Mboned
Internet-Draft
Intended status: Best Current Practice
Expires: September 4, 2018

M. Abrahamsson
T-Systems
T. Chown
Jisc
L. Giuliano
Juniper Networks, Inc.
March 3, 2018

Deprecating ASM for Interdomain Multicast
draft-acg-mboned-deprecate-interdomain-asm-00

Abstract

This document recommends the deprecation of the use of Any-Source Multicast (ASM) for interdomain multicast. It therefore implicitly recommends the use of Source-Specific Multicast (SSM) for interdomain multicast applications, and that hosts and routers that are expected to handle such applications fully support SSM. The recommendations in this document do not preclude the continued use of ASM within a single organisation or domain.

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 September 4, 2018.

Table of Contents

1. Introduction

IP Multicast has been deployed in various forms, both within private networks and on the wider Internet. While a number of service models have been published, and in many cases revised over time, there has been no strong recommendation made on the appropriateness of those models to certain scenarios. This document addresses this gap by making a BCP-level recommendation to deprecate the use of ASM for interdomain multicast, and thus implictly also that all hosts and routers that are expected to support such multicast applications fully support SSM.

This document does not make any statement on the use of ASM within in a single domain or organisation, and therefore does not preclude its use. Indeed, there may be a number of application contexts for which ASM is currently still considered well-suited within a single domain.

2. Multicast routing protocols

The general IP multicast service model [RFC1112] is that sender(s) send to a multicast group address, receivers express an interest in traffic sent to a given multicast group address, and that routers use multicast routing protocols to determine how to deliver traffic from the sender(s) to the receivers.

Two high-level flavours of this service model have evolved over time. In Any-Source Multicast (ASM), any number of sources may transmit multicast packets, and those sources may come and go over the course of a multicast session without being known a priori. In ASM, receivers express interest only in a given multicast group address, and the multicast routing protocol facilitates source discovery at the network layer. In contrast, with Source-Specific Multicast (SSM) the specific source(s) that may send traffic to the group are known in advance, or may be determined during a session, typically through an out-of-band protocol sitting above the network layer. Thus in SSM, receivers express interest in both a multicast group address and specific associated source address(es).

IANA has reserved specific ranges of IPv4 and IPv6 address space for multicast addressing. Guidelines for IPv4 multicast address assignments can be found in [RFC5771], while guidelines for IPv6 multicast address assignments can be found in [RFC2375] and [RFC3307]. The IPv6 multicast address format is described in [RFC4291].

2.1. ASM routing protocols

The most commonly deployed ASM routing protocol is Protocol Independent Multicast - Sparse Mode, or PIM-SM, as detailed in [RFC7761]. PIM-SM, as the name suggests, was designed to be used in scenarios where the subnets with receivers are sparsely distributed throughout the network. Because it does not know sender addresses in advance, PIM-SM uses the concept of a Rendezvous Point (RP) to ’marry up’ senders and receivers, where all routers in a PIM-SM domain are configured to use specific RP(s).

To enable PIM-SM to work between multiple domains, i.e. to allow an RP in one domain to learn the existence of a source in another domain, an inter-RP signalling protocol known as Multicast Source Discovery Protocol (MSDP) [RFC3618] is used. Deployment scenarios for MSDP are given in [RFC4611]. MSDP has remained an Experimental protocol since its publication in 2003, and was not replicated or carried forward for IPv6.

In the absence of MSDP, a new mechanism, Embedded-RP [RFC3956], was defined for IPv6 PIM-SM, which allows routers supporting the protocol to determine the RP for the group without any prior configuration, simply by observing the RP address that is embedded (included) in the IPv6 multicast group address. Embedded-RP allows PIM-SM operation across any IPv6 network in which there is an end-to-end path of routers supporting the protocol.

2.2. SSM Routing protocols

PIM-SSM is detailed in [RFC4607]. In contrast to PIM-SM, PIM-SSM benefits from sender source address(es) being known about in advance, i.e. a given source's IP address is known (by some out of band mechanism), and thus the receiver's router can send a PIM JOIN directly towards the sender, without needing to use an RP.

IPv4 addresses in the 232/8 (232.0.0.0 to 232.255.255.255) range are designated as source-specific multicast (SSM) destination addresses and are reserved for use by source-specific applications and protocols. For IPv6, the address prefix FF3x::/32 is reserved for source-specific multicast use.

3. Discussion

3.1. Observations on ASM and SSM deployments

In enterprise and campus scenarios, ASM in the form of PIM-SM is in relatively common use, and has generally replaced PIM-DM [RFC3973]. The configuration and management of an RP within a single domain is not onerous. However, if interworking with external PIM domains in IPv4 multicast deployments is needed, MSDP is required to exchange information between domain RPs about sources. MSDP remains an Experimental protocol, and can be a complex and fragile protocol to administer and troubleshoot.

PIM-SM is a general purpose protocol that can handle all use cases. In particular, it was designed for cases such as videoconferencing where multiple sources may come and go during a multicast session. But for cases where a single, persistent source is used, and receivers can be configured to know of that source, PIM-SM has unnecessary complexity.

MSDP was not taken forward to IPv6. Instead, IPv6 has Embedded-RP, which allows the RP address for a multicast group to be embedded in the group address, making RP discovery automatic, if all routers on the path between a receiver and a sender support the protocol. Embedded-RP can support lightweight ad-hoc deployments. However, it relies on a single RP for an entire group. Embedded-RP was run

successfully between European and US academic networks during the 6NET project in 2004/05. Its usage generally remains constrained to academic networks.

As stated in RFC 4607, SSM is particularly well-suited to dissemination-style applications with one or more senders whose identities are known (by some mechanism) before the application starts running. PIM-SSM is therefore very well-suited to applications such as classic linear broadcast TV over IP.

SSM requires hosts and their subnet routers using it support the new(er) IGMPv3 [RFC3376] and MLDv2 [RFC3810] protocols. While delayed delivery of support in some OSes has meant that adoption of SSM has also been slower than might have been expected, or hoped, and was a historical reason to use ASM rather than SSM, support for IGMPv3 and MLDv2 is now widespread in common OSes.

3.2. Advantages of SSM for interdomain multicast

A significant benefit of SSM is its reduced complexity through eliminating the network-based source discovery required in ASM. This means there are no RPs, shared trees, Shortest Path Tree (SPT) switchovers, PIM registers, MSDP or data-driven state creation elements to support. SSM is really just a small subset of PIM-SM, plus IGMPv3 / MLDv2.

This reduced complexity makes SSM radically simpler to manage, troubleshoot and operate, particularly for network backbone operators, and this is the main motivation for the recommendation to deprecate the use of ASM in interdomain scenarios. Interdomain ASM is widely viewed as complicated and fragile. By eliminating network-based source discovery for interdomain multicast, the vast majority of the complexity issues go away.

RFC 4607 details many benefits of SSM, including:

> "Elimination of cross-delivery of traffic when two sources simultaneously use the same source-specific destination address;
>
> Avoidance of the need for inter-host coordination when choosing source-specific addresses, as a consequence of the above;
>
> Avoidance of many of the router protocols and algorithms that are needed to provide the ASM service model."

Further discussion can also be found in [RFC3569].

SSM is considered more secure in that it supports access control, i.e. you only get packets from the sources you explicitly ask for, as opposed to ASM where anyone can decide to send traffic to a PIM-SM group address. This topic is expanded upon in [RFC4609].

4. Recommendations

4.1. Deprecating use of ASM for interdomain multicast

This document recommends that the use of ASM is deprecated for interdomain multicast, and thus implictly that hosts and routers that are expected to support such interdomain applications fully support SSM. Best current practices for deploying interdomain multicast using SSM are documented in [RFC8313]

The recommendation applies to the use of ASM between domains where either MSDP (IPv4) or Embedded-RP (IPv6) is required for sharing knowledge of remote sources. It also recommends against the multi-domain use of an ASM group with a single RP in one domain, where multicast tunnels are used between domains.

While MSDP is an Experimental level standard, this document does not propose making MSDP Historic, given its use may be desirable for intradomain multicast use cases.

4.2. Including network support for IGMPv3 / MLDv2

This document recommends that all host and router platforms supporting multicast, and any security appliances that may handle multicat traffic, support IGMPv3 [RFC3376] and MLDv2 [RFC3810]. The updated IPv6 Node Requirements RFC [I-D.ietf-6man-rfc6434-bis] states that MLDv2 support is a MUST in all implementations. Such support is already widespread in common host and router platforms.

Further guidance on IGMPv3 and MLDv2 is given in [RFC4604].

It is sometimes desirable to limit the propagation of multicast messages in a layer 2 network, typically through a layer 2 switch device. In such cases multicast snooping can be used, by which the switch device observes the IGMP/MLD traffic passing through it, and then attempts to make intelligent decisions on which physical ports to forward multicast. Typically, ports that have not expressed an interest in receiving multicast for a given group would not have traffic for that group forwarded through them. Such snooping capability should support IGMPv3 and MLDv2. There is further discussion in [RFC4541].

4.3. Building application support for SSM

There will be a wide range of applications today that only support ASM, whether as software packages, or code embedded in devices such as set-top boxes.

The implicit recommendation to use SSM for interdomain multicast means that applications should use SSM, and operate correctly in an SSM environment, triggering IGMPv3/MLDv2 messages to signal use of SSM.

It is often thought that ASM is required for multicast applications where there are multiple sources. However, RFC 4607 also describes how SSM can be used instead of PIM-SM for multi-party applications:

> "SSM can be used to build multi-source applications where all participants' identities are not known in advance, but the multi-source "rendezvous" functionality does not occur in the network layer in this case. Just like in an application that uses unicast as the underlying transport, this functionality can be implemented by the application or by an application-layer library."

Given all common OSes support SSM, it is then down to the programming language and APIs used as to whether the necessary SSM APIs are available. SSM support is generally quite ubiquitous, with the current exception of websockets used in web-browser based applications.

It is desirable that applications also support appropriate congestion control, as described in [RFC8085], with appropriate codecs, to achieve the necessary rate adaption.

Some useful considerations for multicast applications can still be found in the relatively old [RFC3170].

4.4. Standardising an ASM/SSM protocol mapping mechanism

In the case of existing ASM applications that cannot readily be ported to SSM, it may be possible to use some form of protocol mapping, i.e., to have a mechanism to translate a (*,G) join or leave to a (S,G) join or leave, for a specific source, S. The general challenge in performing such mapping is determining where the configured source address, S, comes from.

There are some existing vendor-specific mechanisms to achieve this function, but none are documented in IETF standards. This appears ti be a useful area for the IETF to work on, but it should be noted that any such effort would only be an interim transition mechanism, and

such mappings do not remove the requirement for applications to be allocated ASM group addresses for the communications.

4.5. Not filtering ASM addressing between domains

A key benefit of SSM is that the multicast application does not need to be allocated a specific multicast group by the network, rather as SSM is inherently source-specific, it can use any group address, G, in the reserved range of IPv4 or IPv6 SSM addresses for its own source address, S.

In principle, if interdomain ASM is deprecated, backbone operators could begin filtering the ranges of group addresses used by ASM. In practice, this is not recommended given there will be a transition period from ASM to SSM, where some form of ASM-SSM mappings may be used, and filtering may preclude such operations.

4.6. Not precluding Intradomain ASM

The use of ASM within a single multicast domain, such as an enterprise or campus, with an RP for the site, is still relatively common today. The operators of such a site may choose to use Anycast-RP [RFC4610] or MSDP for internal RP resilience, at the expense of the extra complexity in managing that configuration.

