White paper
Cisco public

Networking Infrastructure for Telco Edge Data Centers

A Case Study

Contents

1 Introduction to Edge Computing

The objective of edge computing is to provide compute and storage resources closer to the end user, typically deploying them at the network edge. This approach reduces latency, accelerates data processing through localization, and conserves network bandwidth by reducing data travel distances. Edge computing plays an increasingly essential role in supporting scenarios like 5G IoT applications, autonomous driving, industrial automation, and real-time data analytics where minimizing delays and optimizing resource utilization are critical.

In the Service Provider world, Multi-Access Edge Computing (MEC) provides an architectural framework for practical applications of edge computing and is driven by the European Telecommunications Standards Institute (ETSI). MEC solutions are actively used in 4G mobility networks to offer services in the edge and have become increasingly integrated into 5G mobility networks, playing critical roles in facilitating low-latency and mission-critical services. With Virtualization and Cloud Native platforms being increasingly used to deploy 4G and 5G RAN and Packet Core Network Functions, edge computing enables Service Providers to move ahead with network decomposition and network disaggregation and build more agile networks. The edge computing platform can be used in different ways and combinations for deploying:

- Security and network services like firewalls, Dynamic Host Configuration Protocol (DHCP), and DNS
- O-RAN Control and user functions such as gNB-CU-CP, gNB-CU-UP
- Packet Core User Plane Function (UPF)
- Applications

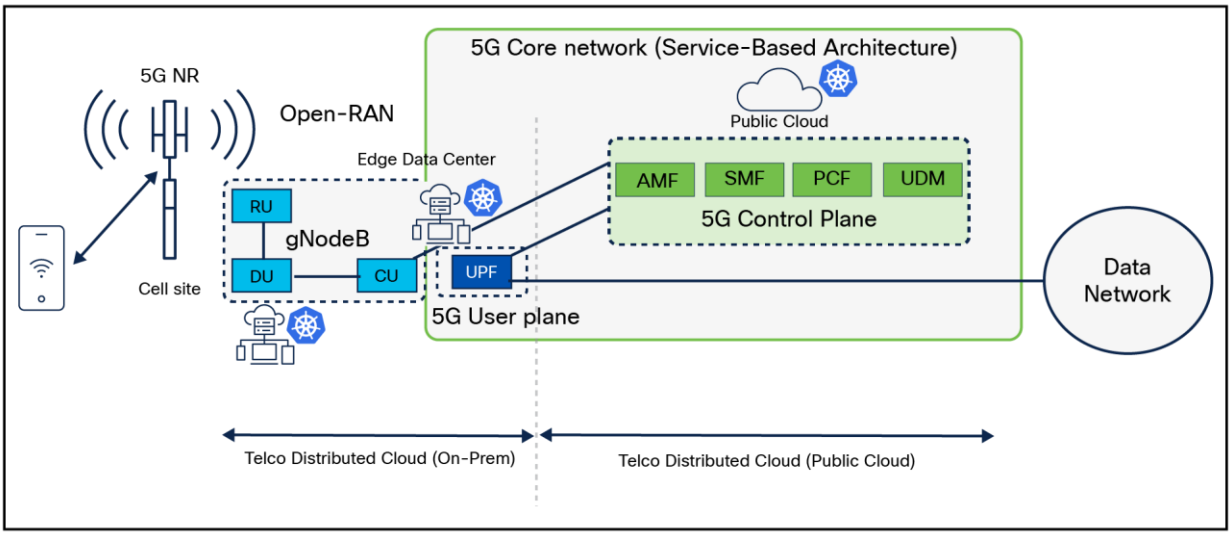

Figure 1.
Distributed Telco Infrastructure to Support 5G

Figure 1 illustrates a hybrid deployment model for 5G and Open-RAN. In this model, the on-premises infrastructure hosts Open-RAN components, including the User Plane Function of the Packet Core, while the remaining Packet Core components are deployed in the Public Cloud. The on-premises infrastructure Data Network Edge Data Center UPF 5G User Plane 5G Core Network (Service Based Architecture) DU RU Cell Site Open-RAN AMF SMF PCF UDM 5G Control Plane CU Public Cloud Telco Distributed Cloud (OnPrem) Telco Distributed Cloud (Public Cloud) 5G NR gNodeB consists of Far Edge Cell Sites and Edge Data Centers, which collectively form the On-Prem Telco Distributed Cloud to support virtualized and cloud native network functions.

The compute and storage infrastructure needed for edge computing is managed as a Distributed Cloud Platform and the network infrastructure needed can vary depending on the requirements. For example, compute nodes can be directly connected to a Transport Router, or to a Software-Defined Networking (SDN) – enabled switching platform. In some cases, integrated edge computing solutions provided by Public Cloud Providers (PCP) can be leveraged.

To support their 5G rollout, Dish Wireless deployed several on-prem Edge Data Centers to serve key markets in their nationwide architecture. These distributed Telco Edge Data Centers host compute infrastructure and services to address the requirements of Open RAN and 5G are interconnected using a high-speed transport backbone.

In this document, we examine the solution for deploying network infrastructure across distributed Telco Edge Data Centers and its integration with the transport domain.

2 The Edge Data Center Network

2.1 Broad Requirements

The services that can be provided by Network Functions hosted in the Edge Data Centers can include:

- DHCP and DNS services for components deployed in the cell sites
- Firewalling to secure connectivity with B2B Partners for 5G Services
- 5G User Plane Functions that are closer to the subscribers and applications
- Open-RAN Control Plane and User Plane functions
- Caching services

The networking infrastructure in the Edge Data Center is required to support the following requirements:

1. Software-Defined Networking (SDN) stack, designed for ease of operation and simplicity in management
2. The ability to easily increase capacity to the Edge Data Center
3. Host Network functions that could be deployed in various form factors including Physical (Physical Network Function or Appliance), Virtual Machines (Virtual Network Function or VNF) and Containers (Container Network Function or CNF)
4. The network functions can be deployed in any rack, can move around in the available compute as required, and support routing adjacency using dynamic routing protocols
5. Seamlessly extend Virtual Routing and Forwarding Tables (VRFs) from the Transport Domain into the Edge Data Center while maintaining network segmentation and providing secure access to the network functions
6. Macro- and Micro-Segmentation to support multi-tenancy and isolation

2.2 Solution

The traditional approach to network deployments, involving the management and monitoring of switches in the data center on a device-by-device basis, becomes inefficient and cumbersome when applied to a distributed telco Edge Data Center model.

The Cisco® ACI portfolio aligns with the distributed telco data centers model and offers solutions to enable the deployment of services across geographically separated edge locations. It provides a fully automated solution to deploy and operationalize a switching fabric to support distributed Edge Data Centers. Furthermore, it utilizes an intent-based approach to define application connectivity in simple terms, which is then translated into networking language (configurations) to configure the switches.

2.3 Picking the Right ACI Solution

2.3.1 Options

Cisco ACI offers users multiple deployment options. These include:

- Cisco ACI Multi-Pod
- Cisco ACI Multi-Site
- Cisco ACI Remote Leaf

A brief and high-level descriptions of these options is provided below:

Pod: A pod is a leaf-and-spine network sharing a common control plane (Intermediate System-to-Intermediate System [ISIS], Border Gateway Protocol [BGP], Council of Oracle Protocol [COOP], etc.). The leaf-and-spine nodes are under the control of an APIC (Application Policy Infrastructure Controller) domain and is a single network fault domain that can be considered as an availability zone.

Remote Leaf: A remote leaf design is where the APIC controllers are located at the main DC, while leaf switches at the remote location (remote leaf), and they logically connect to spines in the main DC over an IP network.

Multi-Pod: A Multi-Pod design consists of a single APIC domain with multiple leaf-and-spine networks (pods) interconnected. As a consequence, a Multi-Pod design is functionally a fabric (an interconnection of availability zones), but it does not represent a single network failure domain, because each pod runs a separate instance of control-plane protocols. A Multi-Pod fabric interconnects different availability zones. Note that any or all these pods can also have remote leafs as part of their inventory.

Multi-Site: A Multi-Site design is the architecture interconnecting multiple APIC cluster domains with their associated pods. A Multi-Site design could also be called a Multi-Fabric design, because it interconnects separate regions (fabrics) each deployed as either a single pod or multiple pods (a MultiPod design).

2.3.2 Considerations

1. The distributed nature of telco Edge Data Centers meant that the network and application connectivity policies required for edge computing are localized. Network functions and applications deployed in these data centers serve traffic from nearby users. As a result, there is no requirement to define network and application policies that interconnect the pods within the fabric or across fabrics. All ACI network and application constructs that are defined remain locally significant to their respective pods.
2. ACI Multi-Pod and ACI Remote Leaf solutions were both considered. While the ACI Remote Leaf solution offers several benefits including a reduced physical footprint and reduced cost (due to lack of spine switches), a few items must be taken into consideration for service provider environments:
    i. Every ACI Remote Leaf must be connected to an Inter-Pod Network (IPN) Router. When more than two switches are required, the number of ports utilized on the PE routers is higher when using a Remote Leaf solution vs. a Multi-Pod based solution. In service provider environments, PE routers typically have a higher per-port cost. The Multi-Pod based solution saves Dish Wireless from using up costly PE ports that may be better allocated for other 5G infrastructure.

    ii. To support and simplify connectivity requirements for Network Functions (NF) that need route peering with the ACI leaf switches, the Floating L3Out feature was selected (covered in more detail later in the document). This feature saves users from having to configure multiple L3Out logical interfaces to maintain routing for NFs that can be provisioned in any available rack. At the time of the design and deployment at Dish Wireless, the Floating L3out feature was not currently supported on ACI Remote Leaf.

2.3.3 The Pick

In the case of Dish Wireless, considering their requirements and the fact that no objects would need to be stretched between Edge Data Centers, the ACI Multi-Pod fabric was selected as the ideal choice to support the deployment and centralized management of distributed pods.

ACI Multi-Pod simplifies operations by enabling a single APIC cluster to manage a set of Pods and create a unified policy domain across all the Pods within a fabric. This ensures consistent end-to-end policy enforcement and isolated failure domains by running separate instances of the fabric control plane protocols (IS-IS, COOP, MP-BGP) within each Pod.

