

MBONED  
Internet-Draft  
Intended status: Informational  
Expires: August 7, 2020

M. McBride  
Futurewei  
O. Komolafe  
Arista Networks  
February 4, 2020

Multicast in the Data Center Overview  
draft-ietf-mboned-dc-deploy-09

## Abstract

The volume and importance of one-to-many traffic patterns in data centers is likely to increase significantly in the future. Reasons for this increase are discussed and then attention is paid to the manner in which this traffic pattern may be judiciously handled in data centers. The intuitive solution of deploying conventional IP multicast within data centers is explored and evaluated. Thereafter, a number of emerging innovative approaches are described before a number of recommendations are made.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 7, 2020.

## Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

<a href="#">1.</a>	Introduction . . . . .	<a href="#">2</a>
<a href="#">1.1.</a>	Requirements Language . . . . .	<a href="#">3</a>
<a href="#">2.</a>	Reasons for increasing one-to-many traffic patterns . . . . .	<a href="#">3</a>
<a href="#">2.1.</a>	Applications . . . . .	<a href="#">3</a>
<a href="#">2.2.</a>	Overlays . . . . .	<a href="#">5</a>
<a href="#">2.3.</a>	Protocols . . . . .	<a href="#">6</a>
<a href="#">2.4.</a>	Summary . . . . .	<a href="#">6</a>
<a href="#">3.</a>	Handling one-to-many traffic using conventional multicast . . . . .	<a href="#">7</a>
<a href="#">3.1.</a>	Layer 3 multicast . . . . .	<a href="#">7</a>
<a href="#">3.2.</a>	Layer 2 multicast . . . . .	<a href="#">7</a>
<a href="#">3.3.</a>	Example use cases . . . . .	<a href="#">9</a>
<a href="#">3.4.</a>	Advantages and disadvantages . . . . .	<a href="#">9</a>
<a href="#">4.</a>	Alternative options for handling one-to-many traffic . . . . .	<a href="#">10</a>
<a href="#">4.1.</a>	Minimizing traffic volumes . . . . .	<a href="#">11</a>
<a href="#">4.2.</a>	Head end replication . . . . .	<a href="#">12</a>
<a href="#">4.3.</a>	Programmable Forwarding Planes . . . . .	<a href="#">12</a>
<a href="#">4.4.</a>	BIER . . . . .	<a href="#">13</a>
<a href="#">4.5.</a>	Segment Routing . . . . .	<a href="#">14</a>
<a href="#">5.</a>	Conclusions . . . . .	<a href="#">15</a>
<a href="#">6.</a>	IANA Considerations . . . . .	<a href="#">15</a>
<a href="#">7.</a>	Security Considerations . . . . .	<a href="#">15</a>
<a href="#">8.</a>	Acknowledgements . . . . .	<a href="#">15</a>
<a href="#">9.</a>	References . . . . .	<a href="#">15</a>
<a href="#">9.1.</a>	Normative References . . . . .	<a href="#">15</a>
<a href="#">9.2.</a>	Informative References . . . . .	<a href="#">16</a>
	Authors' Addresses . . . . .	<a href="#">18</a>

## [1.](#) Introduction

The volume and importance of one-to-many traffic patterns in data centers will likely continue to increase. Reasons for this increase include the nature of the traffic generated by applications hosted in the data center, the need to handle broadcast, unknown unicast and multicast (BUM) traffic within the overlay technologies used to support multi-tenancy at scale, and the use of certain protocols that traditionally require one-to-many control message exchanges.

These trends, allied with the expectation that highly virtualized large-scale data centers must support communication between potentially thousands of participants, may lead to the natural assumption that IP multicast will be widely used in data centers,

specifically given the bandwidth savings it potentially offers. However, such an assumption would be wrong. In fact, there is widespread reluctance to enable conventional IP multicast in data centers for a number of reasons, mostly pertaining to concerns about its scalability and reliability.

This draft discusses some of the main drivers for the increasing volume and importance of one-to-many traffic patterns in data centers. Thereafter, the manner in which conventional IP multicast may be used to handle this traffic pattern is discussed and some of the associated challenges highlighted. Following this discussion, a number of alternative emerging approaches are introduced, before concluding by discussing key trends and making a number of recommendations.