This document does not preclude continued use of ASM in the intradomain scenario. If an organisation, or AS, wishes to use multiple multicast domains within its own network border, that is a choice for that organisation to make, and it may then use MSDP or Embedded-RP internally within its own network.

5. Security Considerations

This document adds no new security considerations. RFC 4609 describes the additional security benefits of using SSM instead of ASM.

6. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed upon publication as an RFC.

7. Acknowledgments

The authors would like to thank members of the IETF mboned WG for discussions on the content of this document, with specific thanks to the following people for their contributions to the document: Hitoshi Asaeda, Dale Carder, Toerless Eckert, Jake Holland, Albert Manfredi, Mike McBride, Per Nihlen, Greg Shepherd, James Stevens, Stig Venaas, Nils Warnke, and Sandy Zhang.

8. References

8.1. Normative References

[RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, DOI 10.17487/RFC1112, August 1989, <https://www.rfc-editor.org/info/rfc1112>.

[RFC2375] Hinden, R. and S. Deering, "IPv6 Multicast Address Assignments", RFC 2375, DOI 10.17487/RFC2375, July 1998, <https://www.rfc-editor.org/info/rfc2375>.

[RFC3170] Quinn, B. and K. Almeroth, "IP Multicast Applications: Challenges and Solutions", RFC 3170, DOI 10.17487/RFC3170, September 2001, <https://www.rfc-editor.org/info/rfc3170>.

[RFC3307] Haberman, B., "Allocation Guidelines for IPv6 Multicast Addresses", RFC 3307, DOI 10.17487/RFC3307, August 2002, <https://www.rfc-editor.org/info/rfc3307>.

[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>.

[RFC3569] Bhattacharyya, S., Ed., "An Overview of Source-Specific Multicast (SSM)", RFC 3569, DOI 10.17487/RFC3569, July 2003, <https://www.rfc-editor.org/info/rfc3569>.

[RFC3618] Fenner, B., Ed. and D. Meyer, Ed., "Multicast Source Discovery Protocol (MSDP)", RFC 3618, DOI 10.17487/RFC3618, October 2003, <https://www.rfc-editor.org/info/rfc3618>.

[RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <https://www.rfc-editor.org/info/rfc3810>.

[RFC3956] Savola, P. and B. Haberman, "Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address", RFC 3956, DOI 10.17487/RFC3956, November 2004, <https://www.rfc-editor.org/info/rfc3956>.

[RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, DOI 10.17487/RFC3973, January 2005, <https://www.rfc-editor.org/info/rfc3973>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <https://www.rfc-editor.org/info/rfc4291>.

[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>.

[RFC4610] Farinacci, D. and Y. Cai, "Anycast-RP Using Protocol Independent Multicast (PIM)", RFC 4610, DOI 10.17487/RFC4610, August 2006, <https://www.rfc-editor.org/info/rfc4610>.

[RFC5771] Cotton, M., Vegoda, L., and D. Meyer, "IANA Guidelines for IPv4 Multicast Address Assignments", BCP 51, RFC 5771, DOI 10.17487/RFC5771, March 2010, <https://www.rfc-editor.org/info/rfc5771>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

8.2. Informative References

[RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, DOI 10.17487/RFC4541, May 2006, <https://www.rfc-editor.org/info/rfc4541>.

[RFC4604] Holbrook, H., Cain, B., and B. Haberman, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast", RFC 4604, DOI 10.17487/RFC4604, August 2006, <https://www.rfc-editor.org/info/rfc4604>.

[RFC4609] Savola, P., Lehtonen, R., and D. Meyer, "Protocol Independent Multicast - Sparse Mode (PIM-SM) Multicast Routing Security Issues and Enhancements", RFC 4609, DOI 10.17487/RFC4609, October 2006, <https://www.rfc-editor.org/info/rfc4609>.

[RFC4611] McBride, M., Meylor, J., and D. Meyer, "Multicast Source Discovery Protocol (MSDP) Deployment Scenarios", BCP 121, RFC 4611, DOI 10.17487/RFC4611, August 2006, <https://www.rfc-editor.org/info/rfc4611>.

[RFC8085] Eggert, L., Fairhurst, G., and G. Shepherd, "UDP Usage Guidelines", BCP 145, RFC 8085, DOI 10.17487/RFC8085, March 2017, <https://www.rfc-editor.org/info/rfc8085>.

[RFC8313] Tarapore, P., Ed., Sayko, R., Shepherd, G., Eckert, T., Ed., and R. Krishnan, "Use of Multicast across Inter-domain Peering Points", BCP 213, RFC 8313, DOI 10.17487/RFC8313, January 2018, <https://www.rfc-editor.org/info/rfc8313>.

[I-D.ietf-6man-rfc6434-bis] Chown, T., Loughney, J., and T. Winters, "IPv6 Node Requirements", draft-ietf-6man-rfc6434-bis-05 (work in progress), February 2018.

Authors' Addresses

Mikael Abrahamsson
T-Systems
Stockholm
Sweden

Email: mikael.abrahamsson@t-systems.se

Tim Chown
Jisc
Lumen House, Library Avenue
Harwell Oxford, Didcot OX11 0SG
United Kingdom

Email: tim.chown@jisc.ac.uk

Lenny Giuliano
Juniper Networks, Inc.
2251 Corporate Park Drive
Hemdon, Virginia 20171
United States

Email: lenny@juniper.net

MBONED
Internet-Draft
Intended status: Informational
Expires: September 1, 2018

M. McBride
Huawei

February 28, 2018

Multicast in the Data Center Overview
draft-ietf-mboned-dc-deploy-02

Abstract

There has been much interest in issues surrounding massive amounts of hosts in the data center. These issues include the prevalent use of IP Multicast within the Data Center. Its important to understand how IP Multicast is being deployed in the Data Center to be able to understand the surrounding issues with doing so. This document provides a quick survey of uses of multicast in the data center and should serve as an aid to further discussion of issues related to large amounts of multicast in the data center.

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 September 1, 2018.

Table of Contents

1. Introduction

Data center servers often use IP Multicast to send data to clients or other application servers. IP Multicast is expected to help conserve bandwidth in the data center and reduce the load on servers. IP Multicast is also a key component in several data center overlay solutions. Increased reliance on multicast, in next generation data centers, requires higher performance and capacity especially from the switches. If multicast is to continue to be used in the data center, it must scale well within and between datacenters. There has been much interest in issues surrounding massive amounts of hosts in the data center. There was a lengthy discussion, in the now closed ARMD WG, involving the issues with address resolution for non ARP/ND multicast traffic in data centers. This document provides a quick survey of multicast in the data center and should serve as an aid to further discussion of issues related to multicast in the data center.

ARP/ND issues are not addressed in this document except to explain how address resolution occurs with multicast.

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. Multicast Applications in the Data Center

There are many data center operators who do not deploy Multicast in their networks for scalability and stability reasons. There are also many operators for whom multicast is a critical protocol within their network and is enabled on their data center switches and routers. For this latter group, there are several uses of multicast in their data centers. An understanding of the uses of that multicast is important in order to properly support these applications in the ever evolving data centers. If, for instance, the majority of the applications are discovering/signaling each other, using multicast, there may be better ways to support them then using multicast. If, however, the multicasting of data is occurring in large volumes, there is a need for good data center overlay multicast support. The applications either fall into the category of those that leverage L2 multicast for discovery or of those that require L3 support and likely span multiple subnets.

2.1. Client-Server Applications

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 are sending multicast streams into the network and the network is forwarding the data across diverse paths.

Windows Media servers send multicast streaming 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.

Market data relies extensively on IP multicast to deliver stock quotes from the data center to a financial services provider and then to the stock analysts. The most critical requirement of a multicast trading floor is that it be highly available. The network must be designed with no single point of failure and in a way the network can respond in a deterministic manner to any failure. Typically redundant servers (in a primary/backup or live live mode) are sending multicast streams into the network and the network is forwarding the

data across diverse paths (when duplicate data is sent by multiple servers).

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. Non Client-Server Multicast Applications

Routers, running Virtual Routing Redundancy Protocol (VRRP), communicate with one another using a multicast address. VRRP packets are sent, encapsulated in IP packets, to 224.0.0.18. A failure to receive a multicast packet from the master router for a period longer than three times the advertisement timer causes the backup routers to assume that the master router is dead. The virtual router then transitions into an unsteady state and an election process is initiated to select the next master router from the backup routers. This is fulfilled through the use of multicast packets. Backup router(s) are only to send multicast packets during an election process.

Overlays may use IP multicast to virtualize L2 multicasts. IP multicast is used to reduce the scope of the L2-over-UDP flooding to only those hosts that have expressed explicit interest in the frames.VXLAN, for instance, is an encapsulation scheme to carry L2 frames over L3 networks. The VXLAN Tunnel End Point (VTEP) encapsulates frames inside an L3 tunnel. VXLANs are identified by a 24 bit VXLAN Network Identifier (VNI). The VTEP maintains a table of known destination MAC addresses, and stores the IP address of the tunnel to the remote VTEP to use for each. Unicast frames, between VMs, are sent directly to the unicast L3 address of the remote VTEP. Multicast frames are sent to a multicast IP group associated with the VNI. Underlying IP Multicast protocols (PIM-SM/SSM/BIDIR) are used to forward multicast data across the overlay.

The Ganglia application relies upon multicast for distributed discovery and monitoring of computing systems such as clusters and grids. It has been used to link clusters across university campuses and can scale to handle clusters with 2000 nodes

Windows Server, cluster node exchange, relies upon the use of multicast heartbeats between servers. Only the other interfaces in the same multicast group use the data. Unlike broadcast, multicast traffic does not need to be flooded throughout the network, reducing the chance that unnecessary CPU cycles are expended filtering traffic on nodes outside the cluster. As the number of nodes increases, the ability to replace several unicast messages with a single multicast message improves node performance and decreases network bandwidth consumption. Multicast messages replace unicast messages in two components of clustering:

- Heartbeats: The clustering failure detection engine is based on a scheme whereby nodes send heartbeat messages to other nodes. Specifically, for each network interface, a node sends a heartbeat message to all other nodes with interfaces on that network. Heartbeat messages are sent every 1.2 seconds. In the common case where each node has an interface on each cluster network, there are N * (N - 1) unicast heartbeats sent per network every 1.2 seconds in an N-node cluster. With multicast heartbeats, the message count drops to N multicast heartbeats per network every 1.2 seconds, because each node sends 1 message instead of N - 1. This represents a reduction in processing cycles on the sending node and a reduction in network bandwidth consumed.

- Regroup: The clustering membership engine executes a regroup protocol during a membership view change. The regroup protocol algorithm assumes the ability to broadcast messages to all cluster nodes. To avoid unnecessary network flooding and to properly authenticate messages, the broadcast primitive is implemented by a sequence of unicast messages. Converting the unicast messages to a single multicast message conserves processing power on the sending node and reduces network bandwidth consumption.

Multicast addresses in the 224.0.0.x range are considered link local multicast addresses. They are used for protocol discovery and are flooded to every port. For example, OSPF uses 224.0.0.5 and 224.0.0.6 for neighbor and DR discovery. These addresses are reserved and will not be constrained by IGMP snooping. These addresses are not to be used by any application.

3. L2 Multicast Protocols in the Data Center

The switches, in between the servers and the routers, rely upon igmp snooping to bound the multicast to the ports leading to interested hosts and to L3 routers. A switch will, by default, flood multicast traffic to all the ports in a broadcast domain (VLAN). IGMP snooping is designed to prevent hosts on a local network from receiving traffic for a multicast group they have not explicitly joined. It

provides switches with a mechanism to prune multicast traffic from links that do not contain a multicast listener (an IGMP client). IGMP snooping is a L2 optimization for L3 IGMP.