With the ACI Multi-Pod fabric, each Edge Data Center would contain ACI Spines and Leafs interconnected with each other and their respective APICs via an InterPod Network (IPN). This architecture is flexible, allowing for the deployment of Remote Leaf switches in the Edge Data Centers should the requirements evolve.

Each ACI pod is deployed in a two-rack system as the initial baseline, as shown in Figure 2. The first two racks have two Cisco Nexus® 9000-based ACI Leaf switches also referred to as Top of Rack (TOR) switches, and their primary role is to provide connectivity to compute and network appliances. Additional racks with two Compute Leafs can be added to scale the footprint.

For connectivity to the Transport domain, a single Leaf switch in each rack is a designated as a Border Leaf. Each Border Leaf connects to the two Provider Edge (PE) routers which are deployed across the first two racks.

Each of the first two racks also contains a single Nexus 9000-based ACI Spine switch. Notice that two Cisco Application Policy Infrastructure Controllers(APIC) are also shown in Figure 2. More is discussed on the APICs and Spines in the next section.

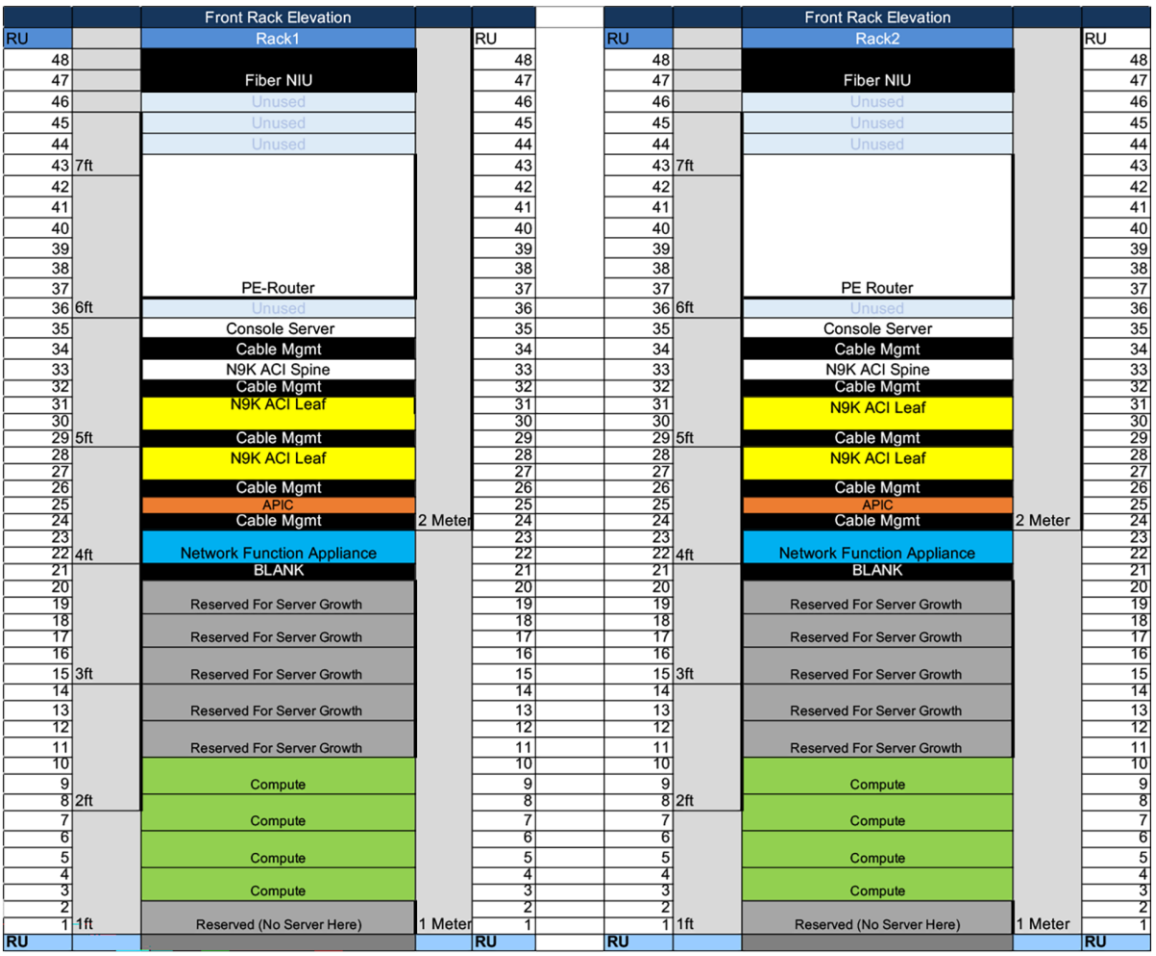

Figure 2.
Rack Elevation for 2-Rack Edge Data Center

3 Nationwide ACI Multi-Pod Fabric Architecture

Edge Data Centers in Dish Wireless' 5G infrastructure are geographically dispersed and responsible for aggregating cell sites, third-party circuits, and providing AWS Direct Connect connectivity. The Edge Data Centers also host critical network functions, including DHCP, DNS, Firewalls, and 5G User Plane Functions.

Based on scale requirements at the time of deployment, a 4-node APIC cluster (3 Active + 1 Standby) was chosen that supports up to 12 ACI Pods per fabric and an aggregate of 80 Leaf switches. The 4-node (3 Active + 1 Standby) APIC cluster provides N+1 redundancy in the case of an APIC failure and gives users the flexibility to scale to 200 ACI Leaf nodes per Multi-Pod fabric by converting the cluster configuration to four Active APIC nodes.

To stay within the verified scale limits of the ACI Multi-pod solution, particularly the maximum number of supported Pods and the 50-millisecond Round-Trip-Time (RTT) latency requirement, a total of five ACI Fabrics were implemented, as illustrated in Figure 3.

The Edge Data Centers assigned to an ACI Fabric are selected based on their relative proximity to each other to meet the latency requirements. Each ACI Fabric corresponds to a geographical region, and all the Edge Data Centers in that region are part of the same ACI Pod.

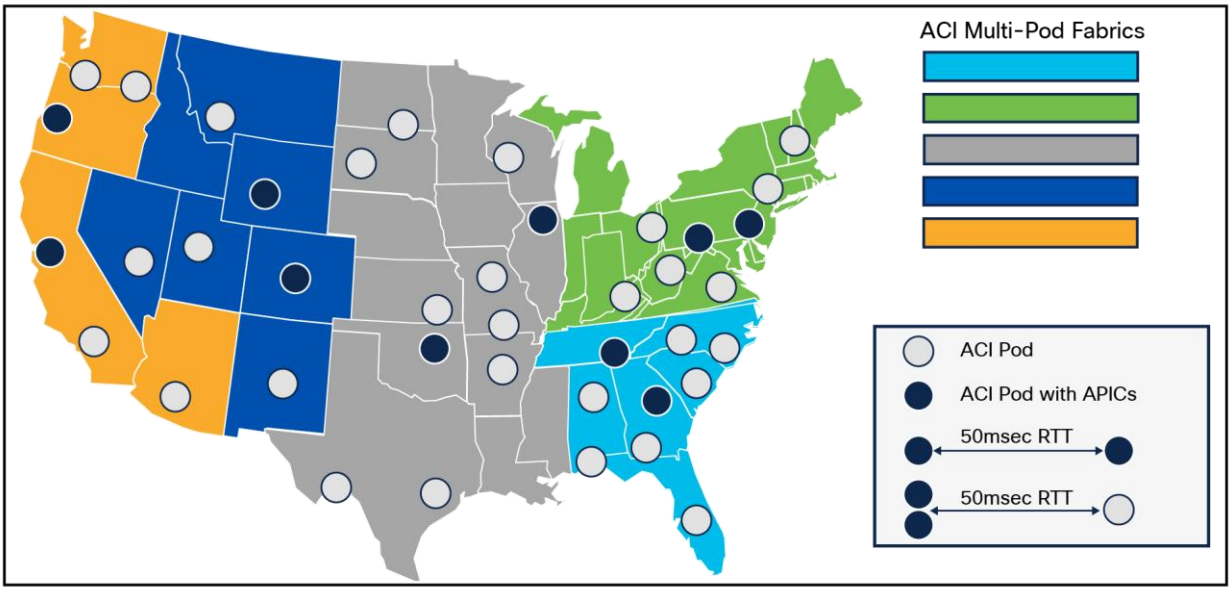

Figure 3.
Nationwide Layout of ACI Multi-Pod Fabrics

In each ACI fabric, two Edge Data Center locations are designated as Primary and Secondary sites where APICs are deployed. The latency between the APIC pods in a region is required to be less than 50 msec Round Trip Time (RTT), and the latency from the other ACI pods that is, non-APIC pods) to both the Primary and Secondary APIC pods is also required to be less than 50 msec RTT.

To provide geo-redundancy and mitigate the risk of an APIC cluster failure in the event of a complete Edge Data Center outage, two Active APICs are deployed in the first site, and one Active plus one Standby APIC are deployed in the secondary site. The rack elevation shown in Figure 2, which includes the APICs, applies only to the Primary and Secondary APIC pods and not for the non-APIC pods.

3.1 Control Plane for ACI Multi-pod Fabrics

In the ACI Multi-pod architecture, to establish the control plane, the member ACI pods are interconnected through an "Inter-Pod Network" (IPN). The IPN is exclusively used Multi-Pod provisioning, Control-Plane, and management. Each ACI Pod connects to the IPN through the spine switches, which are directly connected to the Pod local PE routers, as show in Figure 2.

The Transport backbone, which provides connectivity to all the ACI pods, functions as the Inter-Pod Network (IPN) and provides Layer 3 connectivity services. A common dedicated VRF in the Transport network is assigned for the Inter Pod Network (IPN) across all the ACI Fabrics, as shown in Figure 4. The ACI Spine switches can connect to the Transport network using a dynamic routing protocol, either Open Shortest Path First (OSPF) or BGP.

