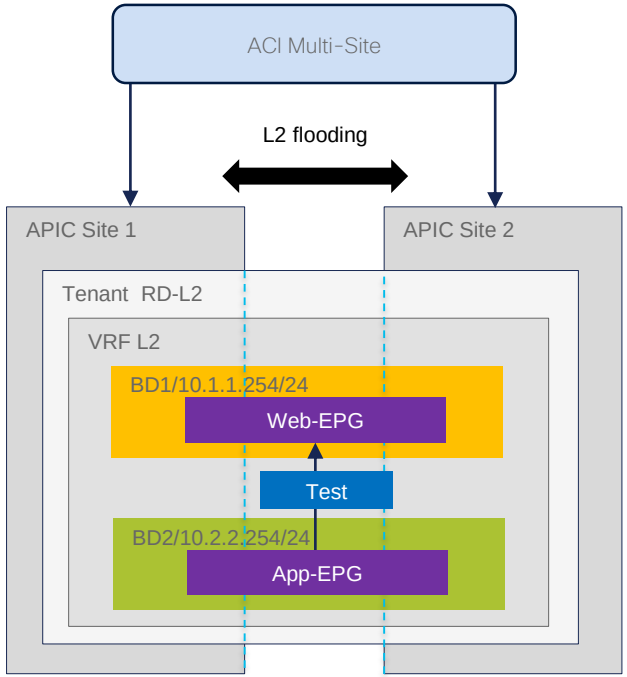
mrib

```
pod35-spine1# show ip mroute 225.0.216.80 vrf overlay-1
IP Multicast Routing Table for VRF "overlay-1"

(*, 225.0.216.80/32), uptime: 2w1d, isis
  Incoming interface: Null, RPF nbr: 0.0.0.0
  Outgoing interface list: (count: 3)
    Tunnel5, uptime: 2w1d
    Ethernet2/2.38, uptime: 2w1d
    Ethernet2/1.37, uptime: 2w1d
```

One spine per site Should have Tunnel Interface as BD GIPo OIL

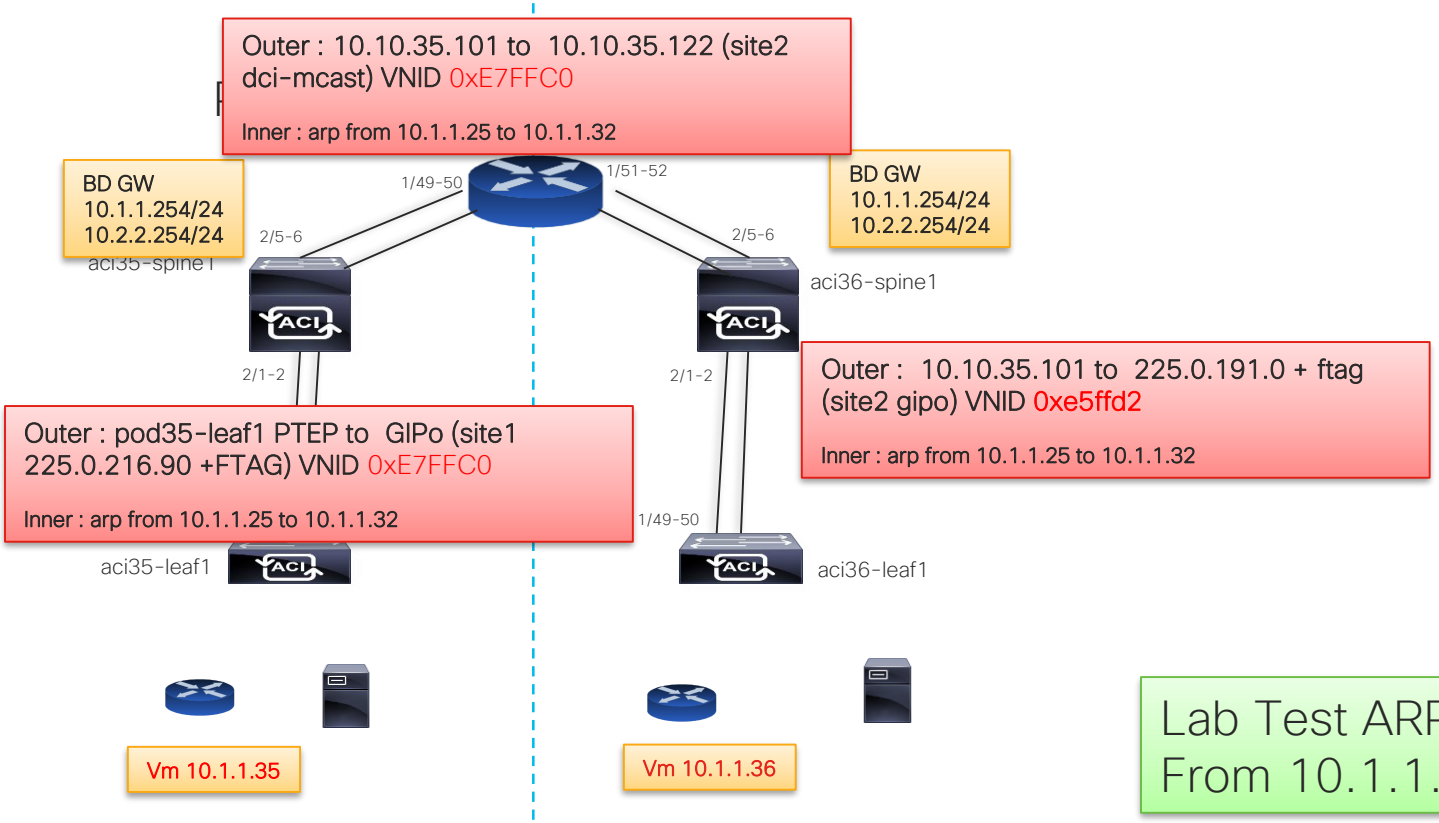
Use case – lab VRF RD-L2:L2





Lab Test ARP broadcast
From 10.1.1.35 to 10.1.1.32

Use case – lab VRF RD-L2:L2



Lab Test ARP broadcast
From 10.1.1.35 to 10.1.1.32

Debugging on Spine TX site – hrep reach (vsh)

```
pod35-spine1# show forwarding  distribution multicast  hrep
MFDM HREP NODE TABLE
--------------------
IP Address: 0xa0a237a      - 10.10.35.122 (remote mcast-hrep address)
Table Id: 4
Flags: 0x0  Type: 1
IfIndex: 0x18010005
Internal BD 0x1001 bd_label 0x0 (hw_label 0x0)
Internal encap 0xb54
Nexthop Information: (num: 2)
Address          Ifindex          Dvif
0xa0a2302        0x1a084025        0x3 (Selected)    - Selected next-hop to reach hrep 10.10.35.2
0xa0a2306        0x1a085026        0x0

pod35-spine1#  show ip route vrf overlay-1 10.10.35.122
IP Route Table for VRF "overlay-1"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.10.35.122/32, ubest/mbest: 2/0
    *via 10.10.35.2, Eth2/5.37, [110/3], 5w2d, ospf-default, intra
    *via 10.10.35.6, Eth2/6.38, [110/3], 5w2d, ospf-default, intra
```

Spine Tx – vsh_lc HREP reachability (mfdm and HAL)

```
module-2# show forwarding  multicast hrep tep_routes

****HREP TEP ROUTES****

-------------------------------------------------------------------------------------------------------------
|     Tep Ip     |     Tep If     |     NH Ip     |     NH If     |    NH dmac     | NH dvif | Vlan Id | Bd Id |
-------------------------------------------------------------------------------------------------------------
|10.10.35.122    |0x18010005      |10.10.35.2     |0x1a084025     |00a6.ca34.101f |3        |2900     |4097   |
-------------------------------------------------------------------------------------------------------------
```

```
module-2# show  platform internal hal  objects mcast hreptep extensions
## Get Extended Objects for mcast hreptep for Asic 0

  OBJECT 0:
Handle                                                  : 38119
tepifindex                                              : 0x18010005
tepipaddr                                               : 10.10.35.122/0
intbdid                                                 : 0x1001
intvlanid                                               : 0xb54
nexthopipaddr                                           : 10.10.35.2/0
nexthopifindex                                          : 0x1a084025
nexthopmacaddr                                          : 00:a6:ca:34:10:1f
```

GIPo route on line card

```
module-2# show forwarding  multicast route group 225.0.216.80 vrf all

  (*, 225.0.216.80/32), RPF Interface: NULL, flags: Dc
    Received Packets: 0 Bytes: 0
    Number of Outgoing Interfaces: 3
    Outgoing Interface List Index: 15
      Ethernet2/1.35 Outgoing Packets:0 Bytes:0
      Ethernet2/2.36 Outgoing Packets:0 Bytes:0
      Tunnel5 Outgoing Packets:0 Bytes:0
```

Elam ingress spine – ingress line card

```
*** Parsed Outer l3 vector
hom_elam_in_l3v_ipv4: 0x2801C103840361444480002
..
  hom_elam_in_l3v_ipv4.da: 0xE100D851    225.0.216.81
  hom_elam_in_l3v_ipv4.sa: 0xA007040    10.0.112.64
*** Parsed Outer l4 vector
hom_elam_in_l4v_tn: 0x39FFF000002000F2
  hom_elam_in_l4v_tn.l4_type: 0x2
  hom_elam_in_l4v_tn.tn_nonce: 0x1
  hom_elam_in_l4v_tn.tn_lsb: 0x1
  hom_elam_in_l4v_tn.tn_nonce_info: 0x8003
  hom_elam_in_l4v_tn.tn_lsb_info: 0x0
  hom_elam_in_l4v_tn.tn_seg_id: 0xE7FFC0
*** Parsed Inner l3 ARP vector
hom_elam_in_l3v_arp:
0x280404800000000000002804048C000410182000201804
  hom_elam_in_l3v_arp.l3_type: 0x4
  hom_elam_in_l3v_arp.pyld_len: 0x0
  hom_elam_in_l3v_arp.etype: 0x806
  hom_elam_in_l3v_arp.pro: 0x800
  hom_elam_in_l3v_arp.hln: 0x6
  hom_elam_in_l3v_arp.pln: 0x4
  hom_elam_in_l3v_arp.op: 0x1
  hom_elam_in_l3v_arp.spa: 0xA010123
  hom_elam_in_l3v_arp.tha: 0x000000000
  hom_elam_in_l3v_arp.tpa: 0xA010120
```

ELAM DCI RX spine

- Elam on rx spine.
- We see vnid translate (before and after)

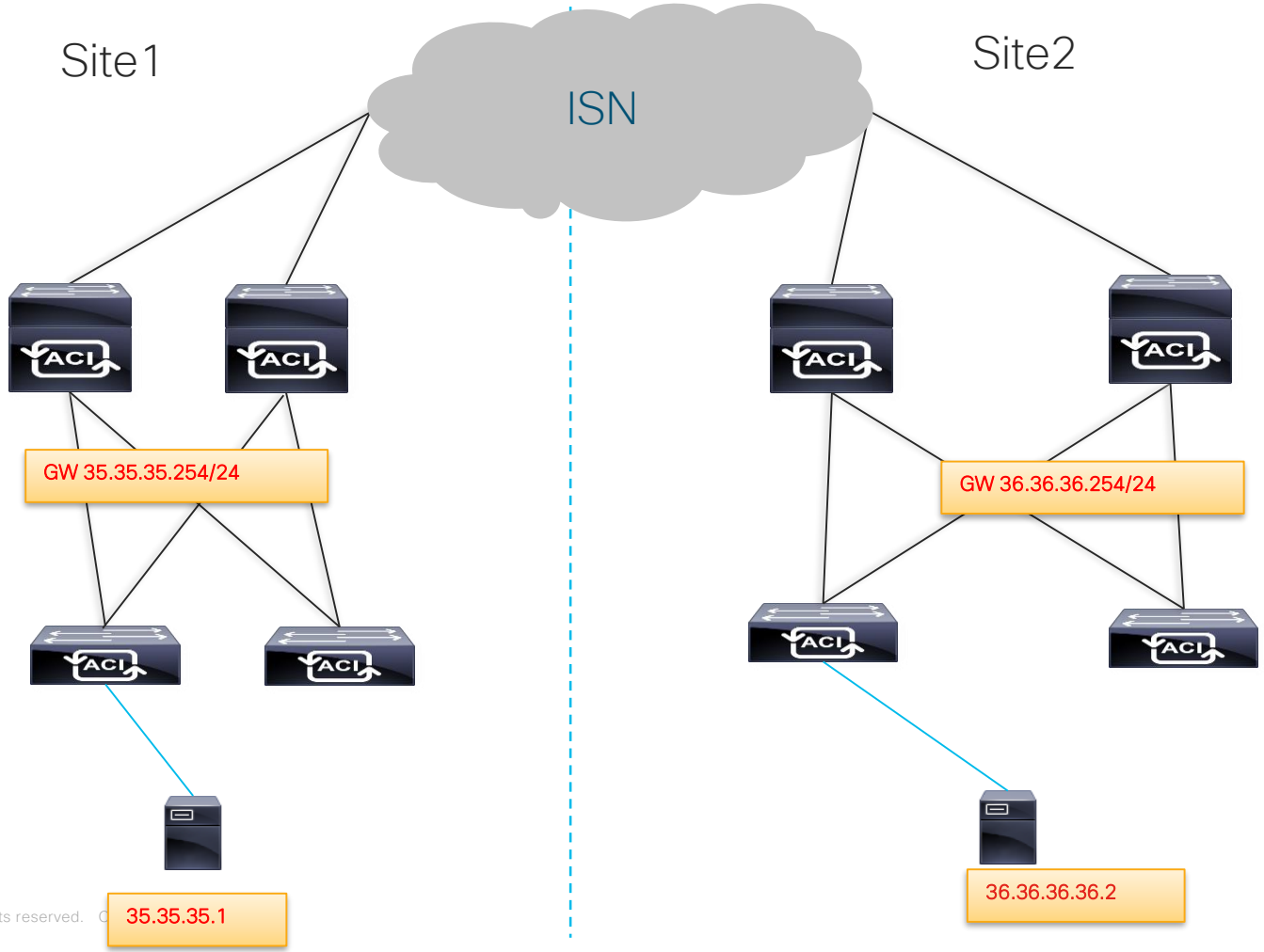
```
*** Parsed Outer l3 vector
hom_elam_in_l3v_ipv4: 0x28288D9428288DE84478002
..
  hom_elam_in_l3v_ipv4.da: 0xA0A237A  10.10.35.122
  hom_elam_in_l3v_ipv4.sa: 0xA0A2365  10.10.35.101

*** Parsed Outer l4 vector
hom_elam_in_l4v_tn: 0x39FFF000002000F2
  hom_elam_in_l4v_tn.tn_nonce_info: 0x8003
  hom_elam_in_l4v_tn.tn_lsb_info: 0x0
  hom_elam_in_l4v_tn.tn_seg_id: 0xE7FFC0

 hom_lurw_vec.info.ifabric_spine.gipo_idx: 0xBF0
 hom_lurw_vec.info.ifabric_spine.is_bd_vnid: 0x1
 hom_lurw_vec.info.ifabric_spine.vnid: 0xE5FFD2
```

MultiSite ARP glean

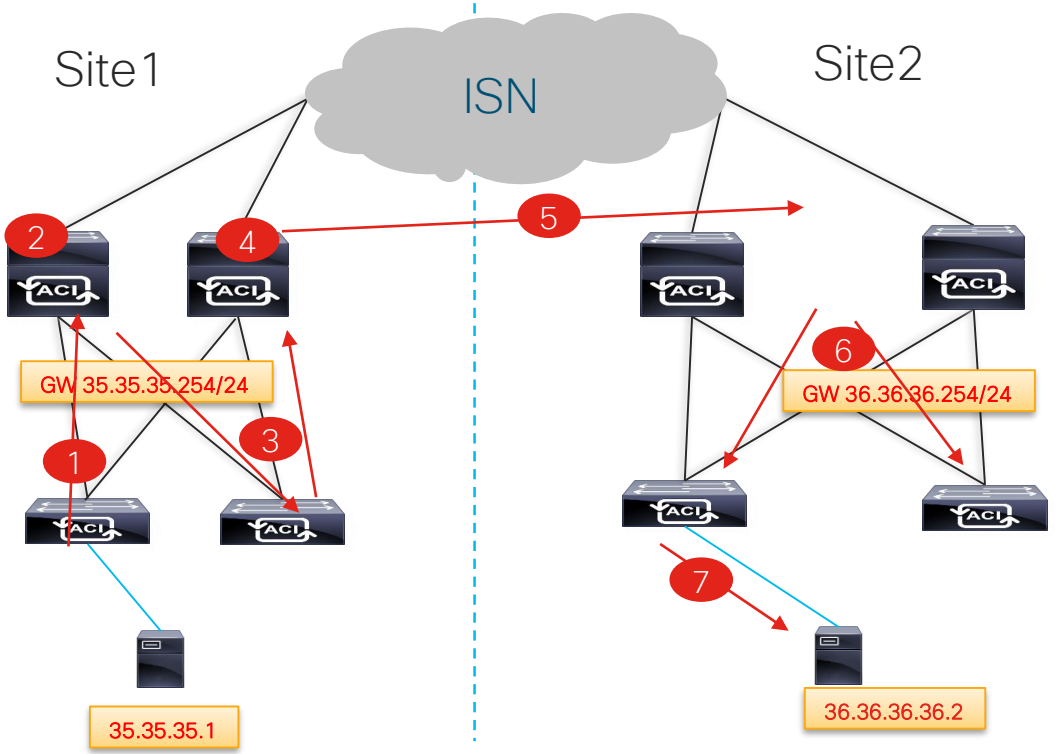
Reminder on ARP glean and multipod glean

- In L3 mode, Spine needs to make a COOP lookup for an EP in a connected BD subnet IP and the EP is unknown
- Spine will generate an ARP glean
- This is flooded in GIPo group 239.255.255.240 within site (or across IPN) and reaches all leaves.
- Ethertype of ARP glean is non standard (etype is FFF2) → you can't get it easily with ELAM using IPv4 or ARP trigger ☹
- Leaves getting it, if the target IP is part of part of one of the BD subnet they have locally, they will generate a real ARP (0x806 etype) to be flooded in the BD
- Before 3.2 THERE was no ARP glean across site in multisite

Topology

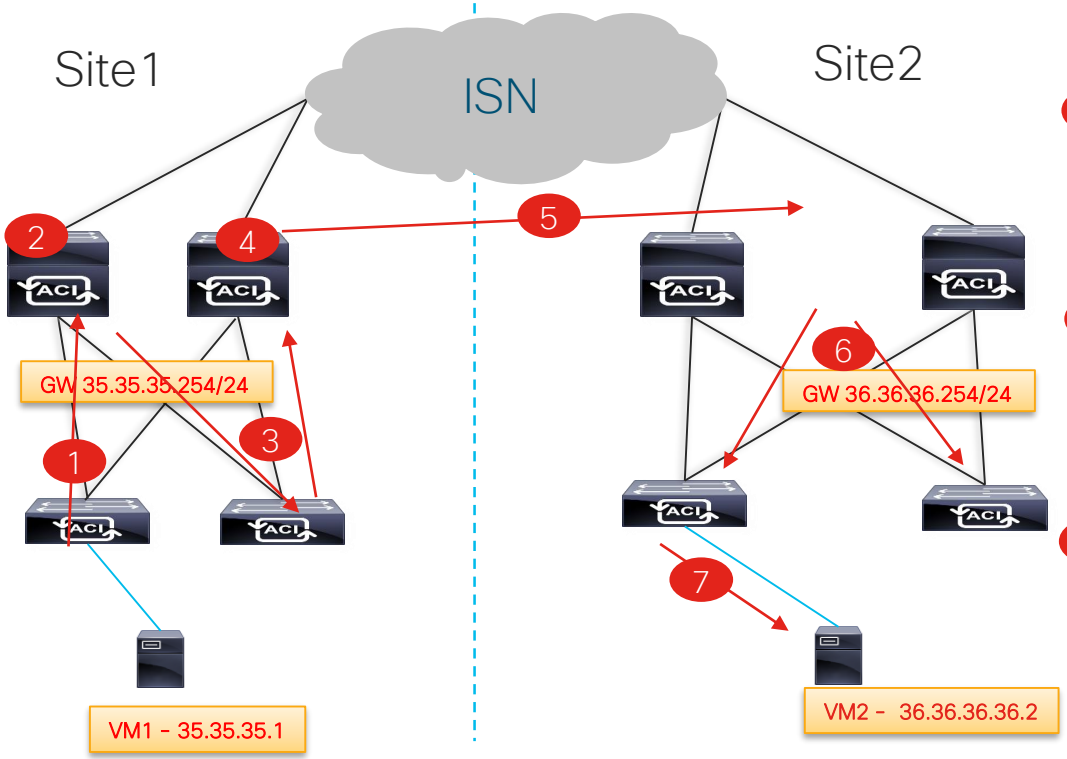


Inter Site Glean – High level



1. IP packet is generated from 35.35.35.1 to 36.36.36.2. Assuming ingress leaf do not know 36.36.36.2, based on its RIB it will send it to anycast spine IPv4
2. The spine that received it will make coop lookup and if 36.36.36.2 is not there (silent host), it will create a MS glean packet.
   - MS glean is the original IP packet encap in Vxlan in VRF VNID with dest ip being 239.255.255.224 (msite glean group)
3. The packet to 239.255.255.224 will be flooded to its GIPo topology in site1 reaching all spines (and leaves)
4. One of the spine in site1 will have in its GIPo topology for 239.255.255.224 a tunnel interface as OIL
5. Packet will be send by that spine (original IP packet encaps in vxlan) with dst IP of outer header being site1 ucast-tep loopback of site2. SRc IP is DP-TEP of site1.
6. One site2 spine gets it and create a usual Mpod ARP glean (flooded to 239.255.255.240) with etype 0xfff2
7. That Mpod glean is flooded in site in GIPo topoloby and each leaves intercept it and send a real ARP packet flooded in the BD on its front panel port

Inter Site Glean – Packet detail



1 Outer : Site1 L1 → SpineAcst

– Inner : IPv4 VM1 to VM2

3 Outer :Site 1 L1 → 239.255.255.224

– Inner IPv4 VM1 to VM2

5 Outer : Site1 ucast TEP → S2 dci-ucast

– inner IPV4 VM1 to VM2

6 Outer : Site1 ucast TEP → 239.255.255.240

– inner Etype 0xfff2 content is ARP with target IP Vm2 – Sender IP 36.36.36.254

7 Real IP packet target IP VM2 – Src IP 36.36.36.254

Forwarding on the TX Site (MSITE Capable Spine)

Ingress LC/FC

- ToR sends the L3 Unicast/L2 ARP Unicast packet towards the spine
- Ingress LC receives the packet with Spine-Proxy as the DIPo
- EP lookup on FC results in a glean-ACL hit, causing a SUP redirect on the FC
- Glean-client on the FC injects an MS-Glean packet, along with existing MP-Glean packet
  - Outer DIPo is a MS-Glean GIPo address, as configured by ISIS
  - Inner ETYPE is preserved, unlike MP-Glean packet which uses a special ETYPE
  - Inner SCLASS is overridden with a reserved value
- MS-Glean GIPo route lookup on the FC results in a Replication List, that contains 1 replication entry each for each of the Remote sites – HREP copy
- For the HREP copy, the packet is sent to Egress LC with some reserved internal vlan representing the corresponding Remote site

Forwarding on the TX Site (MSITE Capable Spine)

- Egress LC
  - Egress LC interprets the internal vlan, and re-writes the outer DIPo with the remote site Unicast ETEP IP
  - Outer Dmac is re-written with that of the selected NH
  - Also, SIPo gets re-written with Unicast ETEP IP of local site
  - HREP copy of the packet is sent towards the IPN

Forwarding on the RX Site (MSITE Capable Spine)

Ingress LC

- HREP packet is received from the IPN with SIPo = Remote site's unicast ETEP, and DIPo = local site's Unicast ETEP IP. This will drive the pkt to DCI Unicast RX path
- Incoming vnid in the pkt is translated to local vnid/BD. BD lookup gives the local sclass
- This packet is then sent to the FC

FC

- Packet follows the regular EP lookup path, which will be miss as the host is silent in the RX site, resulting in MP Glean packet generation
- Glean-client on the RX site will look at the inner sclass in the packet (which is reserved sclass value, programmed at the TX site) – this is to prevent generation of MS Glean copy back to the originating site, that could have caused MS Glean loops across the sites

Site 1 Spine 1 Elam

```
module(DBG-elam-insel6)# set outer ipv4 src_ip 35.35.35.1 dst_ip 36.36.36.2
```

```
*** Parsed Outer l3 vector
  hom_elam_in_l3v_ipv4.da: 0xA00F022    10.0.240.34  anycast-v4 spine
  hom_elam_in_l3v_ipv4.sa: 0xA006040    10.0.96.64   ingress leaf TEP
*** Parsed Outer l4 vector
hom_elam_in_l4v_tn: 0xB000000001000F2
  hom_elam_in_l4v_tn.tn_nonce_info: 0x4003
  hom_elam_in_l4v_tn.tn_lsb_info: 0x0
  hom_elam_in_l4v_tn.tn_seg_id: 0x2C0000
*** Parsed Inner l3 IP vector
hom_elam_in_l3v_ipv6_da_only: 0x8C8C8C040000000000000000000000090909008O4FD002
  hom_elam_in_l3v_ipv6_da_only.ttl: 0x3F
  hom_elam_in_l3v_ipv6_da_only.prot: 0x1
  hom_elam_in_l3v_ipv6_da_only.da: 0x00000000000000000024242402
  hom_elam_in_l3v_ipv6_da_only.sa: 0x23232301
```

Site 1 Spine ELAM after MS glean

```
root@module-2(DBG-elam-insel14)# set inner ipv4 src_ip 35.35.35.1 dst_ip 36.36.36.2
*** Parsed Outer l2 vector
hom_elam_in_l2v_da_sa_qtag0_vntag_cmd: 0x00000E0004002004B000D0D0D0D0D01005E7FFFE0
  hom_elam_in_l2v_da_sa_qtag0_vntag_cmd.da: 0x1005E7FFFE0
  hom_elam_in_l2v_da_sa_qtag0_vntag_cmd.sa: 0xD0D0D0D0D
  hom_elam_in_l2v_da_sa_qtag0_vntag_cmd.qtag0_vld: 0x1
  hom_elam_in_l2v_da_sa_qtag0_vntag_cmd.qtag0_cos: 0x5
*** Parsed Outer l3 vector
hom_elam_in_l3v_ipv4: 0x28018103BFFFFF8047F8052
  hom_elam_in_l3v_ipv4.ttl: 0xFE
  hom_elam_in_l3v_ipv4.prot: 0x11
  hom_elam_in_l3v_ipv4.da: 0xEFFFFFE0      -- 239.255.255.224
  hom_elam_in_l3v_ipv4.sa: 0xA006040       -- 10.00.96.64 Leaf 1 PTEP
*** Parsed Outer l4 vector
hom_elam_in_l4v_tn: 0xB00000010000072
  hom_elam_in_l4v_tn.tn_nonce_info: 0x400001
  hom_elam_in_l4v_tn.tn_lsb_info: 0x0
  hom_elam_in_l4v_tn.tn_seg_id: 0x2C0000
*** Parsed Inner l2 vector
hom_elam_in_l2v_da_sa_qtag0: 0x0000C0C0C0C0CFFFFFFFFFFFF
  hom_elam_in_l2v_da_sa_qtag0.da: 0xFFFFFFFFFFFF
  hom_elam_in_l2v_da_sa_qtag0.sa: 0xC0C0C0C0C
*** Parsed Inner l3 IP vector
hom_elam_in_l3v_ipv6_da_only: 0x8C8C8C040000000000000000000000090900804FD002
hom_elam_in_l3v_ipv6_da_only.ttl: 0x3F
  hom_elam_in_l3v_ipv6_da_only.prot: 0x1
  hom_elam_in_l3v_ipv6_da_only.da: 0x00000000000000024242402       36.36.36.2
  hom_elam_in_l3v_ipv6_da_only.sa: 0x23232301                      35.35.35.1
```

Site 1 Spine control plane

```
pod35-spine1# show  isis  internal mcast routes gipo  | egrep -A 5 "239.255.255.224"
 GIPo: 239.255.255.224 [LOCAL]
    OIF List:
      Ethernet2/1.37
      Ethernet2/2.38
      Tunnel4
pod35-spine1# show ip mroute 239.255.255.224 vrf overlay-1
IP Multicast Routing Table for VRF "overlay-1"

(*, 239.255.255.224/32), Multisite Glean GIPO Route, uptime: 2w0d, isis
  Incoming interface: Null, RPF nbr: 0.0.0.0
  Outgoing interface list: (count: 3)
    Tunnel4, uptime: 1w2d
    Ethernet2/1.37, uptime: 2w0d
    Ethernet2/2.38, uptime: 2w0d
pod35-spine1# show interface tunnel 4
Tunnel4 is up
    MTU 9000 bytes, BW 0 Kbit
    Transport protocol is in VRF "overlay-1"
    Tunnel protocol/transport is ivxlan
    Tunnel source 10.0.96.65/32 (lo0)
    Tunnel destination 10.10.36.101/32
```

There should be one spine Per site with A tunnel OIL for 239.255.255.224.