IGMP snooping, with proxy reporting or report suppression, actively filters IGMP packets in order to reduce load on the multicast router. Joins and leaves heading upstream to the router are filtered so that only the minimal quantity of information is sent. The switch is trying to ensure the router only has a single entry for the group, regardless of how many active listeners there are. 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.

In source-only networks, however, which presumably describes most data center networks, there are no IGMP hosts on switch ports to generate IGMP packets. Switch ports are connected to multicast source ports and multicast router ports. The switch typically learns about multicast groups from the multicast data stream by using a type of source only learning (when only receiving multicast data on the port, no IGMP packets). The switch forwards traffic only to the multicast router ports. When the switch receives traffic for new IP multicast groups, it will typically flood the packets to all ports in the same VLAN. This unnecessary flooding can impact switch performance.

4. L3 Multicast Protocols in the Data Center

There are three flavors of PIM used for Multicast Routing in the Data Center: PIM-SM [RFC4601], PIM-SSM [RFC4607], and PIM-BIDIR [RFC5015]. SSM provides the most efficient forwarding between sources and receivers and is most suitable for one to many types of multicast applications. State is built for each S,G channel therefore the more

sources and groups there are, the more state there is in the network. BIDIR is the most efficient shared tree solution as one tree is built for all S,G's, therefore saving state. But it is not the most efficient in forwarding path between sources and receivers. 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 then will either switch to the SPT (shortest path tree), similar to SSM, or stay on the shared tree (similar to BIDIR). For massive amounts of hosts sending (and receiving) multicast, the shared tree (particularly with PIM-BIDIR) provides the best potential scaling since no matter how many multicast sources exist within a VLAN, the tree number stays the same. IGMP snooping, IGMP proxy, and PIM-BIDIR have the potential to scale to the huge scaling numbers required in a data center.

5. Challenges of using multicast in the Data Center

Data Center environments may create unique challenges for IP Multicast. Data Center networks required a high amount of VM traffic and mobility within and between DC networks. DC networks have large numbers of servers. DC networks are often used with cloud orchestration software. DC networks often use IP Multicast in their unique environments. This section looks at the challenges of using multicast within the challenging data center environment.

When IGMP/MLD Snooping is not implemented, ethernet switches will flood multicast frames out of all switch-ports, which turns the traffic into something more like a broadcast.

VRRP uses multicast heartbeat to communicate between routers. The communication between the host and the default gateway is unicast. The multicast heartbeat can be very chatty when there are thousands of VRRP pairs with sub-second heartbeat calls back and forth.

Link-local multicast should scale well within one IP subnet particularly with a large layer3 domain extending down to the access or aggregation switches. But if multicast traverses beyond one IP subnet, which is necessary for an overlay like VXLAN, you could potentially have scaling concerns. If using a VXLAN overlay, it is necessary to map the L2 multicast in the overlay to L3 multicast in the underlay or do head end replication in the overlay and receive duplicate frames on the first link from the router to the core switch. The solution could be to run potentially thousands of PIM messages to generate/maintain the required multicast state in the IP underlay. The behavior of the upper layer, with respect to broadcast/multicast, affects the choice of head end (*,G) or (S,G) replication in the underlay, which affects the opex and capex of the

entire solution. A VXLAN, with thousands of logical groups, maps to head end replication in the hypervisor or to IGMP from the hypervisor and then PIM between the TOR and CORE 'switches' and the gateway router.

Requiring IP multicast (especially PIM BIDIR) from the network can prove challenging for data center operators especially at the kind of scale that the VXLAN/NVGRE proposals require. This is also true when the L2 topological domain is large and extended all the way to the L3 core. In data centers with highly virtualized servers, even small L2 domains may spread across many server racks (i.e. multiple switches and router ports).

It's not uncommon for there to be 10-20 VMs per server in a virtualized environment. One vendor reported a customer requesting a scale to 400VM's per server. For multicast to be a viable solution in this environment, the network needs to be able to scale to these numbers when these VMs are sending/receiving multicast.

A lot of switching/routing hardware has problems with IP Multicast, particularly with regards to hardware support of PIM-BIDIR.

Sending L2 multicast over a campus or data center backbone, in any sort of significant way, is a new challenge enabled for the first time by overlays. There are interesting challenges when pushing large amounts of multicast traffic through a network, and have thus far been dealt with using purpose-built networks. While the overlay proposals have been careful not to impose new protocol requirements, they have not addressed the issues of performance and scalability, nor the large-scale availability of these protocols.

There is an unnecessary multicast stream flooding problem in the link layer switches between the multicast source and the PIM First Hop Router (FHR). The IGMP-Snooping Switch will forward multicast streams to router ports, and the PIM FHR must receive all multicast streams even if there is no request from receiver. This often leads to waste of switch cache and link bandwidth when the multicast streams are not actually required. [I-D.pim-umf-problem-statement] details the problem and defines design goals for a generic mechanism to restrain the unnecessary multicast stream flooding.

6. Layer 3 / Layer 2 Topological Variations

As discussed in RFC6820, the ARMD problems statement, there are a variety of topological data center variations including L3 to Access Switches, L3 to Aggregation Switches, and L3 in the Core only. Further analysis is needed in order to understand how these variations affect IP Multicast scalability

7. Address Resolution

7.1. Solicited-node Multicast Addresses for IPv6 address resolution

Solicited-node Multicast Addresses are used with IPv6 Neighbor Discovery to provide the same function as the Address Resolution Protocol (ARP) in IPv4. ARP uses broadcasts, to send an ARP Requests, which are received by all end hosts on the local link. Only the host being queried responds. However, the other hosts still have to process and discard the request. With IPv6, a host is required to join a Solicited-Node multicast group for each of its configured unicast or anycast addresses. Because a Solicited-node Multicast Address is a function of the last 24-bits of an IPv6 unicast or anycast address, the number of hosts that are subscribed to each Solicited-node Multicast Address would typically be one (there could be more because the mapping function is not a 1:1 mapping). Compared to ARP in IPv4, a host should not need to be interrupted as often to service Neighbor Solicitation requests.

7.2. Direct Mapping for Multicast address resolution

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 IP address to a multicast MAC address is derived from direct mapping. In IPv4, the mapping is done by assigning the low-order 23 bits of the multicast IP address to fill the low-order 23 bits of the multicast MAC address. When a host joins an IP multicast group, it instructs the data link layer to receive frames that match the MAC address that corresponds to the IP address of the multicast group. The data link layer filters the frames and passes frames with matching destination addresses to the IP module. Since the mapping from multicast IP address to a MAC address ignores 5 bits of the IP address, groups of 32 multicast IP addresses are mapped to the same MAC address. As a result a multicast MAC address cannot be uniquely mapped to a multicast IPv4 address. Planning is required within an organization to select IPv4 groups that are far enough away from each other as to not end up with the same L2 address used. Any multicast address in the [224-239].0.0.x and [224-239].128.0.x ranges should not be considered. 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.

8. IANA Considerations

This memo includes no request to IANA.

9. Security Considerations

No new security considerations result from this document

10. Acknowledgements

The authors would like to thank the many individuals who contributed opinions on the ARMD wg mailing list about this topic: Linda Dunbar, Anoop Ghanwani, Peter Ashwoodsmith, David Allan, Aldrin Isaac, Igor Gashinsky, Michael Smith, Patrick Frejborg, Joel Jaeggli and Thomas Narten.

11. References

11.1. 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>.

11.2. Informative References

[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>.

Author’s Address

Mike McBride
Huawei

Email: michael.mcbride@huawei.com

Internet Area
Internet-Draft
Intended status: Informational
Expires: August 7, 2018

C. Perkins
M. McBride
Futurewei
D. Stanley
HPE
W. Kumari
Google
JC. Zuniga
SIGFOX

February 3, 2018

Multicast Considerations over IEEE 802 Wireless Media
draft-ietf-mboned-ieee802-mcast-problems-01

Abstract

Well-known issues with multicast have prevented the deployment of multicast in 802.11 [dot11], [mc-props], [mc-prob-stmt], and other local-area wireless environments. IETF multicast experts have been meeting together to discuss these issues and provide IEEE updates. The mboned working group is chartered to receive regular reports on the current state of the deployment of multicast technology, create "practice and experience" documents that capture the experience of those who have deployed and are deploying various multicast technologies, and provide feedback to other relevant working groups. This document offers guidance on known limitations and problems with wireless multicast. Also described are various multicast enhancement features that have been specified at IETF and IEEE 802 for wireless media, as well as some operational chioces that can be taken to improve the performace of the network. Finally, some recommendations are provided about the usage and combination of these features and operational choices.

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, 2018.

Table of Contents

1. Introduction

Performance issues have been observed when multicast packet transmissions of IETF protocols are used over IEEE 802 wireless media. Even though enhamcements for multicast transmissions have been designed at both IETF and IEEE 802, incompatibilities still exist between specifications, implementations and configuration choices.

Many IETF protocols depend on multicast/broadcast for delivery of control messages to multiple receivers. Multicast is used for various purposes such as neighborhood discovery, network flooding, address resolution, as well minimizing media occupancy for the transmission of data that is intended for multiple receivers. In addition to protocol use of broadcast/multicast for control messages, more applications, such as push to talk in hospitals, video in enterprises and lectures in Universities, are streaming over wifi. Many types of end devices are increasingly using wifi for their connectivity.

IETF protocols typically rely on network protocol layering in order to reduce or eliminate any dependence of higher level protocols on the specific nature of the MAC layer protocols or the physical media. In the case of multicast transmissions, higher level protocols have traditionally been designed as if transmitting a packet to an IP address had the same cost in interference and network media access, regardless of whether the destination IP address is a unicast address or a multicast or broadcast address. This model was reasonable for networks where the physical medium was wired, like Ethernet. Unfortunately, for many wireless media, the costs to access the medium can be quite different. Multicast over wifi has often been plagued by such poor performance that it is disallowed. Some enhancements have been designed in IETF protocols that are assumed to work primarily over wireless media. However, these enhancements are usually implemented in limited deployments and not widespread on most wireless networks.

IEEE 802 wireless protocols have been designed with certain features to support multicast traffic. For instance, lower modulations are used to transmit multicast frames, so that these can be received by all stations in the cell, regardless of the distance or path attenuation from the base station or access point. However, these lower modulation transmissions occupy the medium longer; they hamper efficient transmission of traffic using higher order modulations to nearby stations. For these and other reasons, IEEE 802 working

groups such as 802.11 have designed features to improve the performance of multicast transmissions at Layer 2 [ietf_802-11]. In addition to protocol design features, certain operational and configuration enhancements can ameliorate the network performance issues created by multicast traffic. as described in Section 5.

In discussing these issues over email, and in a side meeting at IETF 99, it has been generally agreed that these problems will not be fixed anytime soon primarily because it’s expensive to do so and multicast is unreliable. A big problem is that multicast is somewhat a second class citizen, to unicast, over wifi. There are many protocols using multicast and there needs to be something provided in order to make them more reliable. The problem of IPv6 neighbor discovery saturating the wifi link is only part of the problem. Wifi traffic classes may help. We need to determine what problem should be solved by the IETF and what problem should be solved by the IEEE (see Section 8). A "multicast over wifi" IETF mailing list has been formed (mcast-wifi@ietf.org) for further discussion. This draft will be updated according to the current state of discussion.

This Internet Draft details various problems caused by multicast transmission over wireless networks, including high packet error rates, no acknowledgements, and low data rate. It also explains some enhancements that have been designed at IETF and IEEE 802, as well as the operational choices that can be taken, to ameliorate the effects of multicast traffic. Recommendations about how to use and combine these enhancements are also provided.

