Stateless Multicast with Bit Indexed Explicit Replication

IJisbrand Wijnands – Distinguished Engineer
BRKIPM-2239
Agenda

• Introduction
• Solution Overview
• Encapsulation
• Sets and Areas
• Forwarding
• ECMP
• BIER Overlay
• Deployment Scenarios
• Conclusion
Introduction
BIER history

• A team was formed to investigate solutions for multicast in the context of Segment Routing.

• Encoding a Sourced routed Multicast tree path using MPLS labels is difficult.

• The packet header would get very large, and its very hard to parse such header.
The BIER Epiphany

• Only encode the end-receivers (as a Bit Position) in the packet header.
  • Not the intermediate nodes.

• Encode the Bit String in the packet header.
  • Using some sort of encapsulation.

• Create a Bit Forwarding Table on all BIER nodes to allow multicast packet forwarding using the Bit String in the packet.
  • Derived from the RIB, SPF based.

• We call it, Bit Indexed Explicit Replication (BIER).
Tech Fund

• Chambers Tech Fund to prototype the idea Jan 2014.
• Core TF team: Greg Shepherd, Neale Ranns, IJsbrand Wijnands.
• Goal: build a prototype on a real router platform.
• We have a working prototype code on XRVR, VPP, CRS, ASR1K and ASR9K.

Planned Release
• ASR9K - IOS-XR 6.4.1
• ASR1K - IOS-XE 16.6S (July 2017 FCS)
IETF

• The BIER idea was presented in a BOF at the IETF in Hawaii.
  • November 2014.

• A new BIER Working Group has been formed (bier@ietf.org)

• Vendors collaborating (co-authoring) with us;
IETF drafts

- draft-ietf-bier-problem-statement
- draft-ietf-bier-architecture
- draft-ietf-bier-encapsulation-mpls
- draft-ietf-bier-use-cases
- draft-ietf-l3vpn-mvpn-bier
- draft-ietf-ospf-bier-extensions
- draft-ietf-bier-isis-extensions
- draft-xu-idr-bier-extensions
Solution Overview
1. Assign a unique Bit Position from a BitString to each Edge router in the BIER domain.
2. Each Edge floods their Bit Position - Prefix mapping using the IGP (OSPF, ISIS)
1. Assign a unique Bit Position from a BitString to each BFER in the BIER domain.
2. Each BFER floods their Bit Position to BFR-prefix mapping using the IGP (OSPF, ISIS)
Bit Index Forwarding Table

- Based on shortest path route to RID, the Bit Mask Forwarding Table is created
Suppose A learns about D’s interest, in the blue multicast flow. (via BGP, SDN, STATIC, etc…)
Suppose A learns about D and E’s interest, in the blue multicast flow.
(via BGP, SDN, STATIC, etc…)

© 2016 Cisco and/or its affiliates. All rights reserved. Cisco Public
Suppose A leans about D, E and F's interest, in the blue multicast flow. (via BGP, SDN, STATIC, etc…)
Forwarding Packets

• As you can see from the previous slides, the result from the bitwise AND (&) between the Bit Mask in the packet and the Forwarding table is copied in the packet for each neighbor.

• This is the key mechanism to prevent duplication.

• Look at the next slide to see what happens if the bits are not reset

• If the previous bits would not have been reset, E would forward the packet to C and vice versa.
Forwarding Packets
Encapsulation
How many Bits and Where?

• The number of multicast egress routers that can be addressed is depending on the number of Bits that can be included in the BitString

• The BitString length is dependent on the encapsulation type and router platform.

• We’ve analyzed the MPLS option, CRS, ASR1K and ASR9K platform.

• Both these platforms can do 256 bits.

• Other vendors confirmed they can do 256.

• We identified 5 different encoding options, most attractive below;
  1. MPLS, below the bottom label and before IP header.
  2. IPv6, extensions header.
MPLS encapsulation

• The Top Label is allocated by BIER from the downstream platform label space.
• The BIER Header follows directly below the BIER label.
• There is a single BIER label on top, unless the packet is re-encapsulated into a unicast MPLS tunnel.
• The VPN label is allocated from the upstream context label space (optional).
BIER Header

256bits = 32bytes

• [Link](http://www.ietf.org/id/draft-ietf-bier-mpls-encapsulation-01.txt)
Sets and Areas
To increase the scale we group the egress routers in Sets.

Note, Bit Positions 1,2,3 appear in both Sets, and do not overlap due to Sets.

Note, we create different forwarding entries for each Set.
BIER Sets

- There is no topological restriction which set an egress belongs to

<table>
<thead>
<tr>
<th>Set</th>
<th>BM</th>
<th>Nbr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0111</td>
<td>I</td>
</tr>
<tr>
<td>2</td>
<td>0111</td>
<td>I</td>
</tr>
</tbody>
</table>

Note, we create different forwarding entries for each Set
BIER Sets

• If a multicast flow has multiple receivers in different Sets, the packet needs to be replicated multiple times by the ingress router, for each set once.