### [1.1.](#) Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC 2119](#).

## [2.](#) Reasons for increasing one-to-many traffic patterns

### [2.1.](#) Applications

Key trends suggest that the nature of the applications likely to dominate future highly-virtualized multi-tenant data centers will produce large volumes of one-to-many traffic. For example, it is well-known that traffic flows in data centers have evolved from being predominantly North-South (e.g. client-server) to predominantly East-West (e.g. distributed computation). This change has led to the consensus that topologies such as the Leaf/Spine, that are easier to scale in the East-West direction, are better suited to the data center of the future. This increase in East-West traffic flows results from VMs often having to exchange numerous messages between themselves as part of executing a specific workload. For example, a

computational workload could require data, or an executable, to be disseminated to workers distributed throughout the data center which may be subsequently polled for status updates. The emergence of such applications means there is likely to be an increase in one-to-many traffic flows with the increasing dominance of East-West traffic.

The TV broadcast industry is another potential future source of applications with one-to-many traffic patterns in data centers. The requirement for robustness, stability and predicability has meant the TV broadcast industry has traditionally used TV-specific protocols, infrastructure and technologies for transmitting video signals between end points such as cameras, monitors, mixers, graphics

devices and video servers. However, the growing cost and complexity of supporting this approach, especially as the bit rates of the video signals increase due to demand for formats such as 4K-UHD and 8K-UHD, means there is a consensus that the TV broadcast industry will transition from industry-specific transmission formats (e.g. SDI, HD-SDI) over TV-specific infrastructure to using IP-based infrastructure. The development of pertinent standards by the Society of Motion Picture and Television Engineers (SMPTE) [[SMPTE2110](#)], along with the increasing performance of IP routers, means this transition is gathering pace. A possible outcome of this transition will be the building of IP data centers in broadcast plants. Traffic flows in the broadcast industry are frequently one-to-many and so if IP data centers are deployed in broadcast plants, it is imperative that this traffic pattern is supported efficiently in that infrastructure. In fact, a pivotal consideration for broadcasters considering transitioning to IP is the manner in which these one-to-many traffic flows will be managed and monitored in a data center with an IP fabric.

One of the few success stories in using conventional IP multicast has been for disseminating market trading data. For example, IP multicast is commonly used today to deliver stock quotes from stock exchanges to financial service providers and then to the stock analysts or brokerages. It is essential that the network infrastructure delivers very low latency and high throughput, especially given the proliferation of automated and algorithmic trading which means stock analysts or brokerages may gain an edge on competitors simply by receiving an update a few milliseconds earlier. As would be expected, in such deployments reliability is critical.

The network must be designed with no single point of failure and in such a way that it can respond in a deterministic manner to failure. Typically, redundant servers (in a primary/backup or live-live mode) send multicast streams into the network, with diverse paths being used across the network. The stock exchange generating the one-to-many traffic and stock analysts/brokerage that receive the traffic will typically have their own data centers. Therefore, the manner in which one-to-many traffic patterns are handled in these data centers are extremely important, especially given the requirements and constraints mentioned.

Another reason for the growing volume of one-to-many traffic patterns in modern data centers is the increasing adoption of streaming telemetry. This transition is motivated by the observation that traditional poll-based approaches for monitoring network devices are usually inadequate in modern data centers. These approaches typically suffer from poor scalability, extensibility and responsiveness. In contrast, in streaming telemetry, network devices in the data center stream highly-granular real-time updates to a

telemetry collector/database. This collector then collates, normalizes and encodes this data for convenient consumption by monitoring applications. The monitoring applications can subscribe to the notifications of interest, allowing them to gain insight into pertinent state and performance metrics. Thus, the traffic flows associated with streaming telemetry are typically many-to-one between the network devices and the telemetry collector and then one-to-many from the collector to the monitoring applications.

The use of publish and subscribe applications is growing within data centers, contributing to the rising volume of one-to-many traffic flows. Such applications are attractive as they provide a robust low-latency asynchronous messaging service, allowing senders to be decoupled from receivers. The usual approach is for a publisher to create and transmit a message to a specific topic. The publish and subscribe application will retain the message and ensure it is delivered to all subscribers to that topic. The flexibility in the number of publishers and subscribers to a specific topic means such applications cater for one-to-one, one-to-many and many-to-one traffic patterns.