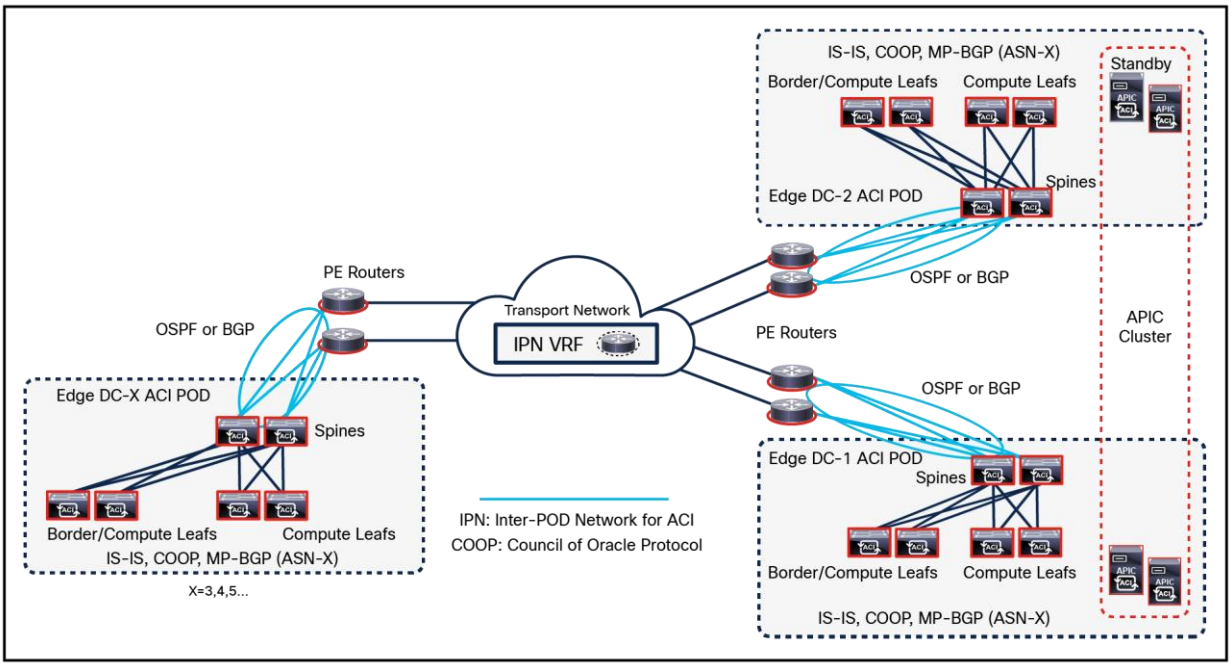

Figure 4.
ACI Multi-POD Fabric Control Plane Integration

The Spines for all Pods within a region peer with each other to establish MP-BGP Ethernet VPN and L3VPN sessions for exchanging endpoint reachability and external routing information. In large MultiPod fabrics with three Pods or more, three External Route Reflector (RR) nodes can be distributed across the pods to optimize the number of BGP sessions between the Spines of the respective ACI pods in the fabric. However, in the case of Dish Wireless, where each Edge Data Center ACI Pod contains two Spines, the reduction in the number of BGP adjacencies was negligible. As a result, the default behavior was used, which involves setting up a full mesh of MP-iBGP sessions between the Spines of each ACI pod within the fabric.

From an ACI policy plane perspective, networking and application policies are localized to each individual ACI Pod. All application traffic, whether inbound or outbound, is expected to utilize the connectivity provided by the directly attached Border Leaf and PE Routers. As such, there is no requirement to stretch any constructs such as Bridge Domains (BD), End Point Groups (EPG), and VRFs across Pods. The IPN is primarily used for Multi-Pod provisioning, Control-Plane traffic, and Pod management and does not carry any spine-to-spine or leaf-to-leaf VXLAN tunnels across the pods. Given the simplified nature of the requirement for inter-pod communication, the full mesh design of BGP sessions is acceptable, with the option to migrate to the route reflector design if necessary.

3.2 SR-MPLS Integration of ACI pods with Transport

Dish Wireless, like many Telco providers, utilizes Segment Routing (SR) and Multiprotocol Label Switching (MPLS) L3VPNs in their transport network. To achieve a unified data plane across the network, they enabled SR/MPLS-based handoff from the ACI Border Leafs to the local PE routers in each Edge Data Center.

Before diving into the SR-MPLS data plane integration, it's important to note the following regarding the overall BGP AS numbering, as shown in Figure 5: Each ACI Multi-Pod fabric is assigned a unique BGP Autonomous System Number (ASN), and all geographically dispersed ACI Pods within an ACI fabric share the same ASN. The shared transport network connected to each Pod/Edge Data Center also has its own unique BGP ASN.

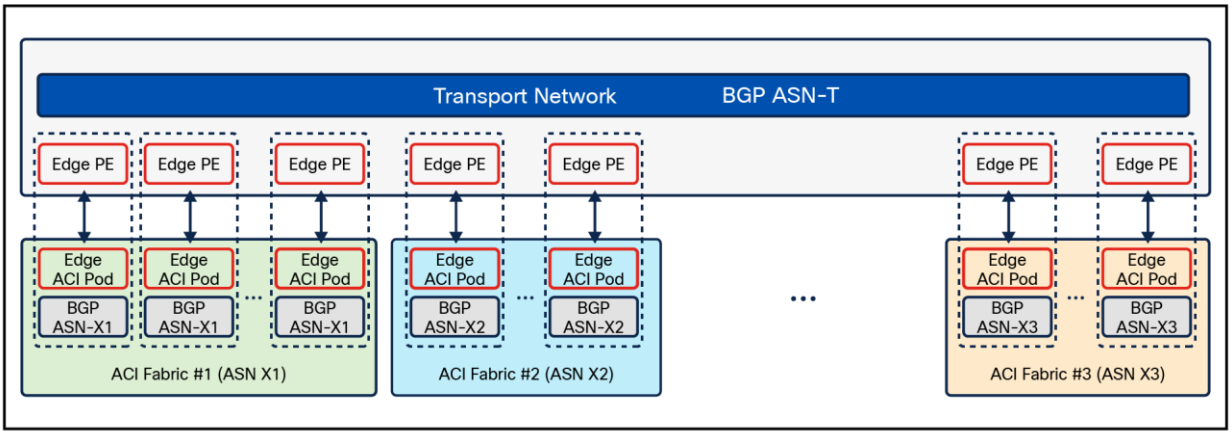

Figure 5.
BGP AS Numbering for ACI Fabrics

As stated earlier, the requirement is for all application traffic entering and exiting a Pod to utilize the links between the Border Leaf switches and the local PE routers that attach to the transport network. Traditionally, the handoff from an ACI Border Leaf to a PE router in a provider network would be achieved using native IP and back-to-back VRF-lite. In an SP environment, this quickly becomes cumbersome as it requires a separate subinterface, additional IP addressing and a routing protocol session to be established for every VRF that needed external connectivity.

With SR-MPLS handoff from ACI, a single BGP EVPN session is used to exchange the information for all prefixes in all the VRF instances, resulting in better scale and automation for both transport and data center environments.

Figure 6 depicts a high-level diagram of the SR/MPLS handoff between multiple ACI Fabrics and the Transport network. The Cisco ACI to SR/MPLS handoff solution uses a standards-based implementation consisting of SR/MPLS, BGP-LU, BGP EVPN, and prefix re-origination between BGP EVPN and VPNv4/v6. The core component of the SR/MPLS handoff solution is the SR-MPLS Infra L3out, which is configured in the infra tenant (where the MPLS forwarding plane sits in the Overlay-1 VRF) and gets deployed on the Border Leaf switches. To achieve faster convergence, users can configure single-hop Bidirectional Forwarding Detection (BFD) for the BGP-LU session and multi-hop BFD for the BGP EVPN session. Please note that the corresponding configuration on the Transport PE routers is applied in the Default or Global VRF.

Each ACI pod contains its own unique SR-MPLS Infra L3Out. Tenant VRFs are attached to the ACI InfraL3out(s) via a SR/MPLS VRF L3Out in order to advertise Tenant/VRF prefixes to the PE routers and importing MPLS VPN prefixes from the PE routers. In Dish Wireless' deployment, the requirement was to connect to a single transport domain and thus only one Infra L3Out per pod was provisioned. The L3out uses fullmesh connectivity between the ACI border leaf switches and the Transport PE Routers to provide redundancy, fast re-convergence as well as maximum bandwidth to/from the Edge Data Center.

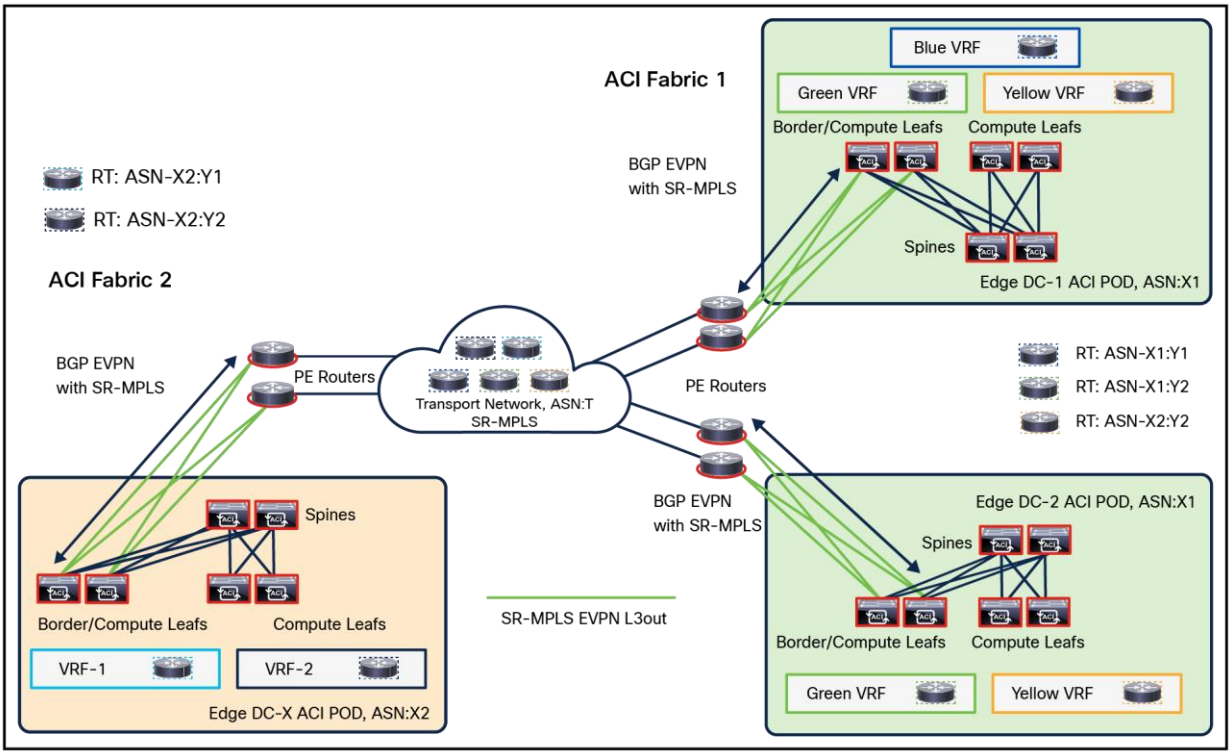

Figure 6.
Data Plane Integration of ACI Pods with Transport

Figure 7 shows a simplistic model in which VRFs from User-defined Tenants leverage SR/MPLS handoff. In the context of a 5G deployment, the VRFs listed below are typically deployed in User Tenants and have corresponding SR/MPLS VRF L3outs defined.