2. Terminology

This document uses the following definitions:

AP
: IEEE 802.11 Access Point.

basic rate
: The "lowest common denominator" data rate at which multicast and broadcast traffic is generally transmitted.

DTIM
: Delivery Traffic Indication Map (DTIM): An information element that advertises whether or not any associated stations have buffered multicast or broadcast frames.

MCS
: Modulation and Coding Scheme.

STA

802.11 station (e.g. handheld device).

TIM
Traffic Indication Map (TIM): An information element that advertises whether or not any associated stations have buffered unicast frames.

3. Identified mulitcast issues

3.1. Issues at Layer 2 and Below

In this section we describe some of the issues related to the use of multicast transmissions over IEEE 802 wireless technologies.

3.1.1. Multicast reliability

Multicast traffic is typically much less reliable than unicast traffic. Since multicast makes point-to-multipoint communications, multiple acknowledgements would be needed to guarantee reception at all recipients. Since typically there are no ACKs for multicast packets, it is not possible for the Access Point (AP) to know whether or not a retransmission is needed. Even in the wired Internet, this characteristic often causes undesirably high error rates. This has contributed to the relatively slow uptake of multicast applications even though the protocols have long been available. The situation for wireless links is much worse, and is quite sensitive to the presence of background traffic. Consequently, there can be a high packet error rate (PER) due to lack of retransmission, and because the sender never backs off. It is not uncommon for there to be a packet loss rate of 5% or more, which is particularly troublesome for video and other environments where high data rates and high reliability are required.

3.1.2. Lower and Variable Data Rate

One big difference between multicast over wired versus multicast over wired is that transmission over wired links often occurs at a fixed rate. Wifi, on the other hand, has a transmission rate which varies depending upon the clients proximity to the AP. The throughput of video flows, and the capacity of the broader wifi network, will change and will impact the ability for QoS solutions to effectively reserve bandwidth and provide admission control.

For wireless stations associated with an Access Points, the power necessary for good reception can vary from station to station. For unicast, the goal is to minimize power requirements while maximizing the data rate to the destination. For multicast, the goal is simply

to maximize the number of receivers that will correctly receive the multicast packet; generally the Access Point has to use a much lower data rate at a power level high enough for even the farthest station to receive the packet. Consequently, the data rate of a video stream, for instance, would be constrained by the environmental considerations of the least reliable receiver associated with the Access Point.

Because more robust modulation and coding schemes (MCSs) have longer range but also lower data rate, multicast / broadcast traffic is generally transmitted at the lowest common denominator rate, also known as the basic rate. Depending on the specific 802.11 technology, and the configured choice for the base data rate for multicast transmission from the Access Point, the amount of additional interference can range from a factor of ten, to a factor thousands for 802.11ac.

Wired multicast also affects wireless LANs when the AP extends the wired segment; in that case, multicast / broadcast frames on the wired LAN side are copied to WLAN. Since broadcast messages are transmitted at the most robust MCS, many large frames are sent at a slow rate over the air.

3.1.3. High Interference

Transmissions at a lower rate require longer occupancy of the wireless medium and thus take away from the airtime of other communications and degrade the overall capacity. Furthermore, transmission at higher power, as is required to reach all multicast clients associated to the AP, proportionately increases the area of interference.

3.1.4. Power-save Effects on Multicast

One of the characteristics of multicast transmission is that every station has to be configured to wake up to receive the multicast, even though the received packet may ultimately be discarded. This process can have a large effect on the power consumption by the multicast receiver station.

Multicast can work poorly with the power-save mechanisms defined in IEEE 802.11e, for the following reasons.

- Clients may be unable to stay in sleep mode due to multicast control packets frequently waking them up.
- Both unicast and multicast traffic can be delayed by power-saving mechanisms.

- o A unicast packet is delayed until a STA wakes up and requests it. Unicast traffic may also be delayed to improve power save, efficiency and increase probability of aggregation.
- o Multicast traffic is delayed in a wireless network if any of the STAs in that network are power savers. All STAs associated to the AP have to be awake at a known time to receive multicast traffic.
- o Packets can also be discarded due to buffer limitations in the AP and non-AP STA.

3.2. Issues at Layer 3 and Above

This section identifies some representative IETF protocols, and describes possible negative effects due to performance degradation when using multicast transmissions for control messages. Common uses of multicast include:

- o Control plane for IPv4 and IPv6
- o ARP and Neighbor Discovery
- o Service discovery
- o Applications (video delivery, stock data etc)
- o Other L3 protocols (non-IP)

3.2.1. IPv4 issues

The following list contains a few representative IPv4 protocols using multicast.

- o ARP
- o DHCP
- o mDNS

After initial configuration, ARP and DHCP occur much less commonly. But service discovery can occur at any time. Apple’s Bonjour protocol, for instance, provides service discovery (for printing) that utilizes multicast. It’s the first thing operators drop. Even if multicast snooping is utilized, many devices register at once using Bonjour, causing serious network degradation.

3.2.2. IPv6 issues

IPv6 makes much more extensive use of multicast, including the following:

- o DHCPv6
- o IPv6 Neighbor Discovery Protocol (NDP) is not very tolerant of packet losses. In particular, the Duplicate Address Detection (DAD) process fails when the owner of an address does not receive the multicast DAD message from another node that wishes to own

that same address. This can result in an address being duplicated in the subnet, breaking a basic assumption of IPv6 connectivity.

- o IPv6 NDP Neighbor Solicitation (NS) messages used in DAD and Address Lookup make use of Link-Scope multicast. In contrast to IPv4, an IPv6 Node will typically use multiple addresses, and may change them often for privacy reasons. This multiplies the impact of multicast messages that are associated to the mobility of a Node. Router advertisement (RA) messages are also periodically multicasted over the Link.
- o Neighbors may be considered lost if several consecutive packets fail.

Address Resolution

Service Discovery

Route Discovery

Decentralized Address Assignment

Geographic routing

3.2.3. MLD issues

Multicast Listener Discovery(MLD) [RFC4541] is often used to identify members of a multicast group that are connected to the ports of a switch. Forwarding multicast frames into a WiFi-enabled area can use such switch support for hardware forwarding state information. However, since IPv6 makes heavy use of multicast, each STA with an IPv6 address will require state on the switch for several and possibly many multicast solicited-node addresses. Multicast addresses that do not have forwarding state installed (perhaps due to hardware memory limitations on the switch) cause frames to be flooded on all ports of the switch.

3.2.4. Spurious Neighbor Discovery

On the Internet there is a "background radiation" of scanning traffic (people scanning for vulnerable machines) and backscatter (responses from spoofed traffic, etc). This means that routers very often receive packets destined for machines whose IP addresses may or may not be in use. In the cases where the IP is assigned to a host, the router broadcasts an ARP request, gets back an ARP reply, and caches it; then traffic can be delivered to the host. When the IP address is not in use, the router broadcasts one (or more) ARP requests, and never gets a reply. This means that it does not populate the ARP cache, and the next time there is traffic for that IP address the router will rebroadcast the ARP requests.

The rate of these ARP requests is proportional to the size of the subnets, the rate of scanning and backscatter, and how long the router keeps state on non-responding ARPs. As it turns out, this rate is inversely proportional to how occupied the subnet is (valid ARPs end up in a cache, stopping the broadcasting; unused IPs never respond, and so cause more broadcasts). Depending on the address space in use, the time of day, how occupied the subnet is, and other unknown factors, on the order of 2000 broadcasts per second have been observed at the IETF NOCs.

On a wired network, there is not a huge difference amongst unicast, multicast and broadcast traffic; but this is not true in the wireless realm. Wireless equipment often is unable to send this amount of broadcast and multicast traffic. Consequently, on the wireless networks, we observe a significant amount of dropped broadcast and multicast packets. This, in turn, means that when a host connects it is often not able to complete DHCP, and IPv6 RAs get dropped, leading to users being unable to use the network.

4. Multicast protocol optimizations

This section lists some optimizations that have been specified in IEEE 802 and IETF that are aimed at reducing or eliminating the issues discussed in Section 3.

4.1. Proxy ARP in 802.11-2012

The AP knows the MAC address and IP address for all associated STAs. In this way, the AP acts as the central "manager" for all the 802.11 STAs in its BSS. Proxy ARP is easy to implement at the AP, and offers the following advantages:

- Reduced broadcast traffic (transmitted at low MCS) on the wireless medium
- STA benefits from extended power save in sleep mode, as ARP requests for STA’s IP address are handled instead by the AP.
- ARP frames are kept off the wireless medium.
- No changes are needed to STA implementation.

Here is the specification language as described in clause 10.23.13 of [dot11-proxyarp]:

> When the AP supports Proxy ARP "[...] the AP shall maintain a Hardware Address to Internet Address mapping for each associated station, and shall update the mapping when the Internet Address of the associated station changes. When the IPv4 address being resolved in the ARP request packet is used by a non-AP STA

currently associated to the BSS, the proxy ARP service shall respond on behalf of the non-AP STA"

4.2. IPv6 Address Registration and Proxy Neighbor Discovery

As used in this section, a Low-Power Wireless Personal Area Network (6LoWPAN) denotes a low power lossy network (LLN) that supports 6LoWPAN Header Compression (HC) [RFC6282]. A 6TiSCH network [I-D.ietf-6tisch-architecture] is an example of a 6LowPAN. In order to control the use of IPv6 multicast over 6LoWPANs, the 6LoWPAN Neighbor Discovery (6LoWPAN ND) [RFC6775] standard defines an address registration mechanism that relies on a central registry to assess address uniqueness, as a substitute to the inefficient Duplicate Address Detection (DAD) mechanism found in the mainstream IPv6 Neighbor Discovery Protocol (NDP) [RFC4861][RFC4862].

The 6lo Working Group is now completing an update [I-D.ietf-6lo-rfc6775-update] to RFC6775. The update enables the registration to a Backbone Router [I-D.ietf-6lo-backbone-router], which proxies for the registered addresses with the mainstream IPv6 NDP running on a high speed aggragating backbone. The update also enables a proxy registration on behalf of the registered node, e.g. by a 6LoWPAN router to which the mobile node is attached.

The general idea behind the backbone router concept is that in a variety of Wireless Local Area Networks (WLANs) and Wireless Personal Area Networks (WPANs), the broadcast/multicast domain should be controlled, and connectivity to a particular link that provides the subnet should be left to Layer-3. The model for the Backbone Router operation is represented in Figure 1.

