BGP Extension for Tunnel Egress Point

draft-hcl-idr-extend-tunnel-egress-point-03

Presenter : Changwang Lin (New H3C Technologies)

Co-Authors : Pengfei Huo (ByteDance)
Gang Chen (ByteDance)
Changwang Lin (New H3C Technologies)
Weiqiang Cheng (China Mobile)
Syed Hasan Raza Naqvi (Broadcom)
Yossi Kikozashvili (DriveNets)

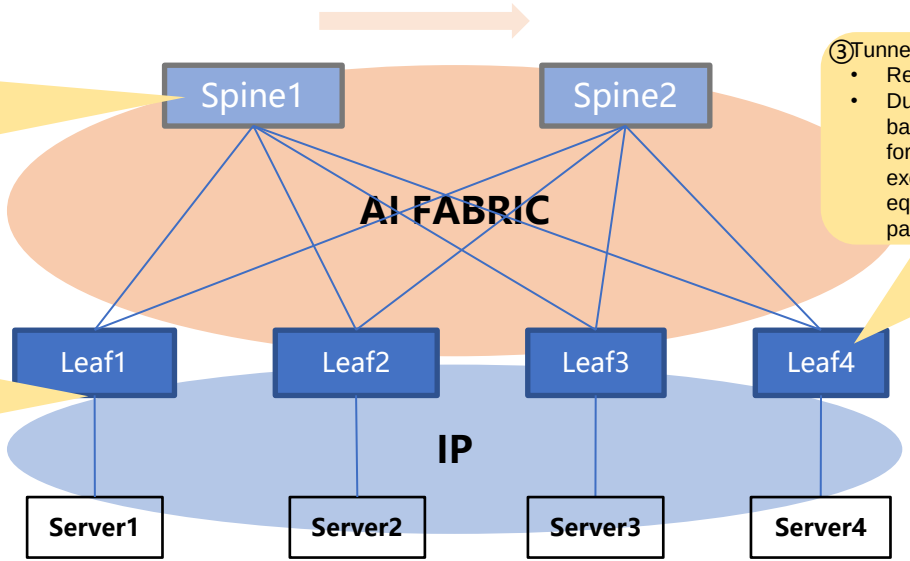
IETF-121

Background

- AI Network Requirements:
  - High network utilization
  - High BW flows & Fewer flows →imperfect load lalancing
  - Unique resiliency requirements
- Load Balancing Techniques:
  - Flow Scheduling: Flexible flow hash Scheduling & congestion-aware adaptive routing
  - Full-scheduling Ethernet:
    - ✓ Spray & Reorder Technology: Distributing packets across multiple paths and reordering them at the destination to ensure correct sequence.
    - ✓ Cell-based DDC Technology: Segment messages into cells and spray them, using non-Ethernet technology for inter-device forwarding.
    - ✓ GSE： Aggregate packets into fixed lengths and spray them, using a GSE extended packet header for ordering.

To enhance network performance using Full-scheduling Ethernet technology, the control plane needs to assist in transmitting certain information. This document proposes extending BGP to carry additional information to support full-scheduling forwarding technology in the data plane.

Requirements for BGP



②AI fabric Forwarder(Maybe two different forwarding behaviors):
- The new data layer forwards based on the outbound port in the new packet header, such as DDC、GSE
- IP forwarding

③Tunnel Egress :
- Reorder packets based on the outgoing interface.
- Due to the AI architecture, forwarding might be based on non-Ethernet technologies, and the forwarding process lacks an Ethernet header exchange step. Therefore, the outbound port equipment needs to reconstruct the Ethernet packet header.

①
Tunnel Ingress :
- Process messages, cutting or aggregation packet, add encapsulation information, assign sequence numbers.
- Packets spraying.

Data plan work step:
①-> ② -> ③

Requirements for BGP
- Carry new routing information:
  - Outgoing interface ID(device ID + port ID)
  - Encapsulation ID

Requirements to Complete This Forwarding Process:
- Tunnel Egress transmit local outgoing interface information to the remote device, allowing the Tunnel Ingress to assign a sequence number based on the it.
- Tunnel Egress generate a locally effective encapsulation information ID for the Ether-packet header and pass it to Tunnel Ingress, who then includes it in the datagram's extension header information.

