draft-sajassi-bess-l3-optimized-irb-03.txt

A. Sajassi (Cisco), C. Wang (Cisco),
K. Ananthamurthy (Cisco),
J. Rabadan (Nokia)

IETF 123, July 2025
Madrid

History

- Rev 00 (Presented in IETF 118)
  - Overall solution
- Rev 01
  - L3-Optimized IRB Option B – ARP Packet Optimization
- Rev 02 (presented in IETF 119)
  - New Caveats and IPv6 ND Message Handling
- Rev 03 (Current)
  - L3-Optimized IRB Option C – BGP Control Plane discovery

EVPN L3-Optimized IRB (Rev 00)

- H1 & H2 are in one subnet (MAC-VRF1), and H3 in a different subnet (MAC-VRF2)
- H1 & H2 is bridged, H1/H2 and H3 is routed
- With L3-Optimized IRB, communication between H1 & H2 is also routed

```
                                               PE2
                            +---------+   +-------------+
               PE1          |         |   | (Mgw, IPgw) |
         +-------------+    |         |   |  IP-VRF1    |
         | (Mgw, IPgw) |    |         |   |     |       |
         |  IP-VRF1    |    |  MPLS/  |---|  MAC-VRF1   |---H2
         |     |       |----|  VXLAN/ |   +-------------+ (M2/IP2)
    H1---|  MAC-VRF1   |    |  NVGRE  |
(M1/IP1) |             |    |         |        PE3
         +-------------+    |         |   +-------------+
                            |         |---| (Mgw, IPgw) |
                            |         |   |   IP-VRF1   |
                            |         |   |      |      |
                            +---------+   |  MAC-VRF2   |---H3
                                          +-------------+ (M3/IP3)
IRB Gateway Anycast MAC: Mgw
IRB Gateway Anycast IP: IPgw

          Figure 1: IRB Model with Distributed IRB Gateways
```

PE devices only need to maintain MAC addresses for locally-connected IP hosts, thus improving MAC scalability of customer bridges and PE devices significantly

L3-Optimized IRB Option A (Rev 00)

- Unconditional ARP Reply and host discovery via data packet gleaning
  - An L3-Optimized PE replies to an ARP Request Message from a locally connected host unconditionally

```
H1                PE1             RR              PE2           PE3       H2
| ARP REQ(BC)    |               |                |             |
|--------------->| RT-2(H1)      |                |             |
|                |-------------->| RT-2(H1)       |             |
|                |               |--------------->|             |
|                |               |---------------------------->|
| ARP REP (UC)   |               |                |             |
|<---------------|               |                |             |
(IP2:Mgw)

ARP REQ (BC): Broadcast ARP REQUEST
ARP REP (UC): Unicast ARP REPLY

                Figure 2: ARP Request from an IP Host
```

L3-Optimized IRB Option B (Rev 01)

- Conditional ARP response and host discovery via ARP Request Optimization
  - An L3-Optimized PE replies to an ARP Request Message from a locally connected host ONLY IF the target host is known to the PE
  - If the target host is unknown, the PE will NOT respond to the ARP Request immediately, instead, the PE re-originates the ARP Request to discover the target host

```
H1              PE1          RR             PE2          PE3      H2
| ARP REQ(BC)    |                           |             |        |
|--------------->|                           |             |        |
|                | H2 unknown, ARP REQ (BC)  |             |        |
|                |-------------------------->|-------------------->|
|                |-------------------------------------->|-->       |
|    ARP REQ(BC) |                           |             |        |
|      <---------|                           |             |        |
ARP REQ (BC): Broadcast ARP REQUEST
ARP REP (UC): Unicast ARP REPLY
```

Figure 5: Host discovery via ARP Request Optimization (Option B)

New Caveats and IPv6 support (Rev 02)

- New Caveats
  - IPv4 Address Conflict Detection (ARP Probe in RFC5227) and Neighbor Unreachability Detection (RFC4861) received from a traditional IRB PE are NOT allowed to reach target host to avoid polluting CE MAC Table.
- MAC-Only and MAC-IP Routes for L3-Optimized IRB mode
  - PE SHOULD NOT advertise MAC-Only routes
  - PE MUST advertise MAC-IP routes with L3-Optimized IRB Flag
  - If a PE chooses to advertise both MAC-Only and MAC-IP routes, then both MAC-Only and MAC-IP routes MUST carry L3-Optimized IRB Flag
- IPv6 ND Message Handling
  - The R (Router), O (Override) and S (Solicited) flags on the NA message from PE to CE, and the RO flags on ARP/ND EC (RFC 9047)

L3-Optimized IRB Option C (Rev 03)

- Host Discovery via BGP Control Plane
  - Option C follows the same procedures as described in Option B, except, if the target host is unknown, the PE originates a Host Discovery Route to discover the target host

```
H1             PE1          RR             PE2          PE3            H2
| ARP REQ(BC)   |                          |             |              |
|-------------->|                          |             |              |
|               | H2 unknown, ARP REQ (BC) |             |              |
|               | If Layer 2 Forwarding is not off, flood across fabric |
|               |------------------------->|--------------------------->|
|               |---------------------------------------->|-->          |
|               | Host Discovery Route     |             | ARP REQ (BC) |
|               |------------------------->|------------>|------------->|
|               | Host Discovery Route     |             | ARP REQ (BC) |
|               |---------------------------------------->|---->        |
|    ARP REQ(BC)|                          |             |              |
|      <--------|                          |             |              |
ARP REQ (BC): Broadcast ARP REQUEST
ARP REP (UC): Unicast ARP REPLY

      Figure 6: Host discovery via BGP Control Plane (Option C)
```

L3-Optimized IRB Option C (Rev 03)

- Host Discovery Route
  - Carrying the Target Host IP Address
  - Carrying the Route-Target of the stretched Layer 2 topology
  - The format of the route will be defined in the future.
- Advantages with Option C
  - Option C utilizes BGP control plane for host discovery, so it gives the flexibility to turn off the Fabric Layer 2 Forwarding entirely
    - No L2 Flooding and no L2 Unicast forwarding in data plane
    - No L2VNI needed in data plane forwarding

L3-Optimized IRB Option C (Rev 03)

- Interop between Option C PE and PEs in other mode
  - When interoperating Non-Option-C PEs, the fabric Layer 2 Forwarding / Flooding must be enabled (Option B/C).
  - The receiving PE is responsible for concurrent local ARP handling.

```
 H1              PE1           RR              PE2              PE3        H2
 |         (Optimized IRB)     |        (Optimized IRB) (Traditional IRB)  |
 |           (Option B/C)      |           (Option B)           |          |
 |               |             |               |                |          |
 | ARP REQ (H2)  |             |               |                |          |
 |-------------->|             |               |                |          |
 |  ARP REQ (H2) |             |               |                |          |
 |         <-----|             |               |                |          |
 |               |ARP REQ(H2)  |               |                |          |
 |               |---------------------------->|--------------------------->|
 |               |ARP REQ(H2) |                |                |          |
 |               |--------------------------------------------->|-->       |
 |               |Host Discovery Route (H2)    |                |          |
 |               |---------------------------->|                |          |
 |               |Host Discovery Route (H2)    |                |          |
 |               |--------------------------------------------->|          |
 |               |             |               |                |          |
```

Next Steps

- Solicit more input on the mailing list
- Asking for WG call

THANK YOU!