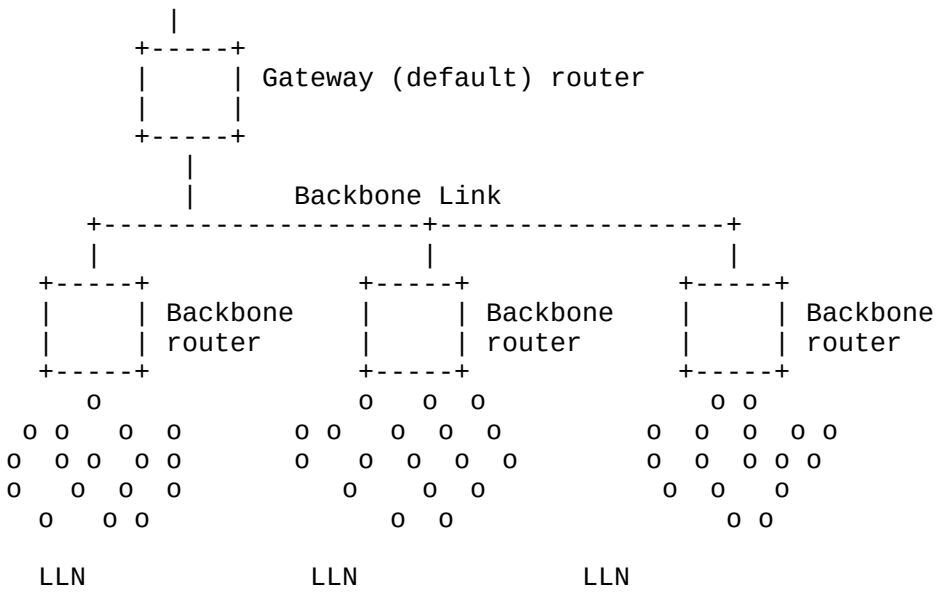

Figure 1: Backbone Link and Backbone Routers

LLN nodes can move freely from an LLN anchored at one IPv6 Backbone Router to an LLN anchored at another Backbone Router on the same backbone, keeping any of the IPv6 addresses they have configured. The Backbone Routers maintain a Binding Table of their Registered Nodes, which serves as a distributed database of all the LLN Nodes. An extension to the Neighbor Discovery Protocol is introduced to exchange that information across the Backbone Link in the reactive fashion of mainstream IPv6 Neighbor Discovery.

RFC6775 and follow-on work (e.g., [I-D.ietf-6lo-ap-nd], are designed to address the needs of LLNs, but the techniques are likely to be valuable on any type of link where sleeping devices are attached, or where the use of broadcast and multicast operations should be limited.

4.3. Buffering to improve Power-Save

Methods have been developed to help save battery life; for example, a device might not wake up when the AP receives a multicast packet. The AP acts on behalf of STAs in various ways. In order to improve the power-saving feature for STAs in its BSS, the AP buffers frames for delivery to the STA at the time when the STA is scheduled for reception. If an AP, for instance, expresses a DTIM of 3 then it will send a multicast packet every 3 packets. But the reality is that most AP’s will send a multicast every 30 packets. For unicast

there's a TIM. But because multicast is going to everyone, the AP sends a broadcast to everyone. DTIM does power management but clients can choose whether or not to wake up or not and whether or not to drop the packet. Unfortunately, without proper administrative control, such clients may no longer be able to determine why their multicast operations do not work.

4.4. IPv6 support in 802.11-2012

IPv6 uses Neighbor Discovery Protocol (NDP) instead of ARP. Every IPv6 node subscribes to a special multicast address for this purpose.

Here is the specification language from clause 10.23.13 of [dot11-proxyarp]:

> "When an IPv6 address is being resolved, the Proxy Neighbor Discovery service shall respond with a Neighbor Advertisement message [...] on behalf of an associated STA to an [ICMPv6] Neighbor Solicitation message [...]. When MAC address mappings change, the AP may send unsolicited Neighbor Advertisement Messages on behalf of a STA."

NDP may be used to request additional information

- Maximum Transmission Unit
- Router Solicitation
- Router Advertisement, etc.

NDP messages are sent as group addressed (broadcast) frames in 802.11. Using the proxy operation helps to keep NDP messages off the wireless medium.

4.5. Conversion of multicast to unicast

It is often possible to transmit multicast control and data messages by using unicast transmissions to each station individually.

4.6. Directed Multicast Service (DMS)

There are situations where more is needed than simply converting multicast to unicast. For these purposes, DMS enables a client to request that the AP transmit multicast group addressed frames destined to the requesting clients as individually addressed frames [i.e., convert multicast to unicast]. Here are some characteristics of DMS:

- Requires 802.11n A-MSDUs

- Individually addressed frames are acknowledged and are buffered for power save clients
- The requesting STA may specify traffic characteristics for DMS traffic
- DMS was defined in IEEE Std 802.11v-2011
- DMS requires changes to both AP and STA implementation.

DMS is not currently implemented in products.

4.7. GroupCast with Retries (GCR)

GCR (defined in [dot11aa]) provides greater reliability by using either unsolicited retries or a block acknowledgement mechanism. GCR increases probability of broadcast frame reception success, but still does not guarantee success.

For the block acknowledgement mechanism, the AP transmits each group addressed frame as conventional group addressed transmission. Retransmissions are group addressed, but hidden from non-11aa clients. A directed block acknowledgement scheme is used to harvest reception status from receivers; retransmissions are based upon these responses.

GCR is suitable for all group sizes including medium to large groups. As the number of devices in the group increases, GCR can send block acknowledgement requests to only a small subset of the group. GCR does require changes to both AP and STA implementation.

GCR may introduce unacceptable latency. After sending a group of data frames to the group, the AP has do the following:

- unicast a Block Ack Request (BAR) to a subset of members.
- wait for the corresponding Block Ack (BA).
- retransmit any missed frames.
- resume other operations which may have been delayed.

This latency may not be acceptable for some traffic.

There are ongoing extensions in 802.11 to improve GCR performance.

- BAR is sent using downlink MU-MIMO (note that downlink MU-MIMO is already specified in 802.11-REVmc 4.3).
- BA is sent using uplink MU-MIMO (which is a .11ax feature).
- Additional 802.11ax extensions are under consideration; see [mc-ack-mux]
- Latency may also be reduced by simultaneously receiving BA information from multiple clients.

5. Operational optimizations

This section lists some operational optimizations that can be implemented when deploying wireless IEEE 802 networks to mitigate the issues discussed in Section 3.

5.1. Mitigating Problems from Spurious Neighbor Discovery

ARP Sponges

An ARP Sponge sits on a network and learn which IPs addresses are actually in use. It also listen for ARP requests, and, if it sees an ARP for an IP address which it believes is not used, it will reply with its own MAC address. This means that the router now has an IP to MAC mapping, which it caches. If that IP is later assigned to an machine (e.g using DHCP), the ARP sponge will see this, and will stop replying for that address. Gratuitous ARPs (or the machine ARPing for its gateway) will replace the sponged address in the router ARP table. This technique is quite effective; but, unfortunately, the ARP sponge daemons were not really designed for this use (the standard one [arpsponge], was designed to deal with the disappearance of participants from an IXP) and so are not optimized for this purpose. We have to run one daemon per subnet, the tuning is tricky (the scanning rate versus the population rate versus retires, etc.) and sometimes the daemons just seem to stop, requiring a restart of the daemon and causing disruption.

Router mitigations

Some routers (often those based on Linux) implement a "negative ARP cache" daemon. Simply put, if the router does not see a reply to an ARP it can be configured to cache this information for some interval. Unfortunately, the core routers which we are using do not support this. When a host connects to network and gets an IP address, it will ARP for its default gateway (the router). The router will update its cache with the IP to host MAC mapping learnt from the request (passive ARP learning).

Firewall unused space

The distribution of users on wireless networks / subnets changes from meeting to meeting (e.g the "IETF-secure" SSID was renamed to "IETF", fewer users use "IETF-legacy", etc). This utilization is difficult to predict ahead of time, but we can monitor the usage as attendees use the different networks. By

configuring multiple DHCP pools per subnet, and enabling them sequentially, we can have a large subnet, but only assign addresses from the lower portions of it. This means that we can apply input IP access lists, which deny traffic to the upper, unused portions. This means that the router does not attempt to forward packets to the unused portions of the subnets, and so does not ARP for it. This method has proven to be very effective, but is somewhat of a blunt axe, is fairly labor intensive, and requires coordination.

Disabling/filtering ARP requests

In general, the router does not need to ARP for hosts; when a host connects, the router can learn the IP to MAC mapping from the ARP request sent by that host. This means that we should be able to disable and / or filter ARP requests from the router. Unfortunately, ARP is a very low level / fundamental part of the IP stack, and is often offloaded from the normal control plane. While many routers can filter layer-2 traffic, this is usually implemented as an input filter and / or has limited ability to filter output broadcast traffic. This means that the simple "just disable ARP or filter it outbound" seems like a really simple (and obvious) solution, but implementations / architectural issues make this difficult or awkward in practice.

NAT

The broadcasts are overwhelmingly being caused by outside scanning / backscatter traffic. This means that, if we were to NAT the entire (or a large portion) of the attendee networks, there would be no NAT translation entries for unused addresses, and so the router would never ARP for them. The IETF NOC has discussed NATing the entire (or large portions) attendee address space, but a: elegance and b: flaming torches and pitchfork concerns means we have not attempted this yet.

Stateful firewalls

Another obvious solution would be to put a stateful firewall between the wireless network and the Internet. This firewall would block incoming traffic not associated with an outbound request. The IETF philosophy has been to have the network as open as possible / honor the end-to-end principle. An attendee on the meeting network should be an Internet host, and should be able to receive unsolicited requests. Unfortunately, keeping the network working and stable is the first priority

and a stateful firewall may be required in order to achieve this.

6. Multicast Considerations for Other Wireless Media

Many of the causes of performance degradation described in earlier sections are also observable for wireless media other than 802.11.

For instance, problems with power save, excess media occupancy, and poor reliability will also affect 802.15.3 and 802.15.4. However, 802.15 media specifications do not include mechanisms similar to those developed for 802.11. In fact, the design philosophy for 802.15 is oriented towards minimality, with the result that many such functions would more likely be relegated to operation within higher layer protocols. This leads to a patchwork of non-interoperable and vendor-specific solutions. See [uli] for some additional discussion, and a proposal for a task group to resolve similar issues, in which the multicast problems might be considered for mitigation.

7. Recommendations

This section will provide some recommendations about the usage and combinations of the multicast enhancements described in Section 4 and Section 5.

(FFS)

8. Discussion Items

This section will suggest some discussion items for further resolution.

The IETF may need to decide that broadcast is more expensive so multicast needs to be sent wired. For example, 802.1ak works on ethernet and wifi. 802.1ak has been pulled into 802.1Q as of 802.1Q-2011. 802.1Q-2014 can be looked at here: http://www.ieee802.org/1/pages/802.1Q-2014.html. If a generic solution is not found, guidelines for multicast over wifi should be established.

To provide an idea going forward, perhaps a reliable registration to Layer-2 multicast groups and a reliable multicast operation at Layer-2 could provide a generic solution. There is no need to support 2^24 groups to get solicited node multicast working: it is possible to simply select a number of trailing bits that make sense for a given network size to limit the amount of unwanted deliveries to reasonable levels. IEEE 802.1, 802.11, and 802.15 should be encouraged to revisit L2 multicast issues. In particular, Wi-Fi provides a broadcast service, not a multicast one; at the PHY level,

all frames are broadcast except in very unusual cases in which special beamforming transmitters are used. Unicast offers the advantage of being much faster (2 orders of magnitude) and much more reliable (L2 ARQ).

9. Security Considerations

This document does not introduce any security mechanisms, and does not have affect existing security mechanisms.

10. IANA Considerations

This document does not specify any IANA actions.

11. Acknowledgements

This document has benefitted from discussions with the following people, in alphabetical order: Pascal Thubert

12. Informative References

[arpsponge] Arien Vijn, Steven Bakker, "Arp Sponge", March 2015.

[dot11] P802.11, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications", March 2012.

[dot11-proxyarp] P802.11, "Proxy ARP in 802.11ax", September 2015.

[dot11aa] P802.11, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 2: MAC Enhancements for Robust Audio Video Streaming", March 2012.

[I-D.ietf-6lo-ap-nd] Thubert, P., Sarikaya, B., and M. Sethi, "Address Protected Neighbor Discovery for Low-power and Lossy Networks", draft-ietf-6lo-ap-nd-05 (work in progress), January 2018.

[I-D.ietf-6lo-backbone-router] Thubert, P., "IPv6 Backbone Router", draft-ietf-6lo-backbone-router-05 (work in progress), January 2018.

