

This Internet-Draft, draft-ietf-mboned-mtrace-v2-08.txt, has expired, and has been deleted from the Internet-Drafts directory. An Internet-Draft expires 185 days from the date that it is posted unless it is replaced by an updated version, or the Secretariat has been notified that the document is under official review by the IESG or has been passed to the RFC Editor for review and/or publication as an RFC. This Internet-Draft was not published as an RFC.

Internet-Drafts are not archival documents, and copies of Internet-Drafts that have been deleted from the directory are not available. The Secretariat does not have any information regarding the future plans of the authors or working group, if applicable, with respect to this deleted Internet-Draft. For more information, or to request a copy of the document, please contact the authors directly.

Draft Authors:

Hitoshi Asaeda<Asaeda@wide.ad.jp>

Tatuya Jinmei<Jinmei\_Tatuya@isc.org>

William Fenner<fenner@fenron.net>

Stephen Casner<casner@packetdesign.com>

Internet Engineering Task Force  
Internet-Draft  
Intended status: Informational  
Expires: January 17, 2013

M. McBride  
H. Lui  
Huawei Technologies  
July 16, 2012

Multicast in the Data Center Overview  
draft-mcbride-armd-mcast-overview-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 <http://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 January 17, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Multicast Applications in the Data Center . . . . .	3
2.1. Client-Server Applications . . . . .	3
2.2. Non Client-Server Multicast Applications . . . . .	4
3. L2 Multicast Protocols in the Data Center . . . . .	6
4. L3 Multicast Protocols in the Data Center . . . . .	7
5. Challenges of using multicast in the Data Center . . . . .	7
6. Layer 3 / Layer 2 Topological Variations . . . . .	9
7. Address Resolution . . . . .	9
7.1. Solicited-node Multicast Addresses for IPv6 address resolution . . . . .	9
7.2. Direct Mapping for Multicast address resolution . . . . .	9
8. Acknowledgements . . . . .	10
9. IANA Considerations . . . . .	10
10. Security Considerations . . . . .	10
11. Informative References . . . . .	10
Authors' Addresses . . . . .	11

## 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 discussion, in ARMD, 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. ARP/ND issues are addressed in [I-D.armd-problem-statement]

## 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 critical 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:

- o 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.
- o 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

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 [I-D.armd-problem-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. 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.

## 9. IANA Considerations

This memo includes no request to IANA.

## 10. Security Considerations

No security considerations at this time.

## 11. Informative References

- [I-D.armd-problem-statement]  
Narten, T., Karir, M., and I. Foo,  
"draft-ietf-armd-problem-statement", February 2012.
- [I-D.pim-umf-problem-statement]  
Zhou, D., Deng, H., Shi, Y., Liu, H., and I. Bhattacharya,  
"draft-dizhou-pim-umf-problem-statement", October 2010.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas,  
"Protocol Independent Multicast - Sparse Mode (PIM-SM):  
Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for  
IP", RFC 4607, August 2006.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano,  
"Bidirectional Protocol Independent Multicast (BIDIR-  
PIM)", RFC 5015, October 2007.

Authors' Addresses

Mike McBride  
Huawei Technologies  
2330 Central Expressway  
Santa Clara, CA 95050  
USA

Email: michael.mcbride@huawei.com

Helen Lui  
Huawei Technologies  
Building Q14, No. 156, Beiqing Rd.  
Beijing, 100095  
China

Email: helen.liu@huawei.com