1. VRF-RAN-F1C [Open RAN Control Plane traffic in Midhaul]
2. VRF-RAN-F1U [Open RAN User Plane traffic in Midhaul]
3. VRF-RAN-OAM [Open RAN Management Plane traffic]
4. VRF-5G-N3U [User Plane traffic in the Packet Core Backhaul]
5. VRF-5G-N6U [User Plane traffic going toward Data Network]
6. VRF-5G-SIGNALING [Packet Core Signaling Traffic]
7. VRF-B2B-PARTNER-DATA [B2B Partner Data Integration]
8. VRF-B2B-PARTNER-SIGNALING [B2B Partner Signaling Traffic Integration]

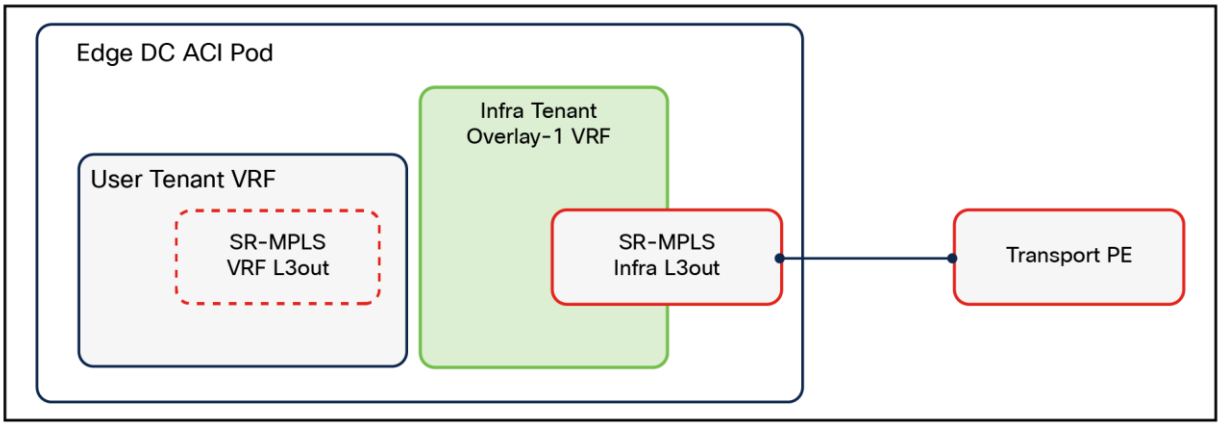

Figure 7.
Mapping of Tenant VRFs to SR-MPLS Handoff

3.3 Considerations for SR-MPLS Handoff

3.3.1 Route Targets

SR/MPLS handoff requires configuring a unique route target per VRF in Cisco ACI and matching route targets on the PE router to ensure that prefixes are accepted on both devices.

In Dish Wireless' environment, ACI Pods across multiple fabrics utilize the same VRFs in shared transport. To meet the requirement of having unique VRF route targets per fabric, Pod-local route targets (as shown earlier in Figure 6) were configured within each ACI VRF that required SR/MPLS connectivity.

These Pod-local VRF route targets were then associated (on the PE router) with the global route targets used in shared transport for each individual VRF. It's important to note that the unique route target per VRF is a requirement within an ACI fabric. Route targets can be reused between ACI fabrics.

3.3.2 Segment-Routing Global Block

The Segment-Routing Global Block (SRGB) is the range of label values reserved for Segment Routing (SR). These values are assigned as Segment IDs (SIDs) to SR-enabled nodes and have global significance throughout a SR domain. SRGB is enabled by default in ACI is set to a range of 16000 to 23999. The SR Global Block in ACI needs to be adjusted to match the PE router and transport SR domain.

3.3.3 Leaf Forwarding Scale Profiles

In a Service Provider environment, a large number of routes/prefixes can be expected to be received on the ACI Border Leaf switches connected to the Transport PE routers. Even with proper summarization, this can eventually lead to Ternary Content Addressable Memory (TCAM) exhaustion on the ACI Leaf switches if not accounted for. On supported platforms, ACI allows the configuration of Forwarding Scale Profiles that allocate TCAM resources to better suit the deployment scenario. In Dish Wireless' deployment, where the number of endpoints is low and the number of prefixes is high, the High LPM Forwarding Scale Profile was deployed on the ACI Leaf switches, allowing for a larger number of prefixes, as shown in Figure 8.

Dual Stack (Default)	High LPM
EP MAC: 24,000	EP MAC: 24,000
EP IPv4: 24,000	EP IPv4: 24,000
EP IPv6: 12,000	EP IPv6: 12,000
LPM: 20,000	LPM: 128,000
Policy: 64,000	Policy: 8,000
Multicast: 8,000	Multicast: 8,000

Figure 8.
Forwarding Scale Profiles for ACI Leaf Switches

Note: The figure above shows the forwarding scale profile selected by the customer for their deployment. For the latest information regarding forwarding scale profiles, please refer to the forwarding scale profile information for the specific Cisco APIC release.

3.3.4 BGP EVPN Max Prefixes

In the SR-MPLS handoff configuration, a single BGP EVPN session in the ACI Infra tenant is used to exchange prefixes for all associated VRF instances. This can lead to large number of prefixes being learned on the SR/MPLS Infra L3out. The default BGP setting imposes a limit of 20K prefixes, beyond which new prefixes are rejected. When the Reject option is deployed, BGP accepts one additional prefix beyond the configured limit, and the APIC raises a fault. If it's anticipated that more than 20K prefixes will be present across all VRF instances, this limit, along with the associated action and fault threshold, can be adjusted by applying a BGP Peer Prefix Policy to the SR-MPLS Infra L3Out.

3.3.5 In-Band Management of ACI Pods

In an ACI Multi-Pod fabric, the mgmt tenant and associated mgmt:inb VRF and bridge domain are stretched across all Pods by default. The ACI network policy definition allows a single SR-MPLS VRF L3Out to be associated at the VRF level, which then propagated to all pods within the mgmt:inb VRF. However, this can result in suboptimal routing, as all subnets used for inband management under the bridge domain are advertised to the transport network from all ACI pods.

Ideally, management traffic from each pod should ideally take the best path via the local PE and transport network. To achieve this, when creating an SR/MPLS VRF L3Out in the mgmt tenant, the Infra L3Out for each pod in the fabric should be added as an additional path. For proper traffic flow, each Infra L3Out added will also need to have a unique outbound route-map matching its site specific in-band management subnet while the inbound route-map matching all prefixes is shared between all SR-MPLS infra L3Out paths as shown in Figure 9.

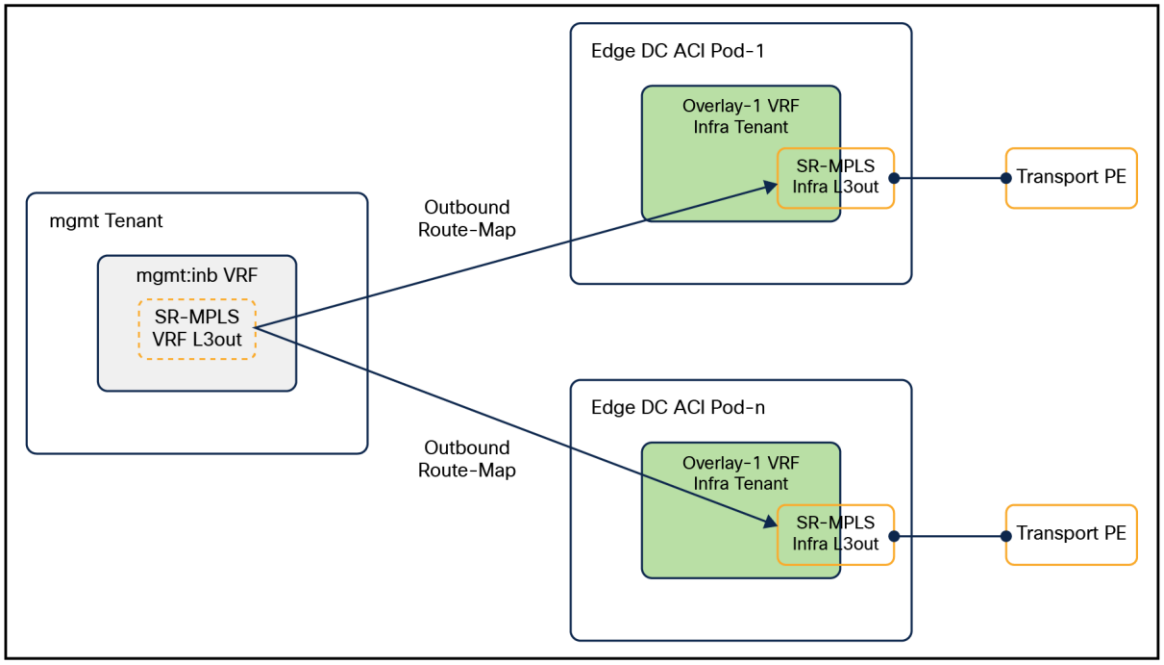

Figure 9.
Inband Mgmt VRF SR-MPLS L3out Policy Definition

4 Connecting Network Functions in ACI to Transport

Most network functions that are deployed in Edge Data Centers require Layer 3 attachment to ACI Leaf switches and end-to-end IP connectivity to external destinations. The following section delves into how these requirements are addressed.

