Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: August 18, 2014
E. Vyncke, Ed.
P. Thubert
E. Levy-Abegnoli
A. Yourtchenko
Cisco
February 14, 2014

# Why Network-Layer Multicast is Not Always Efficient At Datalink Layer
draft-vyncke-6man-mcast-not-efficient-01

## Abstract

Several IETF protocols (IPv6 Neighbor Discovery for example) rely on IP multicast in the hope to be efficient with respect to available bandwidth and to avoid generating interrupts in the network nodes. On some datalink-layer network, for example IEEE 802.11 WiFi, this is not the case because of some limitations in the services offered by the datalink-layer network. This document lists and explains all the potential issues when using network-layer multicast over some datalink-layer networks.

## Status of This Memo



## Copyright Notice

Table of Contents



## 1. Introduction

Several IETF protocols rely on the use of link-local scoped IP multicast in the hope of reducing traffic over the underlying datalink network and generating less operating systems interrupts for the receiving nodes. For example, IPv6 Neighbor Discovery [RFC4861] uses link-local multicast to:

- advertise the presence of a router by sending router advertisement to IPv6 address link-local multicast address (LLMA), ff02::1, whose members are only the IPv6 nodes but per [RFC4291] section 3 those messages must be forwarded on all ports. This IPv6 LLMA is mapped to the Ethernet Multicast Address (EMA) 33:33:00:00:00:01;

- solicit the data-link layer address of an adjacent on-link node by sending a neighbor solicitation to the solicited-node multicast address corresponding to the target address such as ff02:0:0:0:0:1:ffXX:XXXX (where the last 24 bits are the last 24 bits of the target address) as described in [RFC4291]. This IPv6 LLMA is mapped to the EMA 33:33:ff:XX:XX:XX.

## 2.  Issue on Wired Ethernet Network

Most switch vendors implement MLD snooping [RFC4541] in order to forward multicast frames only to switch ports where there is a member of the IPv6 multicast group.  This optimization works by installing hardware forwarding states in the switch.  As there is a finite amount of memory in the switches, especially when the memory is used by the data plane forwarding, there is also a limit to the number of MLD optimization states i.e. a limit to the number of IPv6 multicast groups that can be optimized by the switch; frames destined to groups without such a state are flooded on all ports in the same datalink domain, and generally the use of MLD snooping is reserved to groups with a scope wider than link local.

With IPv6, all nodes have usually at least two IPv6 addresses: a link-local and a global address.  If both addresses are based on EUI-64, then they share the same 24 least-significant bits, hence there is only one solicited-node multicast address per node.  Else, there is a high probability that the 24 least-significant bits are different, hence requiring the membership to two solicited-node multicast addresses.  If a switch uses MLD snooping to install hardware-optimized multicast forwarding states for LLMA, then the switch installs two hardware-optimized states per node as EUI-64 addresses are no more commonly used.  If privacy extension addresses [RFC4941] are used, then every node can have multiple IPv6 global addresses, most of which are not based on EUI-64, a large switch fabric will have to support multiple times more states for multicast EMA than it does for unicast addresses, resulting in an excessive amount of resources in each individual switch to be built at an affordable price.

Therefore, due to cost reason, the multicast optimization by MLD snooping of solicited-node LLMA is disabled on most Ethernet switches.  This means wasting:

- the switch bandwidth as it works as a full-duplex hub;

- the nodes CPU as all nodes will have to receive the multicast frame (if their network adapter is not optimized to support MAC multicast) and quickly drop it.

A special mention must be paid when a layer-2 domain includes legacy devices working on at 10 Mbps half-duplex; for example, in hospitals having old equipments dated back of 1990.  For this case, it takes only 100 300-byte frames per second to already utilize the media to 2.4 % not to mention that the NIC and the processor have to process those frames and that the processor is probably also dated from 1990...

It is unclear what the impact is on virtual machines with different MAC addresses and different IPv6 address connected with a virtual layer-2 switch hosted on a single physical server... The MLD snooping done by the virtual switch will consume CPU by the hypervisor, hence, also reducing the amount of CPU available for the virtual machines.