The Tunnel dest is Dci-ucast of site2

Stats on TX spine

- Gleanc is the process that count glean packet
- Only on Fabric module
- You need to check on each fabric module

```
module-24# show system internal gleanc stats

GLEANC Global and Error Statistics
----------------------------------
Packets received                256852
ARP packets                     1
IP packets                      256850
IPv6 packets                    0
Glean packets                   0
Packets sent                    256851
Msite packets sent              256827

Errors:
------
Invalid packet                  0
Card ID not found               0
Encap failed                    0
Decap failed                    1
Sendmsg failed                  0
Msite encap failed              0
Msite sendmsg failed            0

Decap errors:
-------------
Received transmit header        0
IETH header not found           1
VNTAG header not found          0
IP header not found-1           0
IP header not found-2           0
UDP header not found            0
VxLAN header not found          0
VNID is in BD range             0
Invalid Eth type                0
```

Site 2 – spine coming from ISN

root@module-2(DBG-elam-insel14)# set inner ipv4 src_ip 35.35.35.1 dst_ip 36.36.36.2

```
*** Parsed Outer l3 vector
hom_elam_in_l3v_ipv4: 0x28288D942828919447F4052

  hom_elam_in_l3v_ipv4.ttl: 0xFD
  hom_elam_in_l3v_ipv4.prot: 0x11
  hom_elam_in_l3v_ipv4.da: 0xA0A2465    10.10.36.101  -- site 2 ucast dtep
  hom_elam_in_l3v_ipv4.sa: 0xA0A2365    10.10.35.101  -- site 1 ucast dtep
*** Parsed Outer l4 vector
hom_elam_in_l4v_tn: 0xB00000010000072
  hom_elam_in_l4v_tn.tn_nonce_info: 0x400001
  hom_elam_in_l4v_tn.tn_lsb_info: 0x0
  hom_elam_in_l4v_tn.tn_seg_id: 0x2C0000                  - vnid before rewrite
*** Parsed Inner l3 IP vector
hom_elam_in_l3v_ipv6_da_only: 0x8C8C8C0400000000000000000000000090909008O4FD002
  hom_elam_in_l3v_ipv6_da_only.da: 0x00000000000000024242402
  hom_elam_in_l3v_ipv6_da_only.sa: 0x23232301


 hom_lurw_vec.info.ifabric_spine.sclass: 0x1   - reserved sclass in MS-glean packet
 hom_fpx_lookup_vec.lkup.fibdakey.addr: 0x00000000000000024242402    -- 36.36.36.2
 hom_fpx_lookup_vec.lkup.fibdakey.vrf: 0x278001                     - Local vnid after Vnid rewrite
```

Site 2 – dst leaf

```
module-1(DBG-elam-insel14)# set outer ipv4 dst_ip 239.255.255.240 src_ip 10.10.35.101
```

```
*** Parsed Outer l3 vector
hom_elam_in_l3v_ipv4: 0x28288D97BFFFFFC047F8052
  hom_elam_in_l3v_ipv4.ttl: 0xFE
  hom_elam_in_l3v_ipv4.prot: 0x11
  hom_elam_in_l3v_ipv4.da: 0xEFFFFFF0     225.255.255.240
  hom_elam_in_l3v_ipv4.sa: 0xA0A2365      10.10.35.101
*** Parsed Outer l4 vector
hom_elam_in_l4v_tn: 0x9E0004010004032
  hom_elam_in_l4v_tn.tn_nonce_info: 0x400100
  hom_elam_in_l4v_tn.tn_lsb_info: 0x0
  hom_elam_in_l4v_tn.tn_seg_id: 0x278001
*** Parsed Inner l2 vector
hom_elam_in_l2v_da_sa_qtag0: 0x0000C0C0C0C0CFFFFFFFFFFFF
  hom_elam_in_l2v_da_sa_qtag0.da: 0xFFFFFFFFFFFF
  hom_elam_in_l2v_da_sa_qtag0.sa: 0xC0C0C0C0C0C
  hom_elam_in_l2v_da_sa_qtag0.qtag0_vld: 0x0
  hom_elam_in_l2v_da_sa_qtag0.qtag0_cos: 0x0
  hom_elam_in_l2v_da_sa_qtag0.qtag0_de: 0x0
  hom_elam_in_l2v_da_sa_qtag0.qtag0_vlan: 0x0
*** Parsed Inner l3 IP vector                          Inner is not IP can't be sure this is our frame
hom_elam_in_l3v_ipv6_da_only: 0x00000000000000000045000054000040001AD5CFF0
  hom_elam_in_l3v_ipv6_da_only.l3_type: 0x0
  hom_elam_in_l3v_ipv6_da_only.v6: 0x0
  hom_elam_in_l3v_ipv6_da_only.dscp: 0x3F
  hom_elam_in_l3v_ipv6_da_only.ecn: 0x3
  hom_elam_in_l3v_ipv6_da_only.df: 0x0
  hom_elam_in_l3v_ipv6_da_only.mf: 0x0
  hom_elam_in_l3v_ipv6_da_only.ttl: 0x57
  hom_elam_in_l3v_ipv6_da_only.prot: 0x6B
  hom_elam_in_l3v_ipv6_da_only.da: 0x00000000011400015000001000
  hom_elam_in_l3v_ipv6_da_only.sa: 0x0
```

WE know we RX an mpod glean packet but no way to know if it is the right payload
(as it is not an IP frame

Site 2 – leaf Rx glean ARP

```
pod36-leaf1# tcpdump2 -i tahoe0   | egrep -A 5 -B 12 fff2


14:43:27.119780  Broadcom headers, TX packet 16 bytes
HG_VLAN: 99, DST_MOD: 0, DST_PORT: 0, OPCODE: 2, TC_CLASS: 0
IS-IS, p2p IIH, src-id 4080.000a.0000, length 53
14:43:27.404443  Unknown type of packet , card generation: 1
00:00:00:00:00:00 (oui Ethernet) > 01:01:01:01:01:01 (oui Unknown), ethertype Unknown (0xa00e), length 276:
        0x0000:  0000 0000 0000 0000 0000 77d6 7a64 4141  ..........w.zdAA
        0x0010:  0014 2100 0000 fb18 0010 0000 0f40 0002  ..!..........@..
        0x0020:  4901 8201 40b5 0100 5e7f fff0 000d 0d0d  I...@...^.......
        0x0030:  0d0d 8100 a002 0800 4514 00ce 0000 0000  ........E.......
        0x0040:  fe11 9eab 0a0a 2365 efff fff0 e872 beef  ......#e.....r..    10.10.35.101 -> 239.255.255.240
        0x0050:  00ba 0000 c840 0100 2780 0100 ffff ffff  .....@..'.......
        0x0060:  ffff 000c 0c0c 0c0c fff2 4500 0054 0000  ..........E..T..   Etype fff2
        0x0070:  4000 3f01 ad5f 2323 2301 2424 2402 0800  @.?.._###.$$$...    35.35.35.1 to 36.36.36.2
        0x0080:  8796 057f 44e6 929e e75c 0000 0000 ec35  ....D....\.....5
        0x0090:  0100 0000 0000 1011 1213 1415 1617 1819  ................
        0x00a0:  1a1b 1c1d 1e1f 2021 2223 2425 2627 2829  .......!"#$%&'()
        0x00b0:  2a2b 2c2d 2e2f 3031 3233 3435 3637 0000  *+,-./01234567..
--
```

Egress leaf site 2- Show ip arp internal event-history event

```
pod36-leaf1# show ip arp internal event-history event | more
1) Event:E_DEBUG_DSF, length:143, at 272379 usecs after Fri May 24 13:11:48 2019
    [116] TID 19946:arp_send_request_internal:4687: log_collect_arp_pkt; dip = 36.36.36.2; interface =
Vlan10;iod = 74; Info = Internal Request Do
ne
2) Event:E_DEBUG_DSF, length:138, at 272315 usecs after Fri May 24 13:11:48 2019
    [116] TID 19946:arp_handle_inband_glean:3240: log_collect_arp_glean;dip = 36.36.36.2;interface = Vlan10;info
= Received pkt Fabric-Glean: 1
3) Event:E_DEBUG_DSF, length:163, at 272311 usecs after Fri May 24 13:11:48 2019
    [116] TID 19946:arp_handle_inband_glean:3232: log_collect_arp_glean; dip = 36.36.36.2; interface = Vlan10;
vrf = RD:RD; info = Address in PSVI
 subnet or special VIP
4) Event:E_DEBUG_DSF, length:105, at 272300 usecs after Fri May 24 13:11:48 2019
    [116] TID 19946:arp_objstore_check_if_address_in_subnet:1234: addr 0x24242402 and 242424fe/24 subnet match
5) Event:E_DEBUG_DSF, length:145, at 272126 usecs after Fri May 24 13:11:48 2019
    [116] TID 19946:arp_handle_inband_glean:3101: log_collect_arp_glean;sip = 35.35.35.1;dip = 36.36.36.2;info =
Received glean packet is an IP pa
cket
6) Event:E_DEBUG_DSF, length:177, at 272113 usecs after Fri May 24 13:11:48 2019
    [116] TID 19946:arp_process_receive_packet_msg:7072: log_collect_arp_pkt; filter_id = 5; context_id = 5;
table_id = 5; iod = 0; pkt_len = 170;
 l2_hdr_len = 14,inner_l3_start = 0
```

Transmit ARP to front panel (tcpdump2)

```
pod36-leaf1# tcpdump2 -xxvvi tahoe0 arp
tcpdump2: listening on tahoe0, link-type CISCO_IETH (TX header 16 Bytes), capture size 262144 bytes
14:38:17.301493  Broadcom headers, TX packet 16 bytes
HG_VLAN: 97, DST_MOD: 0, DST_PORT: 0, OPCODE: 4, TC_CLASS: 0
ARP, Ethernet (len 6), IPv4 (len 4), Request who-has 36.36.36.2 (Broadcast) tell 36.36.36.254, length 46
        0x0000:  fb00 0000 0000 0000 0000 0100 0061 8c00
        0x0010:  ffff ffff ffff 0022 bdf8 19ff 0806 0001
        0x0020:  0800 0604 0001 0022 bdf8 19ff 2424 24fe
        0x0030:  ffff ffff ffff 2424 2402 0000 0000 0000
        0x0040:  0000 0000 0000 0000 0000 0000
14:38:18.301719  Broadcom headers, TX packet 16 bytes
HG_VLAN: 97, DST_MOD: 0, DST_PORT: 0, OPCODE: 4, TC_CLASS: 0
ARP, Ethernet (len 6), IPv4 (len 4), Request who-has 36.36.36.2 (Broadcast) tell 36.36.36.254, length 46
        0x0000:  fb00 0000 0000 0000 0000 0100 0061 8c00
        0x0010:  ffff ffff ffff 0022 bdf8 19ff 0806 0001
        0x0020:  0800 0604 0001 0022 bdf8 19ff 2424 24fe
        0x0030:  ffff ffff ffff 2424 2402 0000 0000 0000
        0x0040:  0000 0000 0000 0000 0000 0000
14:38:19.301706  Broadcom headers, TX packet 16 bytes
HG_VLAN: 97, DST_MOD: 0, DST_PORT: 0, OPCODE: 4, TC_CLASS: 0
ARP, Ethernet (len 6), IPv4 (len 4), Request who-has 36.36.36.2 (Broadcast) tell 36.36.36.254, length 46
        0x0000:  fb00 0000 0000 0000 0000 0100 0061 8c00
        0x0010:  ffff ffff ffff 0022 bdf8 19ff 0806 0001
        0x0020:  0800 0604 0001 0022 bdf8 19ff 2424 24fe
        0x0030:  ffff ffff ffff 2424 2402 0000 0000 0000
        0x0040:  0000 0000 0000 0000 0000 0000
```

L3 multicast over multisite (4.1)

Guidelines and Limitations (4.1)

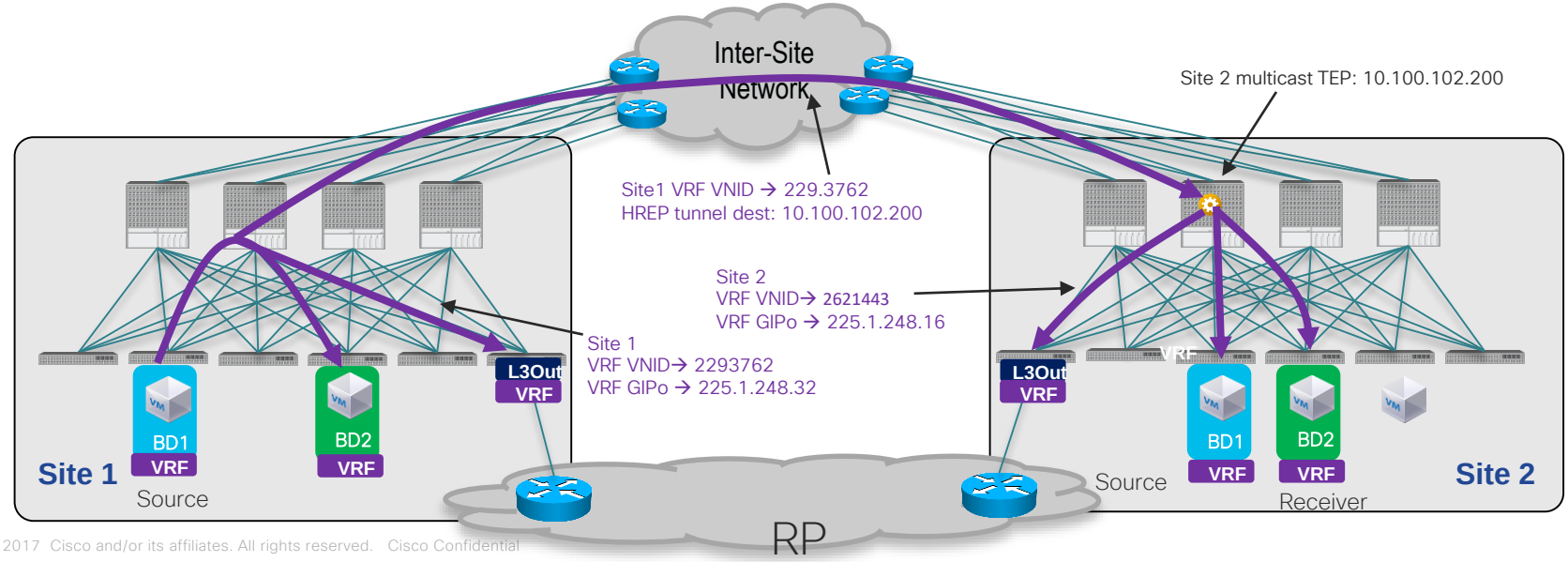
- Each site should have its own L3Out with local Border Leafs. L3Out across sites are not shared.
- All the sites should have common RP and is expected to be reachable independently from each site.
- The source outside is expected to be reachable independently from each site.
- Current solution for multisite works with RP outside the fabric. (RP inside fabric in the future roadmap).
- Multicast traffic is sent to all remote sites that have the VRF deployed, whether there is specific group interest or not. This is similar to BD traffic being sent to remote sites as long as BD is deployed, whether there is group interest or not.

L3 Multicast Multisite overview

- L3 multicast across sites is achieved using VRF GIPo trees in the fabric.
- When VRF is stretched across the sites using MSC and enabled for multicast routing, a VRF GIPo is allocated per VRF and a VRF GIPO tree is formed in the fabric.
- Head End Replication (HREP) tunnels are used for forwarding multicast routed traffic across the sites.
- BDs with L3mcast sources or receivers may or may not be stretched across the sites. The L3Mcast solution still works in either case.

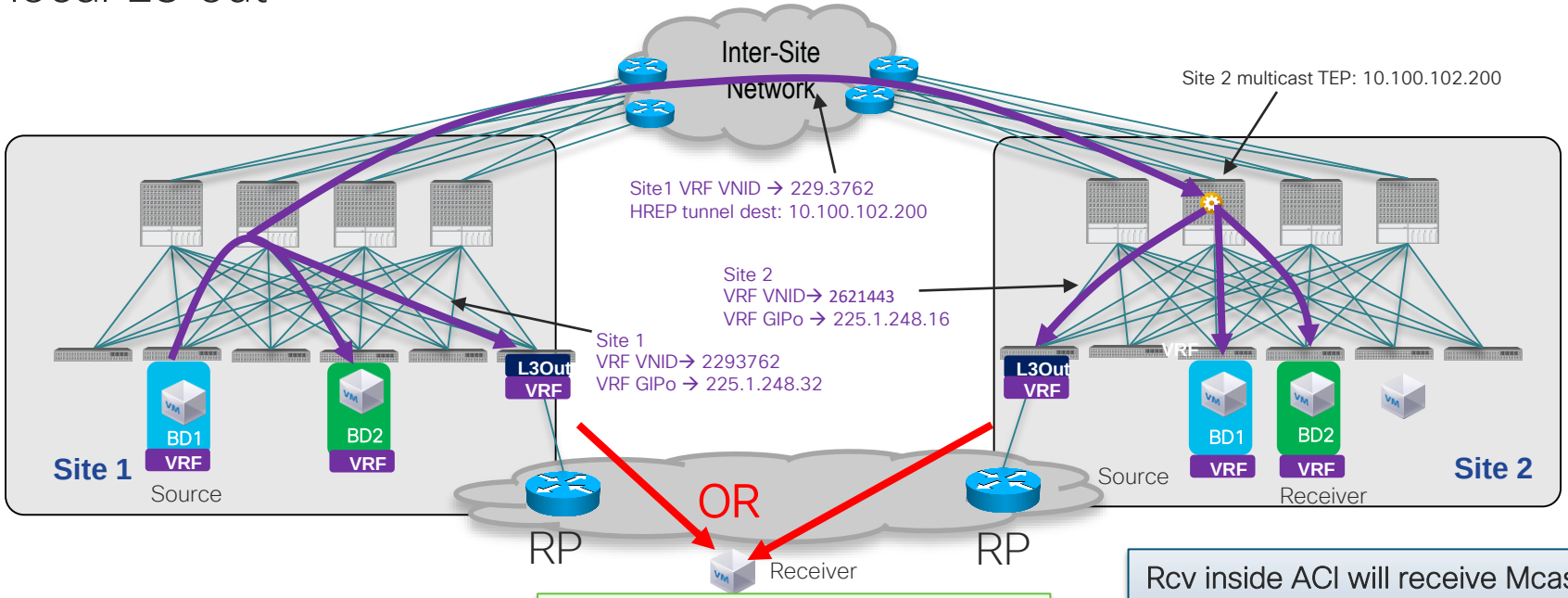
L3 multicast over Multi-site forwarding behavior

- L3 Multicast is always sent to the VRF GIPo within a site (existing behavior)
- Cross site L3 multicast will also be sent in the VRF GIPo. Between sites it is sent over the HREP tunnel to the multicast TEP of the remote site. The VXLAN header will include the source site VRF VNID.
- L3 Multicast at the receiving site will be sent in the VRF GIPo of the receiving site



L3 multicast over Multi-site forwarding behavior

- Site 1 and site2 should have reachability to external RP through their local L3 out



Rcv outside can pull the traffic Either from site 1 or site 2 (depending on where the join lands

Rcv inside ACI will receive Mcast From ACI always (RPF to internal BD subnet

L3 Multicast Multisite overview

- MSC provides a new knob to enable Multicast routing on a stretched VRF.
- When BDs with L3mcast sources are not stretched across the sites, the pervasive subnet information for these BD's need to be pushed to the remote site TORs.
- EPG Pctag/sclass translations need to be pushed to remote site spines.
- A new knob is provided in MSC for the BD's that can have multicast sources to leak this information to remote sites.

L3 Multicast Multisite overview

- For source outside and receiver in a site, the receiver always pulls traffic via site local l3 out (RPF to route on L3 out is always site local). PIM joins to the RP/Source are sent via site local L3 out. Traffic coming in on L3 out on one site is not sent to another sites. This is achieved by a new ACL.
- For Source Inside site1 and Receiver outside, source traffic from site1 to site 2 may go to a receiver via Site2's L3 out. This is because a receiver outside could send a join to either site if the source is inside and the VRF is stretched

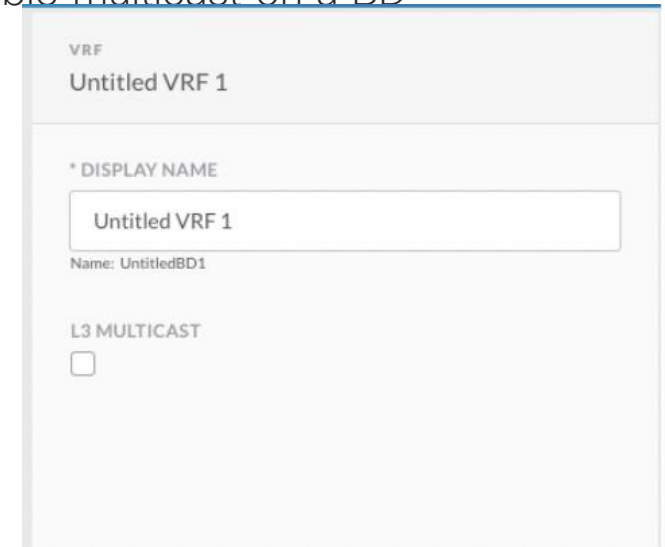
MSC- Multisite Configuration

- New knob in MSC is provided to enable L3 Multicast on a VRF.
- If L3 Multicast knob is enabled, another knob is provided to enable PIM on a BD.
- If PIM is enabled on BD, another knob is provided to enable if Multicast sources are present on the EPG.
- When EPG with a multicast source is not stretched to all sites where VRF is stretched, then MSC creates a shadow EPG on all sites where VRF is stretched.

- When non stretched BD is enabled for multicast routing and EPG under it has multicast source knob enabled, MSC creates a shadow BD and EPG on all sites where VRF is stretched.
- All nodes in the remote sites where the VRF is stretched, APIC downloads shadow BD subnet. Static route with shadow BD subnet gets installed on all the ToRs in remote and local site.

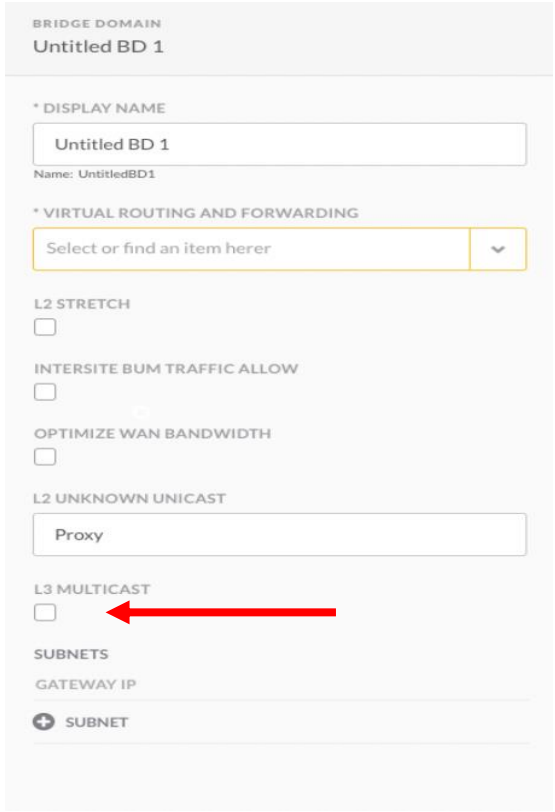
MSC - Multisite Configuration – Step 1

- Enable L3-Multicast on a stretch-VRF. New flag in MSC schema builder page is provided to enable L3 Multicast on a VRF. This will enable PIM on VRF with 1500 MTU on APIC sites.
- Only after multicast is enabled on VRF, user can enable multicast on a BD



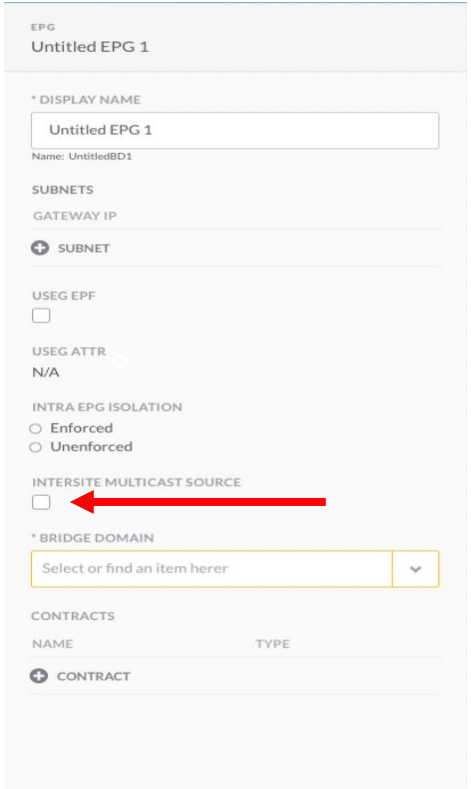
MSC – Multisite Configuration – Step 2

- If L3 Multicast flag is enabled on VRF, another flag is provided to enable L3 Multicast on a BD.
- *If a user tries to enable multicast on a BD who's VRF is not multicast enabled, then MSC UI will disable multicast flag on BD and prompt user to enable multicast on VRF first to use this feature*



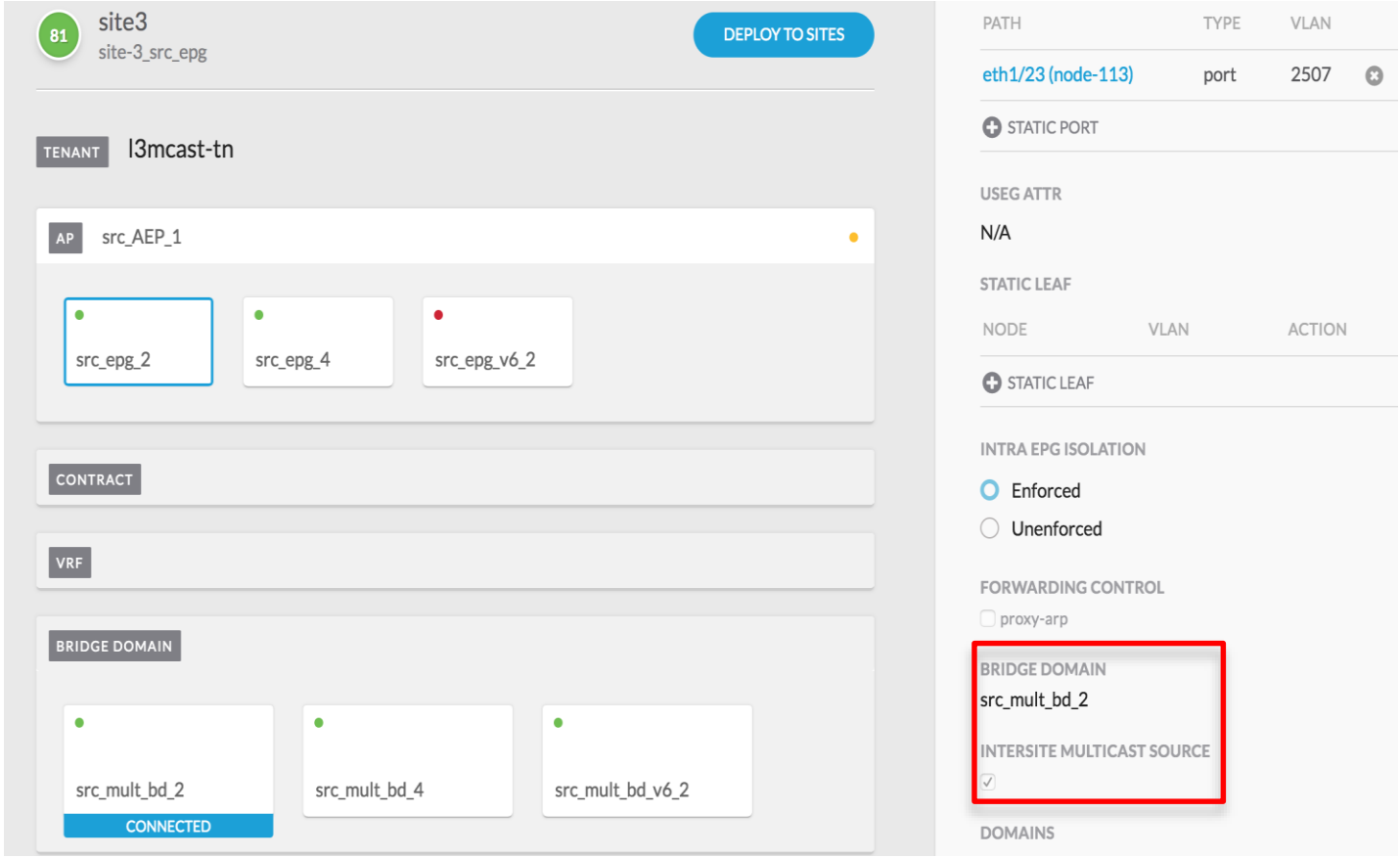
MSC – Multisite Configuration – Step 3

- If L3 Multicast is enabled on BD, another flag is provided to mark an EPG as a Multicast source.
- When EPG with a multicast source is not stretched to all sites where VRF is stretched, then MSO creates a shadow EPG on all sites where VRF is stretched.



APIC

- src_epg_2 is deployed in site 3 with src EPG flag set. BD is not stretched to all the sites



APIC

- BD **src_mult_bd_2** subnet downloaded in site2

In Site 2 where BD not stretched:

```
swmp12-leaf10# sh vlan brief | grep src_mult_bd_2
swmp12-leaf10#
>>>BD not present in the site2

swmp12-leaf10# sh ip route vrf l3mcast-routing-tn:vrf1-en |
grep -A 3 194.169.1
194.169.1.0/24, ubest/mbest: 1/0, attached, direct,
pervasive
    *via 10.5.88.66%overlay-1, [1/0], 1w4d, static
```

In Site 3 where BD is deployed with EPG src flag

```
swmp11-leaf13# sh vlan brief | grep src_mult_bd_2
21 l3mcast-routing-tn:src_mult_bd_2 active   Eth1/23

swmp11-leaf13# sh ip route vrf l3mcast-routing-tn:vrf1-en | grep
194.169.1
194.169.1.0/24, ubest/mbest: 1/0, attached, direct, pervasive
194.169.1.1/32, ubest/mbest: 1/0, attached, pervasive
   *via 194.169.1.1, Vlan21, [1/0], 4d20h, local, local
```

DCI-Mgr remote sclass binding

- Check DCI-Mgr on DCI spine has correctly installed sclass bindings in site 2 for the remote EPG (src_epg_2) in site 3 which has multicast src_epg flag enable

```
swmp12-spine5# show dcimgr repo sclass-maps | grep
16398
------------------------------------------------------
    Remote              |       Local
site Vrf      PcTag     | Vrf      PcTag    Rel-state
------------------------------------------------------
 3   2293760 16398     | 2621440 49166 [formed]
```

fvAEPg	
annotation	orchestrator:msc
childAction	
configIssues	
configSt	applied
descr	
dn	uni/tn-l3mcast-routing-tn/ap-src_AEP_1/epg-src_epg_2
exceptionTag	
extMngdBy	
floodOnEncap	disabled
fwdCtrl	
hasMcastSource	yes
isAttrBasedEPg	no
isSharedSrvMsiteEPg	no
lcOwn	local
matchT	AtleastOne
modTs	2018-05-18T18:38:25.865+00:00
monPolDn	uni/tn-common/monepg-default
name	src_epg_2
nameAlias	
pcEnfPref	enforced
pcTag	16398
prefGrMemb	exclude
prio	unspecified
scope	2293760
shutdown	no

ISIS setup

- Check ISIS sets up VRF GIPO tree with remote site hrep tunnels in OIFlist on the DCI capable spines

```
swmp11-spine5# sh isis internal mcast routes gipo | beg 225.1.248.32
 GIPo: 225.1.248.32 [LOCAL]
   OIF List:
     Ethernet4/1.33
     Ethernet4/20.20(External)
     Ethernet4/27.34
     Tunnel13
     Tunnel15
```

HREP tunnels to remote sites

l3LocalCtxSubstitute

DnName	uni/tn-l3mcast-routing-tn/ctx-vrf1-en
FabEncap	vxlan-2293760
childAction	
dn	topology/pod-1/node-203/sys/inst-overlay-1/localSite-3/localCtxSubstitute-[vxlan-2293760]
lcOwn	local
mcastEncap	225.1.248.32/32
modTs	2018-07-19T20:42:08.584+00:00
status	

ISIS setup

- Check Remote to local vnid mapping on spine so ISIS can add tunnel in the OIFlist

l3RsToLocalCtxSubstitute	
childAction	
dn	topology/pod-1/node-204/sys/inst-overlay-1/remoteSite-2/remoteCtxSubstitute-[vxlan-2621440]/rsToLocalCtxSubstitute
forceResolve	yes
lcOwn	local
modTs	2018-07-19T20:42:46.183+00:00
rType	mo
state	formed
stateQual	none
status	
tCl	l3LocalCtxSubstitute
tDn	topology/pod-1/node-204/sys/inst-overlay-1/localSite-3/localCtxSubstitute-[vxlan-2293760]
tType	mo

Multicast control plane states

- Control plane states are contained within the site.
- BL's across the site don't form PIM neighborship. (through ACI / ISN)
  - Note this is different from single site, where are BL are PIM neighbor together
- PIM hello packets and other control plane SUP generated packets in a site are dropped on the DCI TX spine.
- A new ACL is installed on the TOR for the SUP generated PIM protocol packets which marks the DSCP value (0x39) in the outer vxlan header.
- New ACL on the Spine drops hrep copies going out in RWX for any multicast packets with this DSCP value.