BGP Extensions

BGP Extensions #1 OutInterfaceID ----------→ reorder

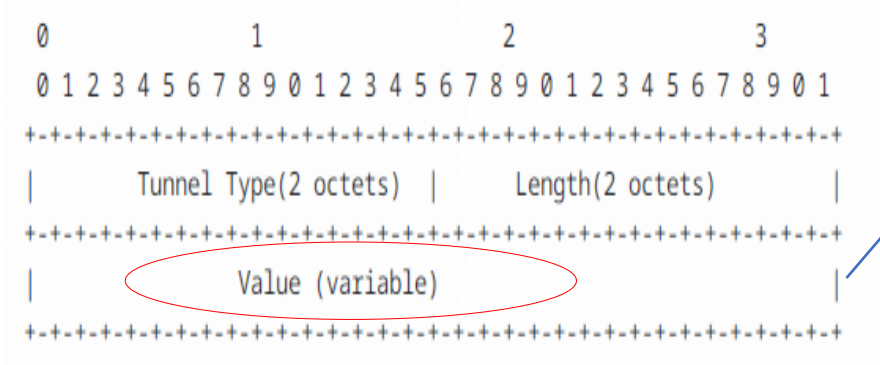
BGP Extensions #2 Encapsulation ID ----------→ re encap

- When the AI fabric is based on IP forwarding, use BGP IPv4/IPv6 address family, carrying only BGP Extensions #1 .

- When using non-IP forwarding in an AI fabric, use EVPN address family, carrying both BGP Extensions #1 And BGP Extensions #2.

BGP Extensions #1 OutInterfaceID

Based on Tunnel Encapsulation TLV(RFC9012)



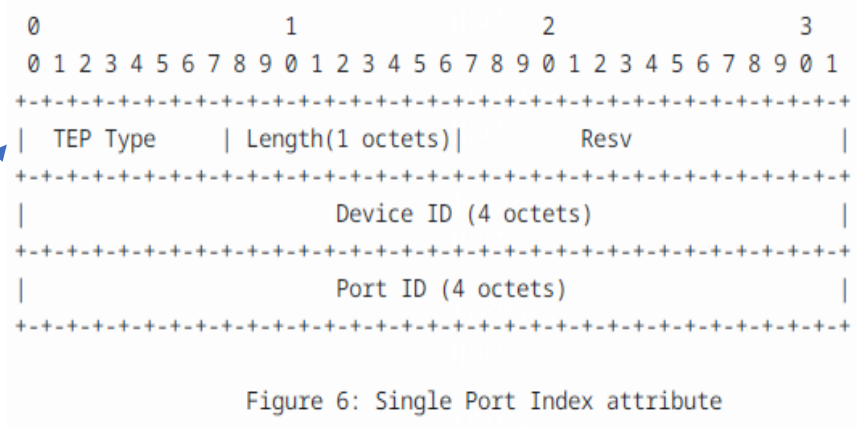
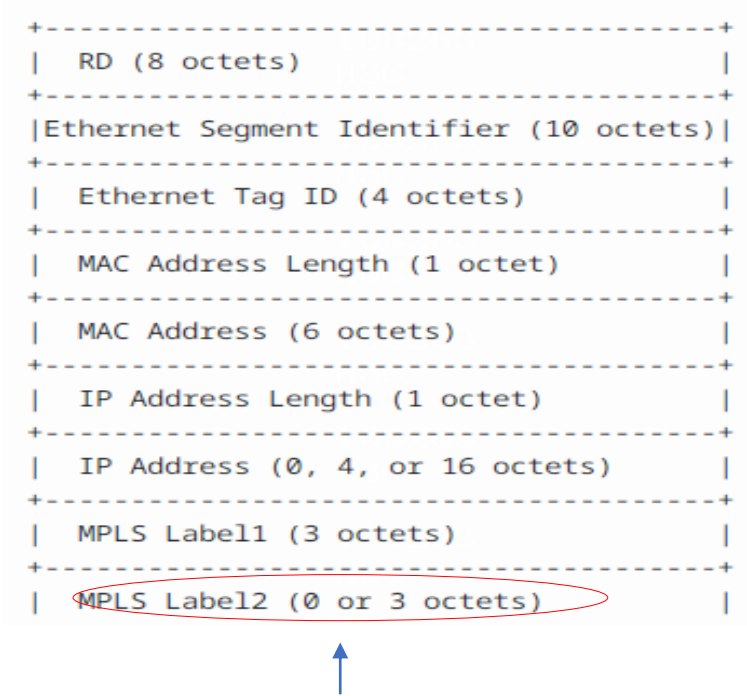


Figure 6: Single Port Index attribute

Add new tunnel type definitions "BGP Tunnel Encapsulation Attribute Tunnel Types“ [IANA-BGP-TUNNEL-ENCAP]

Add a new sub-TLV to carry device ID and port ID.

