BGP Extension for Tunnel Egress Point

draft-hcl-idr-extend-tunnel-egress-point-01

PengFei Huo (ByteDance) (Presenter)
Gang Chen (ByteDance)
Changwang Lin (New H3C Technologies)
Syed Hasan Raza Naqvi (Broadcom)

IETF-120

Introduction

Problem:
Link Congestion Issues in AI Networks

Possible options for the data plane:
Ethernet enhancement
- Per-flow forwarding optimization
- Adaptive Routing

Fully Scheduled Network
- Data traffic spraying
- Packet cell

Solution:
Fully Scheduled Network ,by extending BGP, synchronize tunnel egress point information to provide possibilities for addressing congestion issues in the data plane.

https://datatracker.ietf.org/doc/html/draft-hcl-rtgwg-osf-framework-00

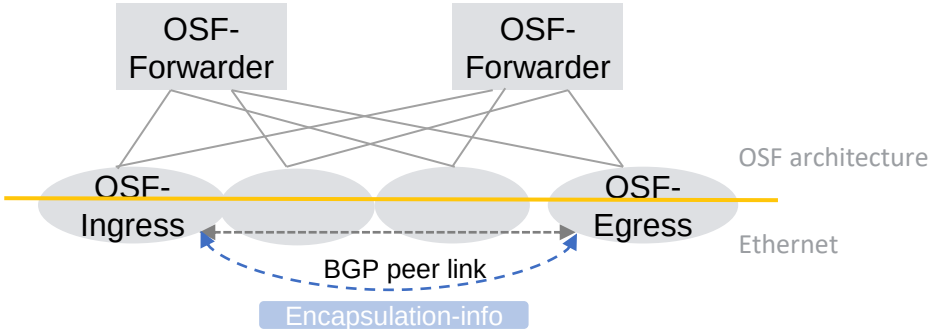
Possible Mechanisms

- **Based on OSF Framework**

In the OSF network architecture, there are two types of device identities, one is the forwarding node, and the other is the encapsulation termination node, which we called " OSF-Forwarder " and “OSF-Ingress/Egress".

- **Extending based on BGP**

Establish BGP neighbor relationships between OSF-Ingress/Egress devices and transmit the necessary encapsulation information for the OSF forwarding layer.



The necessary encapsulation information

Out-of-Order Issues Caused by Data Spraying

- Outbound Interface Information

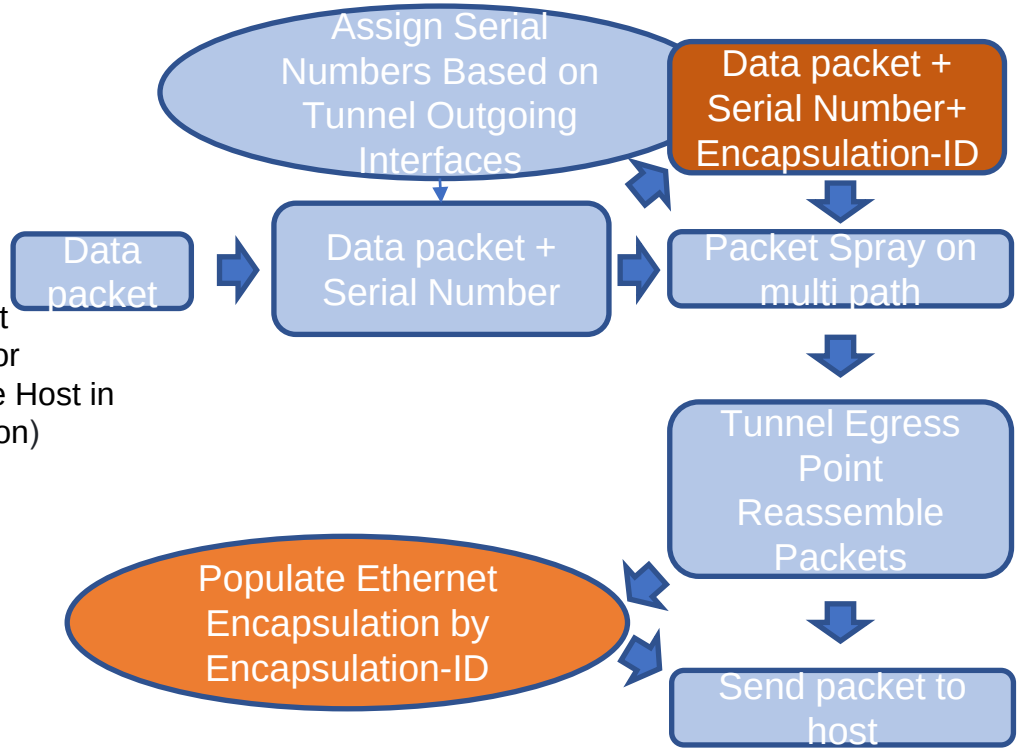
Data Plane Selection within Network Architecture

