Delbecq Adrien
Senior Network SRE @Scaleway

Scaleway’s approach to VXLAN + BGP EVPN Fabric
Summary

- Reminder VXLAN + BGP EVPN
- Fabric Underlay
- Fabric Overlay
- What’s next?
VXLAN – BGP EVPN

VXLAN Terminology

• **VXLAN**: Virtual eXtensible LAN
• **VTEP**: VXLAN Tunnel Endpoint
• **VNI**: VXLAN Network Identifier
• **NVE**: Network Virtual Interface
VXLAN – BGP EVPN

VXLAN Concept

• rfc 7348
• Data-plane technology
• Encapsulate Ethernet on top of UDP
  • Support Bridging & Routing
• Multi – tenant (up to 16M VNI)
• Hardware support
VXLAN – BGP EVPN

VXLAN Encapsulated Frame

Inner Ethernet Frame

14 bytes
Outer Ethernet Header

20 bytes
Outer IP Header

8 bytes
Outer UDP Header

8 bytes
VXLAN Header

Inner Dest MAC
Inner Source MAC
Optional 802.1Q EtherType
Optional Inner 802.1Q EtherType
Original Ethernet Payload
FCS

IP Header Data*  IP Protocol Header Checksum Outer Source IP Outer Dest IP

Source Port Dest Port (8472) UDP Length UDP Checksum

Outer Dest MAC Outer Source MAC Optional 802.1Q EtherType Optional Outer 802.1Q EtherType

VXLAN Flags RSVD VXLAN NI (VNI) RSVD

*IP Header Data = Version, IHL, TOS, Length, ID
VXLAN – BGP EVPN

BGP EVPN Concept

• rfc 8365
• Control plane technology
• Another BGP Address-family
  • from MPLS EVPN (rfc 7432)
  • Support multiple encapsulation
BGP EVPN – route types

- Type 1: Ethernet autodiscovery
- **Type 2**: Host (mac + mac-ip) **routes**
- Type 3: Inclusive Multicast Ethernet tag route
- Type 4: Ethernet Segment Route
- **Type 5**: Ip Prefix Route
- …
VXLAN – BGP EVPN

VXLAN/BGP EVPN – type 2, bridging

hA

bridging

hB

hostA
ip: 2001:db8::1/64
mac: 0050.5600.0001

l2vni

hostB
ip: 2001:db8::2/64
mac: 0050.5600.0002
VXLAN – BGP EVPN

VXLAN/BGP EVPN – routing, type5 / type2 sym model

hA

hostA
ip: 2001:db8::1/64

l3vni

hostB
ip: 2001:db8:0:1::1/64

hB
VXLAN – BGP EVPN

BGP EVPN – type 2, routing, asym model
Fabric Underlay

Layer 1: Remember Clos

- Spines
- Leafs
- 10G
- 100G
- 12
Clos scale: more ingress/egress?
Fabric Underlay

Clos scale: more bandwidth
Fabric Underlay

Grow more with Clos

Super Spine

SS1

SS2

S1.1

S1.2

S2.1

S2.2

L1.1

L1.2

L1.3

L1.4

L2.1

L2.2

L2.3

L2.4
Fabric Underlay

External connectivity?

border leaf

external network
Fabric Underlay

Data Plane: IPv4

- No extended broadcast domain
- IPv6 underlay was not available/ready
- L3 sub-interface everywhere
  - Efficient loop prevention
  - ECMP: 100% bandwidth used
Fabric Underlay

Control Plane: eBGP

- No link-state protocol
  - No OSPF
  - No IS-IS
- iBGP isn’t really good as IGP
- eBGP just fits
  - RFC7938 – draft Lapukhov
- No BFD
Fabric Underlay

Control Plane: eBGP
Fabric Underlay

Addressing plan

• Internet-like addressing plan
  • Use next available prefix
  • No waste

• Topology–driven addressing
  • IP address = function (topology)
  • Human-friendly
Fabric Underlay

Addressing plan: Topology-driven addressing
Fabric Underlay

Management through underlay

• KISS
• Resilient (hello BGP)
• It just works
Fabric Overlay

Everything is now running on overlay

• Adm, bmc (ipmi)
• Public traffic
• VPC (coming soon)
• ...

• Underlay only persists for shelves management
Fabric Overlay

Agnostic spine & superspine

• Spine and superspine are **not** VXLAN aware:
  • KISS
  • Less FIB usage
  • Less features
  • Cheaper
Fabric Overlay

Virtualized Route-reflectors

• Connected on edgeleaves
• Independant from shelves
• Easy to replace with another control plane
  • Cisco xrv, Juniper vRR, Arista vEOS...
  • Bird, FRR...
• HV could handle other services
  • Route-injector
Fabric Overlay

Routing only through type 5

- Type 5 routing only
- Type 2 bridging only: no mix
Fabric Overlay

Host multihoming

- L3 on HV could work
  - But, how to do it with Baremetal services?
  - How to scale bgp sessions number (per vrf)?
- ESI + MC-LAG light= standard
  - But isn’t really plebiscited by vendors
- Anycast VTEP + MC-LAG
  - Non standard
  - It just works
Fabric Overlay

Host multihoming – Anycast VTEP + MC-LAG
Fabric Future

Software VTEP

- Compatible with hw vtep
- Bring your own Control-Plane

- No hardware limits (tcam, fib) ...
- Limited performance (cpu vs asic/fpga)
Scaling – hub&spoke - sharding

- scale limit related to FIB
  - More and more prefixes

- Does all leaf need all routes?
Scaling – hub&spoke

Fabric Future
Fabric Future

Scaling - sharding
Fabric Future

Multi-vendor interoperability

- Cisco – Juniper Interoperability?
  - Bridging OK
  - Routing type 5 OK
  - Routing type 2 KO
    * Cisco use SYM IRB routing with t2
    * Juniper use ASYM IRB routing with t2
Fabric Future

whitebox

• Bring your own Control-Plane

• Standard Linux OS:
  – same automation than on soft VTEP

• Same ASICs (hello Broadcom Trident)

• Cheaper
Thank you

follow me on Linkedin and twitter @Adelbecq