Your Time Is Now
BGP-EVPN and SR DC Fabric
Addressing the evolving Data Center requirements

Ahmad Bilal, Technical Marketing Engineer
Samir Thoria, Distinguished Engineer
BRKSPG-2509
Agenda

- EVPN-SR DC Fabric Introduction
- DC Fabric Building Blocks
- Segment Routing in data center
- EVPN in data center
- Conclusion
Session Non-Objectives

• Following topics are not covered in details in this session
  • NCS5500 Platform
  • Segment Routing deep dive

Related Sessions:
• BRKSPG-2900: Cloud Scale Networking: NCS 5500 and NCS 5000 Series Deepdive
• BRKRST-3122: Segment Routing: Technology deep-dive and advanced use cases
Evolving DC Requirements

• Provide high performance any to any connectivity
• Flexible service/workload placement
• Workload mobility
• Scale
• Traffic engineering
• Efficient bandwidth utilization
• Multi-tenancy with L2 and L3 VPN
EVPN-SR DC Fabric Introduction
EVPN-SR Data Center Fabric

EVPN-SR DC Fabric

BGP EVPN
Segment Routing
IOS-XR
NCS 5500 Platform
Underlay vs. Overlay

**Underlay**
Physical Network
IP/Label Transport
Full bandwidth utilization

**Overlay**
Virtual Network
Provides VPN services
Policy driven
Data Center Underlay – SR/MPLS

- Underlay is routing protocol + end to end topology
- Leaf – Spine Topology
- Uniform Reachability, Deterministic Latency
- High Redundancy: Node/Link Failure
Data Center Overlay – BGP-EVPN/MPLS

- Distribute tenant routes and external network reachability
- Route-Reflectors deployed for scaling purposes
  - Optionally use eBGP for overlay peering
- EVPN next hops are reachable via SR path(s)
End-to-end Unified Control-Plane & Transport

Simplified Protocol Stack

- Single SDN-enabling forwarding
- Common control-plane for L2/L3 VPN
- Simplified Traffic Engineering
- Consistent Data Model & APIs across network

Legacy Protocol Stack

- L2/L3VPN - VPLS, EoMPLS, RSVP-TE
- VPLS, Other Overlays
- L2, STP, IP DC Fabric

A1: Access
C1-3: compute
NCS 5500 in Data Center
NCS 5500 Product Family

- **NCS 5508**
  - Modular 8 slots
  - 13 RU (1/3 rack)
  - 28.8 Tbps @ 7000 W

- **NCS 5516**
  - Modular 16 slots
  - 21 RU (1/2 rack)
  - 57.6 Tbps @ ~18000 W

- **NCS 5501**
  - Fixed 1 RU
  - 800 Gbps @ 243 W

- **NCS 5502**
  - Fixed 2 RU
  - 4.8 Tbps @ 1450 W

NCS 5501

NCS 5502

NCS 5508

NCS 5516
NCS 5500 in Data Center

- High 100G Density
- Low power / Low per port cost
- Medium / High Scale FIB
- Deep Buffers
- SR / MPLS Transport
- EVPN Control Plane
- Traffic Engineering
- L3 Data Center Interconnect (DCI)
IOS-XR in Data Center

Cloud Scale networking operations with XR

- Most dominant & well-known OS in the core/backbone space
- Strong MPLS & SR feature support
- Comprehensive L2/L3 VPN services
- Streaming telemetry support
- Evolved programmability with model driven operations
- Ability to run (host) 3rd party apps
- Common APIs for Unified manageability across the SP portfolio
Segment Routing in Data Center
Segment Routing Overview

• **Source Routing**
  - the source chooses a path and encodes it in the packet header as an ordered list of segments
  - the rest of the network executes the encoded instructions

• **Segment**: an identifier for any type of instruction

• **Forwarding Plane**:  
  - **MPLS**: an ordered list of segments is represented as a stack of labels  
  - **IPv6**: an ordered list of segments is encoded in a routing extension header

• **Multi-Vendor** solution

This presentation: MPLS Data plane
Segment Routing: IGP segments

IGP Prefix Segments
- Shortest-path to the IGP prefix
  - Equal Cost MultiPath (ECMP)-aware
- Global Segment
- Label = 16000 + Index
- Distributed by ISIS/OSPF

IGP Adjacency Segment
- Forward on the IGP adjacency
- Local Segment
- Advertised as label value
- Distributed by ISIS/OSPF

All nodes use default SRGB 16,000 – 23,999
Segment Routing Underlay

Each device in the fabric is assigned a prefix-SID, visible to all other devices.