4.1 Routing Adjacency with Network Functions

To facilitate the flexible deployment of Network Functions in any rack, Cisco ACI's floating L3Out feature can be leveraged to establish routing adjacencies in both virtual and physical domains. The floating L3Out feature enables users to configure an L3Out without specifying any logical interfaces on the leaf. This simplifies configuration and ensures continuous routing adjacency when virtual machines move between hosts attached to the same and/or different leaf nodes. A floating L3Out consists of two main components:

- **Anchor Leaf Nodes:** ACI Leaf switches that establish Layer 3 adjacencies with the network functions are referred to as Anchor Nodes.
- **Non-anchor Leaf Nodes:** The non-anchor leaf nodes do not create any adjacency with the external routers. They act as a "pass-through" for traffic flowing between directly connected network functions and the anchor node.

Figure 10 shows the Layer 3 integration of a pair of Active/Standby virtual firewalls with the ACI pod using a floating L3out. In line with a high-availability design, the active and standby virtual firewalls are deployed in separate racks. The odd numbered leaf switches in each rack serve as the anchor nodes, while the remaining switches in the ACI pod are non-anchor leaf nodes.

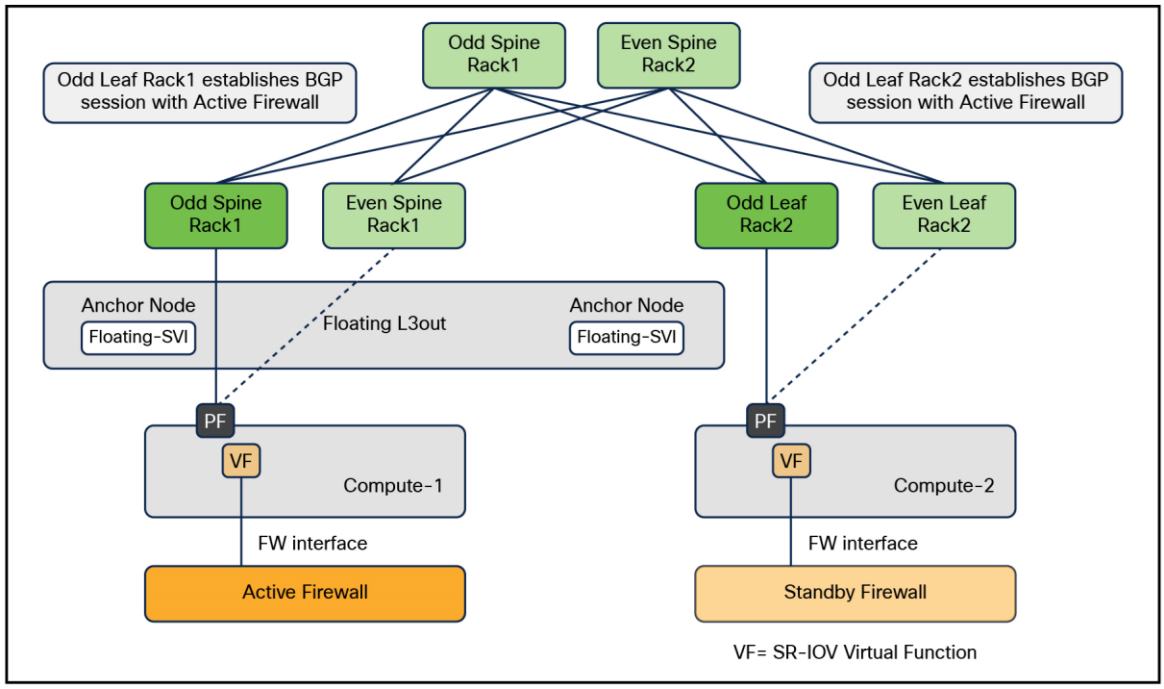

Figure 10.
Floating L3out with Active Standby Virtual Firewall Pair

Both anchor leaf nodes establish BGP sessions with the active firewall in the pair to exchange routes. In the event of a firewall failover from active to standby, the standby firewall assumes the active IP addresses and establishes BGP sessions with the ACI anchor leaf nodes.

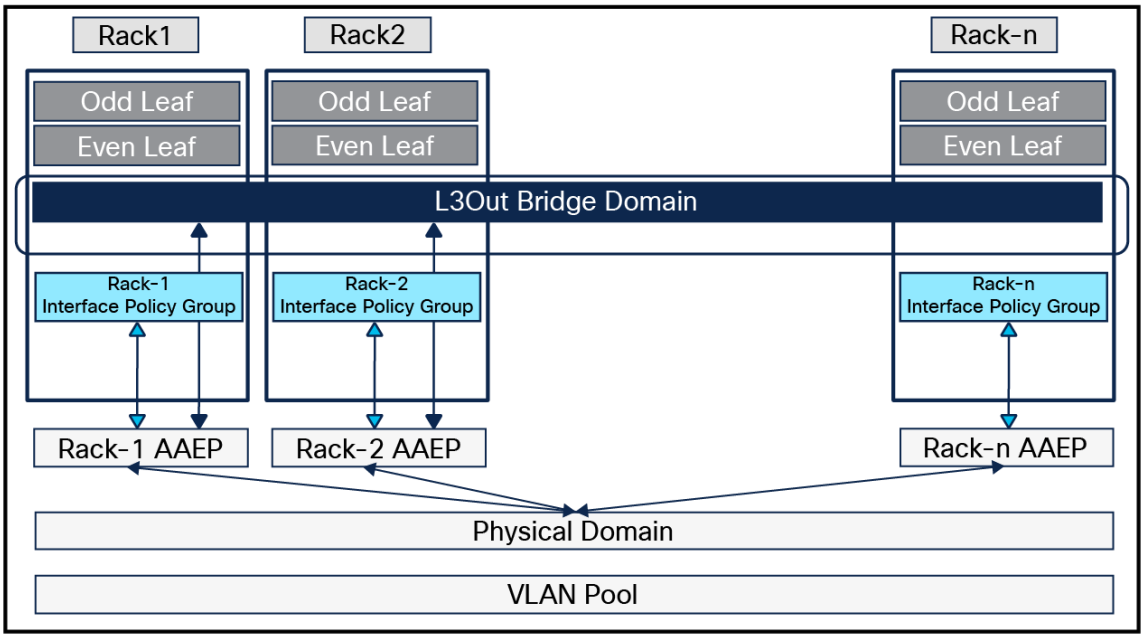

Figure 11.
Layer 2 connectivity inside the ACI Pod for Floating L3out

In the figure above, the creation of a floating L3out deploys a L3Out bridge domain on the anchor and nonanchor leaf switches. Unique VLAN encapsulations for Floating L3out are used in the ACI pod as shown in Figure 11. Irrespective of location of the Network Function, it still can maintain the routing adjacencies with the SVI interfaces deployed on the anchor leaf nodes because of the extension of the L3Out bridge domain across anchor and non-anchor leaf nodes.

Figure 12 shows the Layer 3 integration of a Composite Cloud Native Function (CNF) with the ACI pod. In this case the floating L3out is configured with BGP and static routing protocols with Bidirectional Forwarding Detection (BFD) support. The control plane CNF peers with the Anchor Leafs with BGP and advertises the Loopback of the data plane CNF. The CNFs can be deployed in different racks to achieve high availability or horizontal scaling.

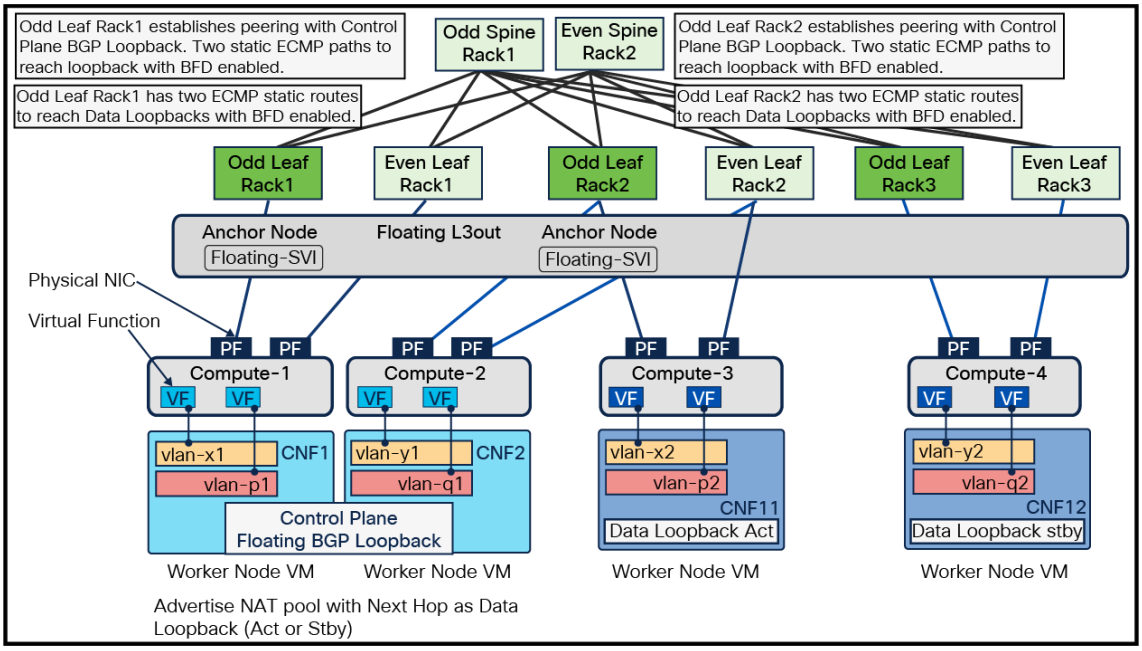

Figure 12.
Floating L3out with Composite Cloud Native Function (CNF)

4.2 Applying Transit Routing in ACI

As covered in the earlier sections, Network Functions (NFs) establish routing adjacencies with the ACI Leaf switches through dynamic routing protocols, and they need to communicate with external destinations outside the ACI Pod. To achieve this, the ACI pod acts as a transit domain, as shown in Figure 13. With the application of Transit Routing, ACI Leafs perform bidirectional redistribution, enabling the exchange of routing information between different routing domains and ensuring IP connectivity. The ACI pod advertises the routes that are learned from one L3Out connection to another L3Out connection on a per-VRF basis.