Multicast control plane states

Site 1 and Site2 BLs do not see each other as PIM neighbor

However each Leaf should have tunnel If to the VRF GIPo

Site 1 BLs

```
swmp11-leaf13# show fabric multicast vrf l3mcast-routing-tn:vrf1-en
Fabric Multicast Enabled VRFs
VRF Name              VRF     Vprime       VN-Seg   VRF   Conv Tunnel
                      ID      If           ID       Role Mode IP
l3mcast-routing-tn:vrf1-en4       Tunnel11     2293760   BL   Fast 113.113.113.1
swmp11-leaf13#

swmp15-leaf9# show fabric multicast vrf l3mcast-routing-tn:vrf1-en
Fabric Multicast Enabled VRFs
VRF Name              VRF     Vprime       VN-Seg   VRF   Conv Tunnel
                      ID      If           ID       Role Mode IP
l3mcast-routing-tn:vrf1-en6       Tunnel13     2293760   BL   Fast 115.115.115.1
```

Site 1 PIM neighbors

```
swmp11-leaf13# sh ip pim neighbor vrf l3mcast-routing-tn:vrf1-en
PIM Neighbor Status for VRF "l3mcast-routing-tn:vrf1-en"
Neighbor       Interface           Uptime    Expires DR     Bidir- BFD
                                                     Priority Capable State
115.115.115.1 Tunnel11           2d20h   00:01:24 1       no    n/a
>>>>>116.116.116.1 in site 2 not in the neighbor list
```

Site 2 BLs

```
swmp12-leaf10# show fabric multicast vrf l3mcast-routing-tn:vrf1-en
Fabric Multicast Enabled VRFs
VRF Name              VRF     Vprime       VN-Seg   VRF   Conv Tunnel
                      ID      If           ID       Role Mode IP
l3mcast-routing-tn:vrf1-en7       Tunnel14     2621440   BL   Fast 116.116.116.1
```

Site 2 PIM neighbors

```
PIM Interface Status for VRF "l3mcast-routing-tn:vrf1-en"
Interface            IP Address      PIM DR Address Neighbor Border
                                                    Count   Interface
Tunnel14             116.116.116.1 116.116.116.1 0       no
>>>>115.115.115.1 in site 1 not in PIM neighbor list
```

Multicast control plane states

- Check ACL installed on the ToR to put DSCP 0x39 in SUP generated PIM packets

```
vsh_lc -c "show platform internal hal tcam ac-tcam">/bootflash/ac_tcam_tor.log

for SUP generated PIM packet:

Index:  643 Table: 0,     QOS, Not Wide,   IPv4, Stats-index: 1286, pkts: 1447, bytes: 96616
                    wide_key: 0x0 / 0x0
                    vec_type: 0x2 / 0x0
                      sup_tx: 0x1 / 0x0
                    l3_prot: 0x67 / 0x0
    Result-index: 1286
               rslt_priority: 0x2
                   stats_vld: 0x1
              qos_map_idx_rw: 0x1
                 qos_map_idx: 0xe8
```

In the output, check entry present for PIM proto 0x67 and sup_tx set to 0x1.

From the output get qos_map_idx 0xe8 and look into the rwqosmaptable to check what is action being taken for this packet

Multicast control plane states

- Check ACL installed on the ToR to put DSCP 0x39 in SUP generated PIM packets.

From the precious slide, qos_map_idx is 0xe8 (232 in dec). Look into the rwqosmaptable with this index and check what is action being taken for this packet.

```
module-1# show platform internal sug table tah_sug_rwx_rwqosmaptable 232
Tgt slot is not 24(0 based)
***** ASIC : 0 *****

============================================================
                                    SLICE : 0
============================================================
ENTRY[000232] = ol_dscp=0x39    ol_dscp_rw=0x1

============================================================
                                    SLICE : 1
============================================================
ENTRY[000232] = ol_dscp=0x39    ol_dscp_rw=0x1
```

Multicast control plane states

- Check ACL on spine FC to kill the HREP copy of any packet with DSCP 0x39 going out to other sites

on Spine FC

```
show platform internal hal tcam rw-tcam

Index:     0 Table: 0,      SUP, Not Wide,    IPv4,
                    wide_key: 0x0 / 0x0
                    vec_type: 0x1 / 0x0
                    bd_label: 0x2 / 0x3ffd
                    l3_dscp: 0x39 / 0x0
   Result index: 0
                         drop: 0x1
```

Multicast data plane

- Traffic coming in on L3 out on one site, is not sent to another sites. This is done with same ACL that is explained in previous slides.
- A new ACL will be installed on the TOR for marking the multicast traffic ingressing on l3out BD from the external router with special DSCP value (0x39) in the outer vxlan header.
- New ACL on the Spine FC will kill hrep copies going out in RWX for any multicast packets with this DSCP value.

Multicast data plane

- Check ACL on ingress ToR where the packets are coming in from L3Out

On Ingress TOR:

```
vsh_lc -c "show platform internal hal tcam ac-tcam">/bootflash/ac_tcam_tor.log

Index:  644 Table: 0,     QOS, Not Wide,   IPv4, Stats-index: 1288, pkts: 871062, bytes: 89264648
                  wide_key: 0x0 / 0x0
                  vec_type: 0x2 / 0x0
                  bd_label: 0x200 / 0xdff
                  l3_da: 0xe0000000 / 0xffffffff
    Result-index: 1288
              rslt_priority: 0x2
                  stats_vld: 0x1
             qos_map_idx_rw: 0x1
                qos_map_idx: 0xe8
```

Bd_label for L3out packets is 0x200 and also l3_da is multicast ip.
We set it to same qos mac 0xe8 to mark DSCP as for PIM pakcet

Multicast data plane

- Check ACL on ingress ToR where the packets are coming in from L3Out

On Ingress TOR:

Take the qos_map_idx and look into the rwqosmaptable to check what is action being taken for this packet

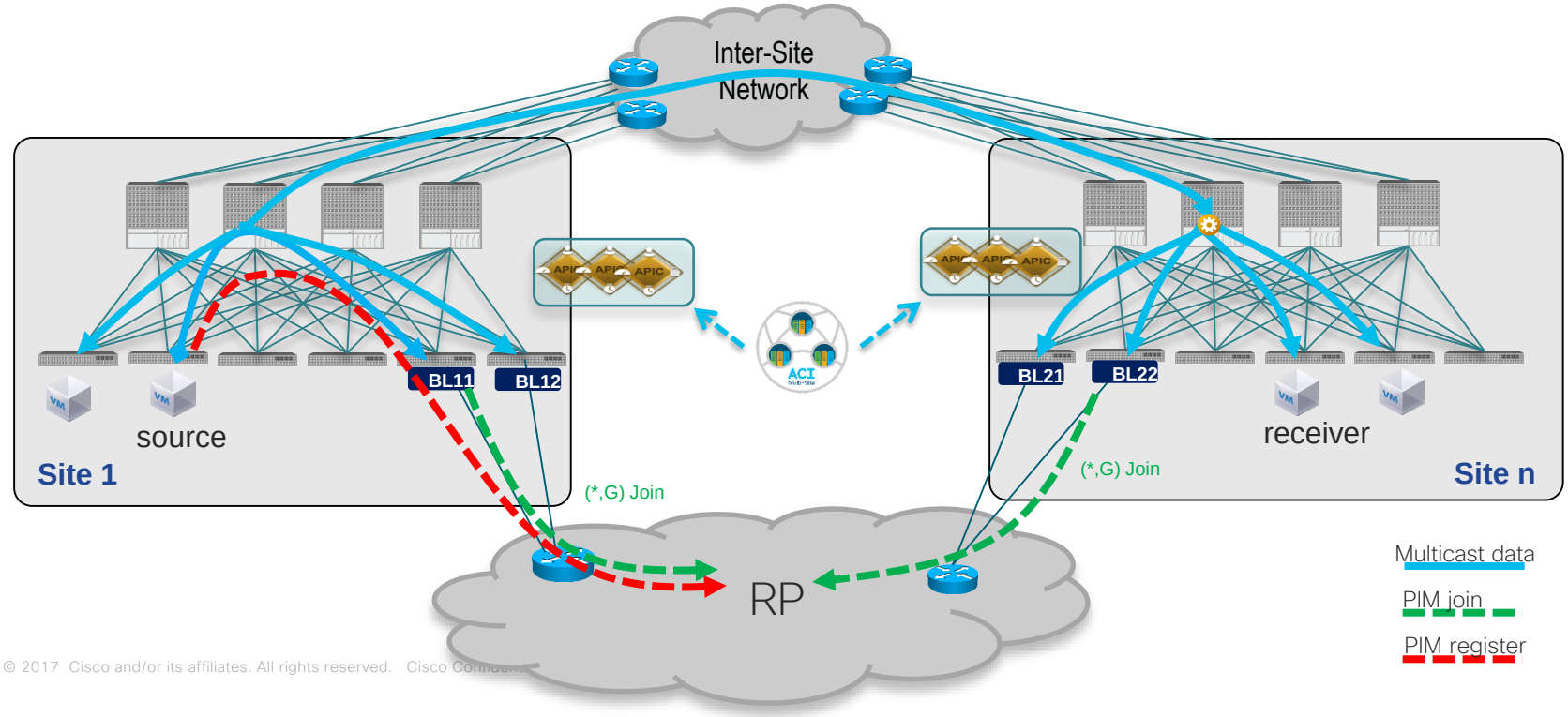
```
module-1# show platform internal sug table tah_sug_rwx_rwqosmaptable 232
Tgt slot is not 24(0 based)
***** ASIC : 0 *****
============================================================
                                    SLICE : 0
============================================================
ENTRY[000232] = ol_dscp=0x39    ol_dscp_rw=0x1
```

- Check ACL on spine FC to kill the HREP copy of any packet with DSCP 0x39 going out to other sites

```
on Spine FC:
show platform internal hal tcam rw-tcam

Index:    0 Table: 0,     SUP, Not Wide,   IPv4,
                  wide_key: 0x0 / 0x0
                  vec_type: 0x1 / 0x0
                  bd_label: 0x2 / 0x3ffd
                  l3_dscp: 0x39 / 0x0
   Result index: 0
                  drop: 0x1
```

Multicast Routing over Multi-site
Sources inside and receivers inside

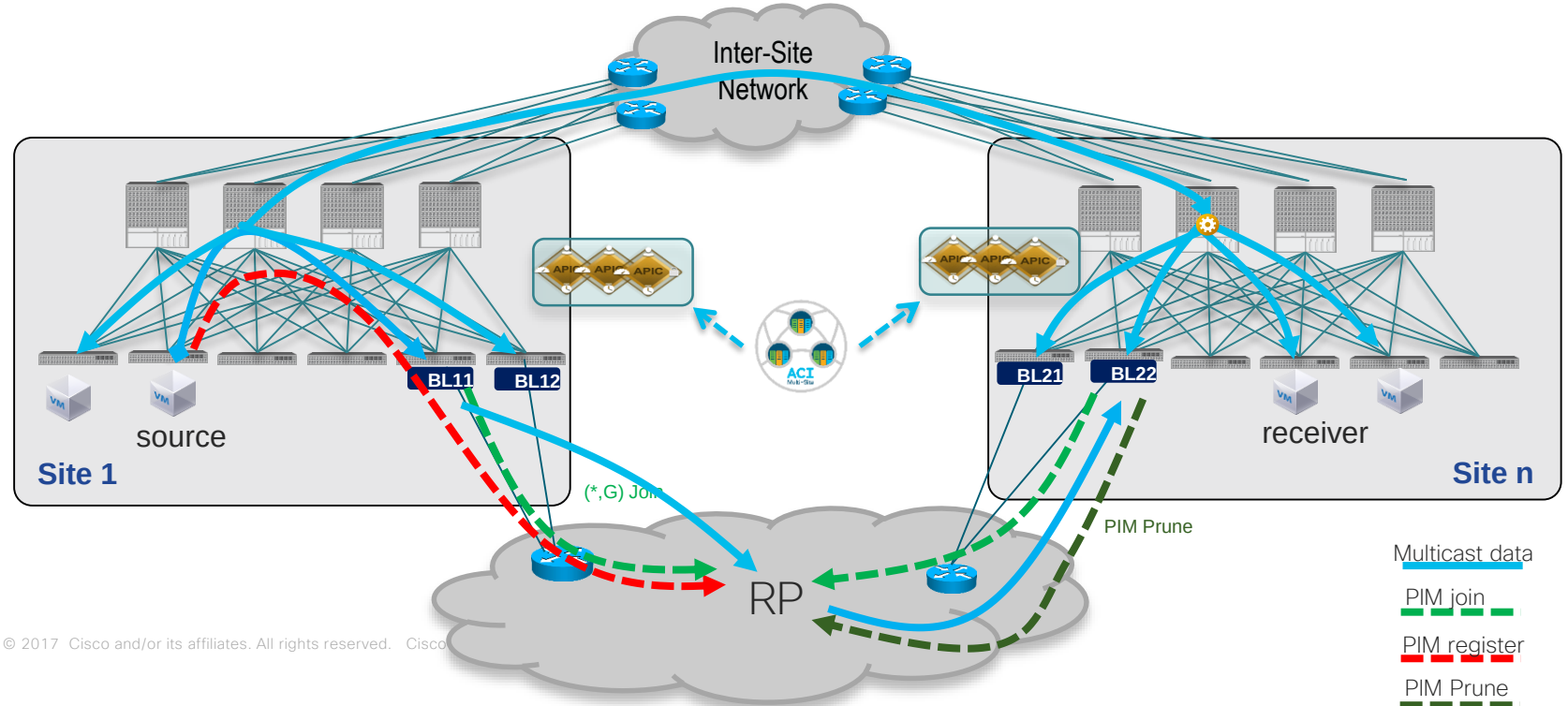
- Source inside and receiver inside will be flooded in the VRF within the site and sent over the HREP tunnel inter-site.



Multicast Routing over Multi-site

Sources inside and receivers inside (cont)

- When RP receives register from source it will forward multicast down the shared tree.
- When BL (BL22) installs (S,G) it will see that source is a pervasive BD and will send prune towards RP.
- Multicast receiver may temporarily receive duplicate packets.



Packet Flows

Source Inside Receiver Inside

1. Source in Site 1 on pervasive BD (Not stretched to site 2) on leaf L11.
2. Receiver R1 is in Site 2 . In site 2 receiver leaf receives igmp join (*,G) and installs (*,G) tree
3. Site 2 BL21 and BL22 receive interest from COOP. BL21 is stripe Owner for multicast group G and responsible for forwarding traffic from outside RP in the fabric. BL21 and BL22 send pim join to RP
4. On source leaf in site 1 when S starts sending traffic, src leaf performs PIM FHR functionality.
5. (S, G) is a miss and DC=1 causes Punt to CPU. Src leaf will send PIM register to RP.
6. RP send PIM join for (S,G) towards source. PIM join for S,G from RP can be targeted to any of the BL (either site 1 or site 2) BL11, BL12, BL21, BL22.
7. From RP source S in Site 1 is reachable via Site 1 – BL11 BL12, Site 2 – BL21 B22 (ECMP).
8. RP sends PIM register stop to source on L11

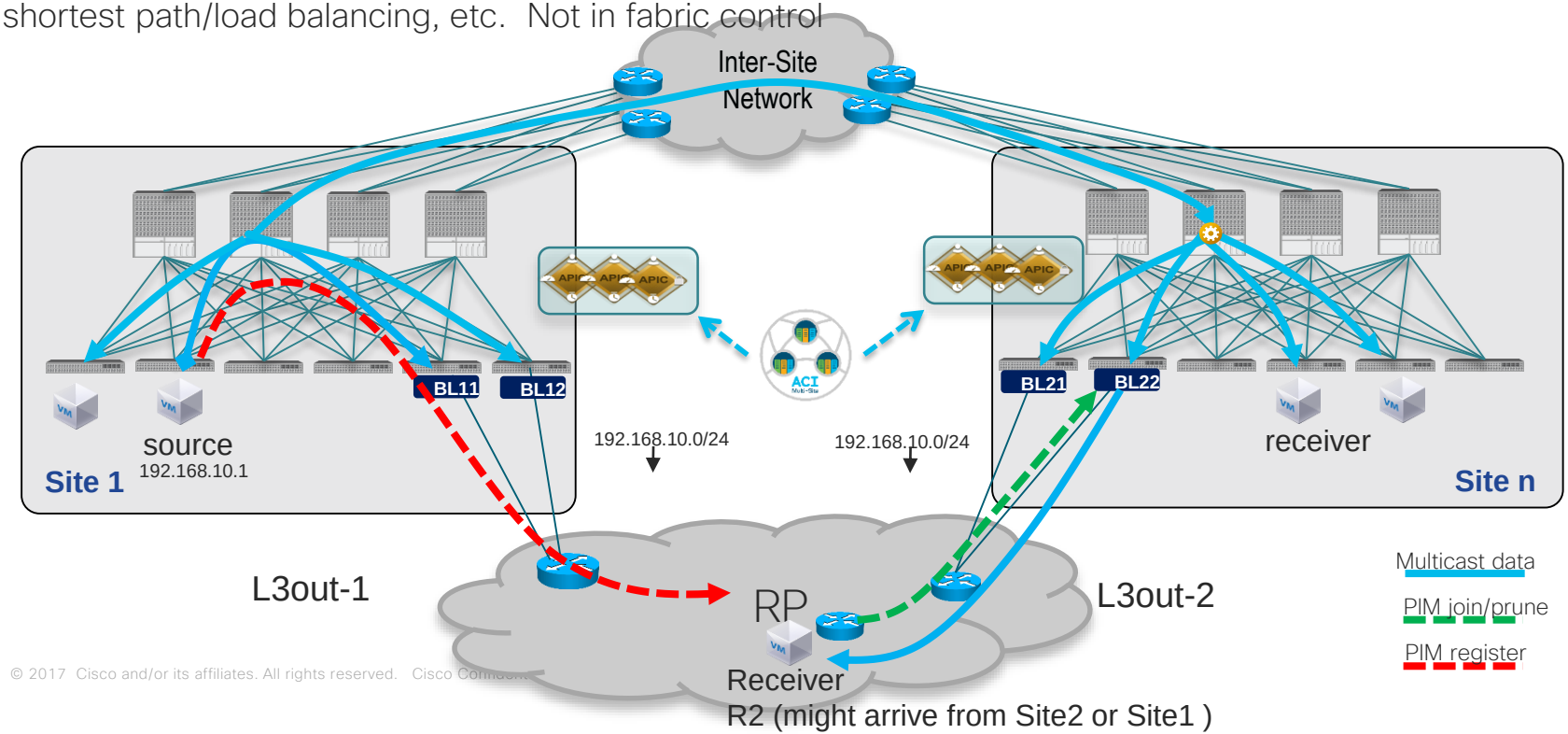
Packet Flows

Source Inside Receiver Inside

10. Now both will happen
    a) Source is sending traffic over IPN on Hrep tunnel to Site 2. This traffic is received on all leaf in Site 2 (src leaf->Site 1 Spine -IPN->S2 Spinte->Site 2 Leaf and BL21, BL22..), OR/AND
    b) Source send traffic to RP which sends traffic to Border leaf BL21 via (*,G). The (*,G) tree on BL21 will also send a copy to CPU to PIM to install (S,G). As soon as PIM get the packet, PIM will know source is pervasive. PIM sends (S,G, rpt) prune to RP.
11. Since 10a, 10b can happen at the same time, it possible for L22 to receive duplicates from IPN or border leaf, when forwarding is happening on (*,G) tree. But as soon as (S,G) is installed in site 2 BL, L22 will receive traffic via IPN.
12. Source S in site 1 to receiver R1 in site 2, traffic is always received via IPN Hrep.

Multicast Routing over Multi-site
Sources inside and receivers outside

- Receiver outside can receive multicast from either site.
- The site where the (S,G) join is received is determined by the external multicast devices based on shortest path/load balancing, etc. Not in fabric control



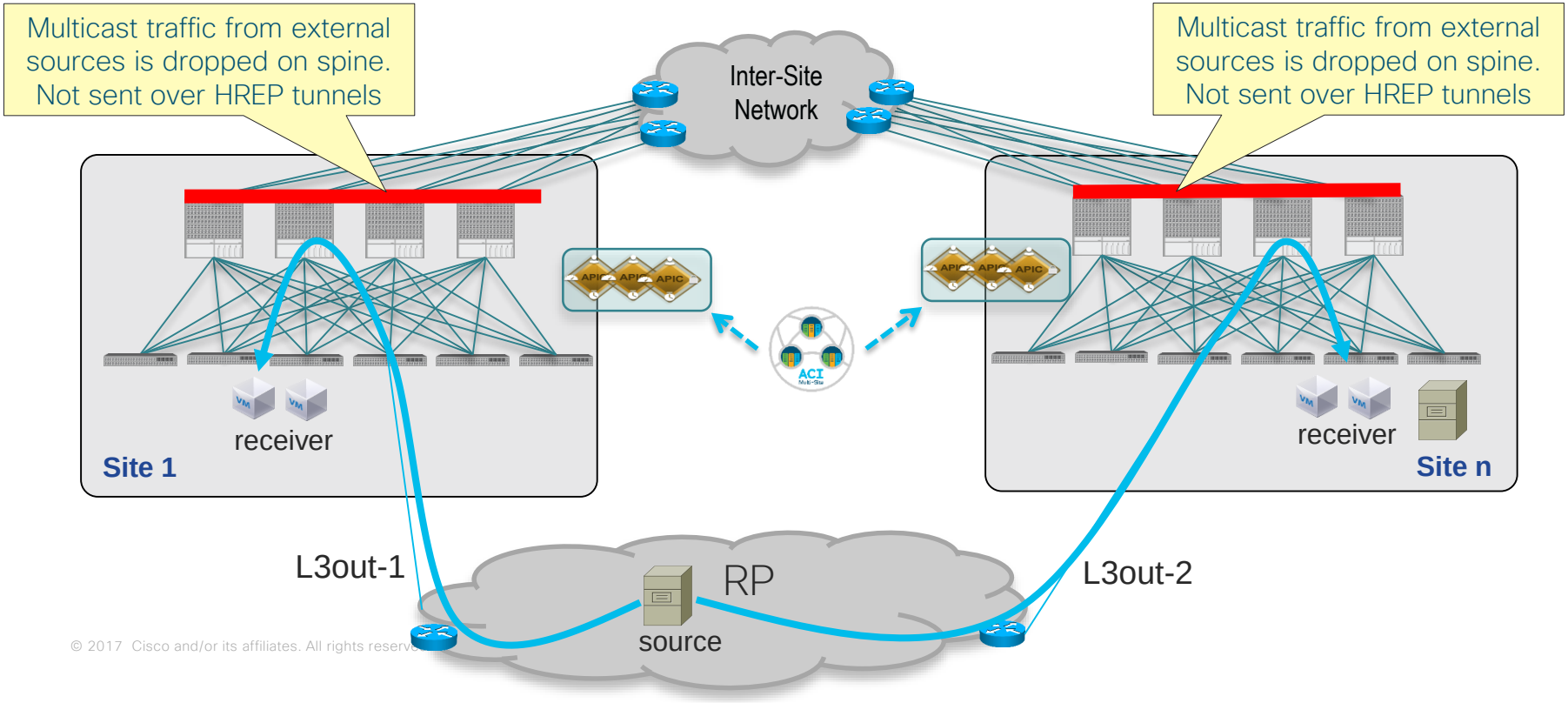
Packet Flows

Source Inside Receiver Outside

1. Source is in site 1 and receiver R2 is outside the fabric.
2. RP receives join from R2 and sets up (*,G) tree.
3. Source in Site 1 on pervasive BD (Not stretched to site 2) on source leaf sends traffic. Ingress src leaf performs PIM FHR functionality.
4. (S, G) is a miss and DC=1 causes Punt to CPU. L11 will send Pim register to RP.
5. RP receives register message from Source Leaf.
6. RP send PIM join for source. PIM join for S,G from RP can be targeted to any of BL11, BL12, BL21, BL22. From RP, source is reachable via BL11 BI12, B21 B22 (ECMP).
7. If PIM (S,G) join comes on Site 1 BL12, Receiver R2 outside the fabric will receive traffic via BL12 (from site 1).
8. If PIM (S,G) join comes on Site 2 BL22, Receiver R2 outside the fabric will receive traffic via BL22 (from site 2).
9. RP sends PIM register stop to source on L11

Multicast Routing over Multi-site
Sources outside receivers inside

- Each site must receive multicast sent from external sources via local L3out



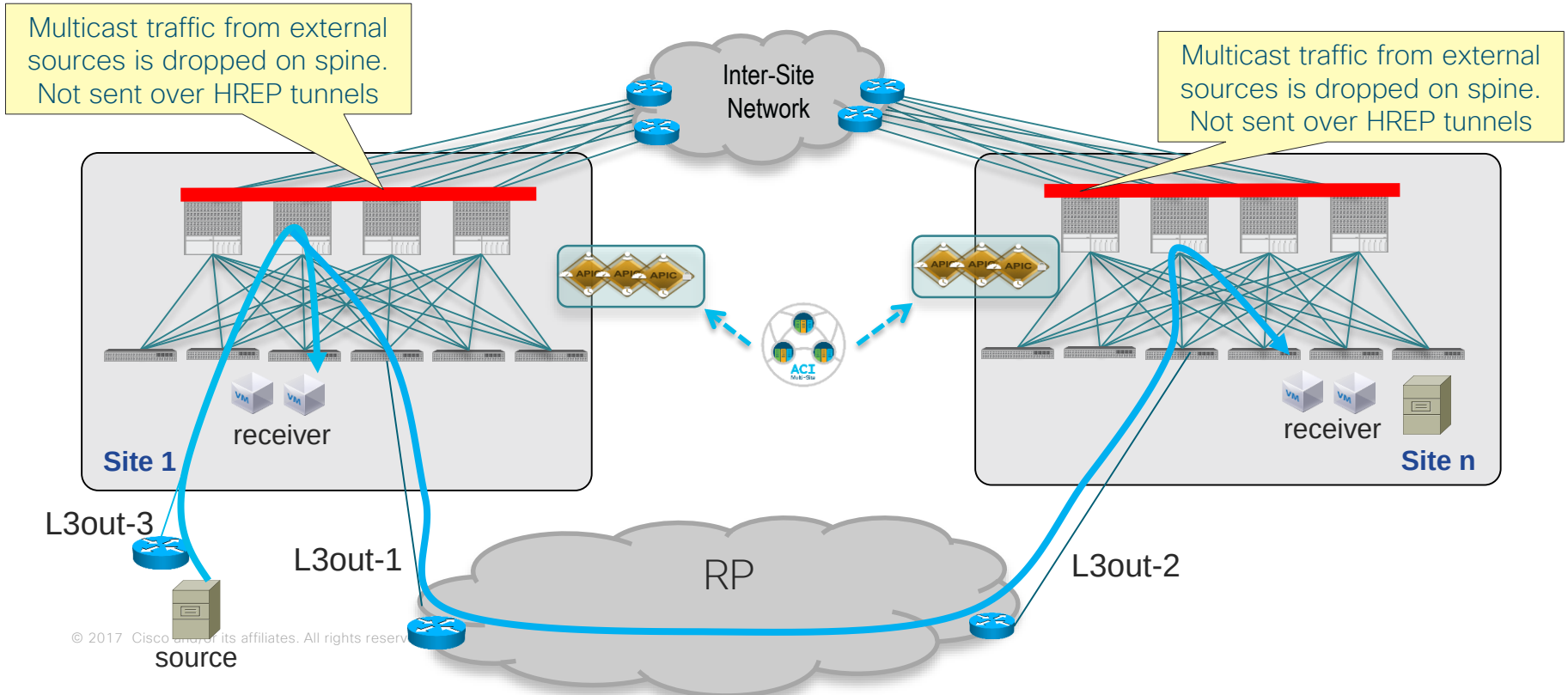
Packet Flows

Source Outside Receiver Inside

1. Source S is outside the fabric.
2. Site 1 receiver leaf receives igmp join (*,G) from R1 and installs (*,G) tree.
3. Site 2 receiver leaf receives igmp join (*,G) from R2 and installs (*,G) tree
4. Site 1 BL11 and BL12 receive interest from COOP. BL11 is stripe Owner for G in site 1. BL11 send pim join to RP.
5. Site 2 BL21 and BL22 receive interest from COOP. BL21 is stripe Owner for G in site 2. BL21 send pim join to RP.
6. Each site pulls traffic from outside Source S using its own L3out. (no stripping across SITE !!)
7. Source start sending traffic. RP gets the traffic and will forward the traffic to Site 1 and Site2 BLs.
8. BL11 and BL12 receive traffic. BL11 (Stripe Winner) will forward the traffic inside Site 1. This traffic will also be marked with a special DSCP 0x39. On BL12 traffic is dropped by (*,G) due to RPF failure. Both BL11 and BL12 will send copy to CPU to PIM. PIM will send pim join to source and send (S,G, rpt) prune to RP
9. BL11 will forward the traffic on VRF GIPO tree.
10. Site 1 Spine drops the IPN HREP copy for the packet with DSCP 0x39.
11. Site 1 receivers will receive the traffic
12. Steps 6-8 are similar for site 2.
13. Site 2 will receive the traffic via it own L3out

Multicast Routing over Multi-site
Sources outside receivers inside with transit case

- Transit multicast use case is supported. One site can be transit for an external source and that multicast flow can arrive at another site via the local L3out. Multicast is not sent over the ISN in this case.

