MP-BGP VxLAN, ACI & Demo

Brian Kvisgaard
System Engineer, CCIE SP #41039
November 2017
Datacenter solutions

Programmable Fabric

VxLAN-BGP EVPN standard-based
Cisco DCNM Automation

Classic Ethernet

Modern NX-OS with enhanced NX-APIs
Automation Ecosystem (Puppet, Chef, Ansible etc.)
Common NX-API across N2K-N9K
Data Center – Growth

ACI / MP-BGP EVPN (VXLAN)
## Why VXLAN Overlay?

<table>
<thead>
<tr>
<th>Design Needs</th>
<th>VXLAN Delivered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any workload anywhere – VLANs limited by L3 boundaries</td>
<td>Any Workload anywhere- across Layer 3 boundaries</td>
</tr>
<tr>
<td>VM Mobility</td>
<td>Seamless VM Mobility</td>
</tr>
<tr>
<td>Scale above 4k Segments (VLAN limitation)</td>
<td>Scale up to 16M segments</td>
</tr>
<tr>
<td>Secure Multi-tenancy</td>
<td>Traffic &amp; Address Isolation</td>
</tr>
</tbody>
</table>

![VXLAN Overlay Diagram]
Challenges with Traditional VXLAN Deployments

Scale and Mobility Limitations

**LIMITED SCALE**
- Flood and learn (BUM) - Inefficient Bandwidth Utilization
- Resource Intensive – Large MAC Tables

**LIMITED WORKLOAD MOBILITY**
- Centralized Gateways – Traffic Hairpinning
- Sub-Optimal Traffic Flow

Barrier for Scaling out Large Data Centers and Cloud Deployments
Next-Gen VXLAN Fabric with BGP-EVPN Control Plane
Delivering Multi-Tenancy and Seamless Host Mobility at Cloud Scale

**In-creased Scale**
Eliminates Flooding
Conversational Learning
Policy-Based Updates

**Optimized Mobility**
Distributed Anycast Gwy
Integrated Routing/Bridging
vPC & ECMP

**Interoperable**
Standards Based
BGP-EVPN VXLAN

**Breaking the Traditional VXLAN Scale Barriers**
## VXLAN: Flood-&-Learn vs EVPN Control Plane

<table>
<thead>
<tr>
<th></th>
<th>Flood-&amp;-Learn</th>
<th>EVPN Control Plane</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overlay Services</strong></td>
<td>L2+L3</td>
<td>L2+L3</td>
</tr>
<tr>
<td><strong>Underlay Network</strong></td>
<td>IP network with ECMP</td>
<td>IP network with ECMP</td>
</tr>
<tr>
<td><strong>Encapsulation</strong></td>
<td>MAC in UDP</td>
<td>MAC in UDP</td>
</tr>
<tr>
<td><strong>Peer Discovery</strong></td>
<td>Data-driven flood-&amp;-learn</td>
<td>MP-BGP</td>
</tr>
<tr>
<td><strong>Peer Authentication</strong></td>
<td>Not available</td>
<td>MP-BGP</td>
</tr>
<tr>
<td><strong>Host Route Learning</strong></td>
<td>Local hosts: Data-driven flood-&amp;-learn</td>
<td>Local Host: Data-driven</td>
</tr>
<tr>
<td></td>
<td>Remote hosts: Data-driven flood-&amp;-learn</td>
<td>replication</td>
</tr>
<tr>
<td><strong>Host Route Distribution</strong></td>
<td>No route distribution.</td>
<td>MP-BGP</td>
</tr>
<tr>
<td><strong>L2/L3 Unicast Forwarding</strong></td>
<td>Unicast encap</td>
<td>Unicast encap</td>
</tr>
<tr>
<td><strong>BUM Traffic forwarding</strong></td>
<td>Multicast replication</td>
<td>Multicast replication</td>
</tr>
<tr>
<td></td>
<td>Unicast/Ingress replication</td>
<td>Unicast/Ingress replication</td>
</tr>
</tbody>
</table>
Design

VXLAN/EVPN Fabric

Spine1

Border Leaf1

Border Leaf2

Leaf

Service Leaf

WAN

vtep

EVPN Route Reflector

Rendezvous Point (Underlay)
Layer 3 Multi Tenancy
Distributed Anycast Gateway – Host Mobility