## [2.2](#). Overlays

Another key contributor to the rise in one-to-many traffic patterns is the proposed architecture for supporting large-scale multi-tenancy in highly virtualized data centers [[RFC8014](#)]. In this architecture, a tenant's VMs are distributed across the data center and are connected by a virtual network known as the overlay network. A number of different technologies have been proposed for realizing the overlay network, including VXLAN [[RFC7348](#)], VXLAN-GPE [[I-D.ietf-nvo3-vxlan-gpe](#)], NVGRE [[RFC7637](#)] and GENEVE [[I-D.ietf-nvo3-geneve](#)]. The often fervent and arguably partisan debate about the relative merits of these overlay technologies belies the fact that, conceptually, it may be said that these overlays simply provide a means to encapsulate and tunnel Ethernet frames from the VMs over the data center IP fabric, thus emulating a Layer 2 segment between the VMs. Consequently, the VMs believe and behave as if they are connected to the tenant's other VMs by a conventional Layer 2 segment, regardless of their physical location within the data center.

Naturally, in a Layer 2 segment, point to multi-point traffic can result from handling BUM (broadcast, unknown unicast and multicast) traffic. And, compounding this issue within data centers, since the tenant's VMs attached to the emulated segment may be dispersed throughout the data center, the BUM traffic may need to traverse the data center fabric.

Hence, regardless of the overlay technology used, due consideration must be given to handling BUM traffic, forcing the data center operator to pay attention to the manner in which one-to-many communication is handled within the data center. And this consideration is likely to become increasingly important with the anticipated rise in the number and importance of overlays. In fact, it may be asserted that the manner in which one-to-many communications arising from overlays is handled is pivotal to the performance and stability of the entire data center network.

### [2.3.](#) Protocols

Conventionally, some key networking protocols used in data centers require one-to-many communications for control messages. Thus, the data center operator must pay due attention to how these control

message exchanges are supported.

For example, ARP [[RFC0826](#)] and ND [[RFC4861](#)] use broadcast and multicast messages within IPv4 and IPv6 networks respectively to discover MAC address to IP address mappings. Furthermore, when these protocols are running within an overlay network, it is essential to ensure the messages are delivered to all the hosts on the emulated Layer 2 segment, regardless of physical location within the data center. The challenges associated with optimally delivering ARP and ND messages in data centers has attracted lots of attention [[RFC6820](#)].

Another example of a protocol that may necessitate having one-to-many traffic flows in the data center is IGMP [[RFC2236](#)], [[RFC3376](#)]. If the VMs attached to the Layer 2 segment wish to join a multicast group they must send IGMP reports in response to queries from the querier. As these devices could be located at different locations within the data center, there is the somewhat ironic prospect of IGMP itself leading to an increase in the volume of one-to-many communications in the data center.

#### [2.4.](#) Summary

[Section 2.1](#), [Section 2.2](#) and [Section 2.3](#) have discussed how the trends in the types of applications, the overlay technologies used and some of the essential networking protocols results in an increase in the volume of one-to-many traffic patterns in modern highly-virtualized data centers. [Section 3](#) explores how such traffic flows may be handled using conventional IP multicast.

### [3.](#) Handling one-to-many traffic using conventional multicast

Faced with ever increasing volumes of one-to-many traffic flows, for the reasons presented in [Section 2](#), it makes sense for a data center operator to explore if and how conventional IP multicast could be deployed within the data center. This section introduces the key protocols, discusses some example use cases where they are deployed in data centers and discusses some of the advantages and

disadvantages of such deployments.

### [3.1.](#) Layer 3 multicast

PIM is the most widely deployed multicast routing protocol and so, unsurprisingly, is the primary multicast routing protocol considered for use in the data center. There are three potential popular modes of PIM that may be used: PIM-SM [[RFC4601](#)], PIM-SSM [[RFC4607](#)] or PIM-BIDIR [[RFC5015](#)]. It may be said that these different modes of PIM tradeoff the optimality of the multicast forwarding tree for the amount of multicast forwarding state that must be maintained at routers. SSM provides the most efficient forwarding between sources and receivers and thus is most suitable for applications with one-to-many traffic patterns. State is built and maintained for each (S,G) flow. Thus, the amount of multicast forwarding state held by routers in the data center is proportional to the number of sources and groups. At the other end of the spectrum, BIDIR is the most efficient shared tree solution as one tree is built for all flows, therefore minimizing the amount of state. This state reduction is at the expense of optimal forwarding path between sources and receivers. This use of a shared tree makes BIDIR particularly well-suited for applications with many-to-many traffic patterns, given that the amount of state is uncorrelated to the number of sources. SSM and BIDIR are optimizations of PIM-SM. PIM-SM is the most widely deployed multicast routing protocol. PIM-SM can also be the most complex. PIM-SM relies upon a RP (Rendezvous Point) to set up the multicast tree and subsequently there is the option of switching to the SPT (shortest path tree), similar to SSM, or staying on the shared tree, similar to BIDIR.