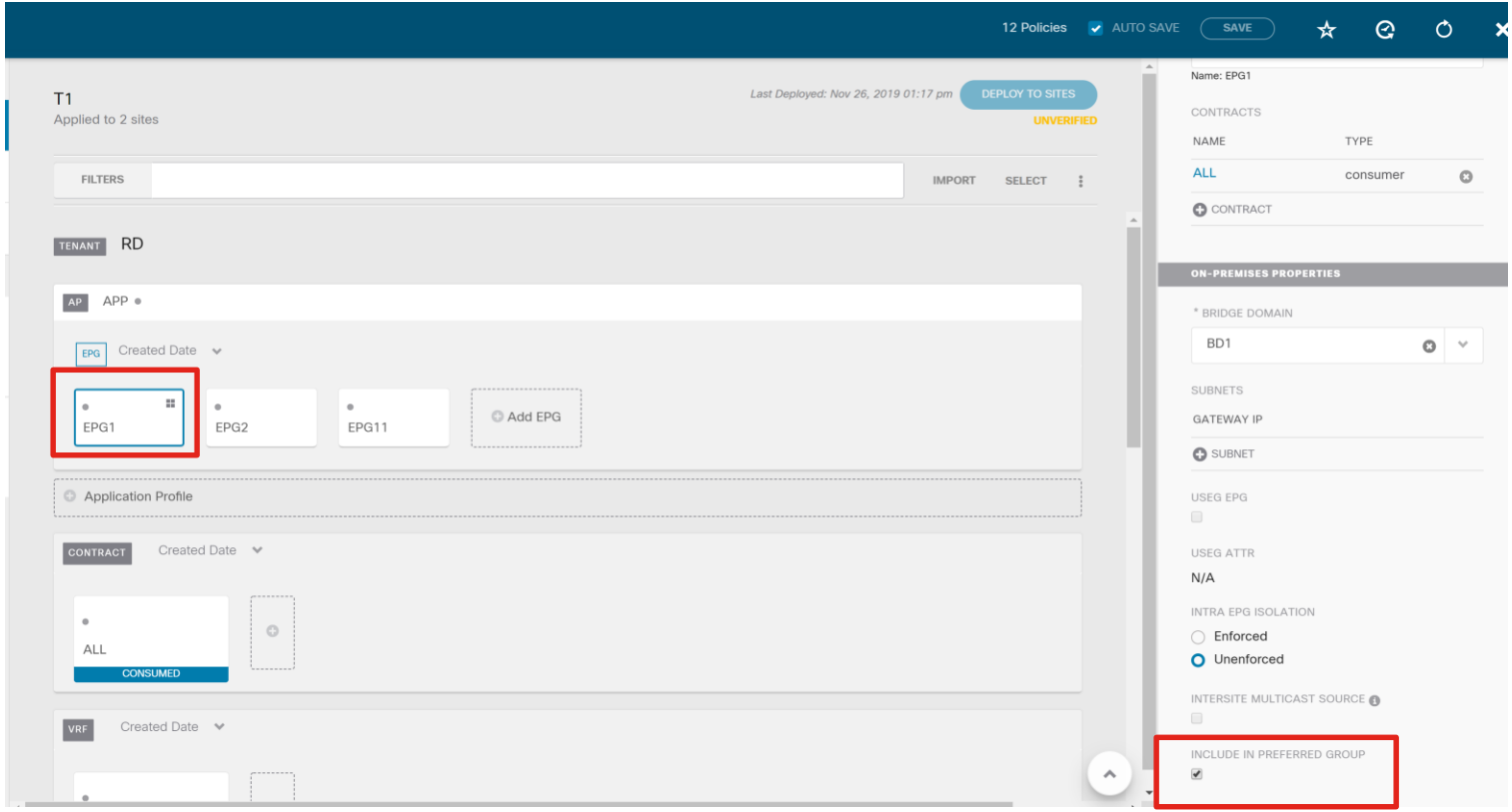


Preferred group and multisite (4.1)

Preferred in multisite

- MSO in 4.1 allows to mark EPG in preferred group
- Marking any EPG of a VRF in preferred group enable preferred group global vrf setting
- Any EPG in preferred has a shadow EPG deployed on any site where the VRF exists (without the need of a cross site contract)

Adding an EPG to preferred group

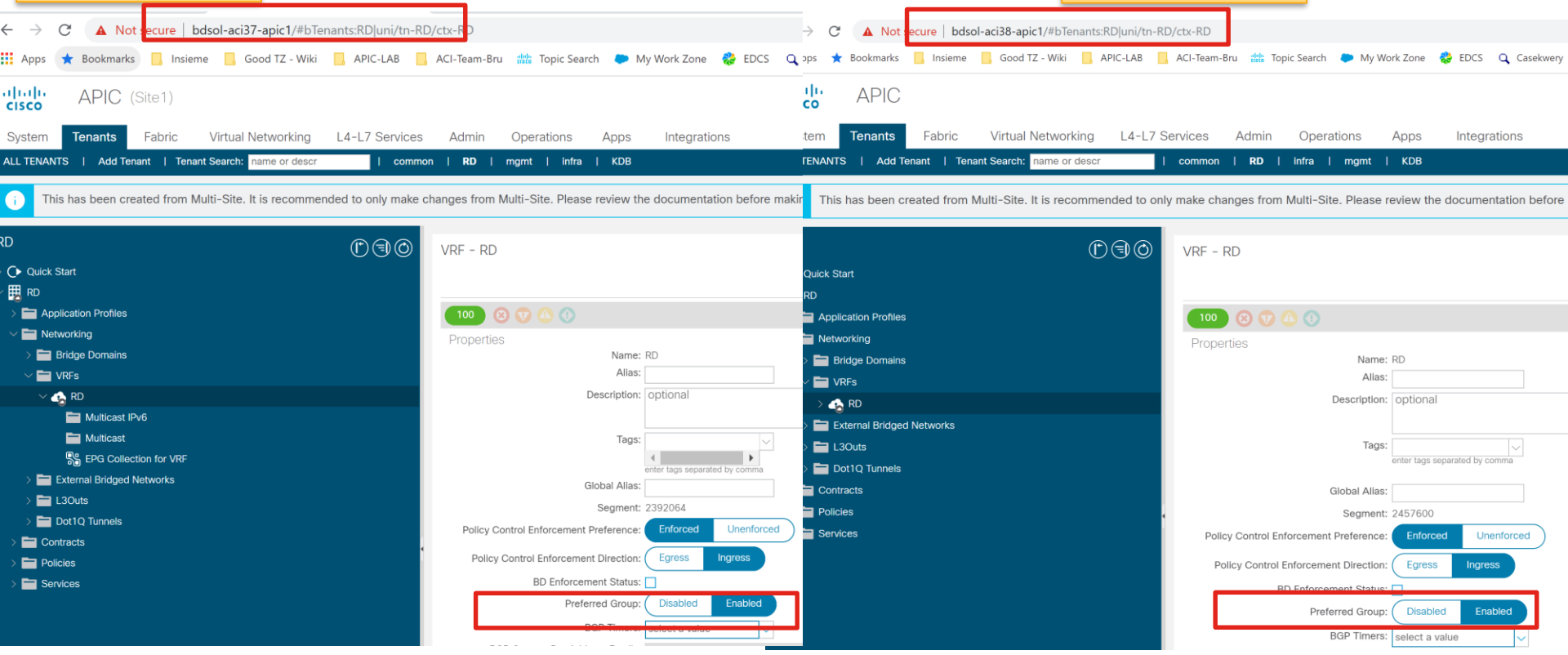


VRF preferred group setting

VRF global setting for preferred is done on both site as soon
As any EPG is marked in the preferred group on MSO

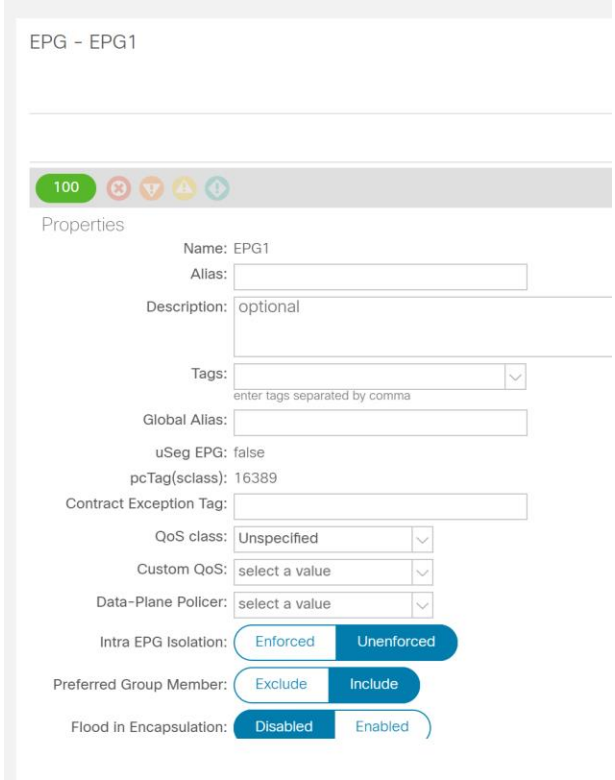
Site-1 APIC

Site-2 APIC



EPG settings for preferred group

Both side should also mark the EPG
As part of the preferred group
(set by MSO)



Example EPG deployed in site-2 only in preferred group

Real EPG in site2

```
bdsol-aci38-apic1# moquery -d uni/tn-RD/ap-test/epg-EPG-site2-only
Total Objects shown: 1

# fv.AEPg
name                : EPG-site2-only
annotation          : orchestrator:msc
configSt            : applied
descr               :
dn                  : uni/tn-RD/ap-test/epg-EPG-site2-only
…
pcEnfPref           : unenforced
pcTag               : 16391
prefGrMemb          : include
prio                : unspecified
rn                  : epg-EPG-site2-only
scope               : 2457600
shutdown            : no
status              :
triggerSt           : triggerable
txId                : 5764607523035213876
uid                 : 15374
```

No remote ID or translation needed in site2

Shadow epg in site 1

```
bdsol-aci37-apic1# moquery -d uni/tn-RD/ap-test/epg-EPG-site2-only
Total Objects shown: 1

# fv.AEPg
name                : EPG-site2-only
annotation          : orchestrator:msc
configSt            : applied
descr               :
dn                  : uni/tn-RD/ap-test/epg-EPG-site2-only
..
pcEnfPref           : unenforced
pcTag               : 49156
prefGrMemb          : include
prio                : unspecified
rn                  : epg-EPG-site2-only
scope               : 2392064
shutdown            : no
status              :
triggerSt           : triggerable
txId                : 5764607523035248870
uid                 : 15374
```

RemoteId and Translation in site1

```
# fv.RemoteId
siteId          : 2
annotation      :
childAction     :
descr           :
dn              : uni/tn-RD/ap-test/epg-EPG-site2-only/stAsc/site-2
extMngdBy       :
lcOwn           : local
modTs           : 2019-11-26T13:27:51.651+00:00
monPolDn        : uni/tn-common/monepg-default

xPcTag : any
Tag    : 16391
       : site-2
```

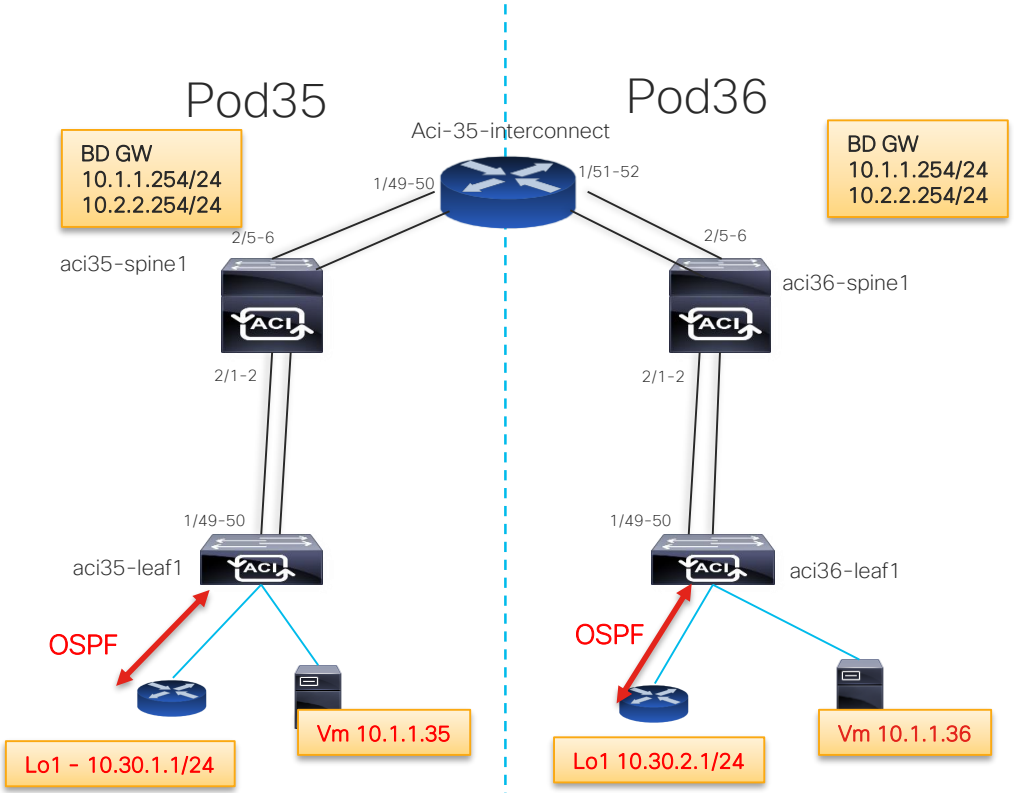
```
S1P1-Spine201# show dcimgr repo sclass-maps  | egrep "Remote|Vrf|2392064.*49156|---"
------------------------------------------------------
     Remote            |        Local
site  Vrf       PcTag  |  Vrf      PcTag     Rel-state
------------------------------------------------------
  2   2457600   16391  |  2392064  49156   [formed]
```

Note on shadow EPG location

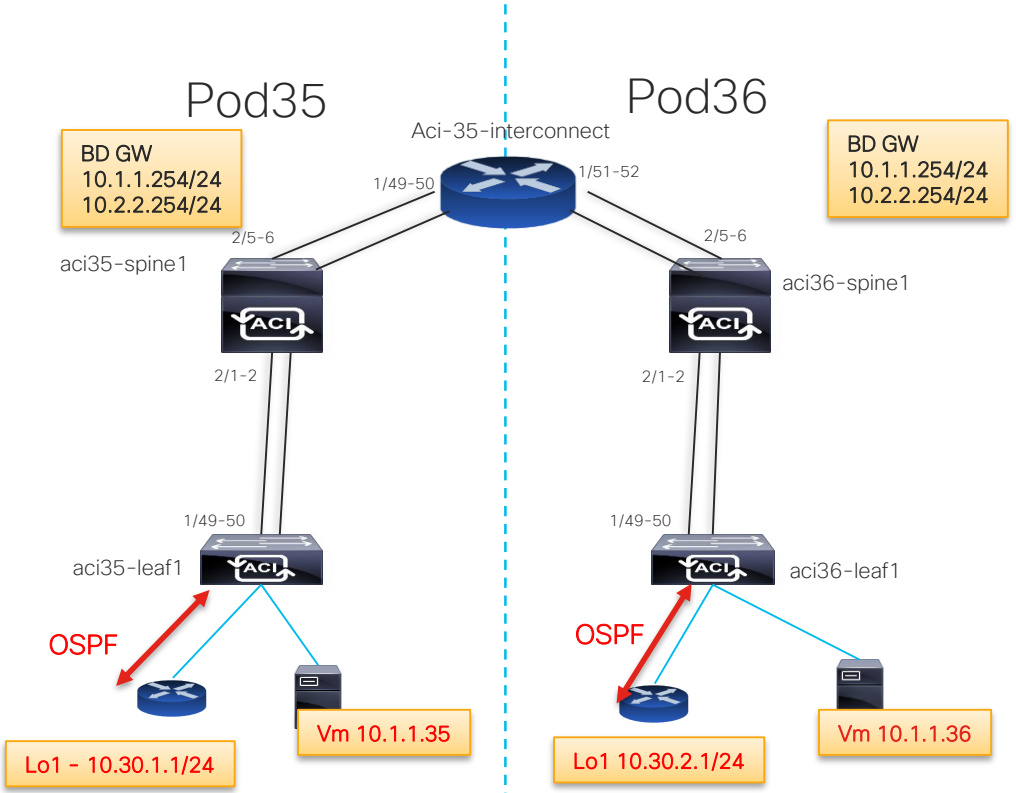
- In the previous example EPG in site-2 only we do not need translation in site 2 !
- Indeed translation is needed for packet from the EPG to somewhere else where the target site must translate the pcTag of src epg set by ingress site to shadow epg value valid in target site
- If EPG is only in site-2 , we will never receive packet with src pcTag on site-2 from ISN (DCN RX) → no need to translate

L3 out and Multisite before 4.2

Use case 3 – lab VRF RD-L2:L2



L3 out – lab VRF RD-L2:L2 (tested with 3.1)



- Working Session :
- in EPG web : 10.1.1.35 to 10.1.1.36
- Local L3 out :
  - 10.1.1.35 to 10.30.1.1
  - 10.1.1.36 to 10.30.2.1
- Non working connection (expected)
  - 10.1.1.35 to 10.30.2.1
  - 10.1.1.36 to 10.30.1.1
- Might or not be Working Direction (return from L3 out):
  - 10.30.2.1 can reach 10.1.1.35
  - 10.30.1.1 can reach 10.1.1.36
- Non working direction (from VM to L3 remote):
  - 10.1.1.35 to 10.30.2.1
  - 10.1.1.36 to 10.30.1.1

Why EP to remote L3 out do not work

- No VPNv4 route exchange across multisite BGP session
- No l2vpn evpn type 5 neither
- Site 2 never got route from Site 1 – L3 out

Only l2vpn evpn capa nego with Peer on the intersite)

```
pod35-spine1# show  bgp l2vpn evpn neigh 10.10.35.112 vrf overlay-1  | egrep -A 1
"capabili"

  Additional Paths capability: advertised received
  Additional Paths Capability Parameters:
  Send capability advertised to Peer for AF:
    L2VPN EVPN
  Receive capability advertised to Peer for AF:
    L2VPN EVPN
```

No l2vpn evpn type5 for subnet neither is advert

```
pod35-spine1# show  bgp l2vpn evpn neigh 10.10.35.112  adver vrf overlay-1  | egrep
"10.10.30"
pod35-spine1#
```

Traffic return from L3 out

- Will work is no subnet configured on l3 out epg (or 0.0.0.0/0)
- Will not work if any subnet is configured under the l3 out with ext subnet for external EPG

- See MSC release notes : https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/aci_multi-site/sw/1x/release_notes/Cisco_ACI_Multi-Site_RN_112.html
- NOTE: The subnet in the L3extInstP must be the same for all inter-related sites (and variable length network masks are not supported).

Why does it work from L3 out to remote EP
Example from 10.30.1.1 to 10.1.1.36

PcTag ingress Site

```
admin@bdsol-aci35-apic1:~> moquery -d  uni/tn-RD-L2/out-L3-35/instP-epg-35 | egrep "dn|pcTag|scope"
dn           : uni/tn-RD-L2/out-L3-35/instP-epg-35
pcTag        : 32772
scope        : 2457600

bdsol-aci35-apic1# moquery -d uni/tn-RD-L2/ctx-L2 | egrep "dn|pcTag"
dn               : uni/tn-RD-L2/ctx-L2
pcTag            : 16386
```

Elam ingress leaf – setting pcTag to vrf pcTag

```
 hom_lurw_vec.info.ifabric_leaf.dclass: 0x1
      hom_lurw_vec.info.ifabric_leaf.sclass: 0x4002
module-1(DBG-elam-insel6)# dec 0x4002
16386
```

Egress spine translate 16386 to 32700

```
pod36-spine1# show dcimgr repo sclass-maps  | egrep "2162688"
------------------------------------------------------------
     Remote             |        Local
site  Vrf       PcTag   |  Vrf      PcTag     Rel-state
------------------------------------------------------------

 1   2457600   16386   |  2162688   32770   [formed]
```

Zoning rule on egress leaf

```
pod36-leaf1# show sys internal policy-mgr stat | egrep "4133"
Rule (4133) DN (sys/actrl/scope-2162688/rule-2162688-s-32770-d-49155-f-default) Ingress: 0, Egress: 0, Pkts: 1475  RevPkts: 0
pod36-leaf1# show zoning-rule | egrep "4133"
4133           32770           49155           default         enabled         2162688         permit
src_dst_any(9)
```

Why would that flow break ?
from L3 out to remote EP
Example from 10.30.1.1 to 10.1.1.36

I added in site1 l3 out subnet 10.30.0.0/16 as ext to ext epg
→ Ingress leaf do not drive pcTag from vrf but from L3 out

```
module-1(DBG-elam-insel6)# show system internal  aclqos prefix
Vrf-Vni VRF-Id Table-Id          Addr                    Class  Shared Remote Complete
2457600 8      0x8           10.30.0.0/16                32772  0      0      No
```

Elam ingress leaf – setting pcTag to vrf pcTag

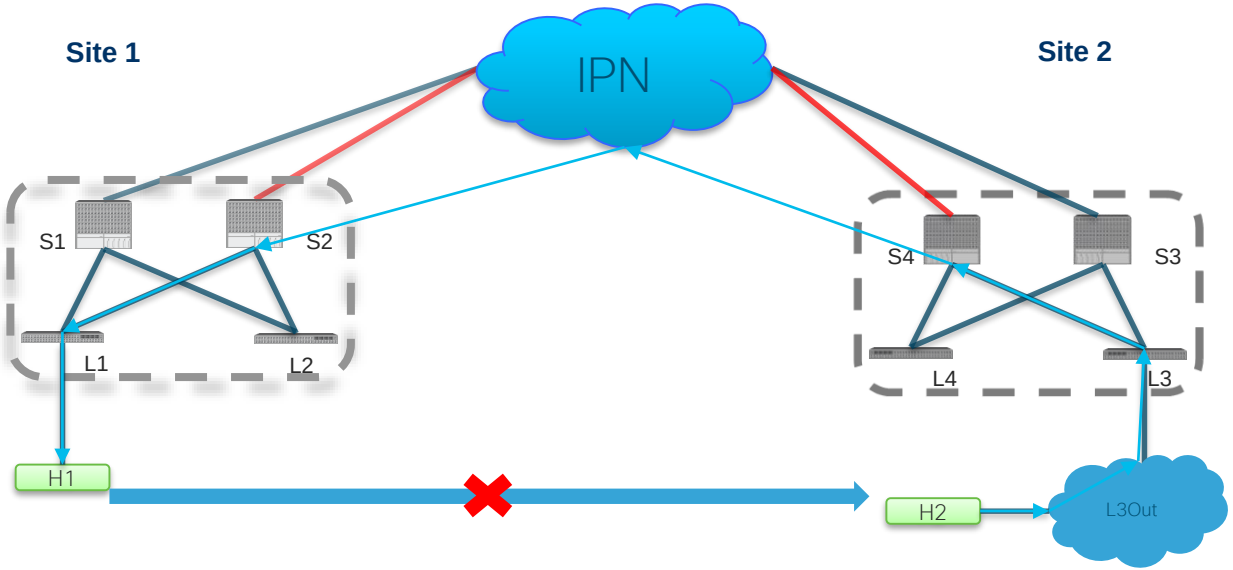
```
   hom_lurw_vec.info.ifabric_leaf.dclass: 0x1
      hom_lurw_vec.info.ifabric_leaf.sclass: 0x8004
module-1(DBG-elam-insel6)# dec 0x8004
32772
```

Egress spine translate 16386 to 32700

```
pod36-spine1# show dcimgr repo sclass-maps  | egrep "2162688"
  1   2457600   16386  |  2162688   32770   [formed]
  1   2457600   32771  |  2162688   49155   [formed]
  1   2457600   49154  |  2162688   16386   [formed]
```

Inter Site L3 out in 4.2

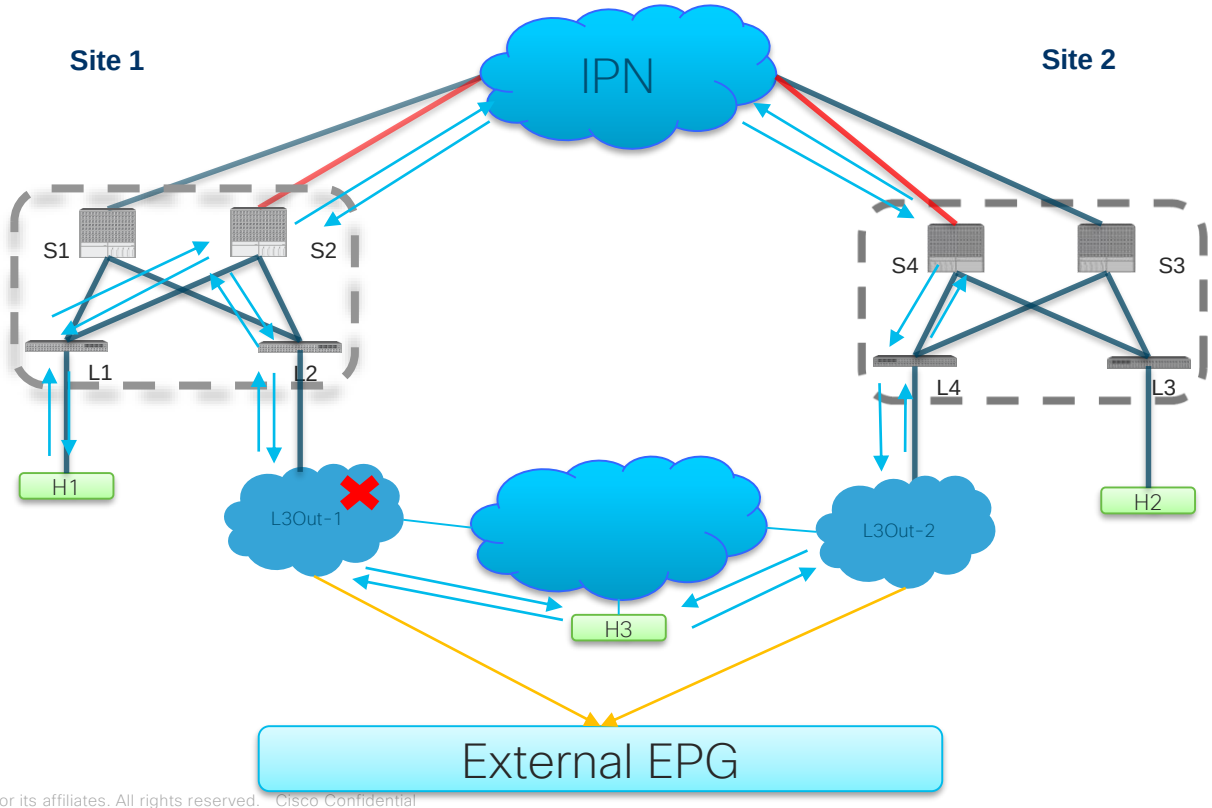
No support for EPG to L3Out across site



With pre 4.2 ACI Multisite design, we allow L3Out to EPG (H2 to H1) traffic across site. L3Out to EPG takes the proxy path to reach the destination EP in the remote site.

EPG to L3Out (H1 to H2) across site is black holed today, reason being, the Wan/L3Out external routes are not programmed across sites.

What is Intersite L3Out?



Design

- In the current pre 4.2 implementation i.e. single site, WAN prefixes from Border Leaf(BL) are carried using BGP L3VPN (VPNv4/6 Address families) within a POD and across PODs. Extending this, support for single L3out will use BGP VPN AFs to carry WAN prefixes across sites.
- In accordance with the design for Msite plus Mpod, BGP Msite Speakers in a site will now reflect the WAN prefixes from local and remote PODS to remote site Speakers. Similarly, speakers will also reflect the remote site VPN prefixes to all Msite forwarders in local and remote PODs within the site. The remote site VPN prefixes are then reflected by the Fabric RRs to all the TORs in the POD.
- In earlier phases of Msite, the VNID translations were being done on the Gen2-3 spines. For WAN prefixes, however, spines are not involved with VNID translations. Instead, the msite VNID translations are done by BGP programming the rewrite VNIDs of remote site routes directly on Leaf.
- There is no sclass translations involved for EPG L3out flow in the msite spines. The policy is always applied in the ingress tor. In this flow, the spines are mere transits.

Intersite L3Out Pre-requisites

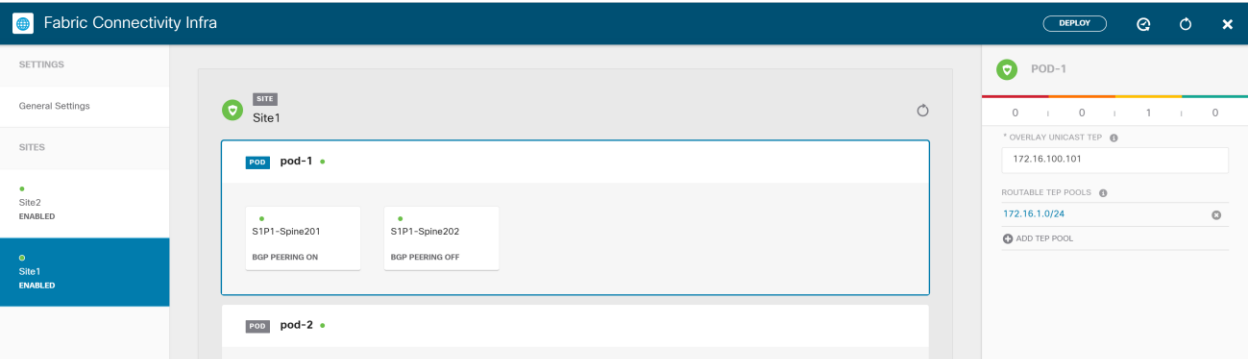
- Basic Multisite requirements
- **BLs needs to be new-gen**
- **The sites where the BL(s) is residing needs to have the Routable Subnets configured.**

  More Information on Routable TEP Pools –

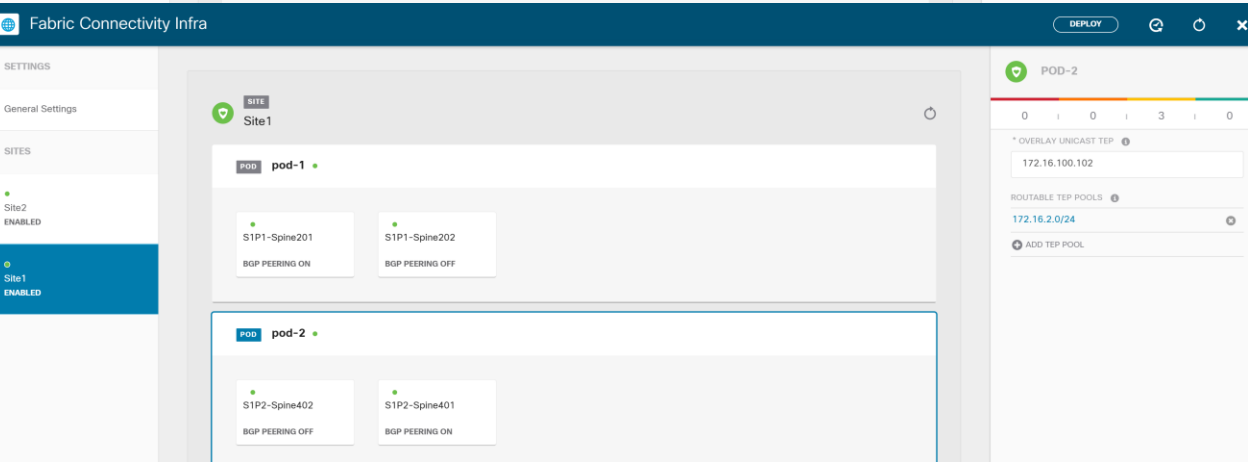
  https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/aci_vpod/installation-upgrade/4-x/Cisco-ACI-Virtual-Pod-Installation-Guide-402/Cisco-ACI-Virtual-Pod-Installation-Guide-402_chapter_010.html

  https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-737909.html

Configure a Routable TEP pool on each Pod of each site – Site 1

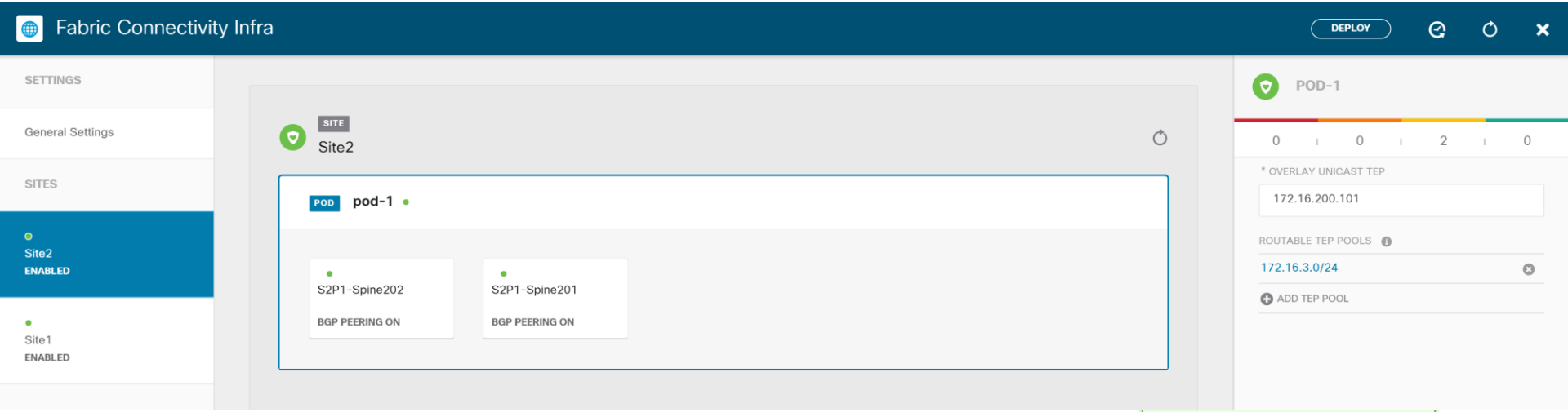


Site 1 – pod 1
172.16.1.0/24



Site 1 – pod 2
172.16.2.0/24

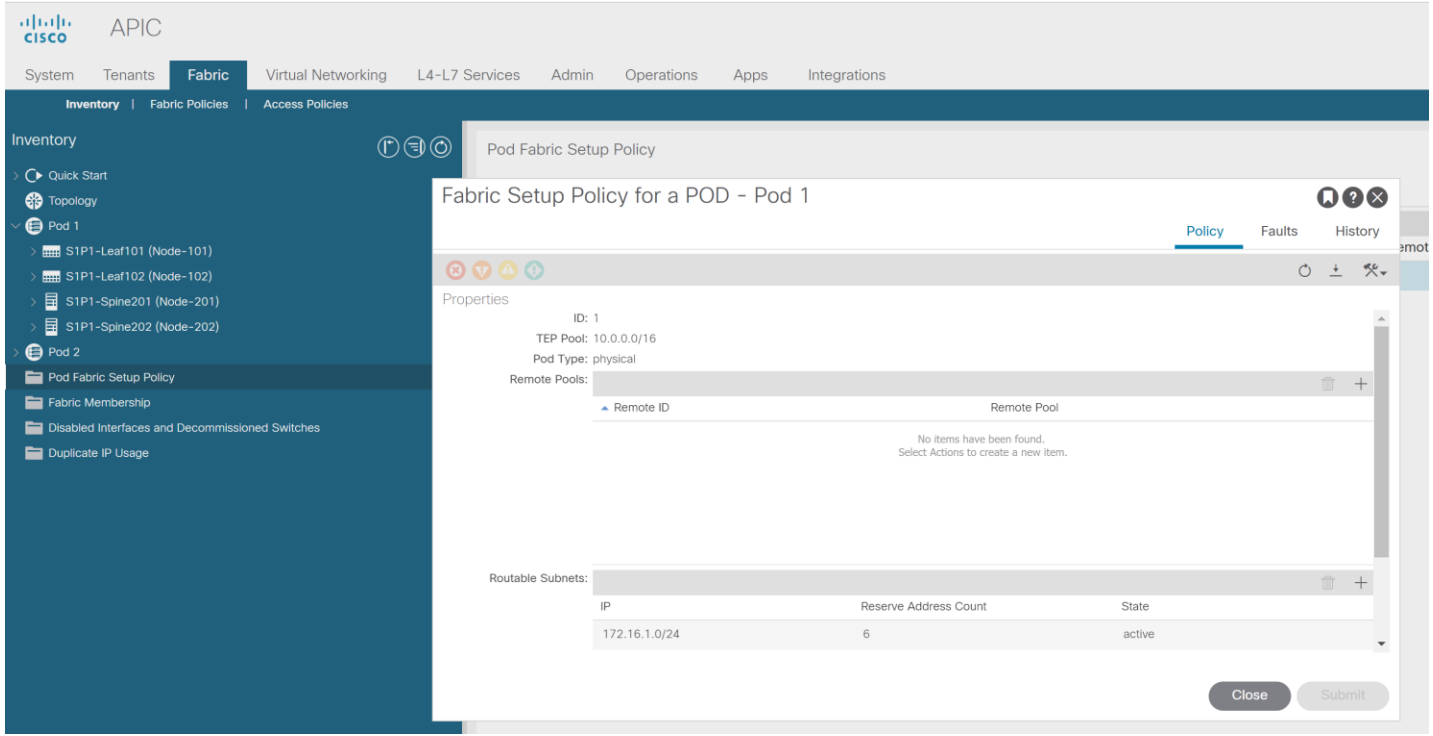
Configure a Routable TEP pool on each Pod of each site – Site 2