Leveraging MLD snooping to save layer-2 switches from flooding link-local multicast messages carries additional challenges.  Unsolicited MLD reports are usually sent once (when link comes up) and not acknowledged.  There exist a retransmission mechanism, but it is not generally deployed, and it does not guarantee that subsequent retransmission won't also get lost.  The switch could easily end up with incomplete forwarding states for a given group, with some of the listeners ports, but not all (much worse than no state at all).  As the switch does not know one of its forwarding entry is incomplete, it can't fall back to broadcasting.  As ordinary MLD routers, the switch could query reports on a periodic basis.  However, it is not practical for layer-2 access switches to send periodic general MLD queries to maintain forwarding states accuracy for at least 2 reasons:

o  The queries must be sourced with a link-local IPv6 address, one per link, and, for many practical reasons, layer-2 switches don't have such address on each link (vlan) they operate on.

o  Since address resolution uses a multicast group, and may happen quite frequently on the link, in order to avoid black holing resolution, the interval for a switch to issue MLD general query would have to be very small (a few seconds).  These MLD queries are themselves sent to a multicast group that all nodes would need to get.  That would completely defeat the purpose of reducing multicast traffic towards end nodes.

## 3.  Issues on IEEE 802.11 Wireless Network

### 3.1.  Multicast over Wireless

Wireless networks are a shared half-duplex media: when one station transmits, then all others must be silent.  A multicast or broadcast transmission from an AP is physically transmitted to all WiFi cliens (STAs) and no other node can use the wireless medium at that time. This is the first issue with the use of wireless for multicast: the medium access behaves as a Ethernet hub.

Depending on distance and radio propagation, different wireless clients may use different transmission encodings and data rates.  A lower data rate effectively locks the medium for a longer time per bit.  In order to reach all nodes, and considering that multicast and broadcast frames are not protected by ARQ (retries), the AP is constrained to transmit all multicast or broadcast frames at the lowest rate possible, which in practice is often translated to rates as low as 1 Mbps or 6 Mbps, even when the unicast rate can reach a hundred of Mbps and above. It results that sending a single multicast frame can consume as much bandwidth as dozens of unicast frames. Table Table 1 provides some example values of the bandwidth used by multicast frames transmitted from the AP (i.e. not counting the original multicast frame transmitted by the WiFi client to the AP when he source is effectively wireless).

| Lowest WiFi rate | Highest WiFi rate | Mcast frame %-age | WiFi Utilization by Mcast |
|---|---|---|---|
| 1 Mbps | 11 Mbps | 1 % | 9 % |
| 6 Mbps | 54 Mbps | 1 % | 9 % |
| 6 Mbps | 54 Mbps | 5 % | 45 % |
| 6 Mbps | 54 Mbps | 10 % | 90 % |

Table 1: Multicast WiFi Usage

If multiple APs cover the same wireless LAN, then the multicast frames must be transmitted by all APs to all their WiFi clients.

Communication of a multicast frame by a WiFi client requires three steps:

1. The WiFi client sends a datalink unicast frame to the AP at its maximum possible rate.

2. The WiFi AP forwards this frame on its wired interface and broadcasts it (as explained above) to all its WiFi clients. If there are multiple APs on the same datalink domain, then, all APs also broadcast this multicast frame to their WiFi clients.

3. A WiFi NIC that implements the STA in the client filters the frames that are effectively expected by this device based on destination address.

Another side effect of multicast frames is that there cannot be an acknowledgement mechanism (ARQ) similar to that used for unicast frame, therefore frames can be missed and NDP does not take this non negligible packet loss into account. This could have a negative impact for Duplicate Address Detection (DAD) if the multicast NS or the multicast NA with override are lost. Assuming a error rate of 8%

of corrupted frame, this means a 8% chance of loosing a complete frame, this means a 16% chance of not detecting a duplicate address.