- Ethernet
- Non-Ethernet

→ Encapsulation ID(Ethernet Encapsulation Required for Forwarding Packets to the Host in the Tunnel Egress Direction)

Control Plane Selection

- BGP ipv4/ipv6 for Ethernet forwarding
- BGP EVPN for Non-Ethernet forwarding



BGP Extensions Encoding TEP(Tunnel Egress Point)

New BGP attributes:

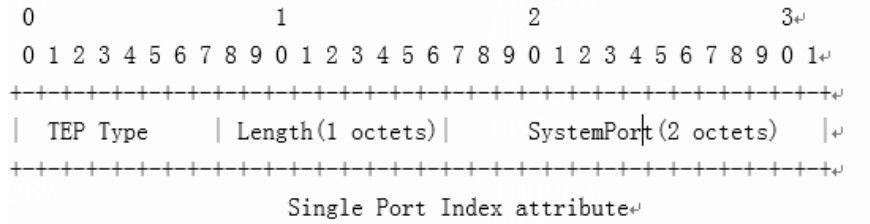
- Single Port Index attribute
- Multiple Port Index attribute
- Encapsulation ID attribute

BGP ipv4/ipv6 Peer:

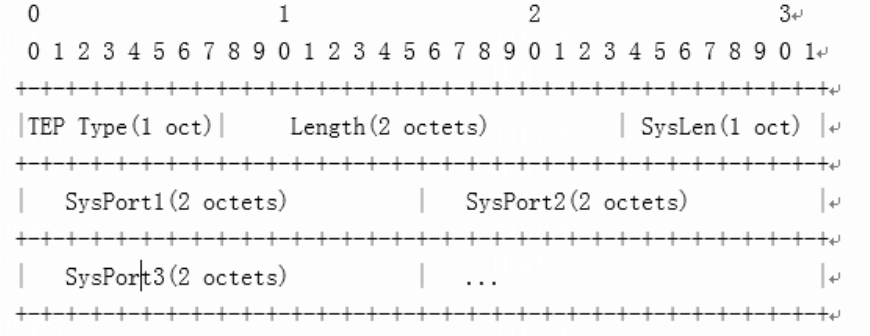
- Carry Single Port Index attribute for layer-three forwarding

BGP EVPN Peer:

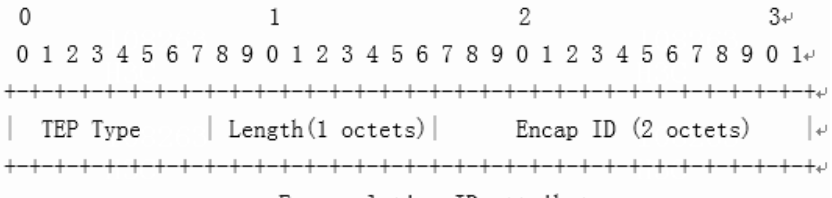
Carry Single Port Index attribute for layer-three/layer-two forwarding
Carry Encapsulation ID attribute for layer-three/layer-two forwarding
Carry Multiple Port Index attribute for layer-two broadcast