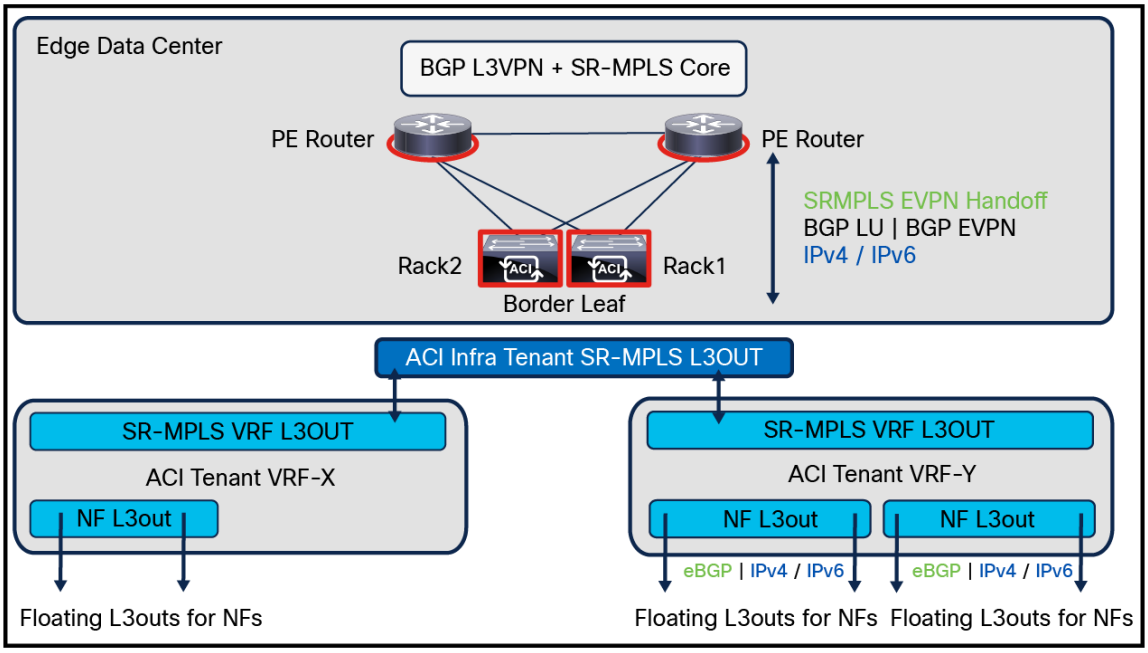

Figure 13.
Application of Transit Routing

Figure 14 shows how a Floating L3Out, along with transit routing, firewalls connectivity toward a B2B Partner. The firewall enables communication between the insecure zone (Partner VRF) and secure zone (Internal VRF). ACI performs two layers of transit routing between the transport domain and the respective zones of the firewall.

Traffic from the B2B Partner enters the network through a partner circuit directly attached to the transport domain in the Partner VRF. This traffic, which is destined for services in the Internal VRF, takes the path through the transport towards the Edge Data Center via the SR/MPLS L3Out to reach the B2B firewall.

Once the traffic enters the ACI Partner VRF using the SR/MPLS VRF L3Out (associated with the Infra L3Out), the traffic is sent through a firewall via an L3Out connected to the firewall's untrusted interface and back into the fabric via a second L3Out attached to the Internal VRF.

Once in the Internal VRF, the B2B Partner traffic can access services hosted in the ACI Internal VRF or use the SR/MPLS VRF L3Out to exit the ACI Pod and access services in the Transport domain Internal VRF.

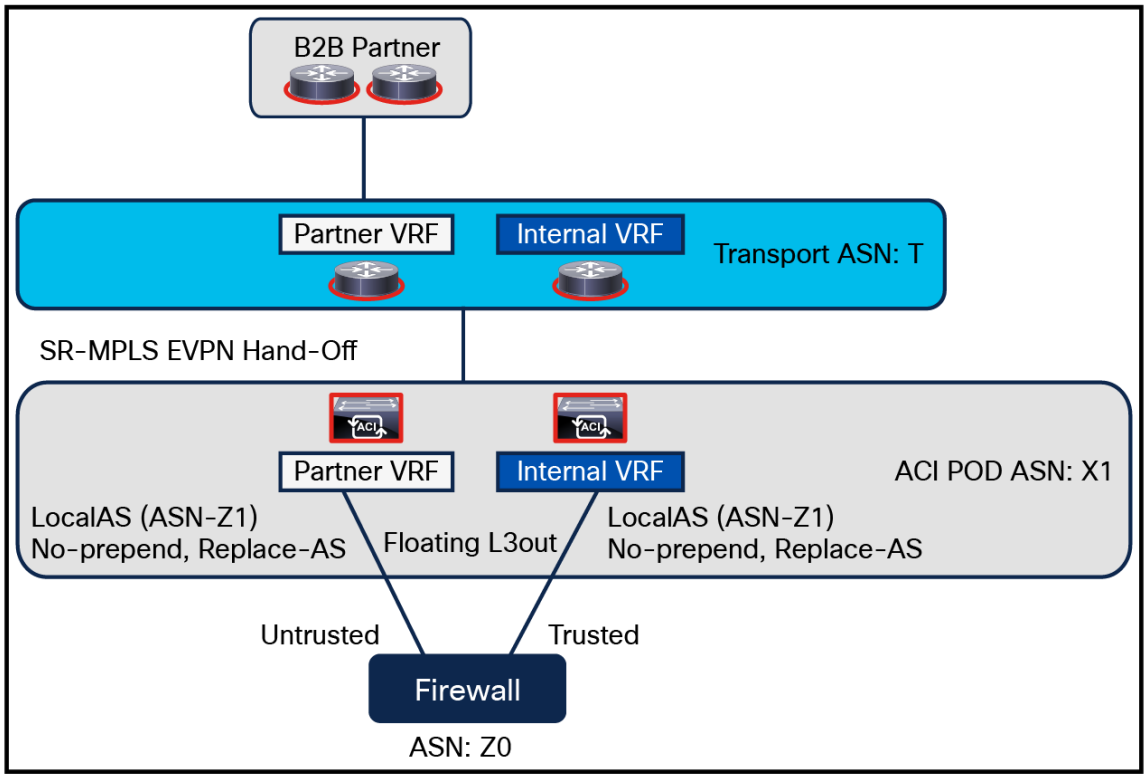

Figure 14.
Transit Routing scenario – Firewall Insertion for B2B Partner connectivity

Another requirement that needs to be addressed is for DHCP and DNS services to be offered in the default or Global VRF of the transport domain. For example, when performing Zero Touch Provisioning, the cell site routers need DHCP services. The DHCP/DNS appliances are deployed in pairs in the ACI pod in a given user defined VRF and use Anycast IPs to provide high availability.

The SR-MPLS handoff provides connectivity to non-default VRFs in the transport domain. To handle Layer 3 connectivity with the default VRF, an additional IP handoff is deployed between the user-defined VRF in the ACI pod and the default VRF in the Transport domain, as shown in Figure 15.

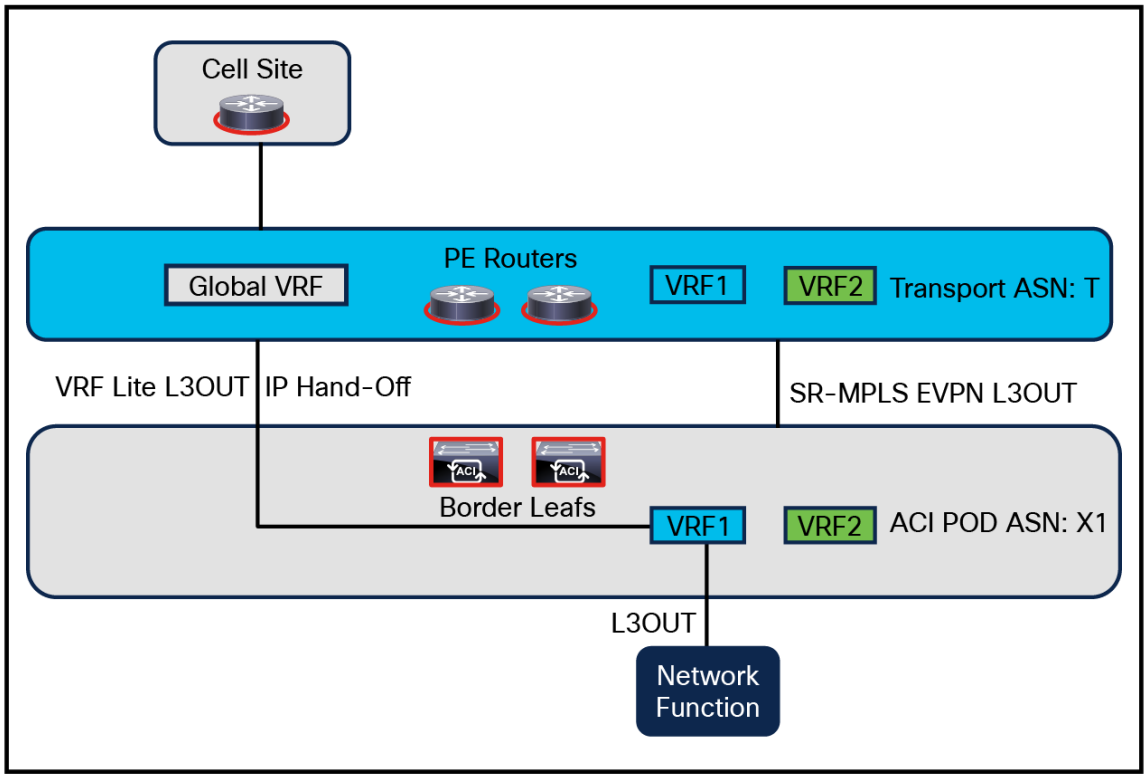

Figure 15.
Co-existence of VRF-Lite and SR-MPLS handoff in same ACI VRF

When DHCP and DNS services are required in another user-defined Tenant/VRF within the ACI pod, route-leaking can be accomplished by exporting/importing ACI Global contracts between the two tenants.