- Host H1 moves to Leaf Switch 3 (L3)
- L3 detects Host 1 and advertises H1 with updated sequence number 1
- L1 sees most recent update and withdraws its route
PBR Support for VXLAN BGP EVPN Fabric

Solution

- Redirect Layer-3 Traffic based on 5-tuple.
- Service Redirection to Load-Balancers and Firewalls.
- Next-Hop can be IPv4/IPv6 Hosts in VRF behind VTEP

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>Next-Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host A</td>
<td>Host B</td>
<td>Firewall</td>
</tr>
<tr>
<td>Host B</td>
<td>Host A</td>
<td>Firewall</td>
</tr>
</tbody>
</table>
Centralized Route Leaking
Extranet Support

Solution

- Use Cases – Shared Services, External Connectivity
- VRF to VRF or VRF to Default
- Centralize Location for leaking routes
Tenant Routed Multicast

VXLAN EVPN

Cisco ASIC – EX/FX

SRC-10
239.10.10.10
10.10.10.100

RCVR-10
10.10.10.10

RCVR-20
10.20.20.20

SRC-99
239.10.10.99
10.30.30.199

RCVR-30
10.30.30.30

RCVR-11
10.10.10.11

RCVR-40
10.40.40.40
VXLAN Operations, Administration & Management (OAM)

Delivering Carrier Grade VXLAN Manageability Solution

- **Ping / Path MTU**
  - Check liveliness of End-host
  - Option to specify Payload Parameters

- **Trace-route / Pathtrace**
  - Trace paths to host and tunnel-endpoint
  - Get path, interface and error statistics along route
  - Specify Payload Parameters for path selection

- **Pro-Active Monitoring**
  - Proactive ping with threshold and state notifications

---

**End-Point Locator / Trace**
- Locate end-host given segment and Device-Id information
VXLAN EVPN Multi-Site

Site 1

Site 2

Border Gateways

Border Gateways

Site 1 VXLAN Tunnel

Overlay Multi-Site

Site 2 VXLAN Tunnel

Scale through Hierarchical Forwarding

Fault Containment

Convergence independent of Network Size

Separate Admin Domains

Single Box

Cisco ASIC Advantage – EX/FX

7.0(3).I7(1)
Multi-Site – Overlay Data Plane

De-capsulation and Re-encapsulation on BGW

Intra-site VXLAN Data Plane

Inter-site VXLAN Data Plane

De-capsulation and Re-encapsulation on BGW

DC Core (Layer-3 Unicast)

Site1

VXLAN EVPN

Spine

VTEP

BGW

VIP1 10.1.1.111

Host1
0000.3010.1101
192.168.10.101

Host2
0000.3020.2101
192.168.20.101

Host3
0000.3010.1102
192.168.10.102

Site2

VXLAN EVPN

Spine

VTEP

BGW

VIP2 10.2.2.222

57
Datacenter solutions

**Application Centric Infrastructure**
- Turnkey integrated solution with security, centralized management, compliance and scale
- Automated application centric-policy model with embedded security
- Broad and deep ecosystem

**Programmable Fabric**
- VxLAN-BGP EVPN standard-based
- Cisco DCNM Automation

**Classic Ethernet**
- Modern NX-OS with enhanced NX-APIs
- Automation Ecosystem (Puppet, Chef, Ansible etc.)
- Common NX-API across N2K-N9K

Commercial, Enterprises, Public sector, Hosters | Service Providers/DoItYourself fabric | Classic Datacenters
Interconnecting ACI Networks Deployment Options

**Single APIC Cluster/Single Fabric**
- Stretched Fabric
- Multi-Pod (from 2.0 release)

**Multiple APIC Clusters/Multiple Fabrics**
- Multi-Fabric (with L2 and L3 DCI)
- Multi-Site (August 17)
ACI Multi-Site Overview

- Separate ACI Fabrics with independent APIC clusters
- Each fabric is considered as a different availability zone
- Scoping of configuration changes
- DR and Active/Active use cases support