### [3.2.](#) Layer 2 multicast

With IPv4 unicast address resolution, the translation of an IP address to a MAC address is done dynamically by ARP. With multicast address resolution, the mapping from a multicast IPv4 address to a multicast MAC address is done by assigning the low-order 23 bits of the multicast IPv4 address to fill the low-order 23 bits of the multicast MAC address. Each IPv4 multicast address has 28 unique bits (the multicast address range is 224.0.0.0/12) therefore mapping a multicast IP address to a MAC address ignores 5 bits of the IP

address. Hence, groups of 32 multicast IP addresses are mapped to



the same MAC address. And so a multicast MAC address cannot be uniquely mapped to a multicast IPv4 address. Therefore, IPv4 multicast addresses must be chosen judiciously in order to avoid unnecessary address aliasing. When sending IPv6 multicast packets on an Ethernet link, the corresponding destination MAC address is a direct mapping of the last 32 bits of the 128 bit IPv6 multicast address into the 48 bit MAC address. It is possible for more than one IPv6 multicast address to map to the same 48 bit MAC address.

The default behaviour of many hosts (and, in fact, routers) is to block multicast traffic. Consequently, when a host wishes to join an IPv4 multicast group, it sends an IGMP [[RFC2236](#)], [[RFC3376](#)] report to the router attached to the Layer 2 segment and also it instructs its data link layer to receive Ethernet frames that match the corresponding MAC address. The data link layer filters the frames, passing those with matching destination addresses to the IP module. Similarly, hosts simply hand the multicast packet for transmission to the data link layer which would add the Layer 2 encapsulation, using the MAC address derived in the manner previously discussed.

When this Ethernet frame with a multicast MAC address is received by a switch configured to forward multicast traffic, the default behaviour is to flood it to all the ports in the Layer 2 segment. Clearly there may not be a receiver for this multicast group present on each port and IGMP snooping is used to avoid sending the frame out of ports without receivers.

A switch running IGMP snooping listens to the IGMP messages exchanged between hosts and the router in order to identify which ports have active receivers for a specific multicast group, allowing the forwarding of multicast frames to be suitably constrained. Normally, the multicast router will generate IGMP queries to which the hosts send IGMP reports in response. However, number of optimizations in which a switch generates IGMP queries (and so appears to be the router from the hosts' perspective) and/or generates IGMP reports (and so appears to be hosts from the router's perspective) are commonly used to improve the performance by reducing the amount of state maintained at the router, suppressing superfluous IGMP messages and improving responsiveness when hosts join/leave the group.

Multicast Listener Discovery (MLD) [[RFC 2710](#)] [[RFC 3810](#)] is used by IPv6 routers for discovering multicast listeners on a directly attached link, performing a similar function to IGMP in IPv4 networks. MLDv1 [[RFC 2710](#)] is similar to IGMPv2 and MLDv2 [[RFC 3810](#)] [[RFC 4604](#)] similar to IGMPv3. However, in contrast to IGMP, MLD does not send its own distinct protocol messages. Rather, MLD is a subprotocol of ICMPv6 [[RFC 4443](#)] and so MLD messages are a subset of

ICMPv6 messages. MLD snooping works similarly to IGMP snooping, described earlier.

### [3.3.](#) Example use cases

A use case where PIM and IGMP are currently used in data centers is to support multicast in VXLAN deployments. In the original VXLAN specification [[RFC7348](#)], a data-driven flood and learn control plane was proposed, requiring the data center IP fabric to support multicast routing. A multicast group is associated with each virtual network, each uniquely identified by its VXLAN network identifiers (VNI). VXLAN tunnel endpoints (VTEPs), typically located in the hypervisor or ToR switch, with local VMs that belong to this VNI would join the multicast group and use it for the exchange of BUM traffic with the other VTEPs. Essentially, the VTEP would encapsulate any BUM traffic from attached VMs in an IP multicast packet, whose destination address is the associated multicast group address, and transmit the packet to the data center fabric. Thus, a multicast routing protocol (typically PIM) must be running in the fabric to maintain a multicast distribution tree per VNI.