5 Infrastructure as Code for ACI Fabrics

In a distributed environment, automating the provisioning and management of network infrastructure through code is essential. There are several tools and integrations suitable to achieve this for the ACI infrastructure deployed in the distributed Edge Data Centers.

Starting with the APIC controller, it serves as a centralized point for programmability, automation, and management for each of the ACI Fabrics. The APIC offers several capabilities, including inventory management, image management, and the implementation of application-centric network policies.

The ACI fabric stores configuration and state information in a hierarchical structure called the Management Information Tree (MIT), which is accessible through an API. The APIC provides a REST API that accepts and returns HTTPS messages containing JavaScript Object Notation (JSON) or Extensible Markup Language (XML) payloads, as demonstrated in Figure 16 and Figure 17.

```
<fvCtx bdEnforcedEnable="no"  dn="uni/tn-SLICE0/ctx-SLICE0-EDG01-VRF01"
       ipDataPlaneLearning="enabled" knwMcastAct="permit" name="SLICE0-EDG01-VRF01"
       pcEnfDir="ingress" pcEnfPref="enforced" >

    <vzAny >
        <vzRsAnyToProv tnVzBrCPName="SLICE0-EDG01-VRF01-CONTRACT" />
        <vzRsAnyToCons tnVzBrCPName="SLICE0-EDG01-VRF01-CONTRACT" />
    </vzAny>

    <bgpRtTargetP af="ipv4-ucast" >
        <bgpRtTarget rt="route-target:as4-nn2:389321:200" type="import"/>
        <bgpRtTarget rt="route-target:as4-nn2:389321:200" type="export"/>
    </bgpRtTargetP>
    <bgpRtTargetP af="ipv6-ucast" >
        <bgpRtTarget rt="route-target:as4-nn2:389321:200" type="export"/>
        <bgpRtTarget rt="route-target:as4-nn2:389321:200" type="import"/>
    </bgpRtTargetP>

</fvCtx>
```

Figure 16.
XML Payload for VRF Definition

```
<l3extOut dn="uni/tn-SLICE0/out-SLICE0-EDG01-VRF01-SR-MPLS-L3OUT"
          enforceRtctrl="export" mplsEnabled="yes" name="SLICE0-EDG01-VRF01-SR-MPLS-L3OUT">

    <l3extRsEctx annotation="" tnFvCtxName="SLICE0-EDG01-VRF01" userdom="all"/>

    <l3extInstP floodOnEncap="disabled" matchT="AtleastOne"
                name="SLICE0-EDG01-VRF01-L3OUT-EPG" pcEnfPref="unenforced" prefGrMemb="exclude">
        <l3extSubnet aggregate="" ip="::/0" scope="import-security"/>
        <l3extSubnet aggregate="" ip="0.0.0.0/0"  scope="import-security"/>
    </l3extInstP>

    <l3extConsLbl name="EDG01-SR-MPLS-PE-L3OUT">
        <l3extRsLblToProfile direction="export"
                             tDn="uni/tn-SLICE0/prof-SLICE0-EDG01-VRF01-L3OUT-EXPORT-ROUTE-MAP"/>
        <l3extRsLblToProfile annotation="" direction="import"
                             tDn="uni/tn-SLICE0/prof-SLICE0-EDG01-VRF01-L3OUT-IMPORT-ROUTE-MAP"/>
        <l3extRsLblToInstP tDn="uni/tn-SLICE0/out-SLICE0-EDG01-VRF01-SR-MPLS-L3OUT/instP-SLICE0-EDG01-VRF01-L3OUT-EPG"/>
    </l3extConsLbl>

</l3extOut>
```

Figure 17.
XML Payload for SR-MPLS VRF L3OUT definition

Using API calls, a higher-level orchestration system can provision and implement changes to the network and application policies on the APIC controller. This following section discusses a few of those options.

5.1 Network Services Orchestrator

Cisco Network Services Orchestrator (NSO) offers a Layered Service Architecture (LSA) with a Customer-Facing Services (CFS) layer that interfaces with OSS/BSS systems and an Resource-Facing Services (RFS) layer responsible for interactions with networking infrastructure. Within the RFS Layer, domain-specific nodes equipped with Network Element Drivers (NED) are available, enabling NSO to serve as a cross-domain orchestrator, capable of managing both Transport and Data Center domains.

Utilizing the Network Element Driver (NED) for APIC and creating automation packages tailored to specific requirements, NSO offers the capability to provision and manage an ACI network and application policies. Figure 18 illustrates the high-level architecture, demonstrating how NSO can orchestrate configuration tasks across the ACI Fabrics that support the distributed Edge Data Center deployment at Dish Wireless.

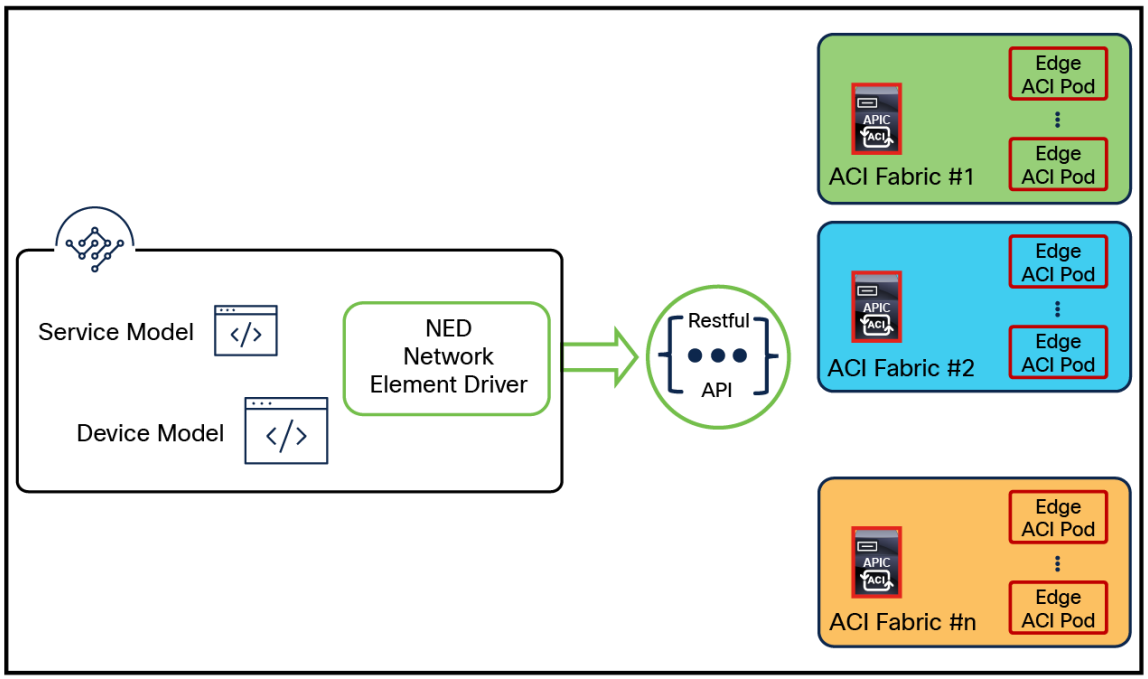

Figure 18.
ACI Network and Application Policy Orchestration with NSO

5.1.1 NSO Core Function Packs

A Core Function Pack (CFP) is an NSO service package designed for specific use cases or sets of use cases. This package can include a bundle of components, such as device drivers, templates, and other CFPs. It's important to note that a CFP is officially released software fully supported by Cisco Technical Assistance Center (TAC). One example of a CFP is the NSO Data Center Software-Defined Core Function Pack (DC-SDN CFP), which offers several capabilities, including:

- Multi-domain orchestration across Data Center and Transport domains
- Provisioning and management of multiple ACI fabrics
- Service Chaining with Policy-Based Routing
- IP and SR/MPLS handoff provisioning

For example, the NSO CFP can be utilized to orchestrate the BGP EVPN SR-MPLS connectivity between the ACI Pods in the Edge Locations and their corresponding PE Routers in the Transport Domain, automating the following tasks:

- Layer 2 and Layer 3 config for underlay BGP-Labeled Unicast
- EVPN Loopback, Transport loopback
- Route Distinguisher, Route Targets, Segment-ID, and Router-id
- BGP EVPN and labeled unicast session
- Single and Multi-hop BFD
- QOS

Additionally, the CFP automates the following on the Cisco PE Routers.

- RT Translation from EVPN to L3VPN
- Mapping of BGP color-community, prefixes, DSCP/EXP to SR policies

In addition to the pre-built NSO Core Function Packages, custom automation packages can be developed to address specific use cases.

5.2 Ansible and/or Terraform

Similar to NSO, the configuration and management of the distributed ACI pods can also be managed by other Infrastructure as Code (IaC) tools such as RedHat Ansible or HashiCorp Terraform, as shown in Figure 19. These IaC tools use Domain-Specific Languages (DSLs) to interact with the APIC controllers and orchestrate configuration across all ACI Fabrics.