- ACI Multi-Site pushes cross-fabric configuration to multiple APIC clusters
- MP-BGP EVPN control plane between sites
- Data Plane VXLAN encapsulation across sites
- End-to-end policy definition and enforcement
ACI Multi-Site Namespace Normalization

- Maintain separate name spaces with ID translation performed on the spine nodes
- Requires specific HW support for this functionality
ACI Multi-Site
Hardware Requirements

- Support all ACI leaf switches (NS, -E, -EX and -FX)
- Only spine nodes with –EX line cards (or newer) to connect to the inter-site network (required FM-E fabrics)
- New FX non modular spine (64 ports) will be supported in Q4CY17 timeframe
- 1st generation spines (including 9336PQ) not supported
  - Can still leverage those for intra-site leaf to leaf communication
ACI Multi-Pod and ACI Multi-Site
What Option to Choose?

Multi-Pod

Pod ‘A’

Pod ‘n’

IPN

APIC Cluster

Multi-Site

Site ‘A’

Site ‘n’

IP

ACI Multi-Site

Operational Simplicity

Feature Richness across Pods

Lower Number of APIC Nodes

Single VMM across Pods

Change Domain Isolation

High Latency across Sites

Fabric Nodes Scale

No Multicast required in the IP Network
ACI Anywhere
Any Workload, Any Location, Any Cloud

Remote PoD
Multi-Pod / Multi-Site
Hybrid Cloud Extension

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

IP
WAN

Remote Location
Public Cloud
On Premise

Security Everywhere
Analytics Everywhere
Policy Everywhere

Any Location
Any Cloud

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere

ACI ANyWHERE
IP WAN
IP WAN

Remote Location
On Premise
Public Cloud

Security Everywhere
Analytics Everywhere
Policy Everywhere
Migrating to ACI – extend your network via Layer 2

Extend the VLAN/EPG

- Single Policy Group (one extended EPG)
- Leverage vPC for interconnect
- BPDU should be enabled on the interconnect ports on the ‘vPC’ domain
TYPICAL 3-TIER APPLICATION

Cisco ACI Policy Model

APPLICATION NETWORK PROFILE

WEB  APPL  DB
SPRITE tenant - migrated

vCenter

L3out
VRF: common

Tennant = common

VRF: VRF_SPRITE
BD: VLAN200
IP Routing: Yes, 10.101.9.1/24

Tennant = SPRITE

WEB client 10.99.2.202

Outside

ANP: WEBAPP

EPG: SHARED
WEB_Frontend 10.101.9.10
MySQL DB 10.101.9.2
SPRITE tenant – Secured Database

Tenant = SPRITE

VRF: VRF_SPRITE (Preferred-Group=Enabled)
BD: VLAN200
IP Routing: Yes, 10.101.9.1/24

Tenant = common

VRF: common

WEB client 10.99.2.202

L3out

vCenter

ANP: WEBAPP

WEB_Frontend 10.101.9.10

MySQL DB 10.101.9.2

tcp/3306

tenant = SPRITE

tenant = common

EPG: Shared

EPG: MYSQL

WEB client 10.99.2.202
SPRITE tenant - Secured

VRF: VRF_SPRITE (Preferred-Group=Enabled)

BD: VLAN200
IP Routing: Yes, 10.101.9.1/24

EPG: WEBSRV
WEB_Frontend 10.101.9.10

EPG: MYSQL
MySQL DB 10.101.9.2

EPG: SHARED
ANP: WEBAPP

WEB client 10.99.2.202

L3out
VRF: common

Outside

Tennant = common

Tennant = SPRITE

vCenter

tcp/https
tcp/3306
Cisco Data Centre Networking Architectures

Application Centric Infrastructure
- VXLAN-based
- Forwarding, Multi-Tenancy & Security
- Turn-key Integrated Controller with Enhanced APIs
- Agnostic Hypervisor integration
- Wide deployment scenarios

Programmable Fabric
- Standards-based
  - VXLAN BGP EVPN
- Forwarding & Multi-Tenancy
- Multi-site support
- Open NX-OS
- DCNM Automation (option)

Classic ethernet
- BYO-Network, scales well
- Simple and well-known deployment
- Open NX-OS
- Enhanced APIs and Automation Ecosystem (DevOps)