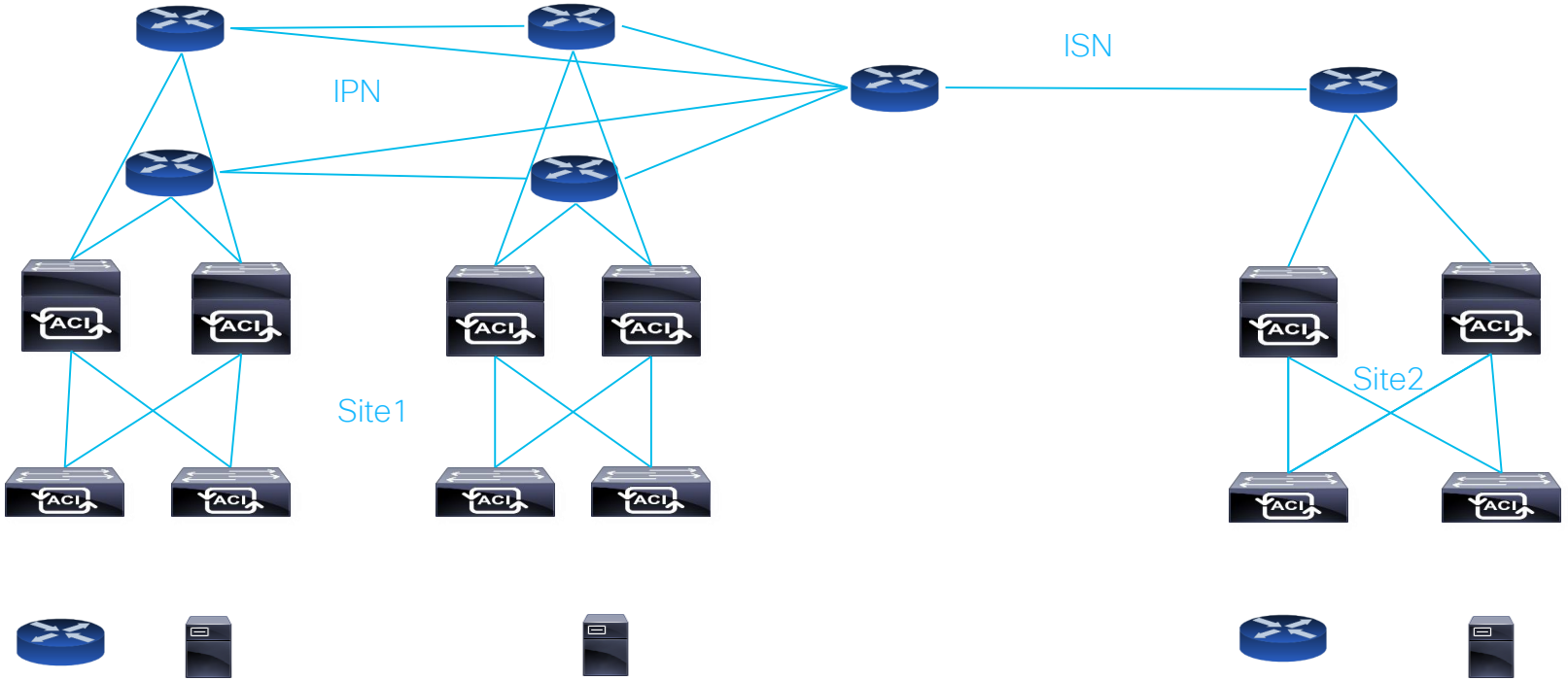
Site 2 – 172.16.3.0/24

Result on the APIC

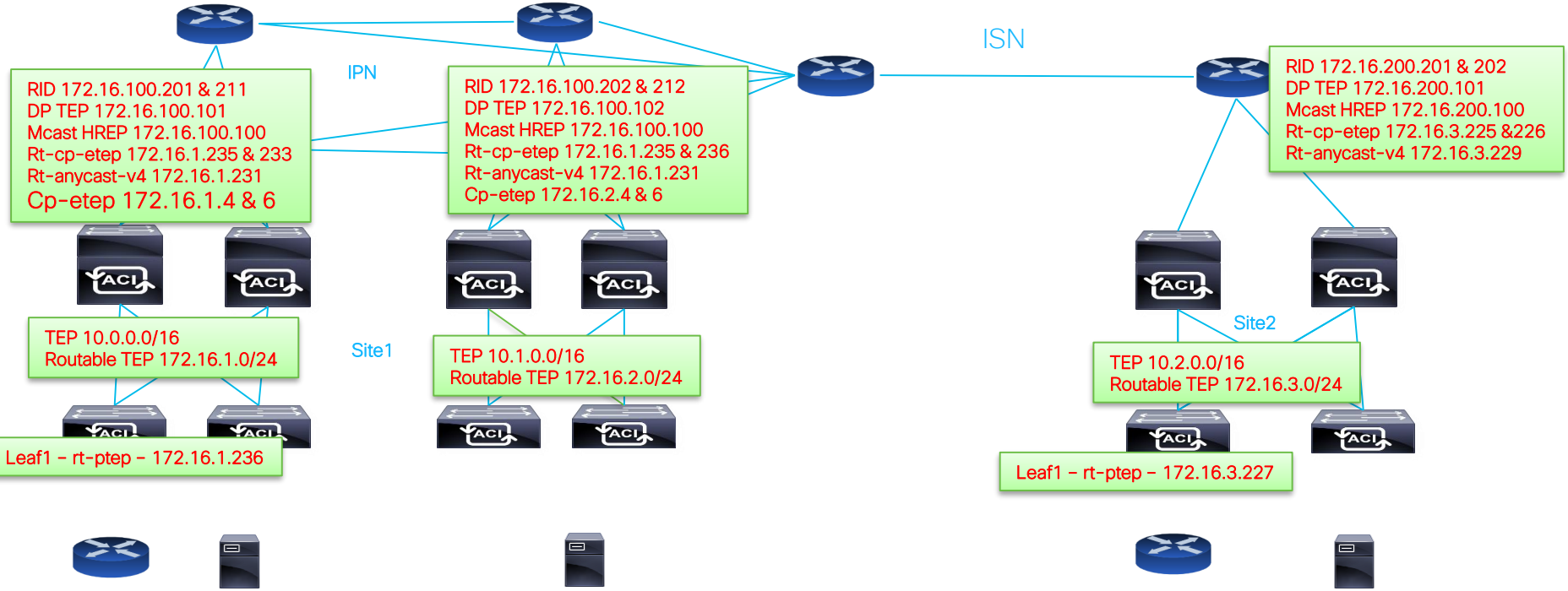
Visible in Fabric – Inventory – Pod Fabric Policy – Routable subnets



Lab setup



Lab setup – infra addresses and loopback



Overlay-1 routing for RTEP

Overlay-1 routing

- All RTEP pool from every site/pod should be injected in overlay-1 routing table (ISIS and OSPF)
- So all devices in network (ISN/IPN/leaf/spine) should have routing to thise tep pool
- This achieved through route-map redistribution between ospf and isis (automatically configured route-map)

Routable TEP pool in the overlay – Site 2 RTEP in site 2 spine

TO OSPF DB

```
S2P1-Spine201# show ip ospf vrf overlay-1 | egrep route-map
 Table-map using route-map interleak_rtmap_infra_prefix_local_
   isis route-map interleak_rtmap_infra_prefix_local_pod_cp_et
   static route-map interleak_rtmap_infra_prefix_local_pteps
   direct route-map interleak_rtmap_infra_prefix_local_pod_dtep_eproxy

S2P1-Spine201# show route-map interleak_rtmap_infra_prefix_local_pod_dtep_eproxy
route-map interleak_rtmap_infra_prefix_local_pod_dtep_eproxy, permit, sequence 1
  Match clauses:
    ip address prefix-lists: infra_prefix_local_ext_routable_loopback
  Set clauses:
S2P1-Spine201# show ip prefix-list  infra_prefix_local_ext_routable_loopback
ip prefix-list infra_prefix_local_ext_routable_loopback: 3 entries
   seq 1 permit 172.16.3.229/32
   seq 2 permit 172.16.3.230/32
   seq 3 permit 172.16.3.225/32

S2P1-Spine201# show route-map  interleak_rtmap_infra_prefix_local_pteps
route-map interleak_rtmap_infra_prefix_local_pteps, permit, sequence 1
  Match clauses:
    ip address prefix-lists: infra_prefix_local_pteps
    ip address prefix-lists: infra_prefix_local_ext_routable_subnet
  Set clauses:
S2P1-Spine201# show ip prefix-list infra_prefix_local_ext_routable_subnet
ip prefix-list infra_prefix_local_ext_routable_subnet: 1 entries
   seq 1 permit 172.16.3.0/24
```

```
S2P1-Spine201# show ip interface vrf  overlay-1
lo13, Interface status: protocol-up/link-up/admin-up, iod: 95, mode:
anycast-v4,external,rt-anycast-v4
  IP address: 172.16.3.229, IP subnet: 172.16.3.229/32
lo14, Interface status: protocol-up/link-up/admin-up, iod: 96, mode:
anycast-v6,external,rt-anycast-v6
  IP address: 172.16.3.230, IP subnet: 172.16.3.230/32
lo15, Interface status: protocol-up/link-up/admin-up, iod: 97, mode:
rt-cp-etep
  IP address: 172.16.3.225, IP subnet: 172.16.3.225/32
```

Spine does inject to ISIS
For Local loopback in RTEP
+ the full RTEP range of the site

OSPF DB on spine site 2

RTEP for site 2 injected to OSPF

```
S2P1-Spine201# show ip ospf database external vrf overlay-1 | egrep 172.16.3
172.16.3.0      172.16.200.201  881         0x8000014d 0x9ec3    0
172.16.3.0      172.16.200.202  874         0x8000014d 0x98c8    0
172.16.3.225    172.16.200.201  871         0x8000014d 0xcbb4    0
172.16.3.226    172.16.200.202  874         0x8000014d 0xbbc2    0
172.16.3.229    172.16.200.201  881         0x8000014d 0xa3d8    0
172.16.3.229    172.16.200.202  884         0x8000014d 0x9ddd    0
172.16.3.230    172.16.200.201  881         0x8000014d 0x99e1    0
172.16.3.230    172.16.200.202  884         0x8000014d 0x93e6    0
```

RTEP for site 1 (pod1 and pod 2 in ospf DB of site 2)

```
S2P1-Spine201# show ip ospf database external vrf overlay-1 | egrep "^172.16.1"
172.16.1.0      172.16.1.4      1319        0x8000064a 0xc42b    0
172.16.1.0      172.16.1.6      1319        0x8000064a 0xb835    0
172.16.1.1      172.16.1.4      1319        0x8000064a 0xba34    0
172.16.1.1      172.16.1.6      1319        0x8000064a 0xae3e    0
172.16.1.2      172.16.1.4      1319        0x80000417 0x1d06    0
172.16.1.2      172.16.1.6      1319        0x80000417 0x1110    0
172.16.1.4      172.16.1.6      1319        0x800003b4 0xc4bd    0
172.16.1.6      172.16.1.4      1319        0x800003b4 0xbcc5    0
172.16.1.231    172.16.1.4      1319        0x80000417 0x221b    0
172.16.1.231    172.16.1.6      1319        0x80000417 0x1625    0
172.16.1.232    172.16.1.4      1319        0x80000417 0x1824    0
172.16.1.232    172.16.1.6      1319        0x80000417 0x0c2e    0
172.16.1.233    172.16.1.6      1319        0x800003b4 0xc9d2    0
172.16.1.235    172.16.1.4      1319        0x800003b4 0xc1da    0
```

```
S2P1-Spine201# show ip ospf database external vrf overlay-1 | egrep "^172.16.2"
172.16.2.0      172.16.2.4      1120        0x80000639 0xd42a    0
172.16.2.0      172.16.2.6      1119        0x80000639 0xc834    0
172.16.2.1      172.16.2.4      1120        0x80000639 0xca33    0
172.16.2.1      172.16.2.6      1119        0x80000639 0xbe3d    0
172.16.2.2      172.16.2.4      1120        0x80000416 0x0d15    0
172.16.2.2      172.16.2.6      1119        0x80000416 0x011f    0
172.16.2.4      172.16.2.6      1119        0x800003b4 0xb2cd    0
172.16.2.6      172.16.2.4      1120        0x800003b4 0xaad5    0
172.16.2.231    172.16.2.4      1120        0x80000416 0x122a    0
172.16.2.231    172.16.2.6      1119        0x80000416 0x0634    0
172.16.2.232    172.16.2.4      1120        0x80000416 0x0833    0
172.16.2.232    172.16.2.6      1119        0x80000416 0xfb3d    0
172.16.2.233    172.16.2.6      1119        0x800003b4 0xb7e2    0
172.16.2.235    172.16.2.4      1120        0x800003b4 0xafea    0
```

Routable TEP pool in the overlay – Site 2 RTEP in site 2 spine From OSPF to ISIS

172.16.1.0 and 172.16.2.0 (RTEP from site1) are injected to ISIS From OSPF in site 2 spine

```
S2P1-Spine201# vsh -c 'show isis protocol vrf overlay-1' | egrep -A 3 Redis
    Redistributing :
      ospf-default    policy interleak_rtmap_infra_prefix_remote_pod_teps
      static          policy interleak_rtmap_infra_prefix_ext_static_routes

S2P1-Spine201# show route-map interleak_rtmap_infra_prefix_remote_pod_teps
route-map interleak_rtmap_infra_prefix_remote_pod_teps, permit, sequence 1
  Match clauses:
    ip address prefix-lists: infra_prefix_ipn_remote_subnets
    ip address prefix-lists: infra_prefix_remote_msite_teps
  Set clauses:
    metric 63
S2P1-Spine201# show ip prefix-list infra_prefix_remote_msite_teps
ip prefix-list infra_prefix_remote_msite_teps: 5 entries
   seq 1 permit 172.16.100.101/32
   seq 2 permit 172.16.100.102/32
   seq 3 permit 172.16.100.100/32
   seq 4 permit 172.16.1.0/24
   seq 5 permit 172.16.2.0/24
```

RTEP in ISIS DB in spine site 2

```
S2P1-Spine201# show isis database detail  vrf overlay-1 | egrep "4..*0x.*|172.16.[1-3]\."
  4090.020a.0000.00-00  0x00000C2E   0xBC39    896        0/0/0/1
    IP Internal    :    172.16.3.228/32  Metric : 1   (I,U)
  4190.020a.0000.00-00* 0x00000C30   0x35A7    683        0/0/0/1
    IP Internal    :    172.16.3.225/32  Metric : 1   (I,U)
    IP Internal    :    172.16.3.229/32  Metric : 1   (I,U)
    IP Internal    :    172.16.3.230/32  Metric : 1   (I,U)
    IP External    :      172.16.2.0/24  Metric : 63  (I,U)
    IP External    :      172.16.1.0/24  Metric : 63  (I,U)
  4290.020a.0000.00-00  0x00000C38   0x34B1    1091       0/0/0/1
    IP Internal    :    172.16.3.226/32  Metric : 1   (I,U)
    IP Internal    :    172.16.3.229/32  Metric : 1   (I,U)
    IP Internal    :    172.16.3.230/32  Metric : 1   (I,U)
    IP External    :      172.16.2.0/24  Metric : 63  (I,U)
    IP External    :      172.16.1.0/24  Metric : 63  (I,U)
  4390.020a.0000.00-00  0x00000C29   0x06D6    948        0/0/0/1
    IP Internal    :    172.16.3.227/32  Metric : 1   (I,U)
  4190.020a.0000.00-00* 0x00000C32   0xE7BC    808        0/0/0/1
    IP Internal    :    172.16.3.225/32  Metric : 1   (I,U)
    IP Internal    :    172.16.3.230/32  Metric : 1   (I,U)
    IP Internal    :    172.16.3.229/32  Metric : 1   (I,U)
    IP External    :      172.16.2.0/24  Metric : 63  (I,U)
    IP External    :      172.16.1.0/24  Metric : 63  (I,U)
S2P1-Spine201#
```

Apic object model

Local External routable subnet on site-2

```
admin@bdsol-aci38-apic1:~> moquery -c fabricExtRoutablePodSubnet
Total Objects shown: 1

# fabric.ExtRoutablePodSubnet
pool                  : 172.16.3.0/24
annotation            : orchestrator:msc
childAction           :
descr                 :
dn                    : uni/controller/setuppol/setupp-1/extrtpodsubnet-[172.16.3.0/24]
extMngdBy             :
lcOwn                 : local
modTs                 : 2019-11-12T15:45:19.738+00:00
name                  :
nameAlias             :
reserveAddressCount   : 0
rn                    : extrtpodsubnet-[172.16.3.0/24]
state                 : active
status                :
uid                   : 15374
```

Pushed by MSC based on MSO config of ext routable subnet pool per Pod

Remote Ext Routable Subnet on site-2

```
admin@bdsol-aci38-apic1:extrtpodsubnet-[172.16.3.0--24]> moquery -c  fv.ExtRoutableRemoteSitePodSubnet
Total Objects shown: 2

# fv.ExtRoutableRemoteSitePodSubnet
pool          : 172.16.1.0/24
annotation    : orchestrator:msc
childAction   :
descr         :
dn            : uni/tn-infra/fabricExtConnP-1/siteConnP-1/extrtremotesitepodsubnet-[172.16.1.0/24]
extMngdBy     :
lcOwn         : local
modTs         : 2019-10-30T17:03:31.911+00:00
monPolDn      : uni/tn-common/monepg-default
name          :
nameAlias     :
podId         : 1
rn            : extrtremotesitepodsubnet-[172.16.1.0/24]
status        :
uid           : 15374

# fv.ExtRoutableRemoteSitePodSubnet
pool          : 172.16.2.0/24
annotation    : orchestrator:msc
childAction   :
descr         :
dn            : uni/tn-infra/fabricExtConnP-1/siteConnP-1/extrtremotesitepodsubnet-[172.16.2.0/24]
extMngdBy     :
lcOwn         : local
modTs         : 2019-10-30T17:03:31.911+00:00
monPolDn      : uni/tn-common/monepg-default
name          :
nameAlias     :
podId         : 2
rn            : extrtremotesitepodsubnet-[172.16.2.0/24]
status        :
uid           : 15374
```

MSO will subscribe tp fabricExternalRoutablePodSubnet to push to other site fvExtRoutableRemoteSitePodSubnet
(awareness of remote site ext subnet)

MO on leaf

Based on previous MO, APIC should configure on each leaf and spine a tunnelCtrlPfxEntry to Allow routers learned from that remote ext routable prefix

```
admin@bdsol-aci38-apic1:extrtpodsubnet-[172.16.3.0--24]> moquery -d  topology/pod-1/node-101/sys/inst-overlay-1/cpfx-[172.16.1.0/24]
Total Objects shown: 1

# tunnel.CtrlPfxEntry
addr          : 172.16.1.0/24
childAction   :
ctrl          : fabric-ext,trusted
descr         :
dn            : topology/pod-1/node-101/sys/inst-overlay-1/cpfx-[172.16.1.0/24]
lcOwn         : local
modTs         : 2019-10-30T17:03:31.973+00:00
name          :
nameAlias     :
rn            : cpfx-[172.16.1.0/24]
status        :
```

Consumer of a Intersite BL

When a new L3 out needs to be stretched to remote site a consumer fvCons for the node fabricExtRoutableNodeDef Is created (consDn is the L3 out in the vrf that needs to be stretched to other site)

```
admin@bdsol-aci38-apic1:extRoutableNodeDef-101> moquery -d uni/controller/extRoutableNodeDef-101
Total Objects shown: 1

# fabric.ExtRoutableNodeDef
id               : 101
childAction      :
configIssues     :
dn               : uni/controller/extRoutableNodeDef-101
exportExtRoutes  : yes
lcOwn            : local
modTs            : 2019-11-12T15:45:19.738+00:00
monPolDn         : uni/fabric/monfab-default
rn               : extRoutableNodeDef-101
status           :
svcId            : 33
```

```
# fv.Cons
consDn       : uni/tn-RD/out-L3out-site2/lnodep-L1/rsnodeL3OutAtt-[topology/pod-1/node-101]
childAction  :
dn           : uni/controller/extRoutableNodeDef-101/cons-[uni/tn-RD/out-L3out-site2/lnodep-L1/rsnodeL3OutAtt-[topology/pod-1/node-101]]
lcOwn        : local
modTs        : 2019-11-18T12:21:53.931+00:00
name         :
nameAlias    :
rn           : cons-[uni/tn-RD/out-L3out-site2/lnodep-L1/rsnodeL3OutAtt-[topology/pod-1/node-101]]
status       :
```

MO – when a L3 out needs to be stretched (contract across site for example)

Site 2 APIC export Route MO

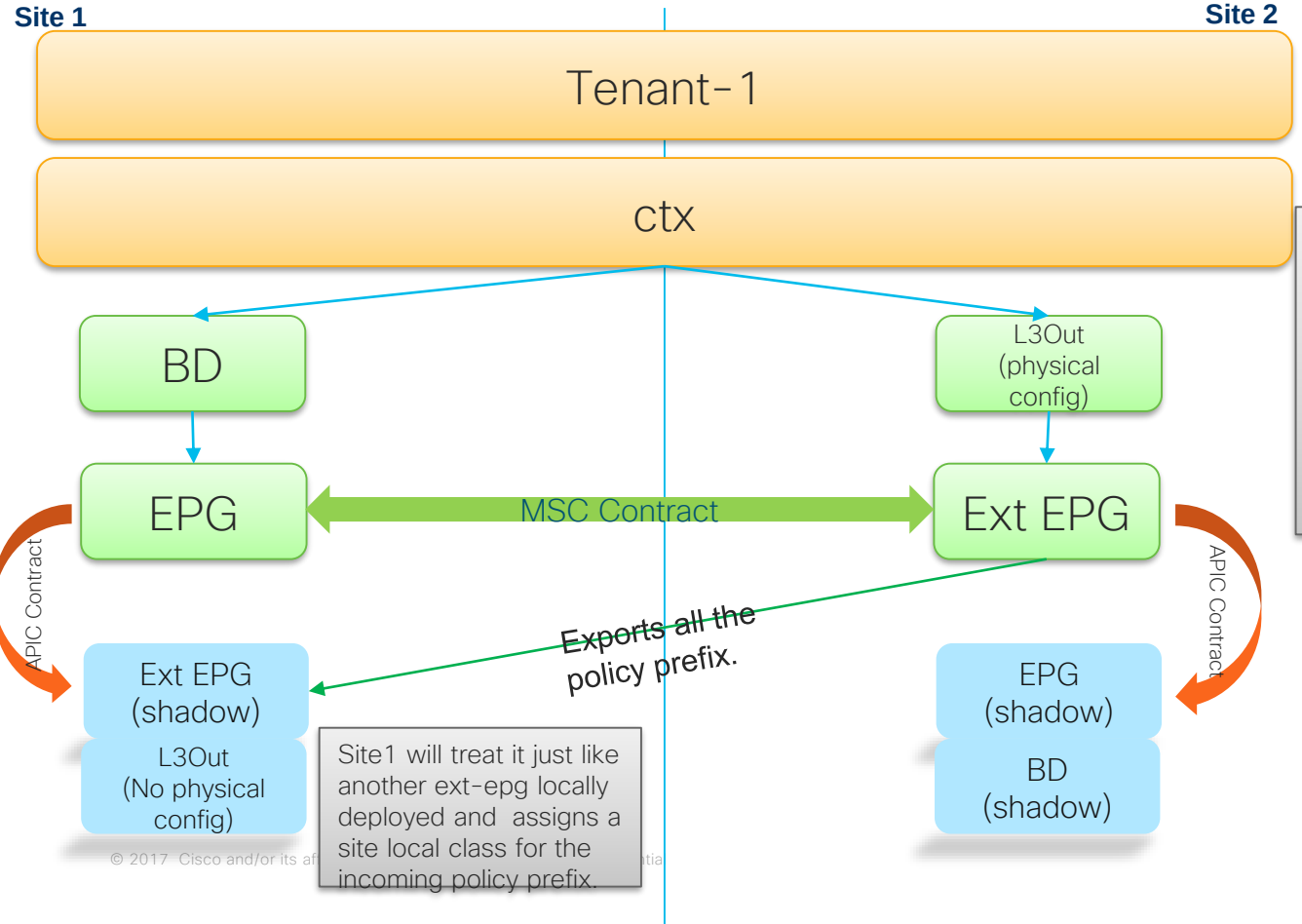
```
# fv.ExportExtRoutes
annotation    :
childAction   :
descr         :
dn            : uni/tn-RD/ctx-RD/stAsc/exportextroutes
extMngdBy     :
lcOwn         : local
modTs         : 2019-11-18T12:48:31.913+00:00
name          :
nameAlias     :
rn            : exportextroutes
status        :
uid           : 15374
```

Site 1 APIC import Route MO From site 2

```
# fv.ImportExtRoutes
annotation    :
childAction   :
descr         :
dn            : uni/tn-RD/ctx-RD/stAsc/site-2/importextroutes
extMngdBy     :
lcOwn         : local
modTs         : 2019-11-18T12:38:23.123+00:00
name          :
nameAlias     :
rn            : importextroutes
status        :
uid           : 15374
```

Apic object model When do we push/create intersite prefix

VRF stretched – Epg Site1 to L3 out Site2



Similar to regular intersite
When cross site contract is adding between EPG and L3 out.
MSO push shadow Ext EPG to site1 and shadow EPG to site2
If BD was local to site1- Shadow BD is pushed to site2 as well

We push l3 out (without Logical node and logical if) and ex epg (l3extInstP)

Site 1 Mo for L3 out in site 2

Shadow L3 out and shadow L3 ext EPG
We laso get l3extSubnet

```
# l3ext.Out
name            : L3out-site2
annotation      : orchestrator:msc
childAction     :
descr           :
dn              : uni/tn-RD/out-L3out-site2
enforceRtctrl   : export
extMngdBy       :
..

# l3ext.InstP
name                  : ExtEPG-Site2
annotation            : orchestrator:msc
childAction           :
configIssues          :
configSt              : applied
descr                 :
dn                    : uni/tn-RD/out-L3out-site2/instP-ExtEPG-Site2
exceptionTag          :
extMngdBy             :
floodOnEncap          : disabled
isSharedSrvMsiteEPg   : no
lcOwn                 : local
matchT                : AtleastOne
mcast                 : no
modTs                 : 2019-11-18T12:38:22.273+00:00
monPolDn              : uni/tn-common/monepg-default
nameAlias             :
pcTag                 : 32772
prefGrMemb            : include

# l3ext.Subnet
dn           : uni/tn-RD/out-L3out-site2/instP-ExtEPG-
Site2/extsubnet-[172.16.3.1/24]
rn           : extsubnet-[172.16.3.1/24]
scope        : import-security
```

Site associated (site2) and info from site2 (remotePcTag from site2 in fvRemoteId

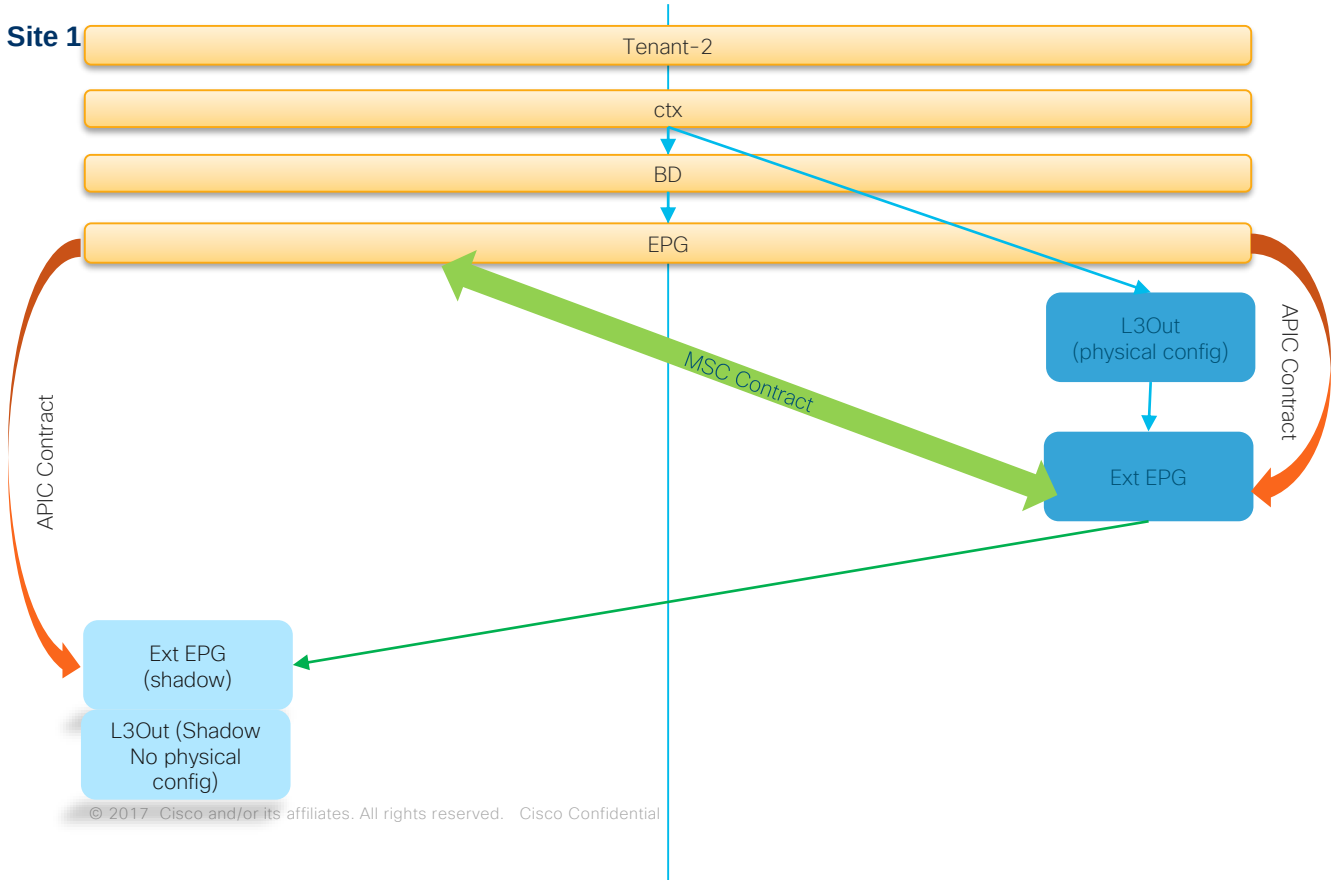
```
# fv.SiteAssociated
annotation    :
childAction   :
descr         :
dn            : uni/tn-RD/out-L3out-site2/instP-ExtEPG-Site2/stAsc
extMngdBy     :
lcOwn         : local
modTs         : 2019-11-18T12:38:22.050+00:00
monPolDn      : uni/tn-common/monepg-default
name          : msc-local
nameAlias     :
ownerKey      :
ownerTag      :
rn            : stAsc
siteId        : 1

# fv.RemoteId
siteId          : 2
annotation      :
childAction     :
descr           :
dn              : uni/tn-RD/out-L3out-site2/instP-ExtEPG-
Site2/stAsc/site-2
extMngdBy       :
lcOwn           : local
modTs           : 2019-11-18T12:38:23.123+00:00
monPolDn        : uni/tn-common/monepg-default
name            :
nameAlias       :
ownerKey        :
ownerTag        :
remoteCtxPcTag  : any
remotePcTag     : 49154
rn              : site-2
status          :
uid             : 15374
```