[I-D.ietf-6lo-rfc6775-update]
Thubert, P., Nordmark, E., Chakrabarti, S., and C. Perkins, "An Update to 6LoWPAN ND", draft-ietf-6lo-rfc6775-update-11 (work in progress), December 2017.

[I-D.ietf-6tisch-architecture]
Thubert, P., "An Architecture for IPv6 over the TSCH mode of IEEE 802.15.4", draft-ietf-6tisch-architecture-13 (work in progress), November 2017.

[ietf_802-11]
Dorothy Stanley, "IEEE 802.11 multicast capabilities", Nov 2015.

[mc-ack-mux]
Yusuke Tanaka et al., "Multiplexing of Acknowledgements for Multicast Transmission", July 2015.

[mc-prob-stmt]
Mikael Abrahamsson and Adrian Stephens, "Multicast on 802.11", March 2015.

[mc-props]
Adrian Stephens, "IEEE 802.11 multicast properties", March 2015.

[RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, DOI 10.17487/RFC4541, May 2006, <https://www.rfc-editor.org/info/rfc4541>.

[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>.

[RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, DOI 10.17487/RFC4862, September 2007, <https://www.rfc-editor.org/info/rfc4862>.

[RFC6282] Hui, J., Ed. and P. Thubert, "Compression Format for IPv6 Datagrams over IEEE 802.15.4-Based Networks", RFC 6282, DOI 10.17487/RFC6282, September 2011, <https://www.rfc-editor.org/info/rfc6282>.

[RFC6775] Shelby, Z., Ed., Chakrabarti, S., Nordmark, E., and C. Bormann, "Neighbor Discovery Optimization for IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs)", RFC 6775, DOI 10.17487/RFC6775, November 2012, <https://www.rfc-editor.org/info/rfc6775>.

[uli] Pat Kinney, "LLC Proposal for 802.15.4", Nov 2015.

Authors' Addresses

Charles E. Perkins
Futurewei Inc.
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1-408-330-4586
Email: charliep@computer.org

Mike McBride
Futurewei Inc.
2330 Central Expressway
Santa Clara, CA 95055
USA

Email: michael.mcbride@huawei.com

Dorothy Stanley
Hewlett Packard Enterprise
2000 North Naperville Rd.
Naperville, IL 60566
USA

Phone: +1 630 979 1572
Email: dstanley@arubanetworks.com

Warren Kumari
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
USA

Email: warren@kumari.net

Juan Carlos Zuniga
SIGFOX
425 rue Jean Rostand
Labege 31670
France

Email: j.c.zuniga@ieee.org

MBONED WG
Internet-Draft
Intended status: Standards Track
Expires: August 25, 2018

Zheng. Zhang
Cui. Wang
ZTE Corporation
Ying. Cheng
China Unicom

February 21, 2018

Multicast YANG Data Model
draft-zhang-mboned-multicast-yang-model-00

Abstract

This document intents to provide a general and all-round multicast YANG data model, which tries to stand at a high level to take full advantages of existed multicast protocol models to control the multicast network, and guides the deployment of multicast service. And also, there will define several possible RPCs about how to interact between multicast YANG data model and multicast protocol models. This multicast YANG data model is mainly used by the management tools run by the network operators in order to manage, monitor and debug the network resources used to deliver multicast service, as well as gathering some data from the network.

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 25, 2018.

Table of Contents

1. Introduction

Currently, there are many multicast protocol YANG models, such as PIM, MLD, and BIER and so on. But all these models are distributed in different working groups as separate files and focus on the protocol itself. Furthermore, they cannot describe a high-level multicast service required by network operators.

This document intents to provide a general and all-round multicast model, which tries to stand at a high level to take full advantages of these aforementioned models to control the multicast network, and guides the deployment of multicast service.

This multicast YANG data model is mainly used by the management tools run by the network operators in order to manage, monitor and debug the network resources used to deliver multicast service, as well as gathering some data from the network.

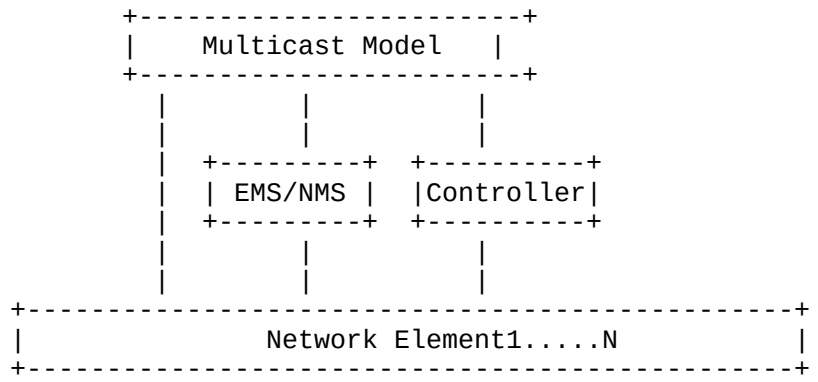

Figure 1: Example usage of Multicast Model

Detailly, in figure 1, there is an example of usage of this multicast model. Network operators can use this model in a controller who is responsible to implement some multicast flows with specific protocols and invoke the corresponding protocols' model to configure the network elements through NETCONF/RESTCONF/CLI. Or network operators can use this model to the EMS/NMS to manage the network elements or configure the network elements directly. For example, a multicast service need to be delopy in a network, supposed that the multicast flow is 239.0.0.0/8, the flow should be transport by BIER technology. Then we use this multicast YANG data model and set the correspond key (239.0.0.0) and associated transport technology with BIER, send the model from controller to every egde node in the network. Then there is an interaction among all the nodes to exchange the multicast flow information. The ingress node will encapsulate the multicast flow with BIER header and send it into the network. Intermediate nodes will forward the flows to all the egress nodes by BIER forwarding.

On the other hand, when the network elements detect failure or some other changes, the network devices can send the affected multicast flows and the associated overlay/ transport/ underlay information to the controller. Then the controller/ EMS/NMS can response immediately due to the failure and distribute new model for the flows to the network nodes quickly. Such as the changing of the failure overlay protocol to another one, as well as transport and underlay protocol.

Specifically, in section 3, it provides a human readability of the whole multicast network through UML like class diagram, which frames different multicast components and correlates them in a readable fashion. Then, based on this UML like class diagram, there is instantiated and detailed YANG model in Section 5.

In other words, this document does not define any specific protocol model, instead, it depends on many existed multicast protocol models and relates several multicast information together to fulfill multicast service.

2. Design of the multicast model

This model includes multicast service keys and three layers: the multicast overlay, the transport layer and the multicast underlay information. Multicast keys include the features of multicast flow, such as(vpnid, multicast source and multicast group) information. In data center network, for fine-grained to gather the nodes belonging to the same virtual network, there may need VNI-related information to assist.

Multicast overlay defines (ingress-node, egress-nodes) nodes information. If the transport layer is BIER, there may define BIER information including (Subdomain, ingress-node BFR-id, egress-nodes BFR-id). If no (ingress-node, egress-nodes) information are defined directly, there may need overlay multicast signaling technology, such as MLD or MVPN, to collect these nodes information.

Multicast transport layer defines the type of transport technologies that can be used to forward multicast flow, including BIER forwarding type, MPLS forwarding type, or PIM forwarding type and so on. One or several transport technologies could be defined at the same time. As for the detailed parameters for each transport technology, this multicast YANG data model can invoke the corresponding protocol model to define them.

Multicast underlay defines the type of underlay technologies, such as OSPF, ISIS, BGP, PIM or BABEL and so on. One or several underlay technologies could be defined at the same time if there is protective requirement. As for the specific parameters for each underlay technology, this multicast YANG data model can depend the corresponding protocol model to configure them as well.

3. UML Class like Diagram for Multicast YANG data Model

The following is a UML like diagram for Multicast YANG data Model.