• Is that a problem? We don’t think so…

• The Set identifier is part of the packet.

• Can be implemented as MPLS label.
BIER Areas

- BIER area’s are like IGP area’s, you only need reachability to the ABR
BIER Areas

• The ABR removes the BIER header from Area 0, and imposes a new BIER header for Area 1 and 2.

• The new BIER header can be determined by a Group/Label lookup.
  • Look for the inner IPv4/6 packet group address, do a lookup in the MFIB
  • Requires flow state on the ABR.

• Similar to Segmented Inter-AS MVPN
Sub-domains
BIER Sub-domains

- Nodes are configured to be in a BIER sub-domain.
- BIER MPLS labels are not installed between sub-domains
BIER Sub-domains

- When IGP changes SPF for B to F, traffic is dropped.
- BIER traffic is never leaked between sub-domains.
- Routers can be part of multiple sub-domains.
- This is a very easy mechanism to force dual plane.
The BFR-id
The BFR-id

• A BFER is uniquely identified by a two tuple {Set ID, Bit Position}

• The number of Bit Positions is depending on the length of the support BitString in the network.

• To make the BFER identifier independent of the BitString length we defined the BFR-id as a number between [1,65535]

• We auto-generate the BFR-id into a {Set ID, Bit Position}, based on the BitString Length.
The BFR-id

• **Formula:**
  
  \[ SI = \frac{\text{BFR-id} - 1}{\text{BitStringLength}} \]
  
  \[ BP = ((\text{BFR-id} - 1) \mod \text{BitStringLength}) + 1 \]

• **Example BFR-id = 129.**
  
  BitString Length 128 -> \{SI = 1, BP = 1\}
  
  BitString Length 256 -> \{SI = 0, BP = 129\}
The BFR-id

• By decoupling the BFR-id from the SetID and Bit Position tuple, the BFR identifier is agnostic to the supported BitString length in the network.

• This is very useful during migration.

• If a network supports multiple BitString lengths, an egress router only needs one BFR-id, and is reachable via each BitString length.
BIER Forwarding

We define two different forwarding methods for BIER.

1. Neighbor based forwarding.
2. Bit Indexed forwarding
BIER Forwarding, neighbor based.

- A packet BitString is matched against each Neighbor in the BFT.
- This model works well for systems that have multi-cores.
- Its works less well for serialized processing as it requires a neighbor walk for each packet.
BIER Forwarding, Bit Indexed

- We translate the BFT neighbor table into a table sorting on Bit Position (so not by neighbor)
- We walk the Bit Mask in the packet and Index into the FIB table.
BIER Forwarding, Bit Indexed

- We walk the Bits in the packet, as soon as we hit a ‘1’, we copy the packet, index into the FIB table with the position of the Bit.
- The Bit Mask entry is reverse ‘&’ with the Bit Mask in the packet.
- This resets the Bits that where processed.
BIER Forwarding, Bit Indexed

- We walk the Bits in the packet, as soon as we hit a ‘1’, we copy the packet, index into the FIB table with the position of the Bit.
- The Bit Mask entry is reverse ‘&’ with the Bit Mask in the packet.
- This resets the Bits that were processed.
BIER Forwarding, Bit Indexed

• Walking the bits in a Bit String takes less clock cycles compared to walking a list of neighbors.

• For that reason its faster to walk the Bit String and index into the neighbor table.

• The table is a NxN bit matrix, where N is the Bit String length.

• Bits that where already processed are reset so we don’t processes them if they appears later in a Bit String. This way we avoid multiple copies being forwarded.
BIER Forwarding, Bit Indexed

- Walking the Bit String in the packet is basically a repeating of ‘Find First Set’ (FFS) Bit operation.
- Could be optimized to record the last position.
- [http://graphics.stanford.edu/~seander/bithacks.html](http://graphics.stanford.edu/~seander/bithacks.html)
  - Checkout Bruijn sequence (Count the consecutive zero bits)
- Is supported in compilers FFS and can be done in HW fairly easily.
ECMP
ECMP

- It is possible the same Bit Position is reachable over different interfaces if there are ECMP paths.

- We distinguish two different ECMP behaviors
  - ECMP via parallel interfaces to a single neighbor.
  - ECMP via different neighbors.

- The handling is different for both cases.

- If the ECMP path is going to a single neighbor, the Bit Mask is the same for each candidate path, no special processing needed.