BGP Extensions #2 Encapsulation ID



The encapsulation ID is carried in the MPLS-LABEL2 field of NRLI.

USE CASE – IP Packet Spray

Control plane: for BGP extension

Use IPv4/IPv6 address family

- LEAF4 BGP route
  Prefix: 200.1.1.1
    TUNE ATTR:
      Device ID:1.1.1.1
      Port ID: 100

- LEAF1 ROUTING-TABLE

prefix	nexthop	Interface	TEP-ID
200.1.1.1/32	Spine1 Spine2	Int1 Int2	Device ID:1.1.1.1 Port ID:100

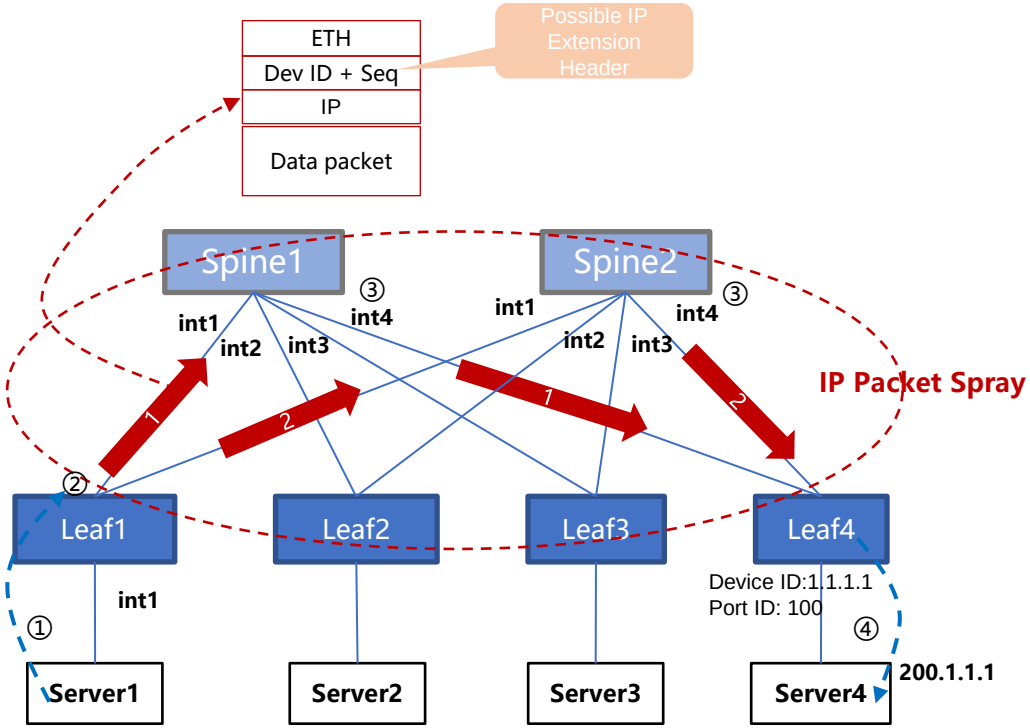
Forwarding step:

①: Calculate the packet sequence number based on port information, fill it and device id into the extension header.

②: Evenly spray the data packets over the ECMP link.

③: Forwarding based on IP.

④: Reorder packets.



USE CASE - DDC

Control plane: : for BGP extension

Use IPv4/IPv6 EVPN address family

- LEAF4 BGP route
  EVPN type-2 route
    Prefix :200.1.1.1
      NLRI:
        MPLS Lable2: Encapsulation ID 1
      TUNE ATTR
        Device ID:1.1.1.1
        Port ID: 100
- LEAF1 ROUTING-TABLE