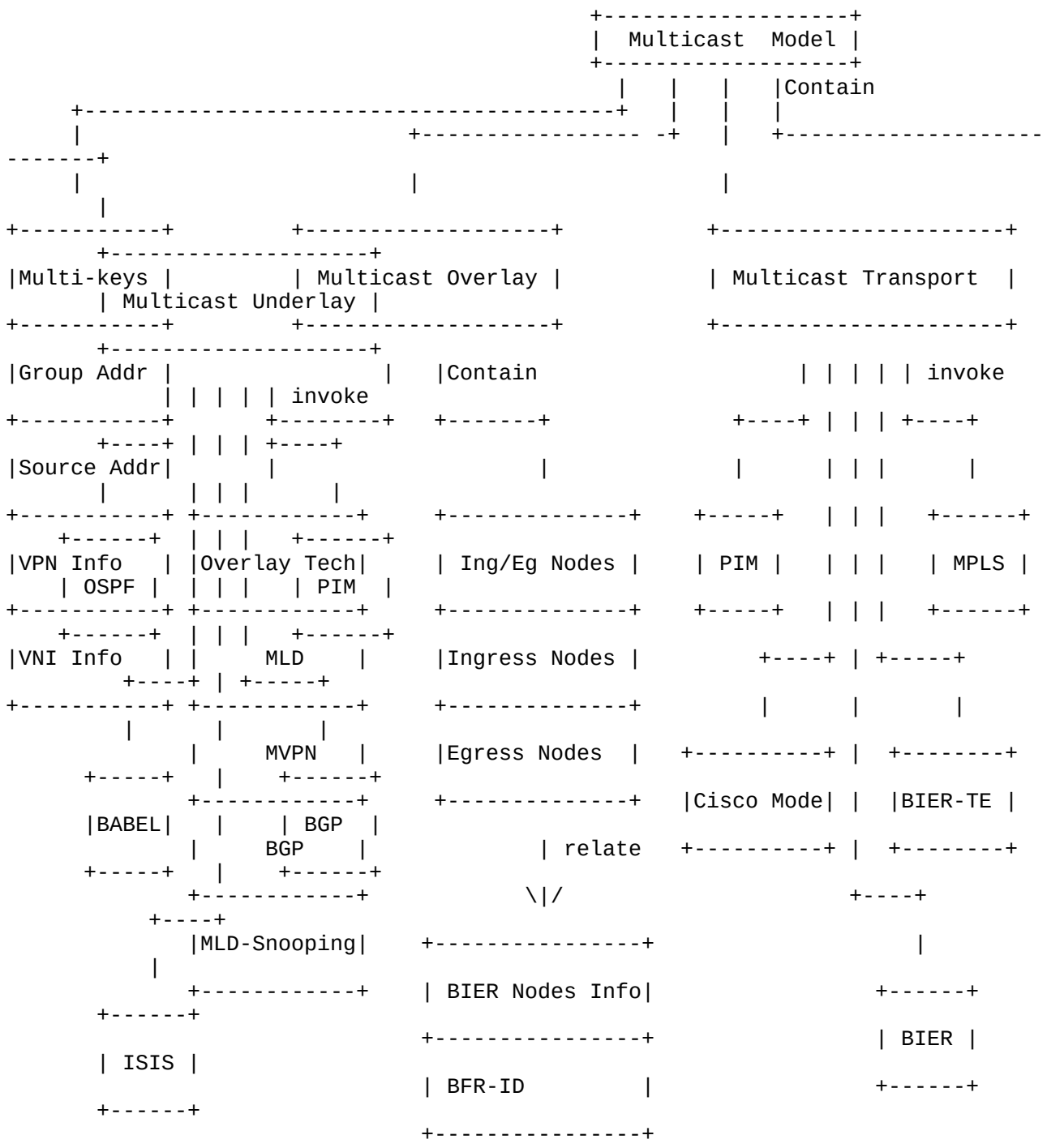

Figure 2: UML like Class Diagram for Multicast YANG data Model

4. Model Structure

```
module: ietf-multicast-model
    +--rw multicast-model
       +--rw multicast-keys* [vpn-rd source-address group-address vni-type vni-v
alue]
          +--rw vpn-rd                rt-types:route-distinguisher
          +--rw source-address        ip-multicast-source-address
          +--rw group-address         rt-types:ip-multicast-group-address
          +--rw vni-type              virtual-type
          +--rw vni-value             uint32
          +--rw multicast-overlay
```

```
|  +--rw ingress-egress
|  |  +--rw ingress-node?   inet:ip-address
|  |  +--rw egress-nodes* [egress-node]
|  |     +--rw egress-node    inet:ip-address
|  +--rw bier-ids
|  |  +--rw sub-domain?     bier:sub-domain-id
|  |  +--rw ingress-node?   bier:bfr-id
```

```
|  |  +--rw egress-nodes* [egress-node]
|  |     +--rw egress-node    bier:bfr-id
|  +--rw overlay-tech-type?   enumeration
+--rw multicast-transport
|  +--rw bier
|  |  +--rw sub-domain?        bier:sub-domain-id
|  |  +--rw (encap-type)?
|  |  |  +--:(mpls)
|  |  |  +--:(eth)
|  |  |  +--:(ipv6)
|  |  +--rw bitstringlength?   bier:bsl
|  |  +--rw set-identifier?    bier:si
|  |  +--rw ecmp?              boolean
|  |  +--rw frr?               boolean
|  +--rw bier-te
|  |  +--rw sub-domain?        bier:sub-domain-id
|  |  +--rw (encap-type)?
|  |  |  +--:(mpls)
|  |  |  +--:(non-mpls)
|  |  +--rw bitstringlength?   bier:bsl
|  |  +--rw set-identifier?    bier:si
|  |  +--rw ecmp?              boolean
|  |  +--rw frr?               boolean
|  +--rw cisco-mode
|  |  +--rw p-group?            rt-types:ip-multicast-group-address
|  |  +--rw graceful-restart?   boolean
|  |  +--rw bfd?                boolean
|  +--rw mpls
|  |  +--rw (mpls-tunnel-type)?
|  |     +--:(mldp)
|  |     |  +--rw mldp-tunnel-id?       uint32
|  |     |  +--rw mldp-frr?             boolean
|  |     |  +--rw mldp-backup-tunnel?   boolean
|  |     +--:(p2mp-te)
|  |        +--rw te-tunnel-id?         uint32
|  |        +--rw te-frr?               boolean
|  |        +--rw te-backup-tunnel?     boolean
|  +--rw pim
|     +--rw graceful-restart?   boolean
|     +--rw bfd?                boolean
+--rw multicast-underlay
   +--rw underlay-requirement?   boolean
   +--rw bgp
   +--rw ospf
   |  +--rw topology-id?   uint8
   +--rw isis
   |  +--rw topology-id?   uint16
   +--rw babel
```

```
  notifications:
    +---n head-end-event
       +--ro event-type?          enumeration
       +--ro multicast-key
       |  +--ro vpn-rd?           rt-types:route-distinguisher
       |  +--ro source-address?   ip-multicast-source-address
       |  +--ro group-address?    rt-types:ip-multicast-group-address
       |  +--ro vni-type?         virtual-type
       |  +--ro vni-value?        uint32
       +--ro overlay-tech-type?   enumeration
       +--ro transport-tech?      enumeration
       +--ro underlay-tech?       enumeration
```