Use to push pcTag translation in site 1
See example later

Shadow EPG config pushed to APIC

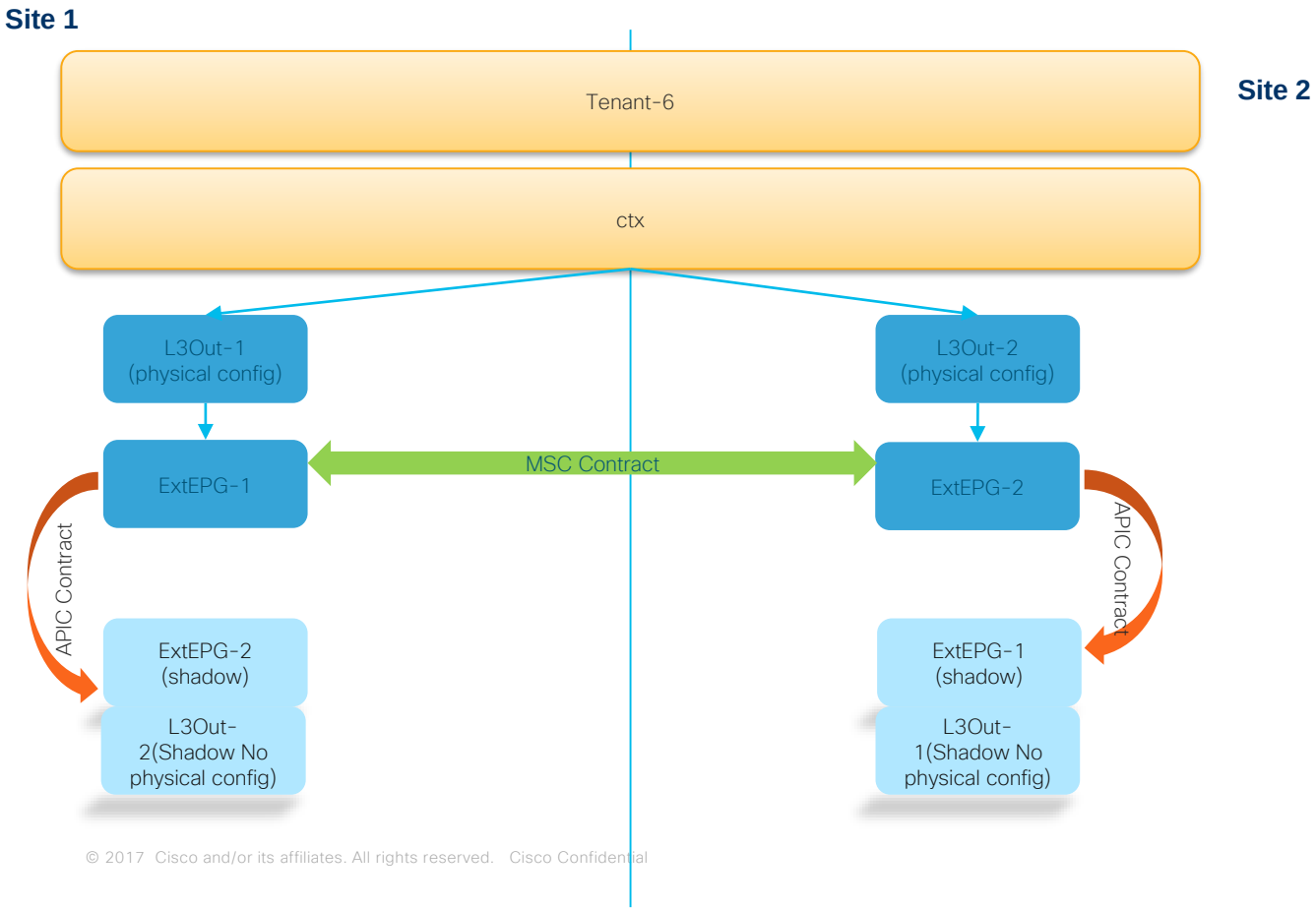
Below l3 out config took on site-1 APIC for Site-2 layer 3
Note :
fvRemoteId for site-2 is present
No Logical node or logical interface, only the External EPG is stretched

VRF/BD and EPG stretched – L3 out in remote site

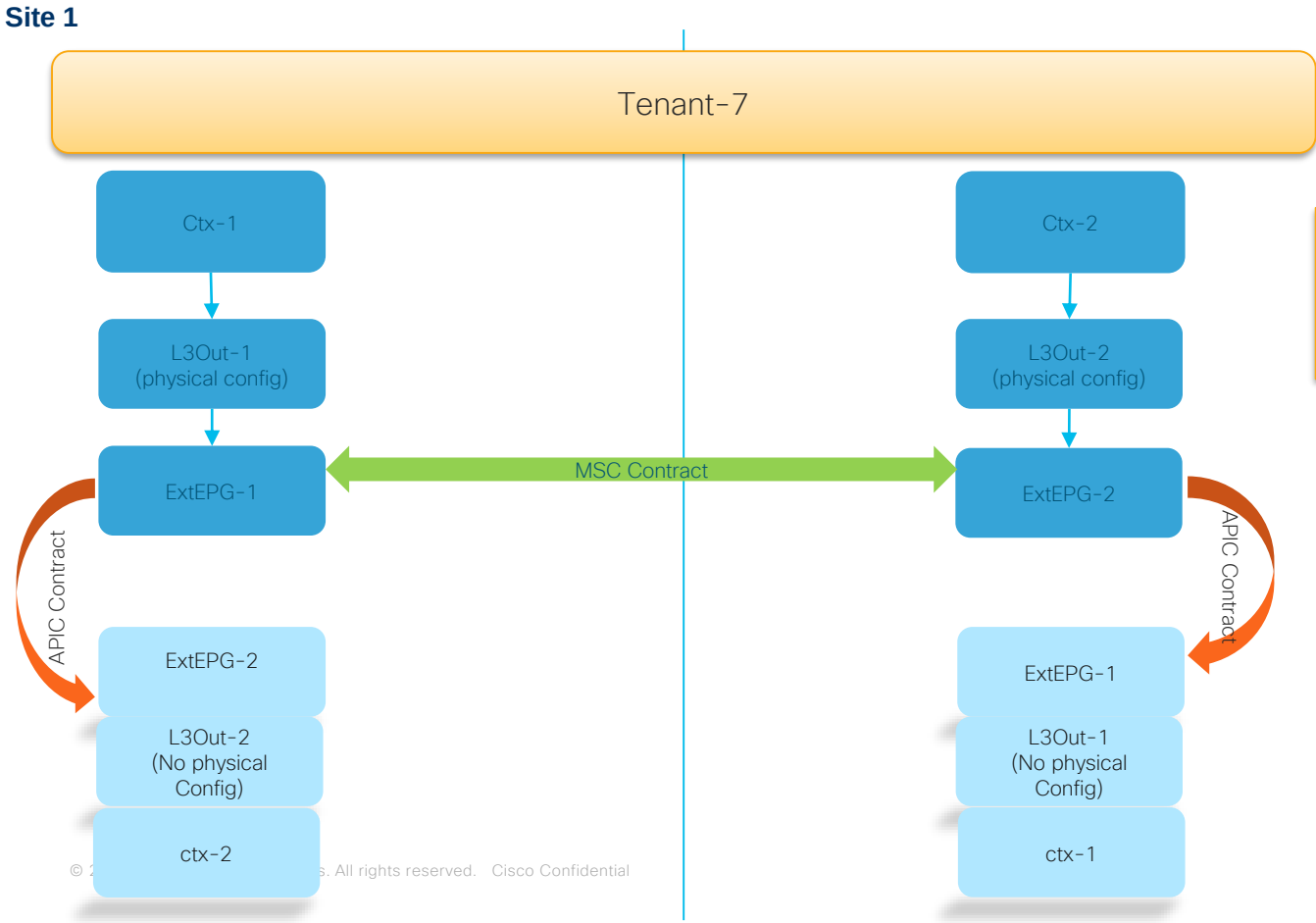


Here only the L3 out is local to site2.
So we push shadow L3 as in first example

VRf stretched L3 transit



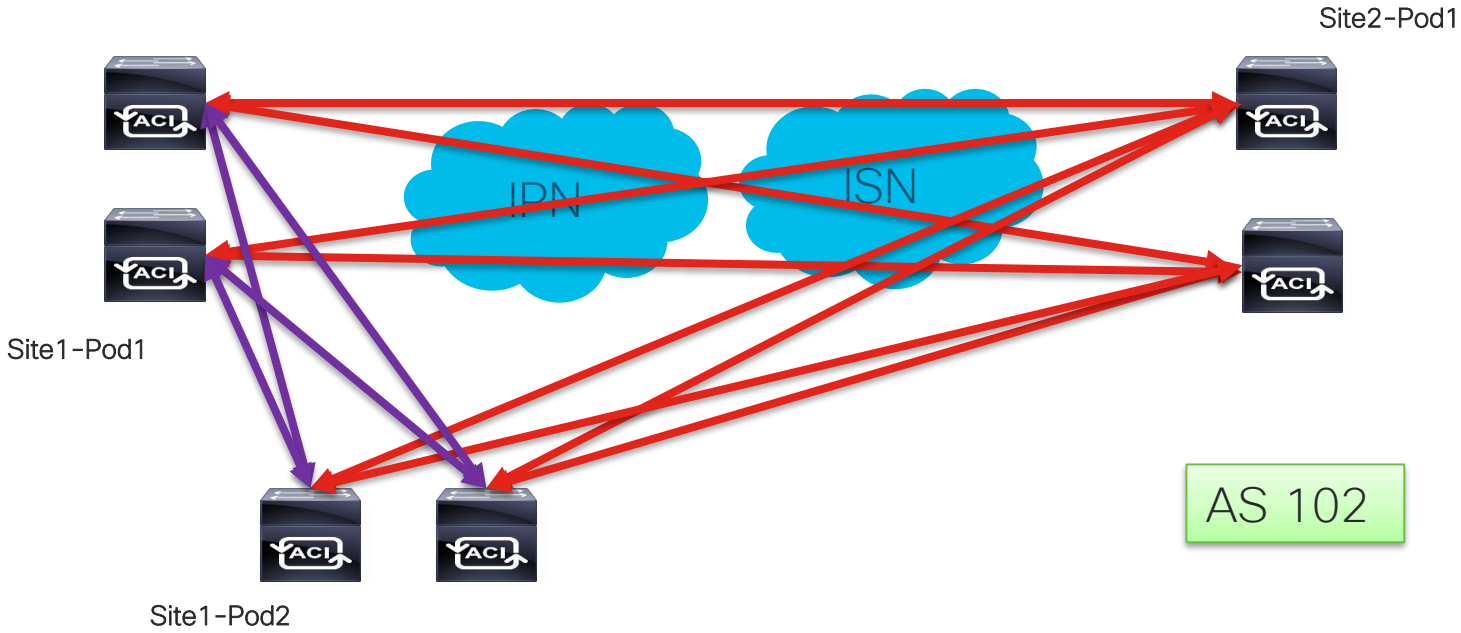
Inter VRF – L3 out transit



Here only the tenant
Is extended
We we will need shadow for
L3 out-ext Epg, and VRF

BGP session vpnv4

Spine to Spine BGP VPNv4 session



Site 1 – BGP VPNv4 session

```
S1P1-Spine201# show bgp vpnv4 unicast summary  vrf overlay-1
BGP summary information for VRF overlay-1, address family VPNv4 Unicast
BGP router identifier 172.16.1.4, local AS number 101
BGP table version is 6672, VPNv4 Unicast config peers 7, capable peers 6
91 network entries and 222 paths using 34648 bytes of memory
BGP attribute entries [31/4588], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [2/8]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.0.184.64     4   101   28797   28903     6672    0    0     2w5d 17     LOCAL LEAF TO POD
10.0.184.67     4   101    7397    7480     6672    0    0    5d02h 13
172.16.2.4      4   101   28809   28850     6672    0    0     2w5d 61     RTEP spine pod 2 site 1 (same AS)
172.16.2.6      4   101   28804   28918     6672    0    0     2w5d 61
172.16.200.201  4   102   25152  246291     6672    0    0 04:20:21 35     CP-ETEP site 2 (AS 102)
172.16.200.202  4   102   25152  246356     6672    0    0 04:18:47 35
```

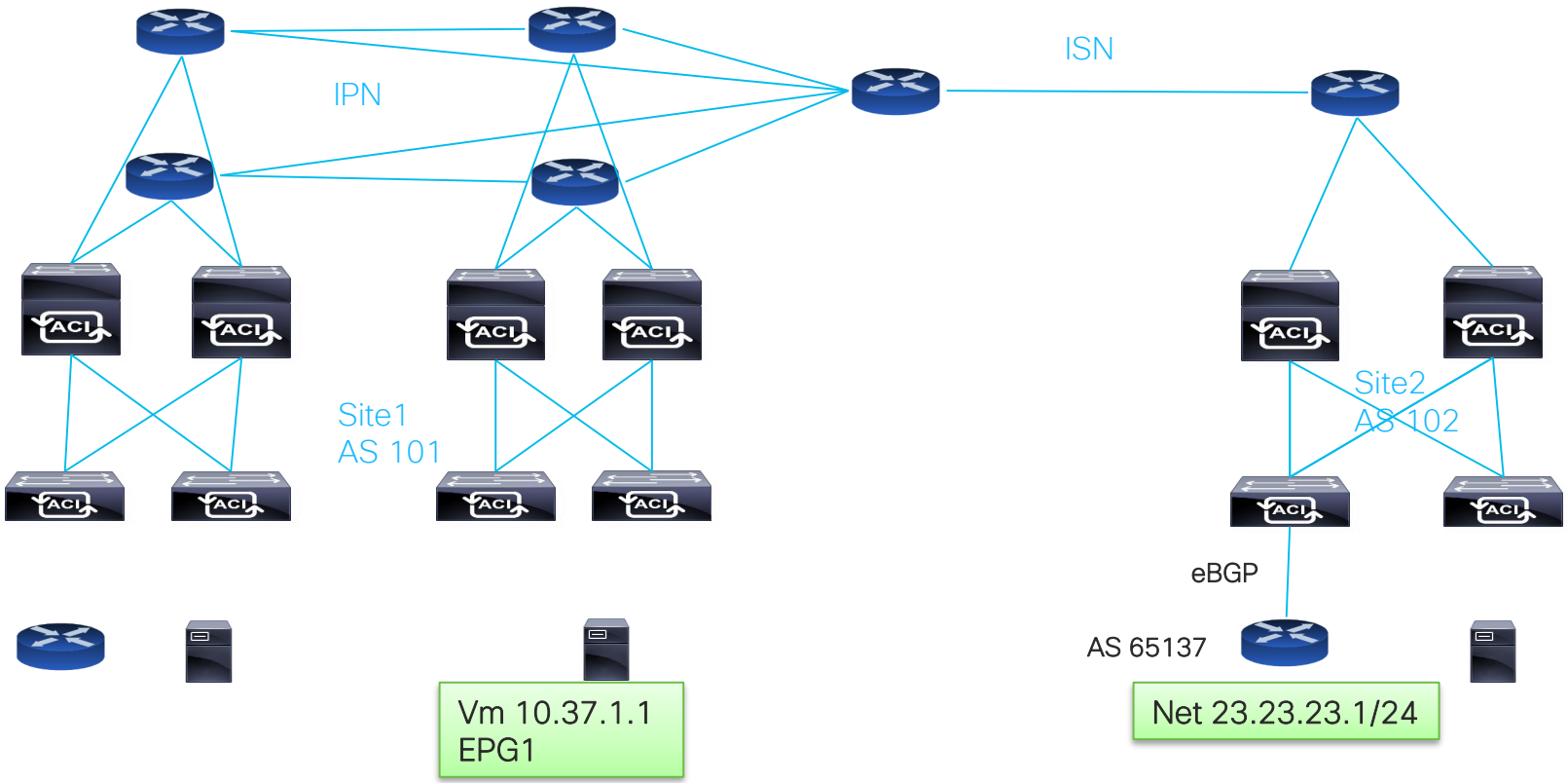
```
S1P2-Spine401# show bgp vpnv4 unicast  summary vrf overlay-1
BGP summary information for VRF overlay-1, address family VPNv4 Unicast
BGP router identifier 172.16.2.4, local AS number 101
BGP table version is 6674, VPNv4 Unicast config peers 7, capable peers 6
91 network entries and 226 paths using 35128 bytes of memory
BGP attribute entries [31/4588], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [2/8]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.1.192.66     4   101   28785   28896     6674    0    0     2w5d 13     LOCAL LEAF TO POD
10.1.192.67     4   101   28788   28908     6674    0    0     2w5d 13
172.16.1.4      4   101   28853   28812     6674    0    0     2w5d 65     RTEP spine pod 1 site 1 (same AS)
172.16.1.6      4   101   28853   28904     6674    0    0     2w5d 65
172.16.200.201  4   102   25191  246368     6674    0    0 04:21:26 35     CP-ETEP site 2 (AS 102)
172.16.200.202  4   102   25192  246293     6674    0    0 04:23:41 35
```

Site 2- BGP VPNv4 session

```
S2P1-Spine201# show bgp vpnv4 unicast summary  vrf overlay-1
BGP summary information for VRF overlay-1, address family VPNv4 Unicast
BGP router identifier 172.16.200.201, local AS number 102
BGP table version is 7726, VPNv4 Unicast config peers 7, capable peers 6
91 network entries and 259 paths using 39088 bytes of memory
BGP attribute entries [19/2812], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [0/0]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.2.144.64     4   102   28795   28899     7726    0    0     2w5d 15     LOCAL LEAF to Site 2
10.2.144.67     4   102   28806   28880     7726    0    0     2w5d 20
172.16.100.201  4   101   25264  163019     7726    0    0 04:25:35 56      Spine in site 1 (Both pod1 and pod2)
172.16.100.202  4   101   25271  163100     7726    0    0 04:23:30 56
172.16.100.211  4   101     311     285     7726    0    0 03:56:38 56
172.16.100.212  4   101     311     285     7726    0    0 03:56:39 56
```

Control Plane for L3 out prefix

Lab setup



Site 2 – BGP prefix (local to site 2)

```
S2P1-Spine201# show bgp vpnv4 unicast 23.23.23.0/24 vrf overlay-1
BGP routing table information for VRF overlay-1, address family VPNv4 Unicast
Route Distinguisher: 101:2457600
BGP routing table entry for 23.23.23.0/24, version 7523 dest ptr 0xaf5aab7c
Paths: (1 available, best #1)
Flags: (0x000002 00000000) on xmit-list, is not in urib, is not in HW
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: internal 0x40000018 0x800040 ref 0 adv path ref 1, path is valid, is best path
  AS-Path: 65137 , path sourced external to AS
    10.2.144.67 (metric 2) from 10.2.144.67 (10.2.144.67)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 1
      Extcommunity:
          RT:102:2457600
          VNID:2457600
          COST:pre-bestpath:168:3221225472

  Path-id 1 advertised to peers:
    10.2.144.64        172.16.100.201     172.16.100.202       172.16.100.211
    172.16.100.212
```

Rx from site2-leaf 1 (10.2.144.67)
Reflected to over leaf in site 2 (10.2.144.64)
And Tx to CP-ETEP of Site 1 Spine (172.16.100.xxx)

Site 2 – spine outbound route-map

```
S2P1-Spine201# show bgp  vpnv4 unicast neighbors   172.16.100.201   vrf overlay-1 | egrep route-map
  Outbound route-map configured is infra-intersite-l3out, handle obtained
  Outbound route-map configured is infra-intersite-l3out, handle obtained
```

```
S2P1-Spine201# show route-map infra-intersite-l3out
route-map infra-intersite-l3out, permit, sequence 1
  Match clauses:
    ip next-hop prefix-lists: IPv4-Node-entry-202
    ipv6 next-hop prefix-lists: IPv6-Node-entry-202
  Set clauses:
    ip next-hop 172.16.3.226
route-map infra-intersite-l3out, permit, sequence 2
  Match clauses:
    ip next-hop prefix-lists: IPv4-Node-entry-101
    ipv6 next-hop prefix-lists: IPv6-Node-entry-101
  Set clauses:
    ip next-hop 172.16.3.227
route-map infra-intersite-l3out, permit, sequence 3
  Match clauses:
    ip next-hop prefix-lists: IPv4-Node-entry-102
    ipv6 next-hop prefix-lists: IPv6-Node-entry-102
  Set clauses:
    ip next-hop 172.16.3.228
route-map infra-intersite-l3out, deny, sequence 999
  Match clauses:
    ip next-hop prefix-lists: infra_prefix_local_pteps_inexact
  Set clauses:
route-map infra-intersite-l3out, permit, sequence 1000
  Match clauses:
  Set clauses:
    ip next-hop unchanged
```

```
S2P1-Spine201# show ip prefix-list IPv4-Node-entry-202
ip prefix-list IPv4-Node-entry-202: 1 entries
   seq 1 permit 10.2.144.66/32
```

```
S2P1-Spine201# show ip prefix-list IPv4-Node-entry-101
ip prefix-list IPv4-Node-entry-101: 1 entries
   seq 1 permit 10.2.144.67/32
```

PTEP of BL

```
S2P1-Leaf101# show ip interface vrf overlay-1  | egrep -A 1 "rt-ptep"
lo2, Interface status: protocol-up/link-up/admin-up, iod: 88, mode: rt-ptep
  IP address: 172.16.3.227, IP subnet: 172.16.3.227/32
```

RTEP of BL

```
S2P1-Spine201# show ip prefix-list IPv4-Node-entry-102
ip prefix-list IPv4-Node-entry-102: 1 entries
   seq 1 permit 10.2.144.64/32
```

Site 1 spine pod 1 – Rx BGP prefix for L3 out in site 2

Those 2 paths are Rx from site 2 spine1 and spine2. the one from spine 1 is chosen as best path and reflected to site1 pod1 leaf + site1 pod 2 routable TEP (iBGP peer)

```
S1P1-Spine201# show bgp vpnv4 unicast   23.23.23.0/24 vrf overlay-1
BGP routing table information for VRF overlay-1, address family VPNv4 Unicast
Route Distinguisher: 101:36012032
BGP routing table entry for 23.23.23.0/24, version 6640 dest ptr 0xaf509a4c
Paths: (4 available, best #4)
Flags: (0x000002 00000000) on xmit-list, is not in urib, is not in HW
Multipath: eBGP iBGP
 Advertised path-id 1
  Path type: external 0x40000028 0x880040 ref 0 adv path ref 1, path is valid, is best
path, remote site path
  AS-Path: 102 65137 , path sourced external to AS
    172.16.3.227 (metric 20) from 172.16.200.201 (172.16.200.201)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 1
      Extcommunity:
          RT:102:36012032
          SOO:102:50331631
          VNID:2457600
          COST:pre-bestpath:166:2684354560
          COST:pre-bestpath:168:3221225472

  Path-id 1 advertised to peers:
    10.0.184.64        10.0.184.67        172.16.2.4         172.16.2.6

 Advertised path-id 3
  Path type: external 0x40000028 0x880040 ref 0 adv path ref 1, path is valid, not bes
reason: newer EBGP path, remote site path
  AS-Path: 102 65137 , path sourced external to AS
    172.16.3.227 (metric 20) from 172.16.200.202 (172.16.200.202)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 1
      Extcommunity:
          RT:102:36012032
          SOO:102:50331631
          VNID:2457600
          COST:pre-bestpath:166:2684354560
          COST:pre-bestpath:168:3221225472
 Path-id 3 advertised to peers:
    10.0.184.64        10.0.184.67
```

Note on RD used

- Site 2 VNID is 2457600 → locally in site 2 we use RD node-id:2457600 (node id is the BL in site2)
- When we send BGP path we send them with different RD
  - 2457600 = 0x258000
  - We prepend the site number : 0x2258000 and convert back to Dec = 36012032
- RD used in node-id:36012032

- Note: WE need this complex RD scheme to ensure uniqueness of RD across the Msite.
  - Pre 4.2 RD is PTEP:ID and we may have same PTEP on multiple side hence
  - 4.2 uses Node:VNID and across site it appends site id to ensure uniqueness

Note on RT used.

- As for RD we need to ensure unique RT so we prepend site-id to the RT
- So locally it is ASN:VNID
- And across site it ASN:site+VNID

Site 1 spine pod 1 – Rx BGP prefix for L3 out in site 2 (cont.)

Path coming back from site1 pod2
Rx on spine pod2 and send to us as iBGP peer

```
  Advertised path-id 2
   Path type: internal 0x40000018 0x880040 ref 0 adv path ref 1, path is valid, not bes
reason: Internal path, remote site path
  AS-Path: 102 65137 , path sourced external to AS
    172.16.3.227 (metric 20) from 172.16.2.4 (172.16.2.4)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 1
      Extcommunity:
          RT:102:36012032
          SOO:102:50331631
          VNID:2457600
          COST:pre-bestpath:165:2415919104
          COST:pre-bestpath:166:2684354560
          COST:pre-bestpath:168:3221225472

 Path-id 2 advertised to peers:
    10.0.184.64        10.0.184.67

 Advertised path-id 4
  Path type: internal 0x40000018 0x880040 ref 0 adv path ref 1, path is valid, not bes
reason: Internal path, remote site path
  AS-Path: 102 65137 , path sourced external to AS
    172.16.3.227 (metric 20) from 172.16.2.6 (172.16.2.6)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 1
      Extcommunity:
          RT:102:36012032
          SOO:102:50331631
          VNID:2457600
          COST:pre-bestpath:165:2415919104
          COST:pre-bestpath:166:2684354560
          COST:pre-bestpath:168:3221225472

 Path-id 4 advertised to peers:
    10.0.184.64        10.0.184.67
```

Site 1 pod1 leaf – Rx BGP path

Note we receive 8 path with same RD. This is because spine In pod-1 didn't act as RR but as simply BGP router getting eBGP Prefix , hence forwarding to each of it neighbor.