Alternatively, rather than setting up a multicast distribution tree per VNI, a tree can be set up whenever hosts within the VNI wish to exchange multicast traffic. For example, whenever a VTEP receives an IGMP report from a locally connected host, it would translate this into a PIM join message which will be propagated into the IP fabric. In order to ensure this join message is sent to the IP fabric rather than over the VXLAN interface (since the VTEP will have a route back to the source of the multicast packet over the VXLAN interface and so would naturally attempt to send the join over this interface) a more specific route back to the source over the IP fabric must be configured. In this approach PIM must be configured on the SVIs associated with the VXLAN interface.

Another use case of PIM and IGMP in data centers is when IPTV servers use multicast to deliver content from the data center to end users. IPTV is typically a one to many application where the hosts are configured for IGMPv3, the switches are configured with IGMP snooping, and the routers are running PIM-SSM mode. Often redundant servers send multicast streams into the network and the network forwards the data across diverse paths.

### [3.4.](#) Advantages and disadvantages

Arguably the biggest advantage of using PIM and IGMP to support one-to-many communication in data centers is that these protocols are

relatively mature. Consequently, PIM is available in most routers and IGMP is supported by most hosts and routers. As such, no

specialized hardware or relatively immature software is involved in using these protocols in data centers. Furthermore, the maturity of these protocols means their behaviour and performance in operational networks is well-understood, with widely available best-practices and deployment guides for optimizing their performance. For these reasons, PIM and IGMP have been used successfully for supporting one-to-many traffic flows within modern data centers, as discussed earlier.

However, somewhat ironically, the relative disadvantages of PIM and IGMP usage in data centers also stem mostly from their maturity. Specifically, these protocols were standardized and implemented long before the highly-virtualized multi-tenant data centers of today existed. Consequently, PIM and IGMP are neither optimally placed to deal with the requirements of one-to-many communication in modern data centers nor to exploit idiosyncrasies of data centers. For example, there may be thousands of VMs participating in a multicast session, with some of these VMs migrating to servers within the data center, new VMs being continually spun up and wishing to join the sessions while all the time other VMs are leaving. In such a scenario, the churn in the PIM and IGMP state machines, the volume of control messages they would generate and the amount of state they would necessitate within routers, especially if they were deployed naively, would be untenable. Furthermore, PIM is a relatively complex protocol. As such, PIM can be challenging to debug even in significantly more benign deployments than those envisaged for future data centers, a fact that has evidently had a dissuasive effect on data center operators considering enabling it within the IP fabric.

#### [4.](#) Alternative options for handling one-to-many traffic

[Section 2](#) has shown that there is likely to be an increasing amount one-to-many communications in data centers for multiple reasons. And [Section 3](#) has discussed how conventional multicast may be used to handle this traffic, presenting some of the associated advantages and disadvantages. Unsurprisingly, as discussed in the remainder of [Section 4](#), there are a number of alternative options of handling this traffic pattern in data centers. Critically, it should be noted that many of these techniques are not mutually-exclusive; in fact many

deployments involve a combination of more than one of these techniques. Furthermore, as will be shown, introducing a centralized controller or a distributed control plane, typically makes these techniques more potent.

#### [4.1](#). Minimizing traffic volumes

If handling one-to-many traffic flows in data centers is considered onerous, then arguably the most intuitive solution is to aim to minimize the volume of said traffic.

It was previously mentioned in [Section 2](#) that the three main contributors to one-to-many traffic in data centers are applications, overlays and protocols. Typically the applications running on VMs are outside the control of the data center operator and thus, relatively speaking, little can be done about the volume of one-to-many traffic generated by applications. Luckily, there is more scope for attempting to reduce the volume of such traffic generated by overlays and protocols. (And often by protocols within overlays.) This reduction is possible by exploiting certain characteristics of data center networks such as a fixed and regular topology, single administrative control, consistent hardware and software, well-known overlay encapsulation endpoints and systematic IP address allocation.

A way of minimizing the amount of one-to-many traffic that traverses the data center fabric is to use a centralized controller. For example, whenever a new VM is instantiated, the hypervisor or encapsulation endpoint can notify a centralized controller of this new MAC address, the associated virtual network, IP address etc. The controller could subsequently distribute this information to every encapsulation endpoint. Consequently, when any endpoint receives an ARP request from a locally attached VM, it could simply consult its local copy of the information distributed by the controller and reply. Thus, the ARP request is suppressed and does not result in one-to-many traffic traversing the data center IP fabric.