Underlay adjacency between the nodes exchanging Prefix-SID (labels).

Segment Routing Underlay

Leaf

Spine
MPLS-SR Data Plane Operations

- Leaf-4 advertises its loopback ipv4 prefix 4.4.4.4/32 with attached prefix-SID 16004
- Spine performs the PHP functionality
Segment Routing in Data Center

- Simplified Traffic Engineering

- Policy driven path selection at the Leaf
- Steer traffic on any path through the network
- No path is signaled
- No LDP and RSVP required
Introduction to BGP EVPN
What is EVPN

- EVPN family introduces next generation solutions for Ethernet services
  - BGP control-plane for Ethernet Segment and MAC distribution learning over MPLS and VXLAN data-plane
  - Same principles and operational experience as in IP VPNs
- No use of Pseudo wires
  - Uses MP2P tunnels for unicast
  - Multi-destination frame delivery via ingress replication (via MP2P tunnels) or LSM
- Multi-vendor solutions
EVPN – Control and Data plane

Control-Plane

- EVPN (MP-BGP)
  - RFC7432

Data-Plane

- Multi-Protocol Label Switching (MPLS)
  - RFC7432
- Provider Backbone Bridges (PBB+MPLS)
  - RFC7623
- Network Virtualization Overlay (VXLAN, NVGRE, MPLSoGRE)
  - draft-ietf-bess-evpn-overlay

LDP, SR or any MPLS transport
BGP EVPN - Ethernet VPN

• Leafs run Multi-Protocol BGP to advertise & learn MAC/IP addresses over the DC Fabric
• MAC/IP addresses are advertised along with an MPLS label to rest of Leafs
BGP EVPN Constructs
**BGP EVPN – EVI**

**EVI**: An EVPN instance extends Layer 2 between the Leafs

- EVI extended over BGP-EVPN Fabric to all the Leafs belonging to the EVI
- Leafs that don’t belong to a specific EVI will not have MAC-VRF for that EVI, providing efficient scalability
BGP EVPN – Ethernet-Segment for Multi-Homing

The bundle on the Leafs connecting to a node should have Identical ES identifier (ESI)

Unique 10-byte global identifier per Ethernet Segment

Ethernet Segment represents a node connected multiple Leaks

Leaf

Spine
BGP EVPN – Host Connectivity Options

- Ethernet Segment Identifier (ESI) ‘0’
- No DF election

- Identical ESI on Leafs
- Identical ESI MAC Address
- Per VLAN DF election

Single Home Device (SHD)

- ESI-0

Multi-home (MHD) All-Active (Per-Flow) LB

- ESI-1

Leaf

Spine

Single homed host

Multi-homing with Link Bundling
EVPN IRB in Data Center

Inter-subnet - Bridged

Intra-subnet - Bridged

BD-1

BD-2

BVI-1
GW MAC

BVI-1
GW MAC

BVI-2
GW MAC

BVI-2
GW MAC

VM

VM

VM

VM
Distributed Anycast Gateway with BGP-EVPN

Optimal intra and inter-subnet connectivity with seamless workload mobility

Distributed Anycast Gateway serves as the gateway for connected hosts

Identical Anycast Gateway Virtual IP and MAC address are configured on all the Leafs

All the BVIs perform active forwarding in contrast to active/standby like FHRP
BGP EVPN in Data Center
Centralized vs. Distributed Routing