- If the ECMP path is to different neighbors, the Bit Mask will be different for each neighbor because the Bit String in the packet is the end result of the lookup in BIER, which is different for each nbr.
Duplicate bit positions need to be resolved, ECMP logic needs to select based on Hash. In the example we selected C.
ECMP

- We distribute the Bit Positions over multiple tables.
- Each Bit Position only appears once in each table.
- Table selection is based on entropy, done before Bit Index lookup.
ECMP

• The number of tables is depending on the number of ECMP paths to different neighbors.

• If the max number of ECMP paths is 4, but there is a Bit Position reachable via 3 paths, this will cause unequal distribution.
BIER Overlay
BIER Overlay

- One of the deployment challenges of BIER is HW support.
- Some routers may be delayed, others may never support it.
- We can run BIER as an overlay using the existing network.
The Next BFR in the path
The Next BFR in the path

• When sending a BIER packet from ingress to egress, some nodes in the path may not support BIER.

• If a node is just pass-through and not doing any replication, it’s ok to skip this node and go to the next BIER node in the path.

• We identified how to modify the SPF algorithm in order to find the next BIER node in the path and tunnel to it (easy with MPLS).

• This solution is beneficial if some routers don’t support BIER, we do expect ingress and egress routers to support BIER.
The next BFR in the path

- With enhanced IGP calculation, A can calculate next BIER node in the path.
- C is installed as an alternate next-hop in the RIB at A.
BIER on a stick
BIER on a stick

- With BIER on a stick, a Fullstack BIER capable router is connected to a non-BIER capable router.
- The BIER router advertises itself through the IGP as being connected to the non-BIER router.
- When calculating the path towards the destination, the SPF is modified such that the BIER router appears to be on the SPF (instead of the non-BIER router)
BIER on a stick

- The SPF on A for E is through B, the modified SPF logic knows G is the BIER router for B.
- G will deal with BIER traffic on behalf of B. Same as D for C
- G and G are Fullstack BIER capable, as control plane and forwarding plane.
- It looks like G is connected to B through a zero cost Tunnel.
BIER on a stick

- A BIER on a stick router can also be used to replace the PE.
- IP traffic tunneled into the BIER router, BIER traffic comes out.
- Same on the egress, BIER traffic is tunneled in, IP traffic comes out.
- This will remove Multicast from the core completely.
BIER on a stick

- The IP multicast traffic on Ingress A is pushed to H. This requires changes on the IP Multicast side (either config or code).
- On the Egress side the RPF instance has to point to the BIER node in order to accept the IP traffic coming down.
Deployment Scenarios
Native BIER
Native BIER

• With Native BIER there is NO PIM involved, just IGMP and BIER.

• The Source and Receiver(s) are connected to BIER router.

• There are no RP’s.

• There is no equivalent of PIM modes, like sparse, ssm, bidir etc..

• We speak of ‘single’ sender and ‘multi’ sender, which is basically the same solution.

• The overlay can be BGP or SDN based.
Native BIER

- E and F announce their Group membership via overlay to all other routers.
- A BIER router connected to the Source can immediately start sending.
Native BIER

- When B learns about a new source, it can immediately start sending.
MVPN over BIER
MVPN over BIER

- BIER replaces PIM, mLDP, RSVP-TE or IR in the core.
- BIER represents a full mesh (P2MP) connectivity between all the PE’s in the network.
- There is no need to explicitly signal any MDT’s (or PMSI’s).
- With MVPN there are many profiles,
  - This is partly due to the tradeoff between ‘State’ and ‘Flooding’.
  - Different C-multicast signaling options.
- MVPN over BIER, there is one profile.
  - BGP for C-multicast signaling.
- No need for Data-MDTs.
MVPN over BIER

- The BGP control plane defined for MVPN can be re-used.
- Big difference, there is no Tree per VPN…!!!
- The BIER packets needs to carry Source ID and upstream VPN context label
Conclusion
Stateless

- There is no Multicast receiver or flow state in the core network (only edge).
  - Imposition of the BIER Header may be done by application, removes state from ingress.
- There is no tree state in the network.
- There is no tree building protocol or logic in the network.
- There is only topology state for the BFER’s, derived from unicast routing.
Scale

- Since there is no flow and tree state, converges as fast as unicast.
- Compared to Ingress Replication, saves 256x (minimum)
Simplicity

- No Reverse Path Forwarding (RPF)
- No Rendezvous Points
- No shared tree / source tree switchover
- No receiver driven tree building
  - BIER is like unicast
  - State is in the packet (like Segment Routing)
Order your BIER online at:

bier@cisco.com
Demo
Complete Your Online Session Evaluation

• Give us your feedback to be entered into a Daily Survey Drawing. A daily winner will receive a $750 Amazon gift card.

• Complete your session surveys through the Cisco Live mobile app or from the Session Catalog on CiscoLive.com/us.

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

• Demos in the Cisco campus
• Walk-in Self-Paced Labs
• Lunch & Learn
• Meet the Engineer 1:1 meetings
• Related sessions
Thank you