January 2022

# TECHNIQUES TO OPTIMIZE SOFTWARE-DEFINED ACCESS LAYER 2 BROADCAST, UNKNOWN-UNICAST AND MULTICAST (BUM) USE-CASES WITH BIT INDEX EXPLICIT REPLICATION (BIER)

Denis Neogi

# TECHNIQUES TO OPTIMIZE SOFTWARE-DEFINED ACCESS LAYER 2 BROADCAST, UNKNOWN-UNICAST AND MULTICAST (BUM) USE-CASES WITH BIT INDEX EXPLICIT REPLICATION (BIER)

AUTHORS:
Denis Neogi

## ABSTRACT

Current implementations involving Broadcast, Unknown-Unicast & Multicast (BUM) traffic in a Software-Defined Access (SDA) fabric typically rely on a traditional multicast model in which each Fabric-Edge device joins a shared Internet Protocol (IP) multicast-group having a pre-defined Multicast Rendezvous Point (RP) node or nodes—effectively establishing one large broadcast domain for all Layer 2 (L2) instances in the fabric. Techniques presented herein provide for the ability to establish a model for assigning shared groups of Virtual Extensible Local Area Network (VxLAN) Virtual Network Instances (VNIs) logically and programmatically on a Fabric-Edge device to a dedicated Bit Index Explicit Replication (BIER) SubDomain based on the provisioning characteristics of the fabric. This will allow the fabric to leverage the benefits of the BIER forwarding plane in order to improve forwarding performance, achieve scalable partitioning of the BUM domains, and reduce control-plane complexity.

## DETAILED DESCRIPTION

Existing deployments of a Software-Defined Access (SDA) fabric can achieve Broadcast, Unknown-Unicast and Multicast (BUM) functionality using a traditional any-source multicast model. Fabric-Edge (FE) devices may join an IP multicast group(s) having a pre-defined Multicast-RP node(s) in a fabric, which establishes the required multicast states in the network for BUM traffic. Failure of the Multicast-RP node may cause extensive re-computation of the multicast states and lead to delayed convergence and outages in the L2 space. In addition, FE devices and their corresponding VxLAN Virtual Network Instances (VNIs) must be carefully and logically mapped to the range of allowed IP multicast groups for BUM use-cases to avoid creating large broadcast domains.



Published by Technical Disclosure Commons, 2022

It is undesirable to unconditionally flood all such packets to all FE devices in fabric, particularly in cases where the FE devices may not even have the specific VNI configured. This unnecessarily consumes bandwidth within the network and ultimately such packets must be discarded upon receipt on the FE devices to avoid leaking packets outside the VNI space.

It is possible to directly apply Bit Index Explicit Replication (BIER) to solve the problems described above and optimize the multicast state and traffic flows in a fabric, however, a complication that remains involves the assignment and scaling of the IP multicast groups.

As with the existing multicast model, there are two options for IP multicast group assignments:

- 1 - 1 Mapping of VNI to IP multicast group: This option achieves complete segmentation per VNI. However, this is not practical as the VNI assignment space can involve up to 16 million variations, while IP multicast groups for BUM use-cases are usually restricted to a small range.
- N - 1 Mapping of VNI to IP multicast group: For this option, N different VNIs are mapped to a common IP multicast group. All traffic sent to this multicast group are unconditionally delivered to all participating edge nodes (even if an edge node is not configured to receive any traffic for a given VNI). This scheme is subject to the limitations noted above in terms of creating a large broadcast domain and sub-optimal usage of network bandwidth.

Both the existing multicast model and the standard BIER model rely on the edge nodes using Protocol Independent Multicast (PIM) and/or Internet Group Management Protocol (IGMP) signaling in order to indicate an intention to 'Join' or 'Leave' a multicast group. Based on this signaling and local device configurations, the multicast infrastructure creates and maintains all the source and group (S, G) entries in the network.