**Distributed Routing**

- Optimized forwarding of east-west traffic
- ARP/MAC state localized to Leafs
  - Helps with horizontal scaling of DC

**Centralized Routing**

- All east->west routed traffic traverses to centralized gateways
- Centralized gateways have full ARP/MAC state in the DC
  - Scale challenge
Integrated Routing and Bridging

**Symmetric IRB**
- Flexible workload placement – any subnet anywhere
- ARP/MAC state localized to Leafs
  - Helps with horizontal scaling of DC

**Asymmetric IRB**
- Egress subnet must be local
- Ingress Leaf needs ARP/MAC state for every egress leaf
  - Limits scale
BGP EVPN All Active per-flow Load balancing

- No dedicated cross link between leafs required
- EVPN based service carving for load balancing of BUM traffic forwarding
- Mass withdraw for faster convergence
BGP EVPN Split Horizon

**Challenge:**
How to prevent flooded traffic from echoing back to a multi-homed Ethernet Segment?
BGP EVPN Designated Forwarder (DF)

**Challenge:**
How to prevent duplicate copies of flooded traffic from being delivered to a multi-homed Ethernet Segment?
BGP EVPN Aliasing

**Challenge:**
How to load-balance traffic towards a multi-homed device across multiple Leafs when MAC addresses are learnt by only a single Leaf?
BGP EVPN MAC Mass-Withdraw

Challenge:
How to inform other Leafs of a failure affecting many MAC addresses quickly while the control-plane re-converges?
BGP EVPN MAC Mobility

**Challenge:**
How to detect the correct location of MAC after the movement of host from one Ethernet Segment to another also called “MAC move”?

Sequence number and Next-Hop value will be changed after the host move

<table>
<thead>
<tr>
<th>MAC</th>
<th>IP</th>
<th>ESI</th>
<th>Seq.</th>
<th>Next-Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC-1</td>
<td>IP-1</td>
<td>0</td>
<td>0</td>
<td>Leaf-1</td>
</tr>
</tbody>
</table>

Host move
BGP EVPN MAC Mobility, continued

Sequence number is incremented and Next-hop is changed to Leaf-3

<table>
<thead>
<tr>
<th>MAC</th>
<th>IP</th>
<th>ESI</th>
<th>Seq.</th>
<th>Next-Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC-1</td>
<td>IP-1</td>
<td>0</td>
<td>1</td>
<td>Leaf-3</td>
</tr>
</tbody>
</table>
Data center interconnect
MPLS Data center with EVPN/SR

- Multi-tenant, scalable, high performance data center
- Provides common operation models across DC & WAN with IOS-XR
- Seamless transport with SR & efficient control plane with EVPN

SR
- Explicit Path Control
- Full path programmability
- TE based on application needs

EVPN
- MPBGP for MAC/IP Distribution
- L2 and L3 VPN services

IOS-XR
- Modular & extensible software
- Automation @ scale
- Visibility & Telemetry
- Strong MPLS & SR support
Complete Your Online Session Evaluation

• Please complete your Online Session Evaluations after each session

• Complete 4 Session Evaluations & the Overall Conference Evaluation (available from Thursday) to receive your Cisco Live T-shirt

• All surveys can be completed via the Cisco Live Mobile App or the Communication Stations

Don’t forget: Cisco Live sessions will be available for viewing on-demand after the event at CiscoLive.com/Online
Continue Your Education

• Walk-in Self-Paced Labs
  • LABSPG-2014: Configuring and Implementing EVPN-SR based Data Center

• Meet the Engineer 1:1 meetings

• Demos in the Cisco campus

• Related sessions
  • BRKSPG-2900: Cloud Scale Networking: NCS 5500 and NCS 5000 Series Deepdive
  • BRKSPG-1001: Designing High Density SP & DC Networks with NCS5500
  • BRKSPG-2404: IOS-XR Platforms: System and Hardware Architectures
  • BRKRST-3122: Segment Routing: Technology deep-dive and advanced use cases
  • BRKSPG-2210: Designing Programmable Access Networks
Thank You