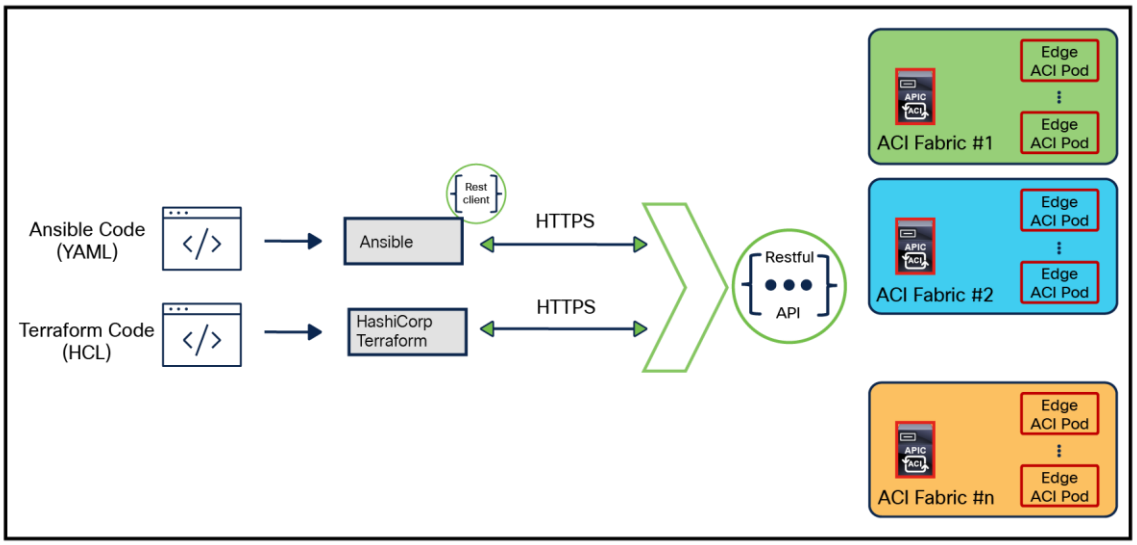

Figure 19.
ACI Network and Application Policy Orchestration with Ansible/Terraform

5.3 Cisco Nexus Dashboard Orchestrator

Let's briefly explore another platform, Cisco Nexus Dashboard Orchestrator (NDO), and how it addresses the configuration and management of distributed ACI pods.

NDO offers consistent network and policy orchestration, providing scalability across multiple data centers through a single pane of glass. Its core function is to interconnect separate Cisco ACI sites, Cisco Cloud ACI sites, and Cisco Nexus Dashboard Fabric Controller (NDFC) sites.

In Dish Wireless' case, there is a need to independently manage each of the ACI Fabrics, which is fully supported by NDO. When creating multi-site application templates, operators have the option to designate the template as "Autonomous." This enables users to associate a template with one or more independently operated sites. Additionally, NDO provides APIs for northbound integration with IaC tools like Ansible, Terraform, and NSO.

The scalability of Nexus Dashboard for managing autonomous sites surpasses the requirements at Dish Wireless. As shown in Figure 20, NDO gives users the option to use a single pane of glass for provisioning ACI policies across all ACI Fabrics.

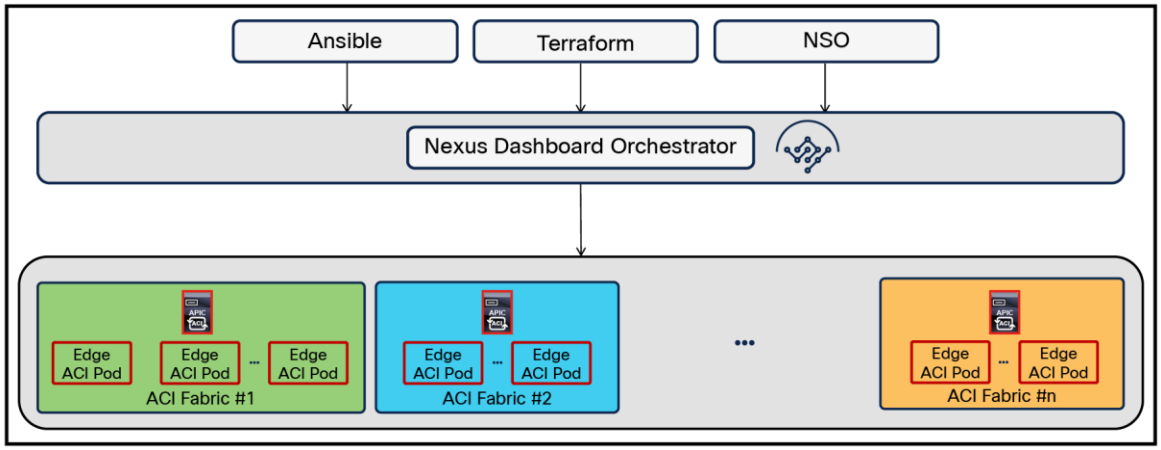

Figure 20.
ACI Network and Application Policy Orchestration with NDO

6 Observability for the ACI Fabrics

For each of the Multi-Pod ACI Fabrics, the respective APIC cluster maintains a representation of the administrative and operational state of the system in the form of a collection of Managed Objects (MOs). The system is comprised of the APIC Controllers and all the fabric switches.

The APIC continuously monitors the system's state and generates faults, events, and logs based on the real-time state of the system. The APIC also gathers statistics (on-demand or continuous) from a variety of sources, including interfaces, VLANs, EPGs, application profiles, ACL rules, tenants, or internal Cisco Application Policy Infrastructure Controller (APIC) processes. The statistics help the operations team with trend analysis and troubleshooting.

Built-in tools in the APIC, such as the Capacity Dashboard, Troubleshooting Wizard, Health Scores, Faults, Events, and Audit records, provide quick and easy ways to validate the health of the fabric at the SDN controller level with rich northbound APIs.

In the Dish Wireless deployment, the operations team can utilize multiple APIC controllers to monitor all ACI pods that support the distributed Edge Data Centers along with the Transport Domain Observability stack.

To unify the monitoring of the transport domain and Edge Data Centers in the same observability platform, the following steps are taken as depicted in Figure 21.

- Syslogs and SNMP Traps from all ACI switches and APIC Controllers are sent to the common Fault Management system.
- Invoke REST API of the APIC controllers from the Performance Monitoring system to fetch statistics of Key Performance Indicators (KPIs).

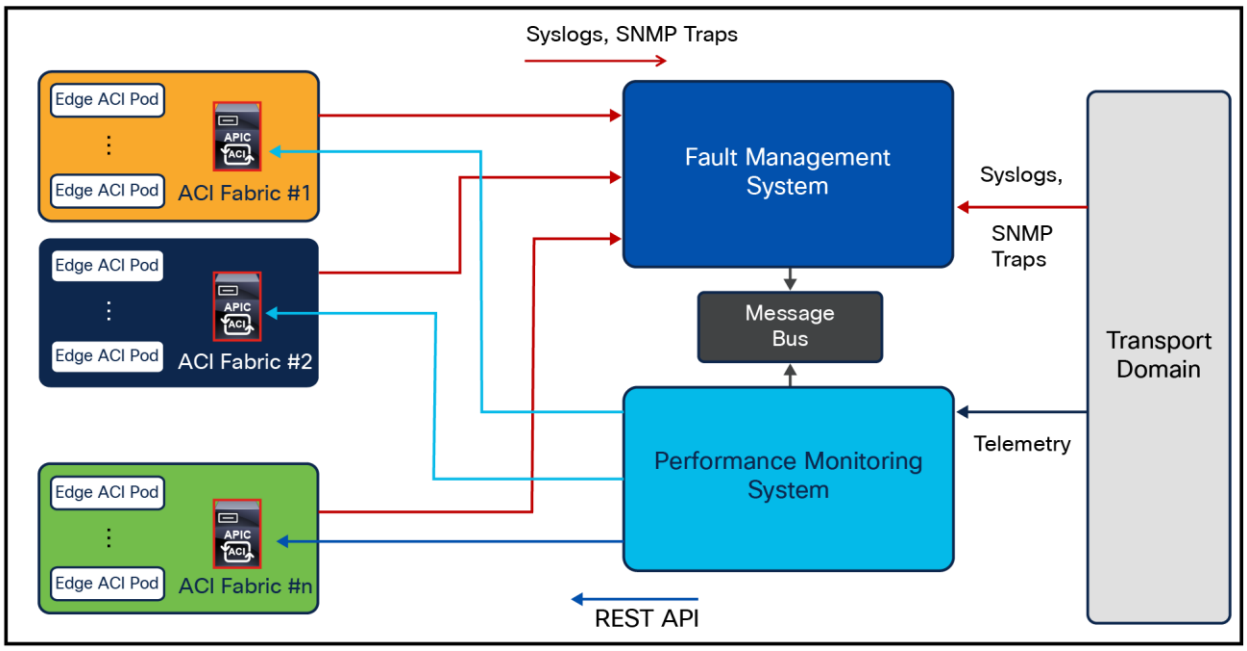

Figure 21.
Extending ACI Monitoring to Common Observability Platform

As a result, the common observability platform can now deliver key capabilities such as:

- The Performance Monitoring system can monitor the KPIs and generate anomaly alerts based on configured thresholds or machine learning.
- The Fault Management system can correlate signals from both the transport and data center domains to diagnose and root cause issues.
- Common point of Integration with ticketing system

7 Summary

Edge Data Centers are a key pillar and building block of Dish Wireless' nationwide 5G infrastructure, playing a vital role in achieving the goals of delivering low latency, high throughput, and security to 5G services. Cisco ACI Multi-Pod technology offers a standardized, repeatable, and resilient architecture for deploying networking infrastructure in these distributed edge data centers. It simplifies operations through centralized management, automation, and real-time network health monitoring.

A summary of key takeaways from this document:

A single APIC cluster manages multiple ACI pods deployed in edge locations, creating a unified policy domain across all pods. This approach offers operational simplification, as the failure domain is localized at the pod level through the isolation of fabric control protocols.

The architecture offers deployment flexibility and agility, allowing for the use of ACI Pods or ACI Remote Leafs in the Edge Data Centers based on specific use cases.

The SR-MPLS handoff capability in ACI allows for using a unified data plane across the transport network. This approach enables Edge Data Centers to connect to the transport network using a single BGP EVPN session, rather than requiring per-VRF configuration. This feature automates the process as the number of VRFs grows to support more 5G services.

The Floating L3OUT capability simplifies the deployment of any Network Function in any rack and facilitates the establishment of routing adjacency with the anchor nodes in the ACI Pod.

The Cisco ACI APIC cluster brings centralized management and automation to all the ACI pods that it manages. It seamlessly integrates with higher level IaC orchestrators such as NSO, Ansible, and Terraform. This integration enables the higher-level IAC orchestrator to provision and manage network and application policies across all ACI Fabrics from a single location. Alternatively, the Nexus Dashboard platform can also be utilized to provision policies across all ACI Fabrics and integrate with various IaC entities.

Authors:

Mehran Chowdhury, Customer Delivery Architect

Asifiqbal Pathan, Principal Architect

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Printed in USA C11-4040318-00 12/23