5. Multicast YANG data Model

```
<CODE BEGINS> file "ietf-multicast-model.yang"
   module ietf-multicast-model {

    yang-version 1.1;

    namespace "urn:ietf:params:xml:ns:yang:ietf-multicast-model";
    prefix multicast-model;

    import ietf-inet-types {
        prefix "inet";
        reference "RFC6991";
    }

    import ietf-routing-types {
        prefix rt-types;
        reference "RFC8294";
    }

    import ietf-bier {
        prefix bier;
    }

    organization " IETF MBONED( MBONE Deployment ) Working Group";
    contact
        "WG List:  <mailto:bier@ietf.org>

         Editor:   Zheng Zhang
                   <mailto:zhang.zheng@zte.com.cn>
         Editor:   Cui Wang
                   <mailto:lindawangjoy@gmail.com>
         Editor:   Ying Cheng
                   <mailto:chengying10@chinaunicom.cn>
        ";
```

```
    description
         "The module defines the YANG definitions for multicast service
          management.

          Copyright (c) 2018 IETF Trust and the persons
          identified as authors of the code.  All rights reserved.

          Redistribution and use in source and binary forms, with or
          without modification, is permitted pursuant to, and subject
          to the license terms contained in, the Simplified BSD License
          set forth in Section 4.c of the IETF Trust's Legal Provisions
          Relating to IETF Documents
          (http://trustee.ietf.org/license-info).
          This version of this YANG module has relationship with overall multica
st
          technologies, such as PIM(RFC7761), BIER(RFC8279), MVPN(RFC6513), and
so
          on; see the RFC itself for full legal notices.";

    revision 2018-02-22 {
        description
         "Initial revision.";
       reference
         "RFC XXXX: A YANG Data Model for multicast YANG.
          RFC 7761: Protocol Independent Multicast - Sparse Mode (PIM-SM):
                    Protocol Specification (Revised).
          RFC 8279: Multicast Using Bit Index Explicit Replication (BIER);
          RFC 6513: Multicast in MPLS/BGP IP VPNs";
    }

/*key*/

    typedef ip-multicast-source-address {
        type union {
            type rt-types:ipv4-multicast-source-address;
            type rt-types:ipv6-multicast-source-address;
        }
        description
            "This type represents a version-neutral IP multicast source
            address.  The format of the textual representation implies
            the IP version.";
        reference
            "RFC8294: Common YANG Data Types for the Routing Area.";
    }

    typedef virtual-type {
        type enumeration {
            enum "vxlan" {
                description "The vxlan type. See more detail in RFC7348.";
            }
```

```
            enum "virtual subnet" {
                description "The nvgre type. See more detail in RFC7637.";
            }
            enum "vni" {
                description "The geneve type. See more detail in [ietf-nvo3-gene
ve].";
            }
        }
        description "The collection of virtual network type.";
    }

    grouping general-multicast-key {
        description "The general multicast keys. They are used to distinguish di
fferent multicast service.";
        leaf vpn-rd {
            type rt-types:route-distinguisher;
            description "A Route Distinguisher used to distinguish routes from d
ifferent MVPNs (RFC 6513).";
            reference
                "RFC8294: Common YANG Data Types for the Routing Area.";
        }
        leaf source-address {
            type ip-multicast-source-address;
            description "The IPv4/IPv6 source address of multicast flow. The val
ue set to zero means that the receiver interests in all source that relevant to
one given group.";
        }
        leaf group-address {
            type rt-types:ip-multicast-group-address;
            description "The IPv4/IPv6 group address of multicast flow. This typ
e represents a version-neutral IP multicast group address. The format of the tex
tual representation implies the IP version.";
            reference
                "RFC8294: Common YANG Data Types for the Routing Area.";
        }
        leaf vni-type {
            type virtual-type;
             description "The type of virtual network identifier. Includes the V
xlan, NVGRE and Geneve. This value and vni-value is used to indicate a specific
virtual multicast service.";
        }
        leaf vni-value {
            type uint32;
            description "The value of Vxlan network identifier, virtual subnet I
D or virtual net identifier. This value and vni-type is used to indicate a speci
fic virtual multicast service.";
        }
    }

/*overlay*/

    grouping overlay-technology {
        leaf overlay-tech-type {
            type enumeration {
                enum mld {
                description "MLD technology is used for multicast overlay. See m
ore detail in [draft-ietf-bier-mld]";
                }
                enum mvpn {
                    description "MVPN technology is used for multicast overlay.
See more detail in RFC6513.";
```

```
                }
                enum bgp {
                    description "BGP technology is used for multicast overlay. S
ee more detail in RFC7716.";
                }
                enum mld-snooping {
                    description "MLD snooping technology is used for multicast o
verlay. See more detail in RFC4541.";
                }
            }
            description "The possible overlay technologies for multicast service
.";
        }
        description "The possible overlay technologies for multicast service.";
    }

    grouping multicast-overlay {
        description "The multicast overlay information, includes ingress node an
d egress nodes' information.";
        container ingress-egress {
            description "The ingress and egress nodes address collection.";
            leaf ingress-node {
                type inet:ip-address;
                description "The ip address of ingress node for one or more mult
icast flow.
                    Or the ingress node of MVPN and BIER. In MVPN, this is the a
ddress of ingress
                    PE; in BIER, this is the BFR-prefix of ingress nodes.";
            }

            list egress-nodes {
                key "egress-node";
                description "The egress multicast nodes of multicast flow.
                    Or the egress node of MVPN and BIER. In MVPN, this is the
                    address of egress PE; in BIER, this is the BFR-prefix of
                    ingress nodes.";

                leaf egress-node {
                    type inet:ip-address;
                    description
                        "The ip-address of egress multicast nodes. See more deta
ils in RFC6513.";
                }
            }
        }

        container bier-ids {
            description "The BFR-ids of ingress and egress BIER nodes for one or
 more multicast flows.";
            leaf sub-domain {
                type bier:sub-domain-id;
                description "The sub-domain that this multicast flow belongs to.
 See more details in RFC8279.";
            }
            leaf ingress-node {
                type bier:bfr-id;
                description "The ingress node of multicast flow. This is the
```

```
                    BFR-id of ingress nodes. See more details in RFC8279.";
            }
            list egress-nodes {
                key "egress-node";
                description "This ID information of one adjacency. See more deta
ils in RFC8279.";

                leaf egress-node {
                    type bier:bfr-id;
                    description "The BFR-ids of egress multicast BIER nodes. See
 more details in RFC8279.";
                }
            }
        }

        uses overlay-technology;
    }


/*transport*/

    grouping transport-pim {
        description "The requirement information of pim transportion. PIM protoc
ol is defined in RFC7761.";
        leaf graceful-restart {
            type boolean;
            description "If the graceful restart function should be supported.";
        }
        leaf bfd {
            type boolean;
            description "If the bfd function should be supported.";
        }
    }

    grouping multicast-transport {
        description "The transport information of multicast service.";
        container bier {
            description "The transport technology is BIER. The BIER technology i
s introduced in RFC8279. The parameter is consistent with the definition in [iet
f-bier-bier-yang].";
            leaf sub-domain {
                type bier:sub-domain-id;
                description "The subdomain id that the multicast flow belongs to
. See more details in RFC8279.";
            }
            choice encap-type {
                case mpls {
                    description "The BIER forwarding depends on mpls. See more d
etails in RFC8296.";
                }
                case eth {
                    description "The BIER forwarding depends on ethernet. See mo
re details in RFC8296.";
                }
                case ipv6 {
                    description "The BIER forwarding depends on IPv6.";
```

```
                }
                description "The encapsulation type in BIER.";
            }
            leaf bitstringlength {
                type bier:bsl;
                description "The bitstringlength used by BIER forwarding. See mo
re details in RFC8279.";
            }
            leaf set-identifier {
                type bier:si;
                description "The set identifier used by the multicast flow. See
more details in RFC8279.";
            }
            leaf ecmp {
                type boolean;
                description "The capability of ECMP. If this value is set to tru
e, ecmp mechanism should be enabled. See more details in RFC8279.";
            }
            leaf frr {
                type boolean;
                description "The capability of fast re-route. If this value is s
et to true, fast re-route mechanism should be enabled. See more details in RFC82
79.";
            }
        }
        container bier-te {
            description "The transport technology is BIER-TE. BIER-TE technology
 is introduced in [ietf-bier-te-arch].";
            leaf sub-domain {
                type bier:sub-domain-id;
                description "The subdomain id that the multicast flow belongs to
. See more details in [ietf-bier-te-arch].";
            }
            choice encap-type {
                case mpls {
                    description "The BIER-TE forwarding depends on mpls. See mor
e details in [ietf-bier-te-arch].";
                }
                case non-mpls {
                    description "The BIER-TE forwarding depends on non-mpls. See
 more details in [ietf-bier-te-arch].";
                }
                description "The encapsulation type in BIER-TE.";
            }
            leaf bitstringlength {
                type bier:bsl;
                description "The bitstringlength used by BIER-TE forwarding. See
 more details in [ietf-bier-te-arch].";
            }
            leaf set-identifier {
                type bier:si;
                description "The set identifier used by the multicast flow, espe
cially in BIER TE. See more details in [ietf-bier-te-arch].";
            }
            leaf ecmp {
                type boolean;
                description "The capability of ECMP. If this value is set to tru
e, ecmp mechanism should be enabled. See more details in [ietf-bier-te-arch].";
            }
            leaf frr {
```

```
                type boolean;
                description "The capability of fast re-route. If this value is s
et to true, fast re-route mechanism should be enabled. See more details in [ietf
-eckert-bier-te-frr].";
            }
        }
        container cisco-mode {
            description "The transport technology is cisco-mode. The Cisco MDT m
ulticast mechanism is defined in RFC6037.";
            leaf p-group {
                type rt-types:ip-multicast-group-address;
                description "The address of p-group. It is used to encapsulate a
nd forward flow according to multicast tree from ingress node to egress nodes.";
            }
            uses transport-pim;
        }
        container mpls {
            description "The transport technology is mpls. MVPN overlay can use
 mpls tunnel technologies to build transport layer. The usage is introduced in RF
C6513.";
            choice mpls-tunnel-type {
                case mldp {
                    description "The mldp tunnel. The protocol detail is defined
 in RFC6388.";
                    leaf mldp-tunnel-id {
                        type uint32;
                        description "The tunnel id that correspond this flow. Th
e detail is defined in RFC6388.";
                    }
                    leaf mldp-frr {
                        type boolean;
                        description "If the fast re-route function should be sup
ported. The detail is defined in RFC6388.";
                    }
                    leaf mldp-backup-tunnel {
                        type boolean;
                        description "If the backup tunnel function should be sup
ported. The detail is defined in RFC6388.";
                    }
                }
                case p2mp-te {
                    description "The p2mp te tunnel. The protocol detail is defi
ned in RFC4875.";
                    leaf te-tunnel-id {
                        type uint32;
                        description "The tunnel id that correspond this flow. Th
e detail is defined in RFC4875.";
                    }
                    leaf te-frr {
                        type boolean;
                        description "If the fast re-route function should be sup
ported. The detail is defined in RFC4875.";
                    }
                    leaf te-backup-tunnel {
                        type boolean;
                        description "If the backup tunnel function should be sup
ported. The detail is defined in RFC4875.";
                    }
                }
                description "The collection types of mpls tunnels";
            }
        }
```

```
        container pim {
            uses transport-pim;
            description "The transport technology is PIM. PIM [RFC7761] is used
commonly in traditional network.";
        }
    }

/*underlay*/

    grouping multicast-underlay {
        description "The underlay information relevant multicast service. Underl
ay protocols are used to build transport layer. It is unnecessary in traditional
 network that use PIM [RFC7761] to build multicast tree. Diversity underlay prot
ocols can be choosed to build BIER transport layer.";
        leaf underlay-requirement {
            type boolean;
            description "If the underlay technology is required.";
        }
        container bgp {
            description "The underlay technology is BGP. BGP protocol RFC4271 sh
ould be triggered to run if BGP is used as underlay protocol.";
        }
        container ospf {
            description "The underlay technology is OSPF. OSPF protocol RFC2328
should be triggered to run if OSPF is used as underlay protocol.";
            leaf topology-id {
                type uint8;
                description "The topology id of ospf instance. The topology id c
an be assigned In some situations. More details is defined in RFC2328.";
            }
        }
        container isis {
            description "The underlay technology is ISIS. ISIS protocol should b
e triggered to run if ISIS is used as underlay protocol. Details is defined in R
FC1195.";
            leaf topology-id {
                type uint16;
                description "The topology id of isis instance. The topology id c
an be assigned In some situations.";
            }
        }
        container babel {
            description "The underlay technology is Babel. Babel protocol should
 be triggered to run if Babel is used as underlay protocol.";
        }
    }

    container multicast-model {
        description "The model of multicast YANG data. Include keys, overlay, tr
ansport and underlay.";

        list multicast-keys{
            key "vpn-rd source-address group-address vni-type vni-value";
            uses general-multicast-key;

            container multicast-overlay {
                description "The overlay information of multicast service. Overl
ay technology is used to exchange multicast flows information. Overlay technolog
y may not be used in SDN controlled completely situation, but it can be used in
partial SDN controlled situation or non-SDN controlled situation. Different over
lay technology can be choosed according to different deploy consideration.";
                uses multicast-overlay;
            }
```

```
            container multicast-transport {
```

```
                description "The transportion of multicast service. Transport pr
otocol is responsible for delivering multicast flows from ingress nodes to egres
s nodes with or without specific encapsulation. Different transport technology c
an be choosed according to different deploy consideration. Once a transport tech
nology is choosed, associated protocol should be triggered to run.";
                uses multicast-transport;
            }
            container multicast-underlay {
                description "The underlay of multicast service. Underlay protoco
l is used to build transport layer. Underlay protocol need not be assigned in or
dinary network since existed underlay protocol fits well, but it can be assigned
 in particular networks for better controlling. Once a underlay technology is ch
oosed, associated protocol should be triggered to run.";
                uses multicast-underlay;
            }
            description "The model of multicast YANG data. Include keys, overlay
, transport and underlay.";
        }
    }

/*Notifications*/

    notification head-end-event {
        leaf event-type {
            type enumeration {
                enum down {
                    description "There is something wrong with head end node, an
d head end node can’t work properlay.";
                }
                enum module-loaded {
                    description "Some new modules that can be used by multicast
flows finish loading.";
                }
                enum module-unloaded {
                    description "Some new modules that can be used by multicast
flows have been unloaded.";
                }
            }
            description "Event type.";
        }
        container multicast-key {
            uses general-multicast-key;
            description "The associated multicast keys that are influenced by he
ad end node failer.";
        }
        uses overlay-technology;

        leaf transport-tech {
            type enumeration {
                enum bier {
                    description "BIER(RFC8279) technology can be used to forward
 multicast flows.";
                }
                enum bier-te {
                    description "BIER-TE(draft-ietf-bier-te-arch) technology can
 be used to forward multicast flows.";
                }
                enum cisco-mode {
                    description "Cisco mode(RFC6037) technology can be used to f
orward multicast flows.";
                }
                enum mldp {
                    description "MLDP(RFC6388) technology can be used to forward
```

```
 multicast flows.";
               }
```

```
                enum p2mp-te {
                    description "P2MP TE(RFC4875) technology can be used to forw
ard multicast flows.";
                }
                enum pim {
                    description "PIM(RFC7761) technology can be used to forward
multicast flows.";
                }
            }
            description "The modules can be used to forward multicast flows.";
        }
        leaf underlay-tech {
            type enumeration {
                enum bgp {
                    description "BGP protocol can be used to build multicast tra
nsport layer.";
                }
                enum ospf {
                    description "OSPF protocol can be used to build multicast tr
ansport layer.";
                }
                enum isis {
                    description "ISIS protocol can be used to build multicast tr
ansport layer.";
                }
                enum babel {
                    description "Babel protocol can be used to build multicast t
ransport layer.";
                }
            }
            description "The modules can be used to build multicast transport la
yer.";
        }
        description "Notification events for the head end nodes. Like head node
failer, overlay/ transport/ underlay module loading/ unloading. And the potentia
l failer about some multicast flows and associated overlay/ transport/ underlay
technologies.";
    }
}
<CODE ENDS>
```

6. Notifications

The defined Notifications include the events of head end nodes. Like head node failer, overlay/ transport/ underlay module loading/ unloading. And the potential failer about some multicast flows and associated overlay/ transport/ underlay technologies.

7. Acknowledgements

The authors would like to thank Stig Venaas, Jake Holland for their valuable comments and suggestions.

8. Normative References

[I-D.ietf-bier-bier-yang]
Chen, R., hu, f., Zhang, Z., dai.xianxian@zte.com.cn, d., and M. Sivakumar, "YANG Data Model for BIER Protocol", draft-ietf-bier-bier-yang-03 (work in progress), February 2018.

[I-D.ietf-bier-te-arch]
Eckert, T., Cauchie, G., Braun, W., and M. Menth, "Traffic Engineering for Bit Index Explicit Replication (BIER-TE)", draft-ietf-bier-te-arch-00 (work in progress), January 2018.

[I-D.ietf-pim-yang]
Liu, X., McAllister, P., Peter, A., Sivakumar, M., Liu, Y., and f. hu, "A YANG data model for Protocol-Independent Multicast (PIM)", draft-ietf-pim-yang-13 (work in progress), January 2018.

[RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, DOI 10.17487/RFC6020, October 2010, <https://www.rfc-editor.org/info/rfc6020>.

[RFC6037] Rosen, E., Ed., Cai, Y., Ed., and IJ. Wijnands, "Cisco Systems' Solution for Multicast in BGP/MPLS IP VPNs", RFC 6037, DOI 10.17487/RFC6037, October 2010, <https://www.rfc-editor.org/info/rfc6037>.

[RFC6087] Bierman, A., "Guidelines for Authors and Reviewers of YANG Data Model Documents", RFC 6087, DOI 10.17487/RFC6087, January 2011, <https://www.rfc-editor.org/info/rfc6087>.

[RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, DOI 10.17487/RFC6241, June 2011, <https://www.rfc-editor.org/info/rfc6241>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/ BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC7223] Bjorklund, M., "A YANG Data Model for Interface Management", RFC 7223, DOI 10.17487/RFC7223, May 2014, <https://www.rfc-editor.org/info/rfc7223>.

[RFC7277] Bjorklund, M., "A YANG Data Model for IP Management", RFC 7277, DOI 10.17487/RFC7277, June 2014, <https://www.rfc-editor.org/info/rfc7277>.

[RFC8022] Lhotka, L. and A. Lindem, "A YANG Data Model for Routing Management", RFC 8022, DOI 10.17487/RFC8022, November 2016, <https://www.rfc-editor.org/info/rfc8022>.

[RFC8177] Lindem, A., Ed., Qu, Y., Yeung, D., Chen, I., and J. Zhang, "YANG Data Model for Key Chains", RFC 8177, DOI 10.17487/RFC8177, June 2017, <https://www.rfc-editor.org/info/rfc8177>.

[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>.

[RFC8294] Liu, X., Qu, Y., Lindem, A., Hopps, C., and L. Berger, "Common YANG Data Types for the Routing Area", RFC 8294, DOI 10.17487/RFC8294, December 2017, <https://www.rfc-editor.org/info/rfc8294>.

Authors' Addresses

Zheng Zhang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: zhang.zheng@zte.com.cn

Cui(Linda) Wang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: lindawangjoy@gmail.com

Ying Cheng
China Unicom
Beijing
China

Email: chengying10@chinaunicom.cn