Another potential option to facilitate IP multicast group assignment is described in the Internet Engineering Task Force (IETF) Draft: https://datatracker.ietf.org/doc/html/draft-wang-bier-vxlan-use-case-02. This potential option utilizes Multicast Listener Discovery (MLD) and IGMP extensions to advertise each edge node's VNI membership in the fabric on a per-edge node and per-VNI basis. However,

this option may become difficult to handle in the control-plane in terms of the volume of advertisements and unique encapsulation strings per VNI as number of edge nodes and VNIs in the fabric are scaled up.

Presented herein is an alternative technique for deploying BIER with SDA or Ethernet Virtual Private Network (EVPN) implementations and translating fabric topology characteristics such that:

- Sets of common VNIs on all edge devices in a fabric are mapped to dedicated BIER SubDomains.
- PIM / IGMP Join / Leave signaling is not required in the fabric in the context of BUM use-cases and, therefore, does not involve IP multicast groups.
- BIER SubDomain memberships are computed and provisioned on the edge devices (via a network controller) and advertised via standardized IGP extensions for BIER.
- BIER forwarding concepts can be leveraged to establish a reduced state multicast forwarding layer in the core fabric network with performance and convergence comparable to unicast forwarding.

As shown below in the sample configuration snippet of Figure 1, the common deployment model is to have all FE devices join a shared IP multicast group (e.g., 239.1.1.1 for all VNIs). During operation, BUM traffic originating from an ingress FE device is to traverse the multicast-tree to reach all other FE devices participating in the shared multicast-group.


*Figure 1: Example Network Configuration Model*

This model can be improved upon by mapping the BUM topology characteristics of the SDA (or EVPN) fabric to a BIER capable multicast domain. For such a solution, various control-plane requirements need to be satisfied. For example, each FE device is to be allocated a unique identifier (ID) in the fabric. In BIER terminology, this will be referred to as the Bit Forwarding Router-ID (BFR-ID). Further, the number of FE devices in the fabric is to be known in order to select a sufficiently large BitStringLength (and SetIndex)

mapping, which identifies each Bit Forwarding Egress Router (BFER) with a unique bit position in the encapsulation string. Additionally, a list of all BUM enabled VNIs programmed on FE devices in the SDA (or EVPN) fabric is to be known.

It should be noted that loopback addresses advertised in the fabric underlay (such as Locator ID Separation Protocol (LISP) Resource Locators (RLOCs) or EVPN Virtual Tunnel Endpoints (VTEPs)) can uniquely identify an FE device in the fabric. The fabric maximum transmission unit (MTU) is also to be set to a sufficiently large value (e.g., 9100 bytes) so that addition of BIER encapsulation to select packet flows (even with larger Bit String Length values) should not result in packet fragmentation and/or be detrimental to fabric performance. Such control-plane requirements can be derived from device configurations while provisioning the fabric and FE devices (e.g., via a network controller, a LISP control-plane node, etc.).

A point of novelty for the technique presented herein rests in the fact that the characteristic of the BUM domains in a fabric can be logically and programmatically modelled as dedicated BIER SubDomains. As a result, there will be no need to explicitly assign or maintain any IP multicast group(s) for the purposes of BUM traffic.

Consequently, FE devices will not need to use PIM / IGMP based Join / Leave signaling in the underlay in order to maintain the multicast state, as this will be handled by explicitly advertising BIER SubDomain memberships via established routing protocol extensions for BIER.

To establish this model, a shared BIER subdomain can be defined in the context of this proposal as a grouping of Fabric-Edges having a common set of BUM-enabled VNI. Consider the following example with 4 FE devices having the following BUM-enabled VNI configuration:

Edge1: VNI1, VNI2, VNI3, VNI4, VNI5, VNI6, VNI7, VNI8

Edge2: VNI1, VNI2, VNI3, VNI4, VNI5

Edge3: VNI1, VNI2, VNI3, - , -

Edge4: VNI1, VNI2, VNI3, - , - , VNI6, VNI7, VNI8

The BIER SubDomains and associated FE devices would therefore be assigned as:

SD1 (VNI1, VNI2, VNI3): Edge1, Edge2, Edge3, Edge4

SD2 (VNI4, VNI5): Edge1, Edge2

SD3 (VNI6, VNI7, VNI8): Edge1, Edge4

Each BFR in the Fabric would therefore need to advertise its BIER SubDomain associations as follows:

Edge1 (BFR1) - SD1, SD2, SD3

Edge2 (BFR2) - SD1, SD2

Edge3 (BFR3) - SD1

Edge4 (BFR4) - SD1, SD3

Ideally, a controller (or mapping system) having full visibility of the fabric deployment could be utilized to enable automation of the entire process of dynamically computing optimal FE groups and applying all necessary configuration to the FEs. In some instances, a network operator may opt to assign a set of VNIs to a BIER SubDomain using any other preferred method, but it may be difficult to ensure this SubDomain grouping is optimal without further re-configurations based on other VNIs being deployed/removed. As long as any given VNI is consistently mapped to same SubDomain ID on all FEs in the Fabric, the packets will be delivered correctly but may not necessarily be in the most optimal way.

From FE perspective at the configuration stage, the FE only needs to know the relationship between a set of VNIs and a BIER SubDomain. All FEs communicating in the context of this shared set of VNIs are to advertise their membership in the selected BIER SubDomain.

The BIER state can be advertised into the fabric by routing protocols supporting BIER extensions (e.g., as described in IETF Request for Comments RFC8444/RFC8401). As a result, the fabric learns the location and reachability for all Edges (BFRs) participating in any given BIER SubDomain and can compute the BIER forwarding state at each node.

A new FE device on-boarded to the fabric (and depending on its BUM Enabled VNI associations) will only affect the BIER SubDomains to which the FE device is to be added. All other BIER SubDomains remain unaffected if there is no change in VNI / FE device memberships.

On an ingress FE device, for the purposes of applying BIER encapsulation to the payload, it is necessary to know the BIER SubDomain to which to send a given packet. Since any given VNI can only participate in one BIER SubDomain, the ingress FE device only needs to know this VNI to BIER SubDomain mapping.

This proposal would also make the BIER SubDomain a configurable attribute of the dynamically allocated L2LISP0.X interface (or any other forwarding construct used for the purposes of VxLAN encapsulation) corresponding to the L2 VNI such that the instance may be moved from group to group based on optimal SubDomain computation. The process of mapping a VNI to a BIER SubDomain on an FE does not involve BIER.

For SDA deployments involving LISP, an L2LISP0.X sub-interface can be configured for every configured L2 VNI. This new BIER SubDomain mapping can be programmed as an attribute of the respective interface. For example, the L2LISP interface to BIER SubDomain mappings could be defined as:

L2LISP0.1, L2LISP0.2, L2LISP0.3 -> SD1

L2LISP0.4, L2LISP0.5 -> SD2

L2LISP0.6, L2LISP0.7, L2LISP0.8 -> SD3

BUM traffic originating from any of these interfaces (or their associated VLANs) would need to broadcast only to FE devices participating in the same BIER SubDomain. The BIER encapsulation for packets by a given FE device would therefore need to specify the SubDomain and encode the list of BFERs (excluding itself) participating in this SubDomain into a corresponding BIER BitString.

The existing model of assigning a common multicast group for a subset of VNIs in an SDA may not be optimal in all cases, as it does not necessarily consider the endpoints of the flows. If even one of the VNIs in the subset is provisioned on an FE, then it will receive all BUM packets for all VNIs in the common multicast group.

In contrast, the approach provided in this proposal provides for partitioning the space based on a group of FEs communicating in the context of a shared set of VNIs.

In this model, the BIER SubDomain becomes the functional equivalent of a multicast group, as it would describe completely all interested recipients of the flow. Therefore, each FE only needs to maintain a single "broadcast" BitString per BIER SubDomain of which it is a member. This "broadcast" BitString at the ingress FE sets the ID of all BFERs (excluding itself) known to participate in that SubDomain as announced by IGP advertisements with BIER extensions.