Alternatively, the functionality supported by the controller can

realized by a distributed control plane. BGP-EVPN [RFC7432, [RFC8365](#)] is the most popular control plane used in data centers. Typically, the encapsulation endpoints will exchange pertinent information with each other by all peering with a BGP route reflector (RR). Thus, information such as local MAC addresses, MAC to IP address mapping, virtual networks identifiers, IP prefixes, and local IGMP group membership can be disseminated. Consequently, for example, ARP requests from local VMs can be suppressed by the encapsulation endpoint using the information learnt from the control plane about the MAC to IP mappings at remote peers. In a similar fashion, encapsulation endpoints can use information gleaned from the BGP-EVPN messages to proxy for both IGMP reports and queries for the attached VMs, thus obviating the need to transmit IGMP messages across the data center fabric.

#### [4.2.](#) Head end replication

A popular option for handling one-to-many traffic patterns in data centers is head end replication (HER). HER means the traffic is duplicated and sent to each end point individually using conventional IP unicast. Obvious disadvantages of HER include traffic duplication and the additional processing burden on the head end. Nevertheless, HER is especially attractive when overlays are in use as the replication can be carried out by the hypervisor or encapsulation end point. Consequently, the VMs and IP fabric are unmodified and unaware of how the traffic is delivered to the multiple end points. Additionally, it is possible to use a number of approaches for constructing and disseminating the list of which endpoints should receive what traffic and so on.

For example, the reluctance of data center operators to enable PIM within the data center fabric means VXLAN is often used with HER. Thus, BUM traffic from each VNI is replicated and sent using unicast to remote VTEPs with VMs in that VNI. The list of remote VTEPs to which the traffic should be sent may be configured manually on the VTEP. Alternatively, the VTEPs may transmit pertinent local state to a centralized controller which in turn sends each VTEP the list of remote VTEPs for each VNI. Lastly, HER also works well when a distributed control plane is used instead of the centralized controller. Again, BGP-EVPN may be used to distribute the

information needed to facilitate HER to the VTEPs.

#### 4.3. Programmable Forwarding Planes

As discussed in [Section 2](#), one of the main functions of PIM is to build and maintain multicast distribution trees. Such a tree indicates the path a specific flow will take through the network. Thus, in routers traversed by the flow, the information from PIM is ultimately used to create a multicast forwarding entry for the specific flow and insert it into the multicast forwarding table. The multicast forwarding table will have entries for each multicast flow traversing the router, with the lookup key usually being a concatenation of the source and group addresses. Critically, each entry will contain information such as the legal input interface for the flow and a list of output interfaces to which matching packets should be replicated.

Viewed in this way, there is nothing remarkable about the multicast forwarding state constructed in routers based on the information gleaned from PIM. And, in fact, it is perfectly feasible to build such state in the absence of PIM. Such prospects have been significantly enhanced with the increasing popularity and performance of network devices with programmable forwarding planes. These

devices are attractive for use in data centers since they are amenable to being programmed by a centralized controller. If such a controller has a global view of the sources and receivers for each multicast flow (which can be provided by the devices attached to the end hosts in the data center communicating with the controller), an accurate representation of data center topology (which is usually well-known), then it can readily compute the multicast forwarding state that must be installed at each router to ensure the one-to-many traffic flow is delivered properly to the correct receivers. All that is needed is an API to program the forwarding planes of all the network devices that need to handle the flow appropriately. Such APIs do in fact exist and so, unsurprisingly, handling one-to-many traffic flows using such an approach is attractive for data centers.

Being able to program the forwarding plane in this manner offers the enticing possibility of introducing novel algorithms and concepts for forwarding multicast traffic in data centers. These schemes typically aim to exploit the idiosyncracies of the data center

network architecture to create ingenious, pithy and elegant encodings of the information needed to facilitate multicast forwarding. Depending on the scheme, this information may be carried in packet headers, stored in the multicast forwarding table in routers or a combination of both. The key characteristic is that the terseness of the forwarding information means the volume of forwarding state is significantly reduced. Additionally, the overhead associated with building and maintaining a multicast forwarding tree has been eliminated. The result of these reductions in the overhead associated with multicast forwarding is a significant and impressive increase in the effective number of multicast flows that can be supported within the data center.

[Shabaz19] is a good example of such an approach and also presents comprehensive discussion of other schemes in the discussion on related work. Although a number of promising schemes have been proposed, no consensus has yet emerged as to which approach is best, and in fact what "best" means. Even if a clear winner were to emerge, it faces significant challenges to gain the vendor and operator buy-in to ensure it is widely deployed in data centers.