For a well-distributed multicast group where relatively few devices actually participate to any given group, there should be no transmission at all if none of the clients expects the multicast destination address, and there should be very few unicast but fast transmissions to the limited set of interest STAs when there is effectively a match in the set of associated devices.  But there is no mechanism in place to ensure that functionality.

## 3.2.  Host Sleep Mode

When a sleeping host wakes up by a user interaction, it cannot determine whether it has moved to another network (SSID are not unique), hence, it has to send a multicast Router Solicitation (which triggers a Router Advertisement message from all adjacent routers) and the mobile host has to do Duplicate Address Detection for its link-local and global addresses, thus means transmitting at least two multicast Neighbour Solicitation messages which will be repeated by the AP to all other WiFi clients.

This process creates a lot of multicast packets:

- one multicast Router Solicitation from the WiFi client, which is received by the AP and if the AP is not optimized, then the Router Solitication is broadcasted again over the wireless link;

- one multicast Neighbor Solitication for the host LLA from the WiFi client, which is received by the AP and if the AP is not optimized, the message is transmitted back over the wireless link;

- per global address (usually 1 or 2 depending on whether privacy extension is active), same behavior as above.

In conclusion and in the good case of not having privacy extension, this means 6 WiFi broadcast packets plus the unicast replies on each wake-up of the device.  Assuming a packet size of 80 bytes, this translates into about 120 bytes to take into account the WiFi frame format which is larger than the usual Ethernet frame, the table Table 2 gives some result of the WiFi utilization just for the multicast part of the wake-up of sleeping devices... This does not take into account the rest of the multicast utilization used by RS, RA, NS, NA, MLD, ... and the associated unicast traffic.

| WiFi Clients | Wake-up Cycle | Mcast packet/sec | Mcast bit/sec | Lowest WiFi Rate | Mcast Utilization |
|---|---|---|---|---|---|
| 100 | 600 sec | 1 | 960 bps | 1 Mbps | 0.1 % |
| 1 000 | 600 sec | 1 | 9600 bps | 1 Mbps | 1.0 % |
| 5 000 | 600 sec | 50 | 48 kbps | 1 Mbps | 4.8 % |
| 5 000 | 300 sec | 100 | 96 kbps | 1 Mbps | 9.6 % |

Table 2: Multicast WiFi Usage by Sleeping Devices

## 3.3. Low Power WiFi Clients

In order to save their batteries, Low Power (LP) hosts go into radio sleep mode until there is a local need to send a wireless frame. Before going into radio sleep mode, the LP hosts signal to the AP that they are going into sleep; this allows the AP to store unicast and multicast frames destined for those sleeping LP clients. LP clients wake up periodically to listen to the WiFi beacon frames transmitted periodically (default every 100 ms) because this beacon frame contains a bit mask (Traffic Indication Map - TIM) indicating for which STA there is waiting unicast traffic and whether there is multicast traffic waiting. If there is multicast traffic waiting, that ALL LP hosts must stay awake to receive all multicast frames sent immediately after by the AP and process them. If there is a bit indicating that unicast traffic is waiting for a specific LP host, then only this LP host will stay awake to poll the AP later to collect its traffic. The TIM maximum length is 2008 bits and the complete beacon frame is less than 300 bytes long.

The table Table 2 indicates the ration of active/sleeping time for LP hosts when multicast is present. In the absence of multicast traffic, the radio is active only 2.4 % of the time while if there are 50 multicast frames of 300 bytes per second, the radio is active 14.4 % of the time, nearly 6 times more often... with a battery life probably reduced by 6...

| Beacon frames/sec | Mcast frames/sec | Mcast frame size (bytes) | Lowest WiFi Rate | Awake time/sec |
|---|---|---|---|---|
| 10 | 0 | 300 bytes | 1 Mbps | 2.4 % |
| 10 | 5 | 300 bytes | 1 Mbps | 3.6 % |
| 10 | 10 | 300 bytes | 1 Mbps | 4.8 % |
| 10 | 50 | 300 bytes | 1 Mbps | 14.4 % |

Table 3: Multicast WiFi Impact on Low Power Hosts

### 3.4. Vendor and Configuration Optimizations

Vendors have noticed the problem and have come with several optimizations such as

- o LP hosts not waking up the main processor when they are not member of the multicast group;

- o APs no transmitting back over radio received Router Sollication multicast messages;

- o ...

