INTERNET-DRAFT
Intended Status: Proposed Standard

Expires: 19 October 2020

Tathagata Nandy
HPE
Nitin Singla
HPE
Utkarsh Srivastava
HPE

April 19, 2020

Multicast Path MTU
draft-nandy-singla-utkarsh-pim-mcast-path-mtu-00

Abstract

Path MTU discovery (rfc1191) is a standard technique to determine the supported MTU between two Internet Protocol (IP) hosts to avoid any fragmentation. In a multicast distribution tree, source will not know where the receivers are located. So the technique used to compute the path MTU for a unicast stream does not work in a multicast network. This document describes a method to discover multicast path MTU with the goal to avoid traffic loss. This solution also aims to solve the problem of traffic loss in for multicast streams because of incorrect MTU setting and no path MTU support for multicast networks.

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 12 October 2020.

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

When one IP host has a large amount of data to send to another host, the data is transmitted as a series of IP datagrams. It is usually preferable that these datagrams be of the largest size that does not require fragmentation anywhere along the path from the source to the destination. (For the case against fragmentation, see [5].) This datagram size is referred to as the Path MTU (PMTU), and it is equal to the minimum of the MTUs of each hop in the path. A shortcoming of the current Internet protocol suite is the lack of a standard mechanism for a host to discover the PMTU of an arbitrary path. Note: The Path MTU is what in [1] is called the "Effective MTU for sending" (EMTU_S). A PMTU is associated with a path, which is a particular combination of IP source and destination address and perhaps a Type-of-service (TOS). The current practice [1] is to use the lesser of 576 and the first-hop MTU as the PMTU for any destination that is not connected to the same network or subnet as the source. In computer networking, multicast is group communication where data transmission is addressed to a group of destination computers simultaneously. Multicast can be one-to-many or many-to-many distribution. Multicast should not be confused with physical layer point-to-multipoint communication. Ethernet frames with a value of 1 in the least-significant bit of the first octet of the destination address are treated as multicast frames and are flooded to all points on the network. This mechanism constitutes multicast at the data link layer. This mechanism is used by IP multicast to achieve one-to-many transmission for IP on Ethernet networks. Modern Ethernet controllers filter received packets to reduce CPU load, by looking up the hash of a multicast destination address in a table, initialized by software, which controls whether a multicast packet is dropped or fully received. IP multicast is a technique for one-to-many communication over an IP network. The destination nodes send Internet Group Management Protocol join and leave messages, for example in the case of IPTV when the user changes from one TV channel to another. Multicast uses network infrastructure efficiently by requiring the source to send a packet only once, even if it needs to be delivered to a large number of receivers. The nodes in the network take care of replicating the packet to reach multiple receivers only when necessary.

2. Conventions used in this document

2.1. Terminology

The reader is assumed to be familiar with the terminology, reference models, and taxonomy defined in [RFC4664] and [RFC4665]. For readability purposes, we repeat some of the terms here. Moreover, we also propose some other terms needed when IP multicast support is discussed.

Multicast domain
: An area in which multicast data is transmitted. In this document, this term has a generic meaning that can refer to Layer-2 and Layer-3. Generally, the Layer-3 multicast domain is determined by the Layer-3 multicast protocol used to establish reachability between all potential receivers in the corresponding domain. The Layer-2 multicast domain can be the same as the Layer-2 broadcast domain (i.e., VLAN), but it may be restricted to being smaller than the Layer-2 broadcast domain if an additional control protocol is used.

PIM-SM
: Protocol Independent Multicast Sparse Mode (PIM-SM) is a family of multicast routing protocols for Internet Protocol (IP) networks that provide one-to-many and many-to-many distribution of data over a LAN, WAN or the Internet. It explicitly builds unidirectional shared trees rooted at a rendezvous point (RP) per group, and optionally creates shortest-path trees per source. PIM-SM uses shared trees by default and implements source-based trees for efficiency; it assumes that no hosts want the multicast traffic unless they specifically ask for it. Senders first send the multicast data to the RP, which in turn sends the data down the shared tree to the receivers.

RP
: Rendezvous Point (RP) is a router in a multicast network domain that acts as a shared root for a multicast shared tree. Any number of routers can be configured to work as RPs and they can be configured to cover different group ranges. An RP acts as the meeting place for sources and receivers of multicast data. In a PIM-SM network, sources must send their traffic to the RP. This traffic is then forwarded to receivers down a shared distribution tree.

2.2. Conventions

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 [RFC2119].

3. Problem Statement

3.1. Motivation

Path MTU discovery computes the lowest MTU supported between two hosts to avoid IP fragmentation. For a unicast packet, source device sends out a packet with Don't Fragment (DF) flag bit set in the IP header [1]. Any device along the path whose MTU is smaller than the packet will drop the packet and send back an ICMP Packet Too Big (Type 2) message containing its MTU, allowing the source host to reduce its Path MTU appropriately. The process is repeated until the MTU is small enough to traverse the entire path

without fragmentation. In a multicast distribution tree, the source does not know the host for a multicast group till the complete multicast tree is built. Hosts in different branches of the tree use IGMP/MLD followed by PIM to become part of the multicast tree. Generally the process starts at the host where it

sends a request to become part of a multicast tree through IGMP joins. The same request is sent to the RP and there by source and group develop a common path. So the technique mentioned above may not work for multicast flows.

3.2. Scalability

Most routers doesn't send ICMP (unreachable; fragmentation needed) messages in response to too-big IPv4 multicast packets with DF-bit set. They're just dropping these packets silently, breaking PMTUD. This is a case of as-per-design feature and is updated in section 7.2 of RFC 1112 that an ICMP error message (Destination Unreachable, Time Exceeded, Parameter Problem, Source Quench, or Redirect) is never generated in response to a datagram destined to an IP host group. The same document also describes why RFC 1112 prohibits sending ICMP error messages in response to multicast datagrams. The processing done on ICMP error replies by the *nix socket API might block the sender socket if an error comes back from a single receiver or if TTL expires when traversing a particularly long branch of the multicast tree, not exactly a good idea in multicast environment.