Prefix	TEP-ID	ENCAP-ID
200.1.1.1/32	Device ID:1.1.1.1 Port ID:100	1

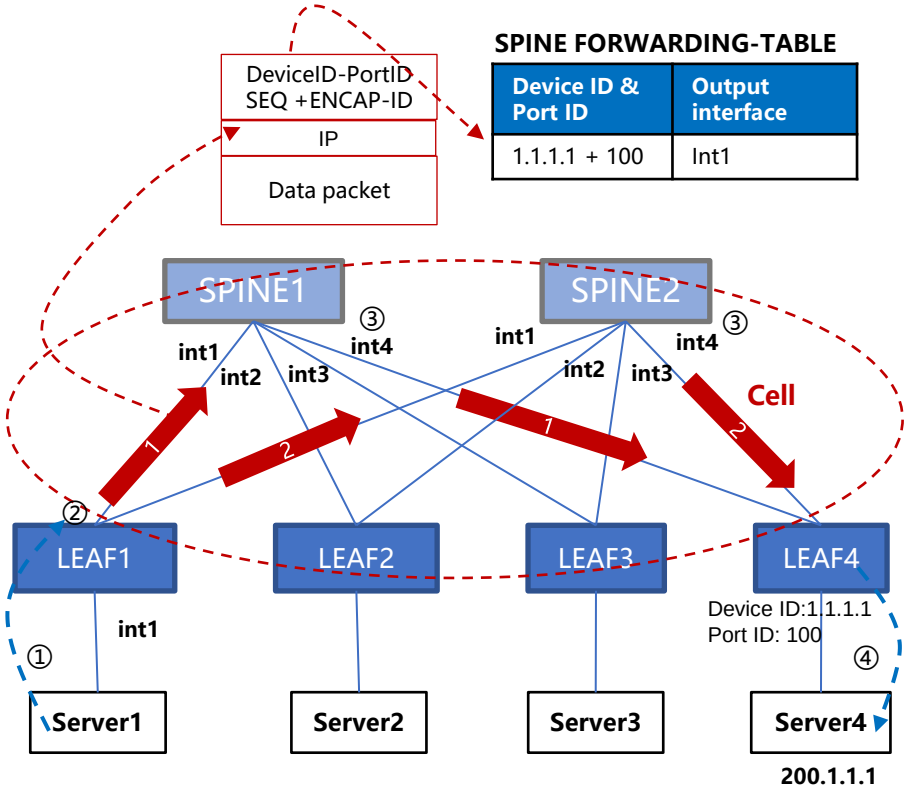
Forwarding step:

①: Calculate the sequence number based on the output port, fill the destination Device ID, PortID, fill ENCAP-ID and sequence number.

②: Split the message into cells, evenly spray the data packets over the ECMP link.

③: Forwarding based on lookup using Device ID & Port ID

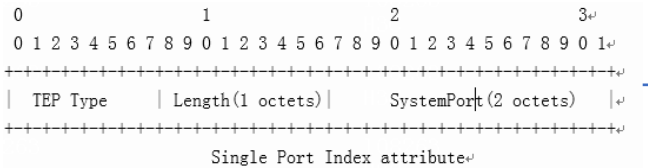
④: Reorder packets, and restore Ethernet frames based on the encapsulation ID.



Update of this draft

Extend the two-byte system port information into four-byte device ID and four-byte port ID.

To enhance scalability



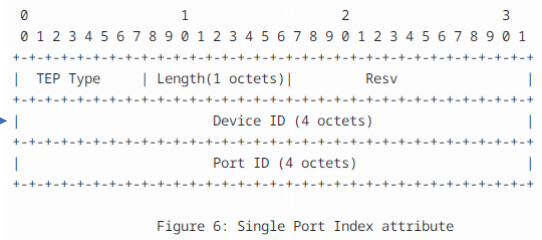
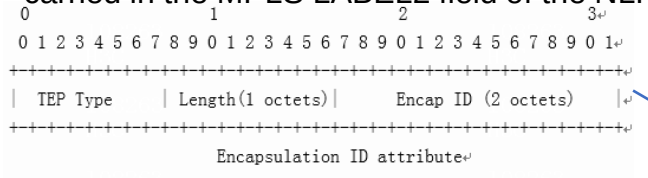


Figure 6: Single Port Index attribute

Change the independent encapsulation information attribute to be carried in the MPLS LABEL2 field of the NLRI.



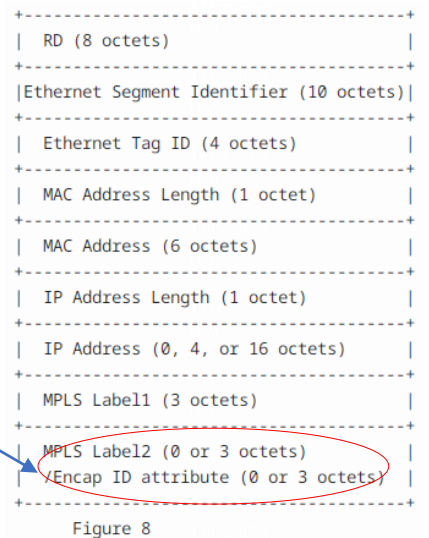


Figure 8

Running Code

Lab Interop-test Status

Currently, we have implemented the BGP functions based on EVPN extensions using the Broadcom JR2 chip and have completed laboratory functionality testing.

By Q4 of this year, the development of devices based on the JR3 chip will also be completed and interoperability testing between different vendors will be conducted in the partner's laboratory.

Currently planned device models for interoperability testing:

- 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

THANKS