We receive 8 path from RD : 101:36012032
They all have BGP NH 172.16.3.227 (leaf in site2)

```
  S1P1-Leaf102# show bgp vpnv4 unicast 23.23.23.0/24 vrf overlay-1 | egrep "Route Distinguisher|Path type|from"
Route Distinguisher: 101:36012032
  Path type: internal 0x40000018 0x880040 ref 0 adv path ref 1, path is valid, not best reason: Router Id, remote site path
    172.16.3.227 (metric 64) from 10.0.184.66 (172.16.1.6)
  Path type: internal 0x40000018 0x880040 ref 0 adv path ref 1, path is valid, not best reason: Router Id, remote site path
    172.16.3.227 (metric 64) from 10.0.184.66 (172.16.1.6)
  Path type: internal 0x40000018 0x880040 ref 0 adv path ref 1, path is valid, not best reason: Router Id, remote site path
    172.16.3.227 (metric 64) from 10.0.184.65 (172.16.1.4)
  Path type: internal 0x40000018 0x880040 ref 0 adv path ref 1, path is valid, not best reason: Router Id, remote site path
    172.16.3.227 (metric 64) from 10.0.184.66 (172.16.1.6)
  Path type: internal 0x40000018 0x880040 ref 1 adv path ref 1, path is valid, is best path, remote site path
    172.16.3.227 (metric 64) from 10.0.184.65 (172.16.1.4)
  Path type: internal 0x40000018 0x880040 ref 0 adv path ref 1, path is valid, not best reason: Router Id, remote site path
    172.16.3.227 (metric 64) from 10.0.184.65 (172.16.1.4)
  Path type: internal 0x40000018 0x880040 ref 0 adv path ref 1, path is valid, not best reason: Router Id, remote site path
    172.16.3.227 (metric 64) from 10.0.184.66 (172.16.1.6)
  Path type: internal 0x40000018 0x880040 ref 0 adv path ref 1, path is valid, not best reason: Neighbor Address, remote site path
    172.16.3.227 (metric 64) from 10.0.184.65 (172.16.1.4)

Route Distinguisher: 102:2392064    (VRF RD:RD)
  Path type: internal 0xc0000018 0x80040 ref 56506 adv path ref 2, path is valid, is best path, remote site path
             Imported from 101:36012032:23.23.23.0/24
    172.16.3.227 (metric 64) from 10.0.184.65 (172.16.1.4)
```

We import the Best path in our own RD (102:2392064 where 23920.64 if site 1 VNID for the VRF)

Site 1 leaf – Best path in Received RD

```
Advertised path-id 1
  Path type: internal 0x40000018 0x880040 ref 1 adv path ref 1, path is valid, is best path, remote site path
  AS-Path: 102 65137 , path sourced external to AS
    172.16.3.227 (metric 64) from 10.0.184.65 (172.16.1.4)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 3
      Extcommunity:
          RT:102:36012032
          SOO:102:50331631
          VNID:2457600
          COST:pre-bestpath:166:2684354560
          COST:pre-bestpath:168:3221225472
```

See RT in Rx path is 102:36012032 .
This will be imported in vrf RD:RD on that leaf (see later slide)

Bgp process importing RT in leaf site1

```
S1P1-Leaf102# show bgp process vrf RD:RD

Information regarding configured VRFs:

BGP Information for VRF RD:RD
VRF Type                        : System
VRF Id                          : 6
VRF state                       : UP
VRF configured                  : yes
VRF refcount                    : 0
VRF VNID                        : 2392064
Router-ID                       : 10.37.1.254
Configured Router-ID            : 0.0.0.0
Confed-ID                       : 0
Cluster-ID                      : 0.0.0.0
MSITE Cluster-ID                : 0.0.0.0
VRF stretched                   : yes
No. of configured peers         : 0
No. of pending config peers     : 0
No. of established peers        : 0
VRF RD                          : 102:2392064
VRF EVPN RD                     : 102:2392064

    Information for address family IPv4 Unicast in VRF RD:RD
    Table Id                    : 6
    Table state                 : UP
    Table refcount              : 9
    Peers      Active-peers    Routes     Paths      Networks   Aggregates
    0          0               9          9          0          0

    Redistribution
        None

    Wait for IGP convergence is not configured
    Export RT list:
        101:2392064
    Import RT list:
        101:2392064
        102:36012032
```

MO involved to get the correct RT export and import configured

```
# bgp.RttEntry
rtt          : route-target:as2-nn4:101:2392064
childAction  :
dn           : topology/pod-1/node-102/sys/bgp/inst/dom-RD:RD/af-ipv4-ucast/rttp-export/ent-route-target:as2-nn4:101:2392064
lcOwn        : local
modTs        : 2019-11-14T12:56:51.915+00:00
rn           : ent-route-target:as2-nn4:101:2392064
status       :

# bgp.RttEntry
rtt          : route-target:as2-nn4:101:2392064
childAction  :
dn           : topology/pod-1/node-102/sys/bgp/inst/dom-RD:RD/af-ipv4-ucast/rttp-import/ent-route-target:as2-nn4:101:2392064
lcOwn        : local
modTs        : 2019-11-14T12:56:51.915+00:00
rn           : ent-route-target:as2-nn4:101:2392064
status       :

# bgp.RttEntry
rtt          : route-target:as2-nn4:102:36012032
childAction  :
dn           : topology/pod-1/node-102/sys/bgp/inst/dom-RD:RD/af-ipv4-ucast/rttp-import/ent-route-target:as2-nn4:102:36012032
lcOwn        : local
modTs        : 2019-11-18T12:38:23.362+00:00
rn           : ent-route-target:as2-nn4:102:36012032
status       :
```

Site 1 leaf import path in local RD

```
BGP routing table information for VRF overlay-1, address family VPNv4 Unicast
Route Distinguisher: 102:2392064    (VRF RD:RD)
BGP routing table entry for 23.23.23.0/24, version 21 dest ptr 0xaed9eed0
Paths: (1 available, best #1)
Flags: (0x08001a 00000000) on xmit-list, is in urib, is best urib route, is in HW
  vpn: version 381, (0x100002) on xmit-list
Multipath: eBGP iBGP

  Advertised path-id 1, VPN AF advertised path-id 1
  Path type: internal 0xc0000018 0x80040 ref 56506 adv path ref 2, path is valid, is best path, remote site path
             Imported from 101:36012032:23.23.23.0/24
  AS-Path: 102 65137 , path sourced external to AS
    172.16.3.227 (metric 64) from 10.0.184.65 (172.16.1.4)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 3
      Extcommunity:
          RT:102:36012032
          SOO:102:50331631
          VNID:2457600
          COST:pre-bestpath:166:2684354560
          COST:pre-bestpath:168:3221225472

  VRF advertise information:
  Path-id 1 not advertised to any peer

  VPN AF advertise information:
  Path-id 1 not advertised to any peer
```

Site 1 leaf – RIB

```
S1P1-Leaf102# vsh -c 'show ip route 23.23.23.0 det vrf RD:RD'
IP Route Table for VRF "RD:RD"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

23.23.23.0/24, ubest/mbest: 1/0
    *via 172.16.3.227%overlay-1, [200/0], 02:17:45, bgp-101, internal, tag 102 (mpls-vpn)
         MPLS[0]: Label=0 E=0 TTL=0 S=0 (VPN)
         client-specific data: 41
         recursive next hop: 172.16.3.227/32%overlay-1
         extended route information: BGP origin AS 65137 BGP peer AS 102 rw-vnid: 0x258000 table-id: 0x6 rw-mac: 0
S1P1-Leaf102#
S1P1-Leaf102#  vsh_lc -c 'dec 0x258000'
2457600
```

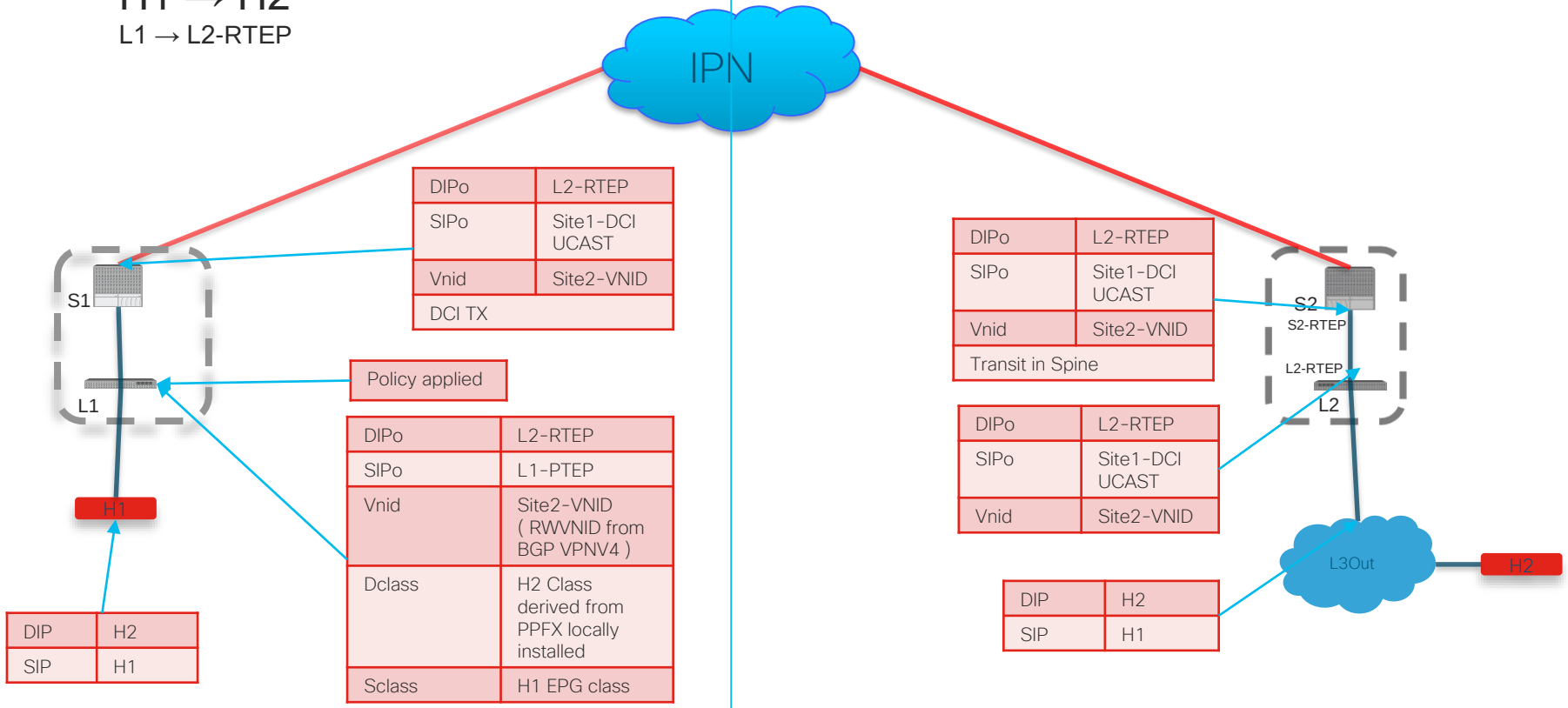
Dataplane

Site 1

Site 2

H1 → H2

L1 → L2-RTEP



Note on Disable Remote EP learning on BL

- Even if disable remote EP learning on BL is not set. We will disable remote IP learning on BL if InterSite L3 out is deployed !

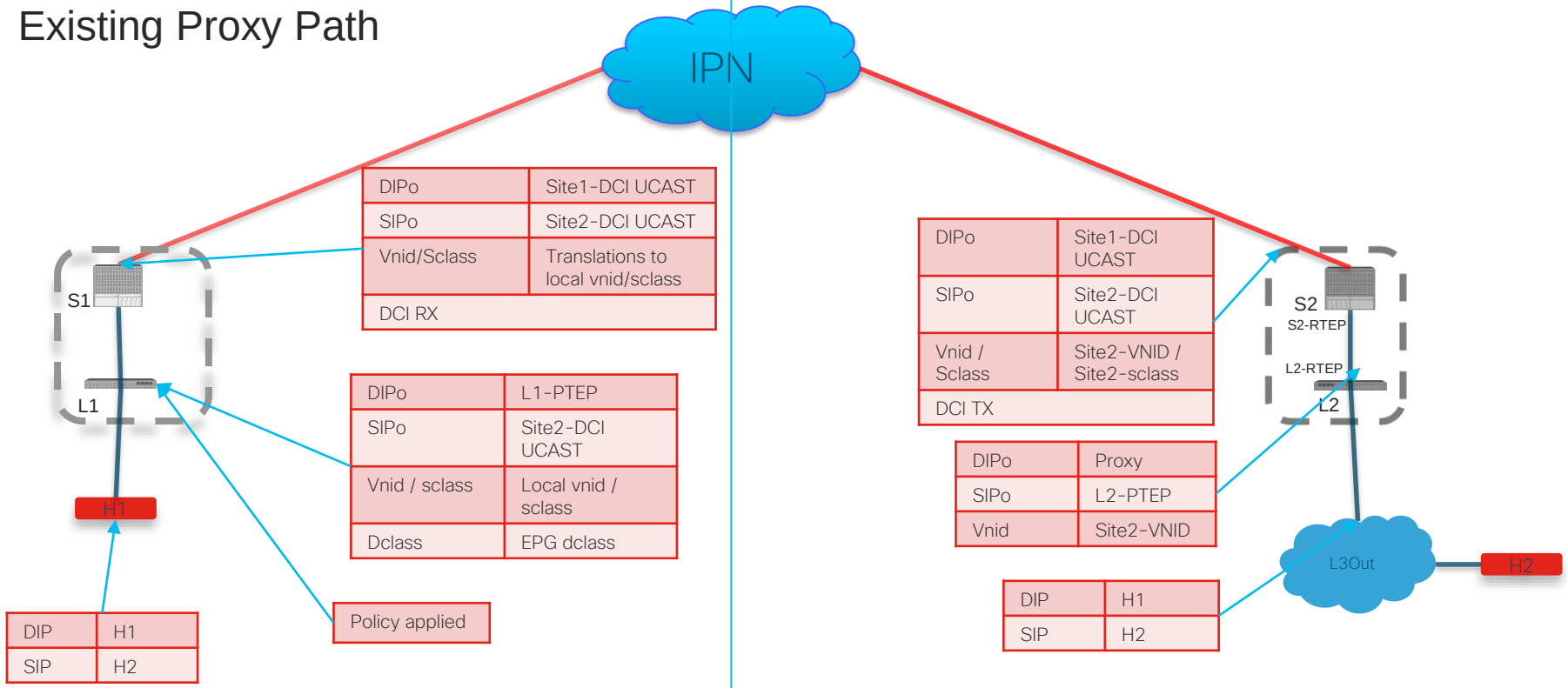
- So return path will always be proxy (L3 out to EP)

```
S2P1-Leaf101# vsh_lc
module-1#  show system internal  epm  vrf  RD:RD detail

VRF RD:RD
vrf type : Tenant
context id : 7 ::: vnid : 2457600
v4_usd_tbl_id: 0x7 :::  v4_tbl_idx : 0x7  ::: v6_tbl_idx :
0x80000007
Scope      : 2457600 ::: Sclass: 16386
EP retention policy valid : Yes
Local EP timeout : 900 ::: Remote EP timeout : 300
EP bounce timeout : 630 ::: EP hold timeout : 300
EP move frequency : 256
Endpoint count : 3
Border Leaf : yes
Learning disabled :no
Learning ip disabled :no
Learning xr ip disabled :no
Learning bl xr ip disabled :yes
Policy mode :Ingress
::::

MAC CKT HT is empty
::::
```

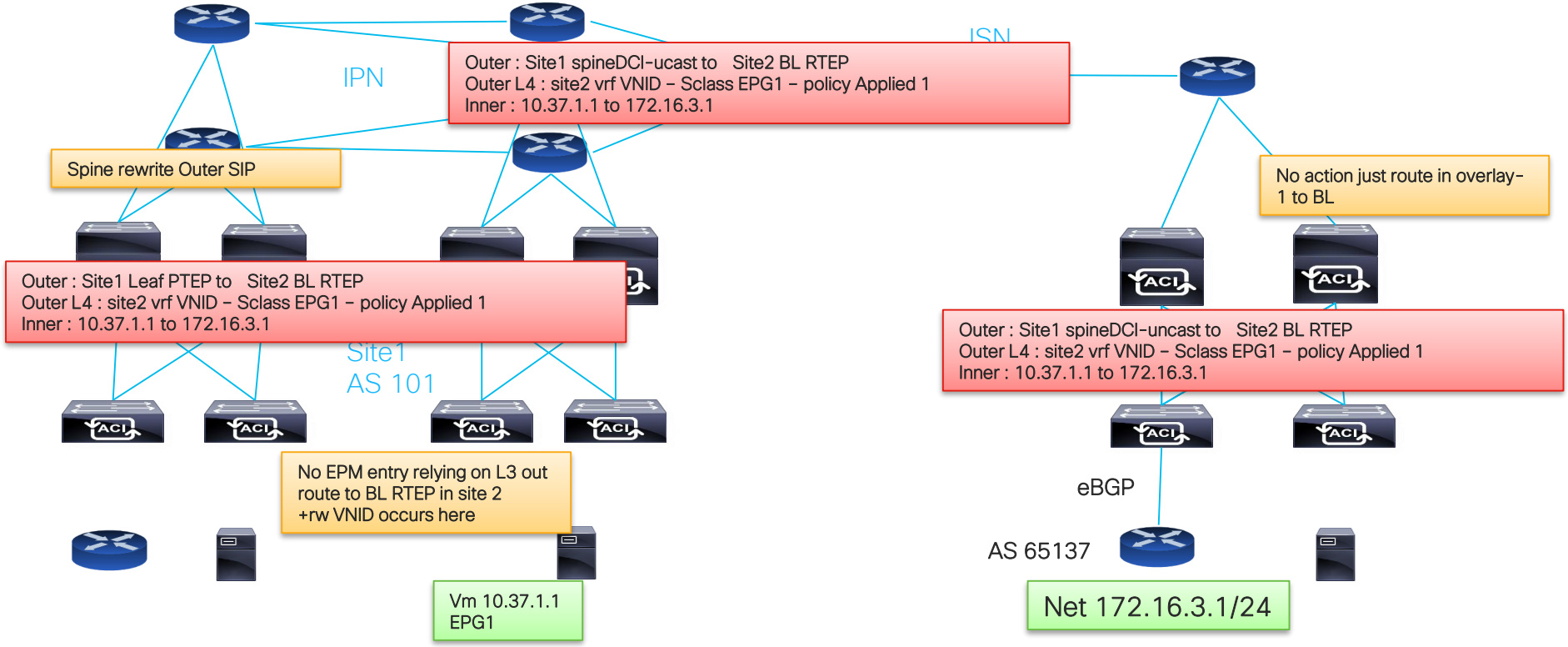
Site 1

Site 2

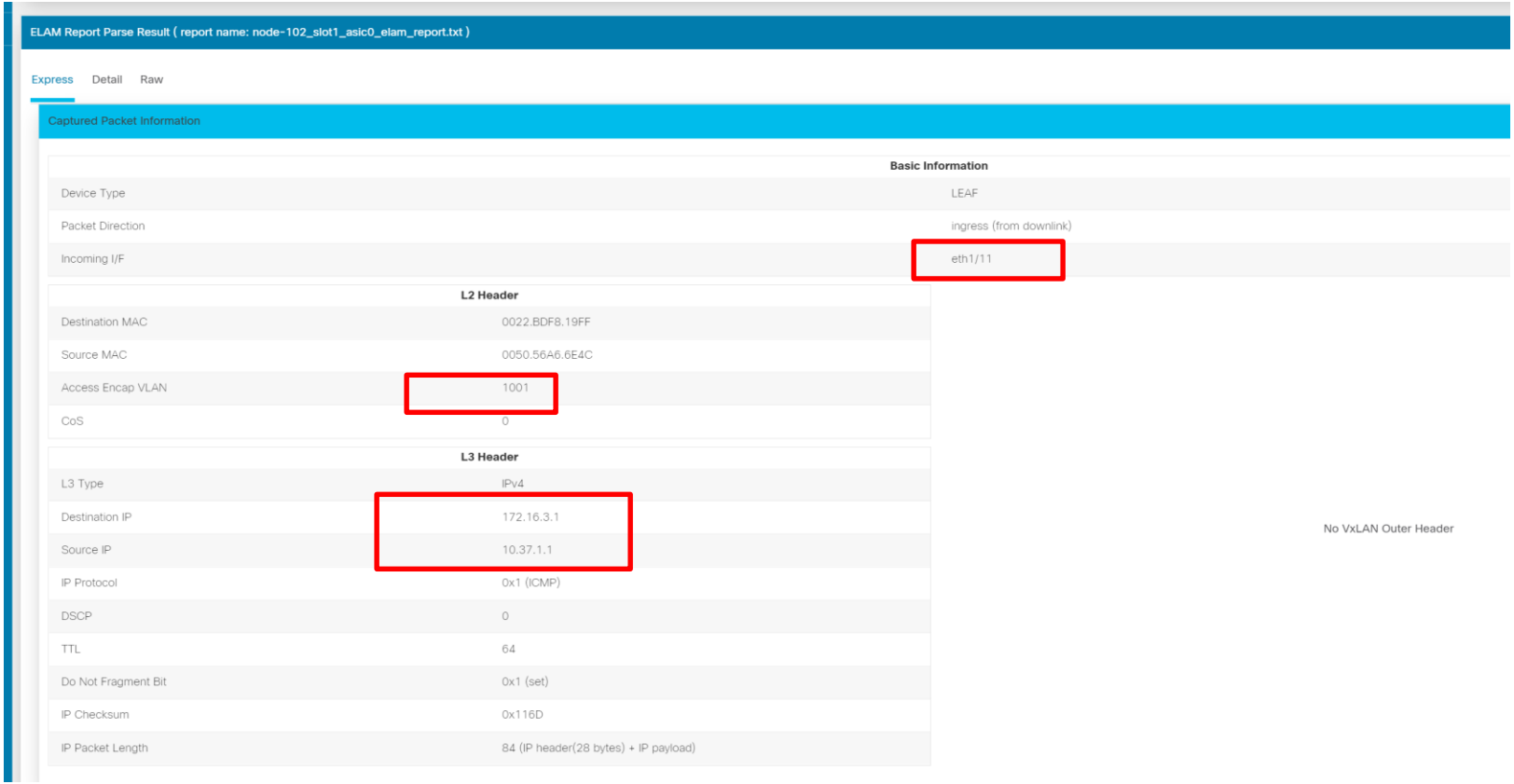
H2 → H1
Existing Proxy Path



Data-Path EP in site 1 to L3 out in site 2



Ingress leaf site 1 – leaf 2



ELAM Report Parse Result (report name: node-102_slot1_asic0_elam_report.txt)

Express Detail Raw

Captured Packet Information

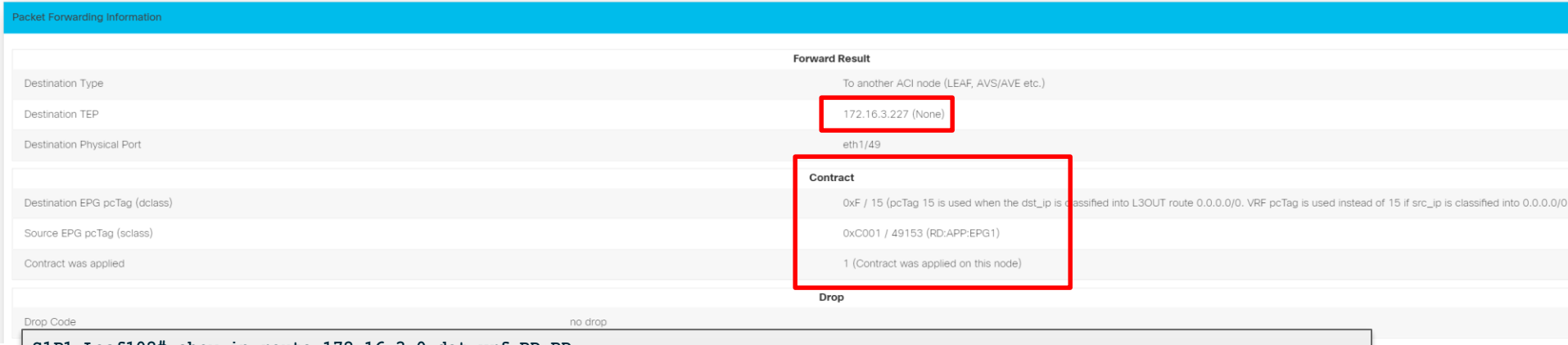
Basic Information	
Device Type	LEAF
Packet Direction	ingress (from downlink)
Incoming I/F	eth1/11

L2 Header	
Destination MAC	0022.BDF8.19FF
Source MAC	0050.56A6.6E4C
Access Encap VLAN	1001
CoS	0

L3 Header	
L3 Type	IPv4
Destination IP	172.16.3.1
Source IP	10.37.1.1
IP Protocol	0x1 (ICMP)
DSCP	0
TTL	64
Do Not Fragment Bit	0x1 (set)
IP Checksum	0x116D
IP Packet Length	84 (IP header(28 bytes) + IP payload)

No VxLAN Outer Header

Pkt Forwarding info in ingress leaf site 1




```
S1P1-Leaf102# show ip route 172.16.3.0 det vrf RD:RD
IP Route Table for VRF "RD:RD"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

172.16.3.0/24, ubest/mbest: 1/0
    *via 172.16.3.227%overlay-1, [200/0], 2d23h, bgp-101, internal, tag 102 (mpls-vpn)
         MPLS[0]: Label=0 E=0 TTL=0 S=0 (VPN)
         client-specific data: 41
         recursive next hop: 172.16.3.227/32%overlay-1
         extended route information: BGP origin AS 65137 BGP peer AS 102 rw-vnid: 0x258000 table-id: 0x6 rw-mac: 0
S1P1-Leaf102#
S1P1-Leaf102# show system internal policy-mgr prefix | egrep "Vrf|0.0.0.0"
Vrf-Vni VRF-Id Table-Id Table-State  VRF-Name                       Addr                                    Class Shared Remote Complete
2392064 6      0x6      Up           RD:RD                                 0.0.0.0/0    15      False  False  False
```

We will Rw VRF send it to
172.16.2.227 which if BL leaf in site2 RTEP
And apply
Contract (dest epg 0x15)

Site 1 - spine

Captured Packet Information

Basic Information	
Device Type	SPINE
Packet Direction	ingress (from leaf/ipn)
Incoming I/F	eth1/2

L2 Header	
Destination MAC	000C.0C0C.0C0C
Source MAC	000C.0C0C.0C0C
Access Encap VLAN	No VLAN Tag
CoS	No VLAN Tag (= No CoS)

Outer L2 Header	
Destination MAC	000D.0D0D.0D0D
Source MAC	000C.0C0C.0C0C
Access Encap VLAN	2
CoS	0

L3 Header	
L3 Type	IPv4
Destination IP	172.16.3.1
Source IP	10.37.1.1
IP Protocol	0x1 (ICMP)
DSCP	0
TTL	63
Do Not Fragment Bit	0x1 (0x1)

Outer L3 Header	
L3 Type	IPv4
Destination IP	172.16.3.227 (null)
Source IP	10.0.184.67 (S1P1-Leaf102)
IP Protocol	0x11 (UDP)
DSCP	0
TTL	32
Do not Fragment Bit	0x0 (0x0)

Outer L4 Header	
L4 Type	iVxLAN
DL (Don't Learn Bit)	0 (not set)
Src Policy Applied Bit	1 (Contract was applied on the previous node)
Dst Policy Applied Bit	1 (Contract was applied on the previous node)
Source EPG (sclass / src pcTag)	0xc001 / 49153 (null)
VRF/BD VNID	0x258000 / 2457600 (null)

We can verify
VRF VNID was indeed rewritten to 0x258000 = 2457600
Site 2 vrf vnid :

```
bdsol-aci38-apic1# moquery -d uni/tn-RD/ctx-RD  | egrep scope
scope                : 2457600
```

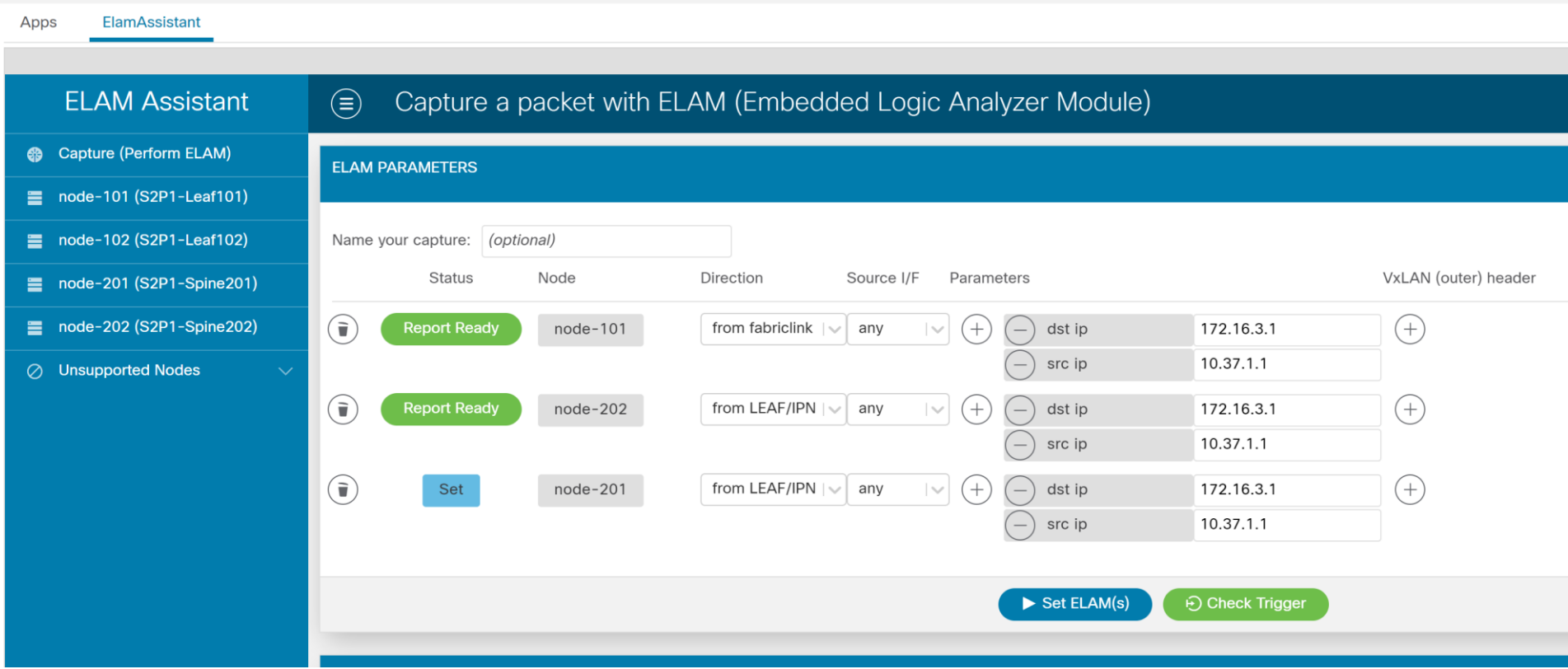
Site 1 – spine Packet forwarding

Packet Forwarding Information

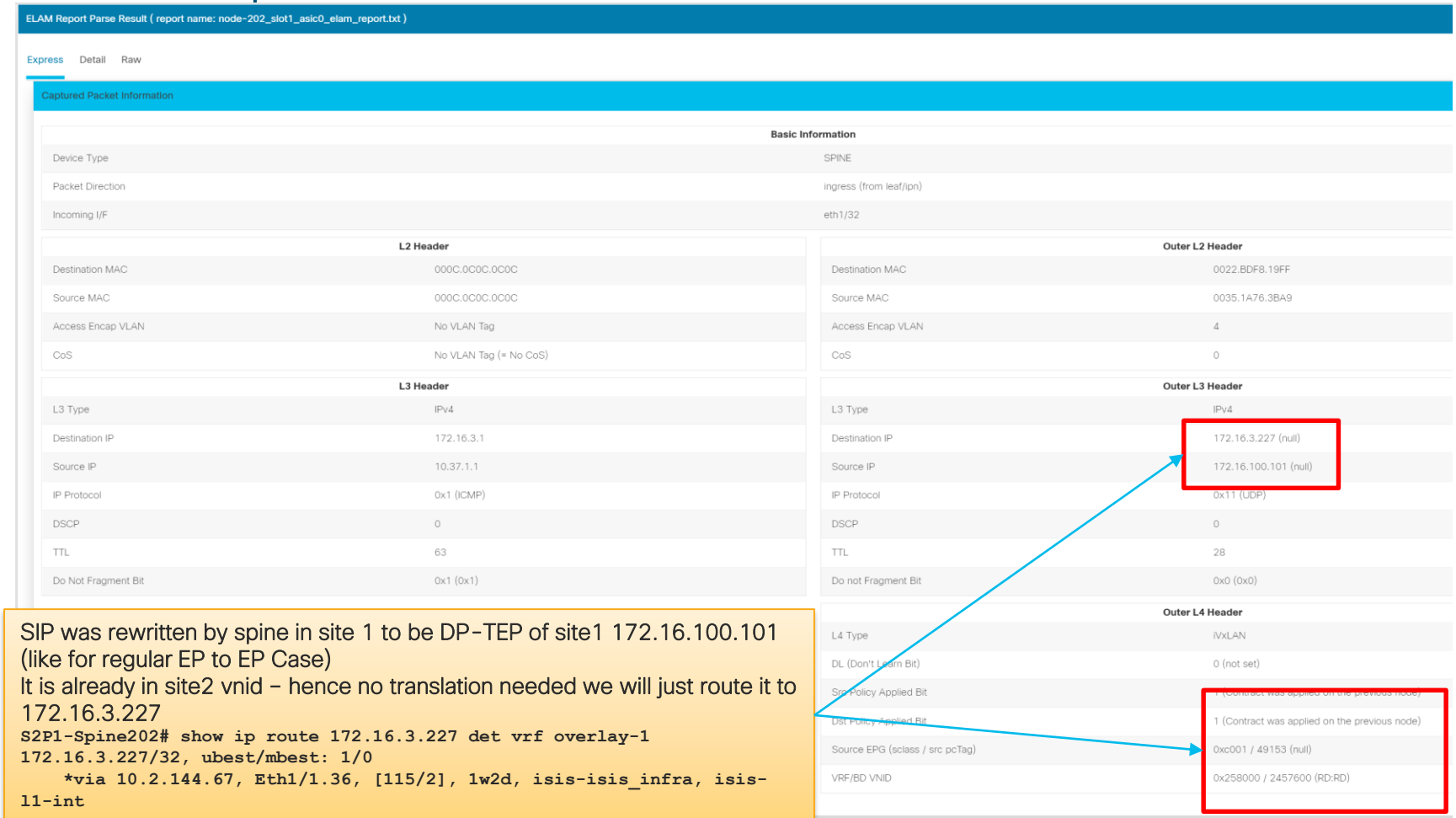
	Forward Result
Destination Type	To Proxy TEP (Spine Proxy)
Destination Fabric Card	eth1/29
	Contract
No contract is applied on SPINE	
	Drop
Drop Code	no drop

```
S1P1-Spine201# show lldp nei  | egrep "1\/29"
IPN1                 Eth1/29         120        BR          Ethernet1/53
S1P1-Spine201#
```

ELAM Assistant Site2



ELAM spine site2 from ISN



SIP was rewritten by spine in site 1 to be DP-TEP of site1 172.16.100.101 (like for regular EP to EP Case)

It is already in site2 vnid – hence no translation needed we will just route it to 172.16.3.227

```
S2P1-Spine202# show ip route 172.16.3.227 det vrf overlay-1
172.16.3.227/32, ubest/mbest: 1/0
    *via 10.2.144.67, Eth1/1.36, [115/2], 1w2d, isis-isis_infra, isis-
l1-int
```

Site 2 – egress leaf 1

ELAM Report Parse Result (report name: node-101_slot1_asic0_elam_report.txt)

Express Detail Raw

Captured Packet Information

Basic Information	
Device Type	LEAF
Packet Direction	egress (from fabriclink)
Incoming I/F	eth1/50

L2 Header	
Destination MAC	000C.0C0C.0C0C
Source MAC	000C.0C0C.0C0C
Access Encap VLAN	No VLAN Tag
CoS	No VLAN Tag (= No CoS)

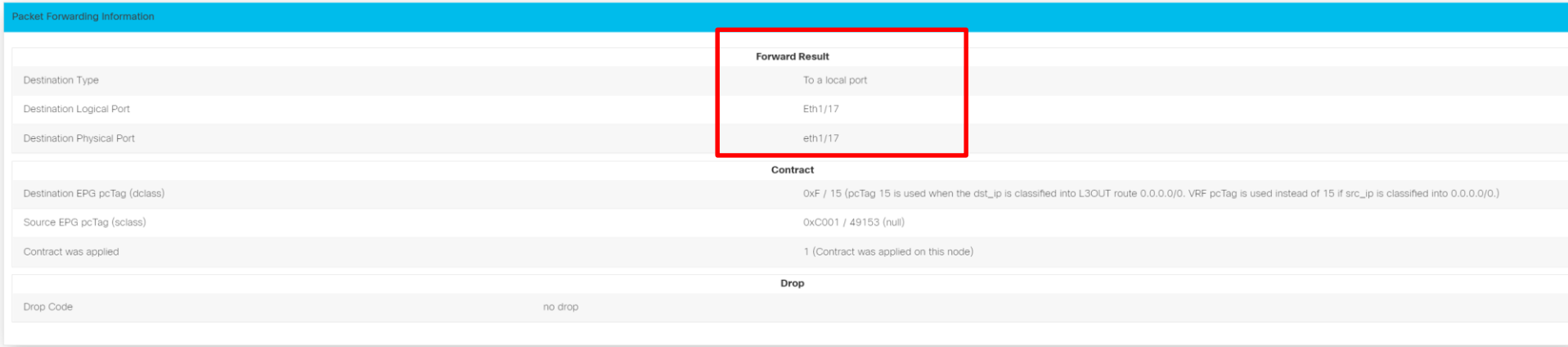
L3 Header	
L3 Type	IPv4
Destination IP	172.16.3.1
Source IP	10.37.1.1
IP Protocol	0x1 (ICMP)
DSCP	0
TTL	63
Do Not Fragment Bit	0x1 (0x1)

Outer L2 Header	
Destination MAC	000C.0C0C.0C0C
Source MAC	000D.0D0D.0D0D
Access Encap VLAN	2
CoS	0

Outer L3 Header	
L3 Type	IPv4
Destination IP	172.16.3.227 (null)
Source IP	172.16.100.101 (null)
IP Protocol	0x11 (UDP)
DSCP	0
TTL	27
Do not Fragment Bit	0x0 (0x0)

Outer L4 Header	
L4 Type	iVxLAN
DL (Don't Learn Bit)	0 (not set)
Src Policy Applied Bit	1 (Contract was applied on the previous node)
Dst Policy Applied Bit	1 (Contract was applied on the previous node)
Source EPG (sclass / src pcTag)	0xc001 / 49153 (null)
VRF/BD VNID	0x258000 / 2457600 (RD:RD)

Egress leaf site 2 sending traffic to L3 out



Packet Forwarding Information

Forward Result	
Destination Type	To a local port
Destination Logical Port	Eth1/17
Destination Physical Port	eth1/17
Contract	
Destination EPG pcTag (dclass)	0xF / 15 (pcTag 15 is used when the dst_ip is classified into L3OUT route 0.0.0.0/0. VRF pcTag is used instead of 15 if src_ip is classified into 0.0.0.0/0.)
Source EPG pcTag (sclass)	0xC001 / 49153 (null)
Contract was applied	1 (Contract was applied on this node)
Drop	
Drop Code	no drop

ECMP

- If there are multiple BLs in a site (in Site2) advertising the same route via L3Out, NBL (in Site1) will have ECMP towards this route.