Single Port Index attribute



Multiple Port Index attribute

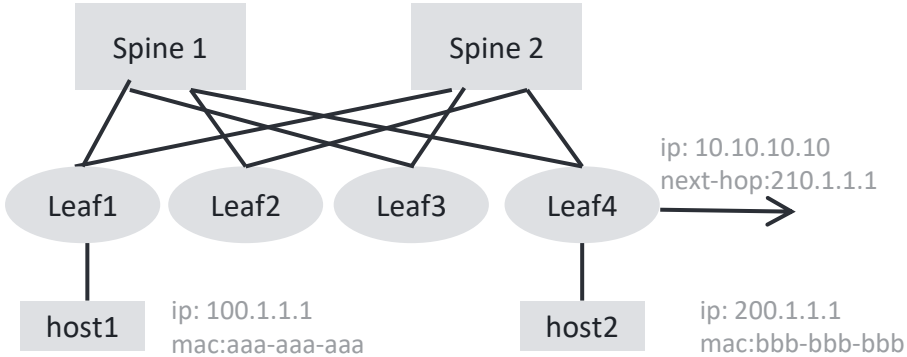


Encapsulation ID attribute

USE CASE

EVPN L3 networking:

- Establish Full Mesh BGP Neighbor Relationships Among Leaf Nodes
- Leaf4 Collects Local Routing Information and Outgoing Interface Information
- Synchronize These Routing Information to Other BGP Neighbors
- Leaf1 Forms Tunnel Route



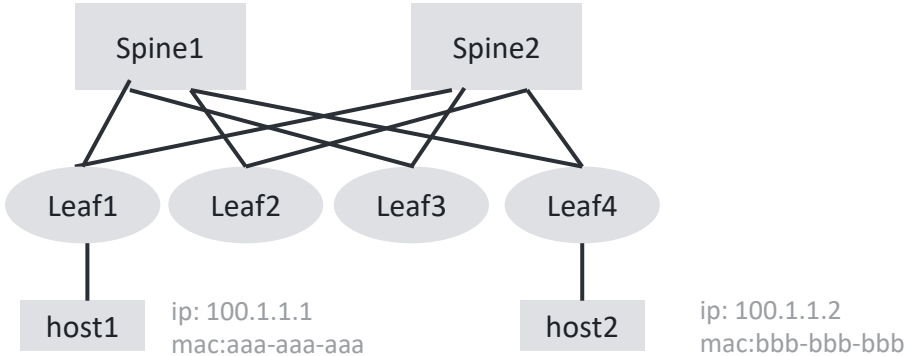
Leaf 1Routing-table

prefix	nexthop	out-interface	encapsulation information
100.1.1.1/32	100.1.1.1	dev1 + port 1	encap-id
200.1.1.1/32	200.1.1.1	dev2 + port 1	encap-id
10.1.1.0/24	10.1.1.1	intreace 1	NA

USE CASE

EVPN L2 networking:

- MAC learning and table synchronization
- Unicast traffic interconnection
- Establishment of broadcast domain
- Interconnection of broadcast traffic
- Integrated Routing and Bridging (IRB) forwarding



Leaf 1 L2-Routing-table

prefix	mac	out-interface	encapsulation information
100.1.1.2	bbb-bbb-bbb	dev4+port 1	encap-id
100.1.1.1	aaa-aaa-aaa	intreace 1	NA

Running Code

Lab Interop-test Status

Hardware and software development is underway, with plans to conduct interoperability testing in ByteDance's laboratory in 2024:

- H3C: S12500AI-96B-NCFK,S12500AI-18D48B-NCPK,S12500AI-36DH20EP-NCPN,S12500AI-NCFN
- Drivenets:ASA926-18XKE-O-AC,AS9936-128D-O-AC
- Ruijie : RG-S6940-36QC20F4,RG-X112-128F4

Next Step

- Any questions or comments are Welcomed
- Continue improving the OSF network architecture