AP can also work in 'AP isolation mode' where there is no direct traffic between WiFi clients, this mode has a positive side-effect when a WiFi client transmits a multicast frame as this frame is transmitted at the highest possible rate over the WiFi medium and the AP will not re-transmit if back to all other WiFi clients at the lowest rate.

### 3.5. Even Unicast NDP is not Optimum

While this is not directly related to the subject of this document, it is worth mentioning anyway as this is important for devices running on battery.

NDP cache needs to be maintained by refreshing the neighbor cache for entries which are in the STALE state.  This requires yet another Neighbor Solicitation / Neighbor Advertisement round.  Even if the destination IP and MAC addresses are unicast, this traffic is generated and again wakes up mobile devices.

## 4. Measuring the Amount of IPv6 Multicast

There are basically three ways to measure the amount of IPv6 multicast traffic:

- o sniffing the traffic and generating statistics, somehow an overkill:
- o exporting IPfix data and doing aggregation on the ff02::/16 link-local multicast prefix
- o using SNMP to query on the AP the IP-MIB [RFC4293] with commands such as:
    - * snmpwalk -c private -v 1 udp6:[2001:db8::1] -Ci -m IP-MIB ifDesc: to get the interface names and index;
    - * snmpwalk -c private -v 1 udp6:[2001:db8::1] -Ci -m IP-MIB ipIfStatsOutTransmits.ipv6: to get the global transmit counters (i.e. unicast and multicast as there is no broadcast in IPv6);
    - * snmpwalk -c private -v 1 udp6:[2001:db8::1] -Ci -m IP-MIB ipIfStatsOutMcastPkts.ipv6: to get the multicast packet counter.

## 5. Acknowledgements

The authors would like to thank Norman Finn, Michel Fontaine, Steve Simlo, Ole Troan, and Stig Venaas for their suggestions and comments.

## 6. IANA Considerations

This memo includes no request to IANA.

## 7. Security Considerations

The only security considerations about this document is that by forcing a lot of traffic to be multicast, then, a denial of service (DoS) attack could be mounted on available bandwidth and battery of some network nodes.

## 8. Informative References

[RFC4291]  Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

[RFC4293]  Routhier, S., "Management Information Base for the Internet Protocol (IP)", RFC 4293, April 2006.

[RFC4541]  Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, May 2006.

[RFC4861]  Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[RFC4941]  Narten, T., Draves, R., and S. Krishnan, "Privacy Extensions for Stateless Address Autoconfiguration in IPv6", RFC 4941, September 2007.

[packet_loss]
Department of Computer Sciences, University of Wisconsin Madison, USA, "Diagnosing Wireless Packet Losses in 802.11: Separating Collision from Weak Signal", <http://pages.cs.wisc.edu/~suman/pubs/diagnose.pdf>.

Authors' Addresses

Eric Vyncke (editor)
Cisco
De Kleetlaan, 6A
Diegem  1831
BE

Phone: +32 2 778 4677
Email: evyncke@cisco.com

Pascal Thubert
Cisco
Batiment D, 45 Allee des Ormes
MOUGINS, PROVENCE-ALPES-COTE D'AZUR  06250
France

Email: pthubert@cisco.com

Eric Levy-Abegnoli
Cisco
Batiment D, 45 Allee des Ormes
MOUGINS, PROVENCE-ALPES-COTE D'AZUR  06250
France

Email: elevyabe@cisco.com

Andrew Yourtchenko
Cisco
De Kleetlaan, 6A
Diegem  1831
BE

Phone: +32 2 704 5494
Email: ayourtch@cisco.com