The configuration stage described above is what determines the relationship between VNIs and BIER SubDomains. This relationship may change dynamically based on the configurational state of the network. In the scope of a BIER SubDomain at any given FE, the addition/removal of a recipient of the multicast flow only involves set/unset of a bit (corresponding to the BFER) in the BitString.

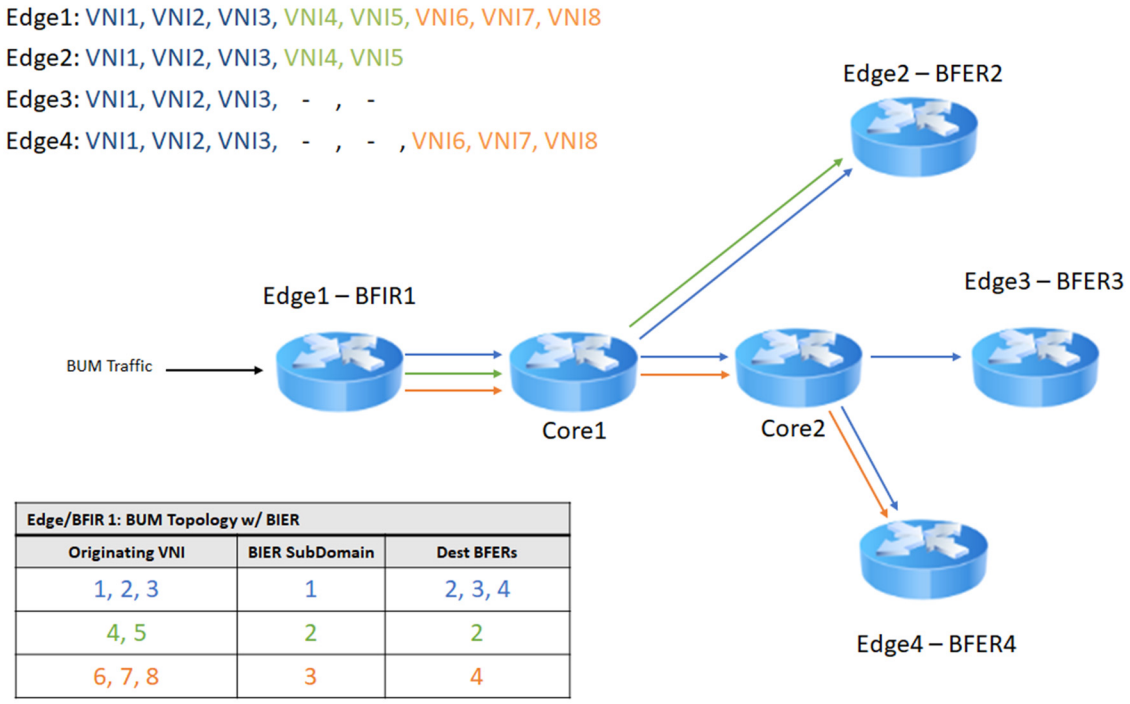
To complete the forwarding state, the FE then also computes the BIER forwarding chains in the Forwarding Information Base (FIB) based on reachability information derived from same IGP advertisement. This uses the standard process as described in the BIER forwarding specifications (i.e., as prescribed per RFC 8279 and RFC 8296).

All FEs should advertise the same BFR-ID for all SubDomains in which it participates (including the default SubDomain 0). Since all FEs in the fabric participate in SubDomain 0, the forwarding optimization would be to share and use the forwarding chains from SubDomain 0 across all other SubDomains at the FE. In this manner, forwarding chains only need to be recomputed if an FE is added/removed from the fabric. At forwarding time, the data-plane can derive the SubDomain BitString from the originating L2LISP0.X interface, apply the encapsulation, and forward packets along the shared forwarding chains.

It should be noted that once all FE + BIER SubDomain associations are advertised and known, this BIER encapsulation becomes constant, as there are no group specific receivers to encode, rather it is always a broadcast within any given SubDomain. As the payload of the BIER packet would be User Datagram Protocol (UDP) + VxLAN

encapsulated data, the protocol field in the BIER header would specify 'UDP' and the VxLAN header would specify the originating L2 VNI.