#### 4.4. BIER

As discussed in [Section 3.4](#), PIM and IGMP face potential scalability challenges when deployed in data centers. These challenges are typically due to the requirement to build and maintain a distribution tree and the requirement to hold per-flow state in routers. Bit Index Explicit Replication (BIER) [[RFC 8279](#)] is a new multicast forwarding paradigm that avoids these two requirements.

When a multicast packet enters a BIER domain, the ingress router, known as the Bit-Forwarding Ingress Router (BFIR), adds a BIER header to the packet. This header contains a bit string in which each bit maps to an egress router, known as Bit-Forwarding Egress Router (BFER). If a bit is set, then the packet should be forwarded to the associated BFER. The routers within the BIER domain, Bit-Forwarding Routers (BFRs), use the BIER header in the packet and information in the Bit Index Forwarding Table (BIFT) to carry out simple bit-wise operations to determine how the packet should be replicated optimally so it reaches all the appropriate BFERs.

BIER is deemed to be attractive for facilitating one-to-many communications in data centers [[I-D.ietf-bier-use-cases](#)]. The BFIRs are the encapsulation endpoints in the deployment envisioned with overlay networks. So knowledge about the actual multicast groups does not reside in the data center fabric, improving the scalability compared to conventional IP multicast. Additionally, a centralized controller or a BGP-EVPN control plane may be used with BIER to ensure the BFIR have the required information. A challenge associated with using BIER is that it requires changes to the forwarding behaviour of the routers used in the data center IP fabric.

#### [4.5.](#) Segment Routing

Segment Routing (SR) [[RFC8402](#)] is a manifestation of the source routing paradigm, so called as the path a packet takes through a network is determined at the source. The source encodes this information in the packet header as a sequence of instructions. These instructions are followed by intermediate routers, ultimately resulting in the delivery of the packet to the desired destination. In SR, the instructions are known as segments and a number of different kinds of segments have been defined. Each segment has an identifier (SID) which is distributed throughout the network by newly defined extensions to standard routing protocols. Thus, using this information, sources are able to determine the exact sequence of segments to encode into the packet. The manner in which these instructions are encoded depends on the underlying data plane. Segment Routing can be applied to the MPLS and IPv6 data planes. In the former, the list of segments is represented by the label stack and in the latter it is represented as an IPv6 routing extension header. Advantages of segment routing include the reduction in the amount of forwarding state routers need to hold and the removal of the need to run a signaling protocol, thus improving the network scalability while reducing the operational complexity.

The advantages of segment routing and the ability to run it over an unmodified MPLS data plane means that one of its anticipated use

cases is in BGP-based large-scale data centers [[RFC7938](#)]. The exact manner in which multicast traffic will be handled in SR has not yet been standardized, with a number of different options being considered. For example, since with the MPLS data plane, segments



are simply encoded as a label stack, then the protocols traditionally used to create point-to-multipoint LSPs could be reused to allow SR to support one-to-many traffic flows. Alternatively, a special SID may be defined for a multicast distribution tree, with a centralized controller being used to program routers appropriately to ensure the traffic is delivered to the desired destinations, while avoiding the costly process of building and maintaining a multicast distribution tree.

## 5. Conclusions

As the volume and importance of one-to-many traffic in data centers increases, conventional IP multicast is likely to become increasingly unattractive for deployment in data centers for a number of reasons, mostly pertaining to its relatively poor scalability and inability to exploit characteristics of data center network architectures. Hence, even though IGMP/MLD is likely to remain the most popular manner in which end hosts signal interest in joining a multicast group, it is unlikely that this multicast traffic will be transported over the data center IP fabric using a multicast distribution tree built and maintained by PIM in the future. Rather, approaches which exploit idiosyncracies of data center network architectures are better placed to deliver one-to-many traffic in data centers, especially when judiciously combined with a centralized controller and/or a distributed control plane, particularly one based on BGP-EVPN.

## 6. IANA Considerations

This memo includes no request to IANA.