4. Multicast Data Path

The multicast Stream between a Source and a Host for a particular Group uses the following path.

1. Source Router sends PIM Register Packets to the Rendezvous Point (RP) Router with the Source encapsulated in it. This is a Unicast Packet.

2. Host Router Sends PIM Joins to the RP and from there the Source and the Core based tree is built.

4.1 First hop Source router and rendezvous point pre-Registration

For the network segment between the first hop router and the PIM Rendezvous point (RP), multicast data packets are encapsulated into PIM register messages. PIM Register messages are unicast messages and the standard Path MTU discovery technique will work for this segment.

4.2 Multicast Flow and PMTU

For other segments in the network, data will be sent as multicast packets and the following sequence is used to determine the path MTU for different branches in the multicast tree:

1. A new multicast flow received on any router will not have any match in the multicast routing table and hence it is treated as unknown multicast flow. Such streams are copied to CPU to program the flows in HW.

2. When the Packet is processed by multicast process to program an unknown flow it computes the Outgoing interfaces list (Olist) for the flow based on IGMP/MLD joins or PIM joins from downstream Routers.

3. The proposal is for each interfaces in the Olist, an additional check is performed where the MTU supported on the interface is compared with the size of the multicast data packet. If the packet size is greater than the supported MTU, an ICMP Fragmentation Needed (Type 3, Code 4) message containing its MTU, allowing the source DR to re-compute MTU appropriately. This is done irrespective of whether DF bit is set or not.

4. An error message will be logged in each of the Routers performing this check. Optionally an SNMP trap can also be send. This would lead the admin to either change the MTU of the Interfaces for the Multicast Data to go through or the Source DR to fragment and send the Data.

5. Optionally as per implementation, some routers can program the Mroute Entry with Error displaying that the packets might be dropped because of large size. This could be implementation specific.

6. Optionally, in all the Routers where this check is performed, the unknown Multicast Data packet can be programmed as a bridge entry in Hardware such that no further packets reach the CPU.

7. This computation is done at the Connection establishment phase itself for the PIM-SM network such that the Mroute Entry is never programmed in Hardware without the MTU computation.

4.3 Last Hop Router to the Host MTU

The Host sends IGMP Joins to join a particular group and when unknown multicast is received at the router, it would compute the MTU for those joined paths and would send an ICMP error packet back to the source if there is a violation.

1. Source host will learn about the lowest MTU supported among all the branches of the multicast tree and uses the updates the size of the datagrams accordingly.

2. This path is same as the previous section only, the only difference is that Joins are not PIM Joins but IGMP Joins.

5 IANA Considerations

This memo includes no request to IANA.

6 Security Considerations

This Path MTU Discovery mechanism makes possible two denial-of-service attacks, both based on a malicious party sending false Datagram Too Big messages to an Internet host. In the first attack, the false message indicates a PMTU much smaller than reality. This should not entirely stop data flow, since the victim

host should never set its PMTU estimate below the absolute minimum, but at 8 octets of IP data per datagram, progress could be slow. In the other attack, the false message indicates a PMTU greater than reality. If believed, this could cause temporary blockage as

the victim sends datagrams that will be dropped by some router. Within one round-trip time, the host would discover its mistake (receiving Datagram Too Big messages from that router), but frequent repetition of this attack could cause lots of datagrams to be dropped. A host, however, should never raise its estimate of the PMTU based on a Datagram Too Big message, so should not be vulnerable to this attack. A malicious party could also cause problems if it could stop a victim from receiving legitimate Datagram Too Big messages, but in this case there are simpler denial-of-service attacks available. In another case if the packets are always rejected because of higher MTU and the sender does not change the packet size or the admin does not adjust the MTU, there is a risk of a DOS attack on the Switch sending the ICMP Error packet. Multicast packet send at high rate can consume the CPU resources of all the Routers implementing the PMTU for Multicast.

7 References

7.1 Normative References

[1] J. Mogul, S. Deering. Path MTU Discovery. RFC 1191, DECWRL and Stanford University, November, 1990.

[2] J. Postel, INTERNET CONTROL MESSAGE PROTOCOL. RFC 791, ISI, September 1981.

7.2 Informative References

[3] <https://blog.ipspace.net/2015/09/path-mtu-discovery-doesnt-work-with-ip.html>

[4] <https://en.wikipedia.org/wiki/Multicast>

[5] <https://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ip-multicast/whitepaper_c11-508498.html>

8 Acknowledgments

The authors thank the contributors of [RFC1191] and RFC{5501] since the structure and content of this document were, for some sections, largely inspired from it. The authors also thank Mark Pearson and others for their valuable reviews and feedback.

9 Authors' Addresses

Tathagata Nandy
Hewlett Packard India Software Operations Pvt. Ltd.
Survey # 192, Whitefield Road,
Mahadevapura Post, Bangalore 560048. India
Phone: (+91) 9611895857
EMail: tathagata.nandy@hpe.com

Nitin Singla
Hewlett Packard India Software Operations Pvt. Ltd.
Survey # 192, Whitefield Road,
Mahadevapura Post, Bangalore 560048. India
Phone: (+91)7411937209
EMail: singla@hpe.com

Utkarsh Srivasta
Hewlett Packard India Software Operations Pvt. Ltd.
Survey # 192, Whitefield Road,
Mahadevapura Post, Bangalore 560048. India
Phone: (+91)7411937209
EMail: usrivastava@hpe.com