MBONED
Internet-Draft
Intended status: Informational
Expires: August 11, 2019
M. McBride
Huawei
O. Komolafe
Arista Networks
February 07, 2019

Multicast in the Data Center Overview
draft-ietf-mboned-dc-deploy-04

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 judiously 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

Copyright Notice

Table of Contents

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

1. Introduction

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 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 future highly virtualized 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 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 cameras, studios, mixers, encoders, servers etc. 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 SMPTE, 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 the stock exchange to financial services provider and then to the stock analysts or brokerages. The network must be designed with no single point of failure and in such a way that the network can respond in a deterministic manner to any 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. Another critical requirement is reliability and traceability; regulatory and legal requirements means that the producer of the marketing data must know exactly where the flow was sent and be able to prove conclusively that the data was received within agreed SLAs. 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.

Many data center cloud providers provide publish and subscribe applications. There can be numerous publishers and subscribers and many message channels within a data center. With publish and subscribe servers, a separate message is sent to each subscriber of a publication. With multicast publish/subscribe, only one message is sent, regardless of the number of subscribers. In a publish/subscribe system, client applications, some of which are publishers and some of which are subscribers, are connected to a network of message brokers that receive publications on a number of topics, and send the publications on to the subscribers for those topics. The more subscribers there are in the publish/subscribe system, the greater the improvement to network utilization there might be with multicast.

2.2. Overlays

The proposed architecture for supporting large-scale multi-tenancy in highly virtualized data centers [RFC8014] consists of a tenant's VMs distributed across the data center 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 typically 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 consider the manner in which one-to-many communication is handled within the IP fabric.

2.3. Protocols

Conventionally, some key networking protocols used in data centers require one-to-many communication. For example, ARP and ND 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, then it 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]. Popular approaches in use mostly seek to exploit characteristics of data center networks to avoid having to broadcast/multicast these messages, as discussed in Section 4.1.

There are networking protocols that are being modified/developed to specifically target working in a data center CLOS environment. BGP has been extended to work in these type of DC environments and well supports multicast. RIFT (Routing in Fat Trees) is a new protocol

being developed to work efficiently in DC CLOS environments and also is being specified to support multicast addressing and forwarding.

3. Handling one-to-many traffic using conventional multicast

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 flavours 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 (S,G)s, 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 still 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 meaning a a multicast MAC address cannot be uniquely mapped to a multicast IPv4 address. Therefore, planning is required within an organization to choose IPv4 multicast addresses judiciously in order to avoid 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.

IGMP snooping, with proxy reporting or report suppression, actively filters IGMP packets in order to reduce load on the multicast router by ensuring only the minimal quantity of information is sent. The switch is trying to ensure the router has only a single entry for the group, regardless of the number of active listeners. If there are two active listeners in a group and the first one leaves, then the switch determines that the router does not need this information since it does not affect the status of the group from the router's point of view. However the next time there is a routine query from the router the switch will forward the reply from the remaining host, to prevent the router from believing there are no active listeners. It follows that in active IGMP snooping, the router will generally only know about the most recently joined member of the group.

In order for IGMP and thus IGMP snooping to function, a multicast router must exist on the network and generate IGMP queries. The tables (holding the member ports for each multicast group) created for snooping are associated with the querier. Without a querier the tables are not created and snooping will not work. Furthermore, IGMP general queries must be unconditionally forwarded by all switches involved in IGMP snooping. Some IGMP snooping implementations include full querier capability. Others are able to proxy and retransmit queries from the multicast router.

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, 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 is forwards the data across diverse paths.

Windows Media servers send multicast streams to clients. Windows Media Services streams to an IP multicast address and all clients subscribe to the IP address to receive the same stream. This allows a single stream to be played simultaneously by multiple clients and thus reducing bandwidth utilization.

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 them 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.

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 characteristics and 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.

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. And Section 3 has discussed how conventional multicast may be used to handle this traffic. Having said that, there are a number of alternative options of handling this traffic pattern in data centers, as discussed in the subsequent section. 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, makes these techniques more potent.

4.1. Minimizing traffic volumes

If handling one-to-many traffic in data centers can be challenging then arguably the most intuitive solution is to aim to minimize the volume of such traffic.

It was previously mentioned in Section 2 that the three main causes of one-to-many traffic in data centers are applications, overlays and protocols. While, relatively speaking, little can be done about the volume of one-to-many traffic generated by applications, 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: fixed and regular topology, owned and exclusively controlled by single organization, well-known overlay encapsulation endpoints etc.

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 about local MAC addresses, MAC to IP address mapping, virtual networks identifiers etc can be disseminated. Consequently, ARP requests from local VMs can be suppressed by the encapsulation endpoint.

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 and IGMP 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 appropriate 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 faciliate HER to the VTEPs.

4.3. 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 ceneters [I-D.ietf-bier-use-cases]. The deployment envisioned with overlay networks is that the the encapsulation endpoints would be the BFIR. 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, unlike most of the other approaches discussed in this draft, it requires

changes to the forwarding behaviour of the routers used in the data center IP fabric.

4.4. Segment Routing

Segment Routing (SR) [I-D.ietf-spring-segment-routing] adopts the the source routing paradigm in which the manner in which a packet traverses a network is determined by an ordered list of instructions. These instructions are known as segments may have a local semantic to an SR node or global within an SR domain. SR allows enforcing a flow through any topological path while maintaining per-flow state only at the ingress node to the SR domain. 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 a routing extension header. Use-cases are described in [I-D.ietf-spring-segment-routing] and are being considered in the context of BGP-based large-scale data-center (DC) design [RFC7938].

Multicast in SR continues to be discussed in a variety of drafts and working groups. The SPRING WG has not yet been chartered to work on Multicast in SR. Multicast can include locally allocating a Segment Identifier (SID) to existing replication solutions, such as PIM, mLDP, P2MP RSVP-TE and BIER. It may also be that a new way to signal and install trees in SR is developed without creating state in the network.

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 its inherent 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 by PIM. Rather, approaches which exploit characteristics of data center network architectures (e.g. fixed and regular topology, owned and exclusively controlled by single organization, well-known overlay encapsulation endpoints etc.) 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-06 (work in progress), January 2018.

[I-D.ietf-nvo3-geneve]
Gross, J., Ganga, I., and T. Sridhar, "Geneve: Generic Network Virtualization Encapsulation", draft-ietf-nvo3-geneve-06 (work in progress), March 2018.

[I-D.ietf-nvo3-vxlan-gpe]
Maino, F., Kreeger, L., and U. Elzur, "Generic Protocol Extension for VXLAN", draft-ietf-nvo3-vxlan-gpe-06 (work in progress), April 2018.

[I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-15 (work in progress), January 2018.

[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>.

[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 (NVO3)", 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>.

[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>.

Authors' Addresses

Mike McBride
Huawei

Email: michael.mcbride@huawei.com

Olufemi Komolafe
Arista Networks

Email: femi@arista.com