## 7. Security Considerations

No new security considerations result from this document

## 8. Acknowledgements

## 9. References

### 9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

## 9.2. Informative References

- [I-D.ietf-bier-use-cases]  
Kumar, N., Asati, R., Chen, M., Xu, X., Dolganow, A., Przygienda, T., Gulko, A., Robinson, D., Arya, V., and C. Bestler, "BIER Use Cases", [draft-ietf-bier-use-cases-09](#) (work in progress), January 2019.
- [I-D.ietf-nvo3-geneve]  
Gross, J., Ganga, I., and T. Sridhar, "Geneve: Generic Network Virtualization Encapsulation", [draft-ietf-nvo3-geneve-13](#) (work in progress), March 2019.
- [I-D.ietf-nvo3-vxlan-gpe]  
Maino, F., Kreeger, L., and U. Elzur, "Generic Protocol Extension for VXLAN", [draft-ietf-nvo3-vxlan-gpe-07](#) (work in progress), April 2019.
- [RFC0826] Plummer, D., "An Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardware", STD 37, [RFC 826](#), DOI 10.17487/RFC0826, November 1982, <<https://www.rfc-editor.org/info/rfc826>>.
- [RFC2236] Fenner, W., "Internet Group Management Protocol, Version 2", [RFC 2236](#), DOI 10.17487/RFC2236, November 1997, <<https://www.rfc-editor.org/info/rfc2236>>.
- [RFC2710] Deering, S., Fenner, W., and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6", [RFC 2710](#), DOI 10.17487/RFC2710, October 1999, <<https://www.rfc-editor.org/info/rfc2710>>.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", [RFC 3376](#), DOI 10.17487/RFC3376, October 2002, <<https://www.rfc-editor.org/info/rfc3376>>.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", [RFC 4601](#), DOI 10.17487/RFC4601, August 2006, <<https://www.rfc-editor.org/info/rfc4601>>.
- [RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", [RFC 4607](#), DOI 10.17487/RFC4607, August 2006, <<https://www.rfc-editor.org/info/rfc4607>>.

Internet-Draft

Multicast in the Data Center

February 2020

- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", [RFC 4861](#), DOI 10.17487/RFC4861, September 2007, <<https://www.rfc-editor.org/info/rfc4861>>.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", [RFC 5015](#), DOI 10.17487/RFC5015, October 2007, <<https://www.rfc-editor.org/info/rfc5015>>.
- [RFC6820] Narten, T., Karir, M., and I. Foo, "Address Resolution Problems in Large Data Center Networks", [RFC 6820](#), DOI 10.17487/RFC6820, January 2013, <<https://www.rfc-editor.org/info/rfc6820>>.
- [RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", [RFC 7348](#), DOI 10.17487/RFC7348, August 2014, <<https://www.rfc-editor.org/info/rfc7348>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", [RFC 7432](#), DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC7637] Garg, P., Ed. and Y. Wang, Ed., "NVGRE: Network Virtualization Using Generic Routing Encapsulation", [RFC 7637](#), DOI 10.17487/RFC7637, September 2015, <<https://www.rfc-editor.org/info/rfc7637>>.
- [RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", [RFC 7938](#), DOI 10.17487/RFC7938, August 2016, <<https://www.rfc-editor.org/info/rfc7938>>.
- [RFC8014] Black, D., Hudson, J., Kreeger, L., Lasserre, M., and T. Narten, "An Architecture for Data-Center Network Virtualization over Layer 3 (NV03)", [RFC 8014](#),

DOI 10.17487/RFC8014, December 2016,  
<<https://www.rfc-editor.org/info/rfc8014>>.

- [RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", [RFC 8279](#), DOI 10.17487/RFC8279, November 2017, <<https://www.rfc-editor.org/info/rfc8279>>.

McBride & Komolafe

Expires August 7, 2020

[Page 17]

---

Internet-Draft

Multicast in the Data Center

February 2020

- [RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", [RFC 8365](#), DOI 10.17487/RFC8365, March 2018, <<https://www.rfc-editor.org/info/rfc8365>>.

- [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", [RFC 8402](#), DOI 10.17487/RFC8402, July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.

- [Shabaz19] Shabaz, M., Suresh, L., Rexford, J., Feamster, N., Rottenstreich, O., and M. Hira, "Elmo: Source Routed Multicast for Public Clouds", ACM SIGCOMM 2019 Conference (SIGCOMM '19) ACM, DOI 10.1145/3341302.3342066, August 2019.

- [SMPTE2110] "SMPTE2110 Standards Suite", <<http://www.smpte.org/st-2110>>.

#### Authors' Addresses

Mike McBride  
Futurewei

Email: [michael.mcbride@futurewei.com](mailto:michael.mcbride@futurewei.com)

Olufemi Komolafe  
Arista Networks

Email: [femi@arista.com](mailto:femi@arista.com)

McBride & Komolafe

Expires August 7, 2020

[Page 18]