Figure 2, below, illustrates an example packet flow involving the various BIER subdomain assignments of this example.




| Edge/BFIR 1: BUM Topology w/ BIER | | |
|---|---|---|
| Originating VNI | BIER SubDomain | Dest BFERs |
| 1, 2, 3 | 1 | 2, 3, 4 |
| 4, 5 | 2 | 2 |
| 6, 7, 8 | 3 | 4 |

*Figure 2: Example Packet Flow*

With the BIER encapsulation applied, the packet can be delivered to all BFERs indicated in the BIER BitString (for the corresponding BIER SubDomain) following the established BIER and unicast forwarding states in the network (i.e., as prescribed per RFC 8279 and RFC 8296).

When packet is received at the BFERs, the BIER header can be removed (as packet is indicated for-us with our BFR-ID set in the BitString) and further UDP/VxLAN decapsulation can be performed before transmitting the inner payload. It should be noted that it is guaranteed that the packet will never need to be discarded as the target VNI must exist on the Egress Edge devices (BFERs) as advertised by the routing protocol.

The above example illustrates a scenario with logical partitions into multiple BIER SubDomains based on common sets of BUM enabled VNIs. However, in deployments where all/most BUM enabled VNIs may be shared between edge devices in the fabric, the

controller/user may choose to segment the L2 broadcast-domains based on user-defined policies and configurations. The end-result in either scenario is a reduction in the overall volume of BUM traffic within the fabric, as packets can be delivered much more selectively and only to registered FE devices.

For use-cases involving deployment of a new VNI at a given FE (assuming the same VNI will also be deployed at several other FEs in the fabric), two potential scenarios may be considered.

(1) There already exists a BIER SubDomain for BUM traffic between all the same FEs. In this case, when the VNI is provisioned at the FEs, the VNI can be mapped to the existing SubDomain. No further IGP advertisements are needed in this scenario as the existing BitString / forwarding chains for the SubDomain already contains all the information needed to deliver the traffic to all interested recipients; or

(2) If no such SubDomain currently exists, then a new SubDomain can be allocated and the new VNI can be mapped to the new SubDomain on all relevant FEs. The FEs in this new SubDomain can then advertise their membership via the BIER IGP extensions, thus allowing computation of the BitString.

For example, with reference to the example as illustrated in Figure 2, if VNI4 is deployed on FE3 and FE4, then scenario (1) is implemented and VNI4 should be merged into SubDomain1 on all FEs using VNI4. In another example, if VNI4 is deployed to FE3 only, then scenario (2) is implemented and VNI4 should be moved to some newly allocated SubDomainX on all FEs using VNI4

In summary, the novelty of this proposal to partition the broadcast domains in a fabric as dedicated BIER SubDomains, when coupled with the BIER forwarding plane may provide for the ability to achieve various improvements over conventional multicast implementations. For example, scalable partitioning and creation of BUM traffic groups may minimize bandwidth usage by BUM traffic flows and associated multicast state advertisements in the fabric. Further, the techniques presented herein are not dependent on IP multicast groups, multicast-RP nodes, or associated PIM/IGMP signaling

mechanisms as all BIER specific states can be computed and provisioned to the fabric by a network controller. No reserved multicast groups are used in the SDA for BUM use-cases and therefore, no associated multicast protocol states or signaling mechanisms are used in order to implement the techniques prescribed herein.

Computations of optimal FE grouping and VNI assignments to BIER SubDomains can be automated in order to ensure packet flows are delivered only to interested recipients for any given VNI. Further, optimized traffic flows can be provided within the core network such that packets may only reach edge devices participating in the same BIER SubDomain (i.e., shared set of VNIs) and BIER packets may be replicated only on ports where needed in order to reach a target BFER(s). Accordingly, techniques of this proposal may provide a level of network performance/convergence that is comparable to unicast forwarding.