  If there are multiple BLs in multiple sites (Site2, Site3) advertising the same route via L3Out, NBL
  (in Site1) will have ECMP towards this route across the sites.

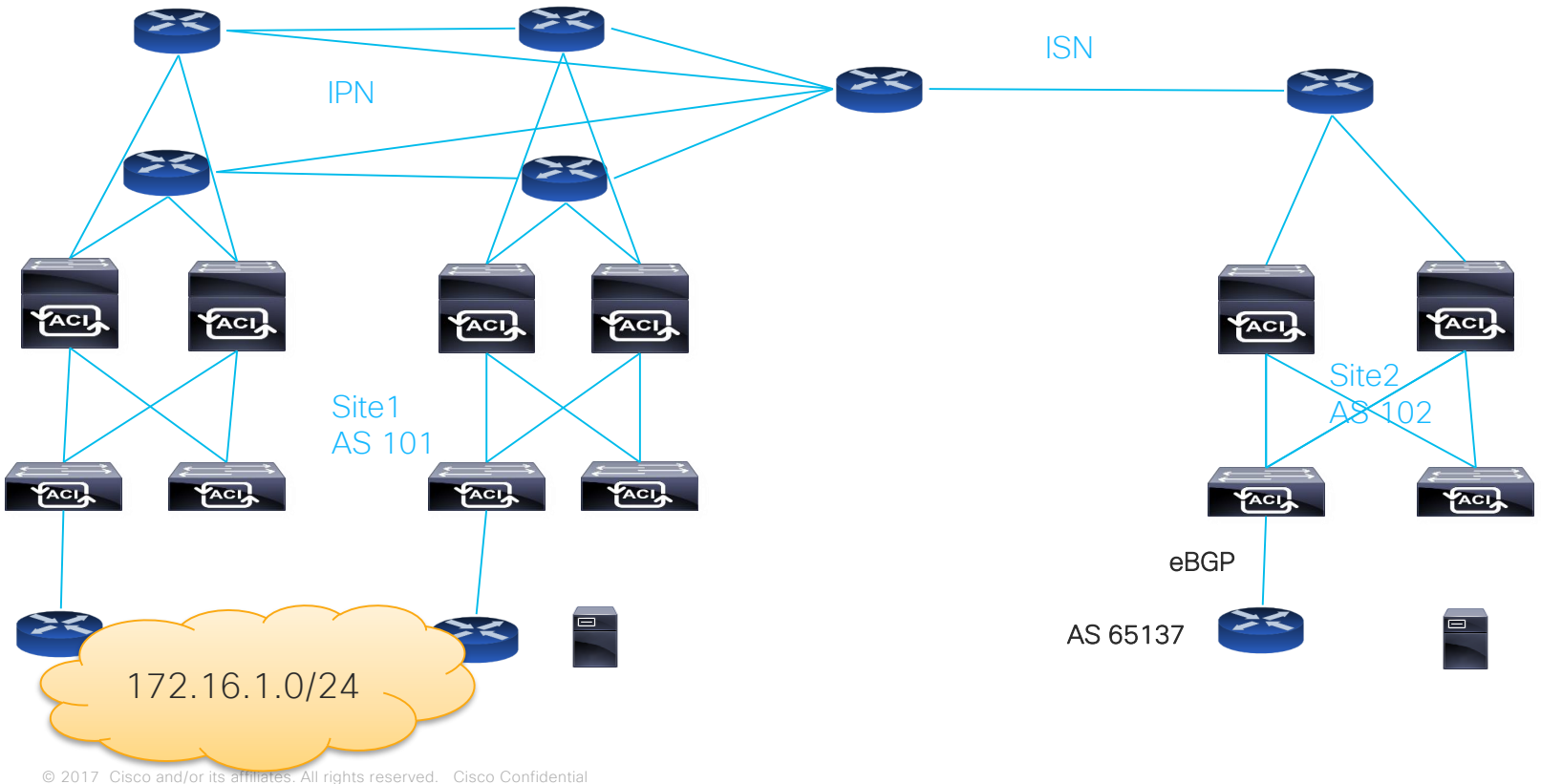
  A site (Site1) with multipod having BLs in multiple pods advertising the same route via L3Out, NBL in the same site (Site1) will always prefer the local site BL. When all the BL's in the site (Site1) goes down, then remote site path will be installed.

Tunnel

- Dynamic Tunnels
  Dynamic tunnels will be created between NBL to BL-RTEP, BL to BL-RTEP across sites for data plane traffic.
- Below tunnel on site 1 leaf to R-TEP of site 2 (172.16.3.x)

```
S1P1-Leaf101# show system internal epm  interface all | egrep "Tunn.*172.16.3"
 Tunnel15         0x1801000f UP    No  172.16.3.227    0000.0000.0000 0
 Tunnel16         0x18010010 UP    No  172.16.3.228    0000.0000.0000 0
```

ECMP example



Routing Table

Site1-Pod1-Leaf1 – BL hence prefer eBGP router

```
172.16.1.0/24, ubest/mbest: 1/0
    *via 10.37.100.2%RD:RD, [20/0], 01w02d, bgp-101, external, tag 651
         recursive next hop: 10.37.100.2/32%RD:RD
```

Site2 -Leaf1 (all leaf) – GET ECMP routes to both Site1 BL RTEP

```
172.16.1.0/24, ubest/mbest: 2/0
    *via 172.16.1.236%overlay-1, [200/0], 01w01d, bgp-102, internal, tag 101
         recursive next hop: 172.16.1.236/32%overlay-1
    *via 172.16.2.234%overlay-1, [200/0], 00:13:39, bgp-102, internal, tag 101
         recursive next hop: 172.16.2.234/32%overlay-1
```

Site1-Pod1-Leaf2 – Regular Mpod – Prefer Local pod TEP

```
172.16.1.0/24, ubest/mbest: 1/0
    *via 10.0.184.64%overlay-1, [200/0], 01w02d, bgp-101, internal, tag 65137
         recursive next hop: 10.0.184.64/32%overlay-1
```

Site1-Pod2-Leaf1 – BL hence prefer eBGP route

```
172.16.1.0/24, ubest/mbest: 1/0
    *via 10.37.100.10%RD:RD, [20/0], 00:23:31, bgp-101, external, tag 65137
         recursive next hop: 10.37.100.10/32%RD:RD
```

Site1-Pod2-Leaf2 - Regular Mpod – Prefer Local pod TEP

```
172.16.1.0/24, ubest/mbest: 1/0
    *via 10.1.192.66%overlay-1, [200/0], 00:12:53, bgp-101, internal, tag 65137
         recursive next hop: 10.1.192.66/32%overlay-1
```

Note only leaf on remote site Sees the prefix with NH being the RTEP. Indeed NH is set in outbound Route-map of spine towards ISN

BGP path – site 1- Pod1

Site1 – Pod1 –leaf1 BL – 2path
- Internal path from pod2-BL
- External path from L3 out (preferred)

```
 VPN AF advertised path-id 2
  Path type: internal 0xc0000018 0x40 ref 56506 adv path ref 1, path
is valid, not best reason: Internal path
             Imported from 301:2392064:172.16.1.0/24
  AS-Path: 65137 , path sourced external to AS
    10.1.192.66 (metric 64) from 10.0.184.65 (172.16.1.4)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 2
      Extcommunity:
          RT:101:2392064
          VNID:2392064
          COST:pre-bestpath:165:2415919104
          COST:pre-bestpath:168:3221225472
      Originator: 10.1.192.66 Cluster list: 172.16.1.4 172.16.2.1

  Advertised path-id 1, VPN AF advertised path-id 1
  Path type: external 0x28 0x0 ref 0 adv path ref 2, path is valid,
is best path
  AS-Path: 65137 , path sourced external to AS
    10.37.100.2 (metric 0) from 10.37.100.2 (10.37.0.101)
      Origin IGP, MED not set, localpref 100, weight 0
      Extcommunity:
          RT:101:2392064
          VNID:2392064
```

Site1 – Pod1 –leaf2 non-BL – 2 iBGP path
- From Remote pod BL
- From local pod BL (preferred)

```
  VPN AF advertised path-id 2
  Path type: internal 0xc0000018 0x40 ref 56506 adv path ref 1, path is
valid, not best reason: NH metric
             Imported from 301:2392064:172.16.1.0/24
  AS-Path: 65137 , path sourced external to AS
    10.1.192.66 (metric 64) from 10.0.184.65 (172.16.1.4)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 2
      Extcommunity:
          RT:101:2392064
          VNID:2392064
          COST:pre-bestpath:165:2415919104
          COST:pre-bestpath:168:3221225472
      Originator: 10.1.192.66 Cluster list: 172.16.1.4 172.16.2.1

  Advertised path-id 1, VPN AF advertised path-id 1
  Path type: internal 0xc0000018 0x40 ref 56506 adv path ref 2, path is
valid, is best path
             Imported from 101:2392064:172.16.1.0/24
  AS-Path: 65137 , path sourced external to AS
    10.0.184.64 (metric 3) from 10.0.184.65 (172.16.1.4)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 1
      Extcommunity:
          RT:101:2392064
          VNID:2392064
          COST:pre-bestpath:168:3221225472
      Originator: 10.0.184.64 Cluster list: 172.16.1.4
```

Note : site 1 –pod2 would show similar info(not shown in slide)

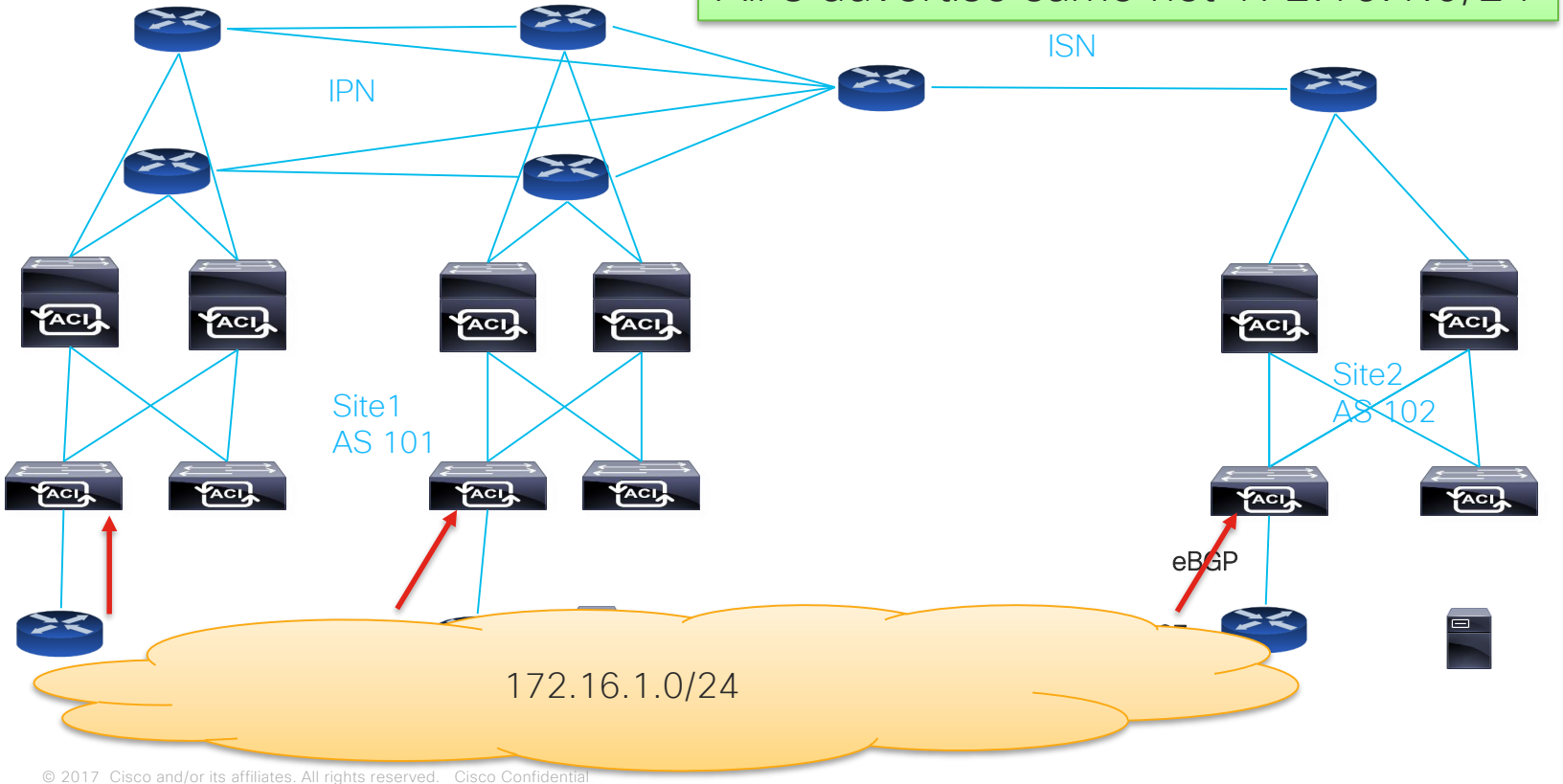
BGP path – site 2

Site2 -Leaf1 (all leaf) – GET ECMP routes to both Site1 BL RTEP

```
S2P1-Leaf101# show bgp vpnv4 unicast 172.16.1.0/24 vrf RD:RD
BGP routing table information for VRF overlay-1, address family VPNv4 Unicast
Route Distinguisher: 101:2457600    (VRF RD:RD)
BGP routing table entry for 172.16.1.0/24, version 41 dest ptr 0xc64b3c30

  Advertised path-id 1, VPN AF advertised path-id 1
  Path type: internal 0xc0000018 0x80040 ref 56506 adv path ref 2, path is valid, is best path, remote site path
             Imported from 301:19169280:172.16.1.0/24
  AS-Path: 101 65137 , path sourced external to AS
    172.16.2.234 (metric 64) from 10.2.144.65 (172.16.200.201)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 4
      Extcommunity:
          RT:101:19169280
          SOO:101:33554415
          VNID:2392064
          COST:pre-bestpath:165:2415919104
          COST:pre-bestpath:166:2684354560
          COST:pre-bestpath:168:3221225472

  VPN AF advertised path-id 2
  Path type: internal 0xc0020018 0x80040 ref 56506 adv path ref 1, path is valid, not best reason: Neighbor Address, remote site path, multipath
             Imported from 101:19169280:172.16.1.0/24
  AS-Path: 101 65137 , path sourced external to AS
    172.16.1.236 (metric 64) from 10.2.144.65 (172.16.200.201)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Received path-id 4
      Extcommunity:
          RT:101:19169280
          SOO:101:33554415
          VNID:2392064
          COST:pre-bestpath:165:2415919104
          COST:pre-bestpath:166:2684354560
          COST:pre-bestpath:168:3221225472
```

ECMP example

Adding a 3rd External L3 out in site-2
All 3 advertise same net 172.16.1.0/24



Routing Table

Site1-Pod1-Leaf1 – BL hence prefer eBGP router

```
172.16.1.0/24, ubest/mbest: 1/0
    *via 10.37.100.2%RD:RD, [20/0], 01w02d, bgp-101, external, tag 65137
         recursive next hop: 10.37.100.2/32%RD:RD
```

Site2 -Leaf1 (BL) – GET eBGP route from L3 out

```
172.16.1.0/24, ubest/mbest: 1/0
    *via 10.37.100.6%RD:RD, [20/0], 00:00:55, bgp-102, external, tag 65137
         recursive next hop: 10.37.100.6/32%RD:RD
```

Site1-Pod1-Leaf2 – Regular Mpod – Prefer Local pod TEP

```
172.16.1.0/24, ubest/mbest: 1/0
    *via 10.0.184.64%overlay-1, [200/0], 01w02d, bgp-101, internal, t
         recursive next hop: 10.0.184.64/32%overlay-1
```

Site2 -Leaf2 (non-BL) – only route to local Site BL TEP

```
172.16.1.0/24, ubest/mbest: 1/0
    *via 10.2.144.67%overlay-1, [200/0], 00:03:24, bgp-102, internal, tag 65137
         recursive next hop: 10.2.144.67/32%overlay-1
```

Site1-Pod2-Leaf1 – BL hence prefer eBGP route

```
172.16.1.0/24, ubest/mbest: 1/0
    *via 10.37.100.10%RD:RD, [20/0], 00:23:31, bgp-101, external, tag 65137
         recursive next hop: 10.37.100.10/32%RD:RD
```

Site1-Pod2-Leaf2 – Regular Mpod – Prefer Local pod TEP

```
172.16.1.0/24, ubest/mbest: 1/0
    *via 10.1.192.66%overlay-1, [200/0], 00:12:53, bgp-101, internal, tag 65137
         recursive next hop: 10.1.192.66/32%overlay-1
```

BGP path – site 2 – NonBL

We are getting 3 path as expected (one from each BL

```
VPN AF advertised path-id 3
 Path type: internal 0xc0000018 0x80040 ref 56506 adv path ref 1, path is valid, not best reason: AS Path, remote site path
            Imported from 301:19169280:172.16.1.0/24
 AS-Path: 101 65137 , path sourced external to AS
   172.16.2.234 (metric 64) from 10.2.144.65 (172.16.200.201)
     Origin IGP, MED not set, localpref 100, weight 0
     Received label 0
     Received path-id 4
     Extcommunity:
         RT:101:19169280
         SOO:101:33554415
         VNID:2392064
         COST:pre-bestpath:165:2415919104
         COST:pre-bestpath:166:2684354560
         COST:pre-bestpath:168:3221225472

 VPN AF advertised path-id 2
 Path type: internal 0xc0000018 0x80040 ref 56506 adv path ref 1, path is valid, not best reason: Neighbor Address, remote site path
            Imported from 101:19169280:172.16.1.0/24
 AS-Path: 101 65137 , path sourced external to AS
   172.16.1.236 (metric 64) from 10.2.144.65 (172.16.200.201)
     Origin IGP, MED not set, localpref 100, weight 0
     Received label 0
     Received path-id 4
     Extcommunity:
         RT:101:19169280
         SOO:101:33554415
         VNID:2392064
         COST:pre-bestpath:165:2415919104
         COST:pre-bestpath:166:2684354560
         COST:pre-bestpath:168:3221225472

 Advertised path-id 1, VPN AF advertised path-id 1
 Path type: internal 0xc0000018 0x40 ref 56506 adv path ref 2, path is valid, is best path
            Imported from 101:2457600:172.16.1.0/24
 AS-Path: 65137 , path sourced external to AS
   10.2.144.67 (metric 3) from 10.2.144.65 (172.16.200.201)
     Origin IGP, MED not set, localpref 100, weight 0
     Received label 0
     Received path-id 1
     Extcommunity:
         RT:102:2457600
         VNID:2457600
         COST:pre-bestpath:168:3221225472
```

Path from site1-pod 2

Path from site1- Pod

Path local from site 2 - preferred

Criteria used to exclude path 1 and
Is misleading in the command outp
Actually we pick up criteria on lowe
IGP metric (3 for local site path Vs
for remote site path.

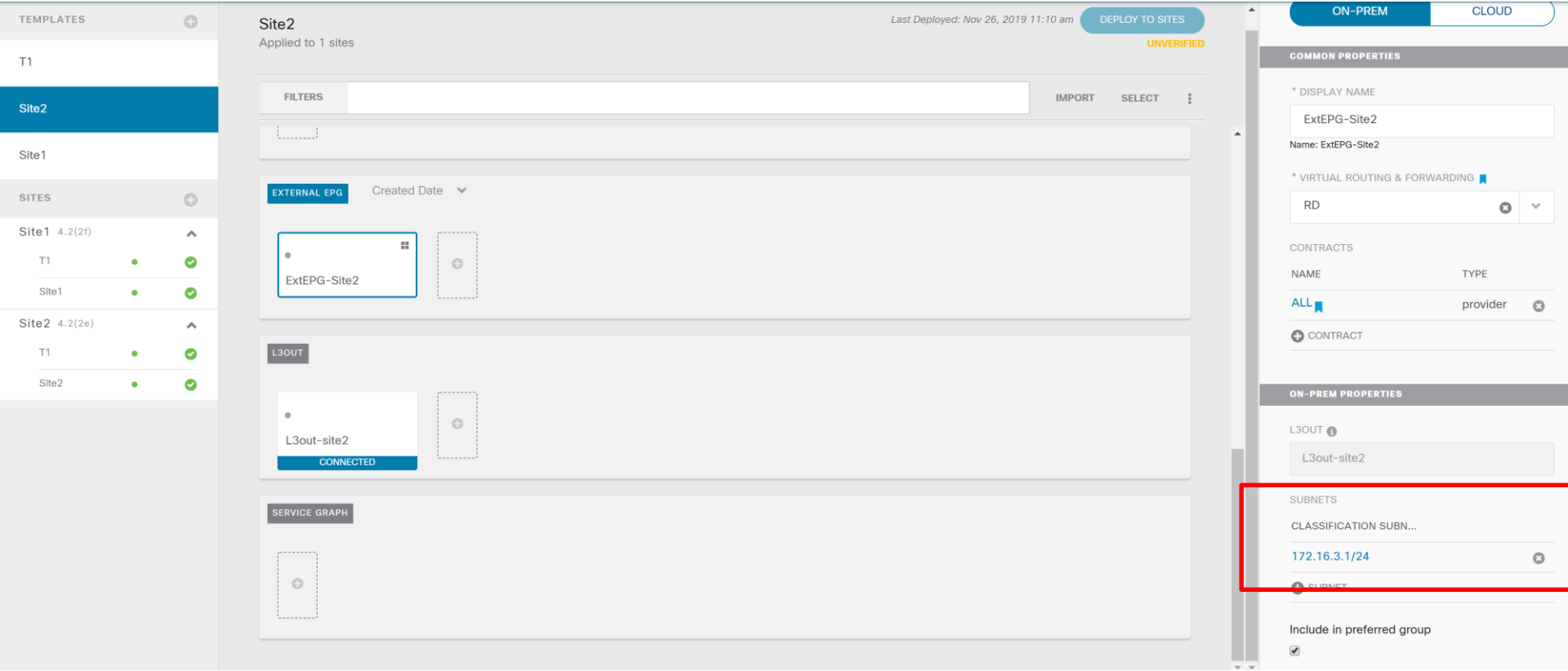
Contract enforcement Specific Prefix

L3 out prefix (import-security) with intersite l3 out

Epg To layer 3 out path

- EPG site 1 to L3 out site 2
- In that direction Policy is always enforced on ingress server leaf as policymgr prefix is propagated to both sites
- NO NEED for pcTag translation every thing happen in src site between server epg pcTag and Local pcTAg of remote L3 out

Add a subnet in the l3 out epg for site-2 layer 3 out



Policymgr prefix and zoning-rule

```
S1P1-Leaf102# show system internal policy-mgr prefix
Requested prefix data

Vrf-Vni VRF-Id Table-Id Table-State  VRF-Name                        Addr                                  Class Shared Remote Complete
======= ====== ========== =======  ============================== ===================================== ====== ====== ====== ========
2392064 6      0x6        Up       RD:RD                                    0.0.0.0/0                   15     False  False  False
2392064 6      0x80000006 Up       RD:RD                                        ::/0                    15     False  False  False
2392064 6      0x6        Up       RD:RD                                172.16.3.1/24                   32772  False  True   False

S1P1-Leaf102# show zoning-rule scope 2392064
+---------+--------+--------+----------+----------------+---------+---------+------+----------+----------------------+
| Rule ID | SrcEPG | DstEPG | FilterID |      Dir       |  operSt |  Scope  | Name |  Action  |       Priority       |
+---------+--------+--------+----------+----------------+---------+---------+------+----------+----------------------+
|   4105  |   0    | 16387  | implicit |    uni-dir     | enabled | 2392064 |      |  permit  |   any_dest_any(16)   |
|   4106  |   0    |   0    | implicit |    uni-dir     | enabled | 2392064 |      | deny,log |   any_any_any(21)    |
|   4107  |   0    |   0    | implarp  |    uni-dir     | enabled | 2392064 |      |  permit  |  any_any_filter(17)  |
|   4108  |   0    |   15   | implicit |    uni-dir     | enabled | 2392064 |      | deny,log | any_vrf_any_deny(22) |
|   4109  | 49153  | 32772  |    8     |     bi-dir     | enabled | 2392064 | ALL  |  permit  |    fully_qual(7)     |
|   4110  | 32772  | 49153  |    8     | uni-dir-ignore | enabled | 2392064 | ALL  |  permit  |    fully_qual(7)     |
|   4113  | 49153  |   15   |    8     |    uni-dir     | enabled | 2392064 | ALL  |  permit  |    fully_qual(7)     |
|   4114  | 32770  | 49153  |    8     |    uni-dir     | enabled | 2392064 | ALL  |  permit  |    fully_qual(7)     |
|   4112  | 49153  | 32771  |    8     |     bi-dir     | enabled | 2392064 | ALL  |  permit  |    fully_qual(7)     |
|   4111  | 32771  | 49153  |    8     | uni-dir-ignore | enabled | 2392064 | ALL  |  permit  |    fully_qual(7)     |
+---------+--------+--------+----------+----------------+---------+---------+------+----------+----------------------+
```

ELAM assistant ingress server leaf site -1

Packet Forwarding Information

Forward Result	
Destination Type	To another ACI node (LEAF, AVS/AVE etc.)
Destination TEP	172.16.3.227 (None)
Destination Physical Port	eth1/49
Contract	
Destination EPG pcTag (dclass)	0x8004 / 32772 (L3OUT RD:L3out-site2:ExtEPG-Site2)
Source EPG pcTag (sclass)	0xC001 / 49153 (RD:APP:EPG1)
Contract was applied	1 (Contract was applied on this node)
Drop	
Drop Code	no drop

L3 out prefix (import-security) with intersite l3 out

Layer 3 out site-2 to EPG site-1

- BL have disable remote IP learning on BL set per intersite feature, hence it takes proxy path and ingress BL will never enforce policy
- NEED pcTag translate in egress spine site and policy enforced in egress leaf
- Site 2 BL ingress leaf set the following pcTag
  - Sclass – site 2 local value for l3 out site 2
  - Dclass – 1 (proxy-path)
- Site 1 spine (egress site spine)
  - Does sclass translate in vxlan header from site2 local to site1 local value for site2 L3 out
- Policy is enforce in site 1 server leaf using
  - Sclass – Site1 local value for l3out site-2
  - Dclass – Epg pcTag

Ingress layer-3 BL in site-2

Packet Forwarding Information

Forward Result	
Destination Type	To another ACI node (LEAF, AVS/AVE etc.)
Destination TEP	10.2.104.64 (IPv4 Spine-Proxy)
Destination Physical Port	eth1/49

Contract	
Destination EPG pcTag (dclass)	0x1 / 1 (pcTag 1 is to ignore contract for special packets such as Spine-Proxy, ARP, Multicast etc..)
Source EPG pcTag (sclass)	0xC002 / 49154 (null)
Contract was applied	0 (Contract was not applied on this node)

Drop	
Drop Code	no drop

Server leaf site-1 (L3 out site2 to epg site1)

Outer L4 Header	
L4 Type	iVxLAN
DL (Don't Learn Bit)	1 (set)
Src Policy Applied Bit	0 (Contract has yet to be applied)
Dst Policy Applied Bit	0 (Contract has yet to be applied)
Source EPG (sclass / src pcTag)	0x8004 / 32772 (L3OUT RD:L3out-site2:ExtEPG-Site2)
VRF/BD VNID	0x248000 / 2392064 (RD:RD)

Packet Forwarding Information

Forward Result	
Destination Type	To a local port
Destination Logical Port	Eth1/11
Destination Physical Port	eth1/11

Contract	
Destination EPG pcTag (dclass)	0xC001 / 49153 (RD:APP:EPG1)
Source EPG pcTag (sclass)	0x8004 / 32772 (L3OUT RD:L3out-site2:ExtEPG-Site2)
Contract was applied	1 (Contract was applied on this node)

Drop

Egress spine (spine site-1)

As on regular intersite communication dcimgr on spine
Gets the translation from 49154 (l3 out site2) to 32772 (shadow epg value)

```
S1P1-Spine201# show dcimgr repo sclass-maps  | egrep "Remote|Vrf|2392064|---"
------------------------------------------------------------
     Remote             |        Local
site  Vrf       PcTag   |  Vrf       PcTag     Rel-state
------------------------------------------------------------
  2   2457600   16386   |  2392064   32770    [formed]
  2   2457600   16389   |  2392064   49153    [formed]
  2   2457600   16388   |  2392064   32771    [formed]
  2   2457600   49154   |  2392064   32772    [formed]
  2   2457600   16390   |  2392064   16388    [formed]
```