INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: 18 July 2021

Tathagata Nandy
HPE
Utkarsh Srivastava
HPE
January 18, 2021

Multicast Path MTU
draft-nandy-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.

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.

PIM-SSM
PIM source-specific multicast (SSM) uses a subset of PIM sparse mode and IGMP version 3 (IGMPv3) to allow a client to receive multicast traffic directly from the source. PIM SSM uses the PIM sparse-mode functionality to create an SPT between the receiver and the source, but builds the SPT without the help of an RP.

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 Path MTU

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

1. The Sender device connected Router, periodically sends probe messages for a well-known Multicast Group that falls in the PIM-SSM range. The probe packet here is nothing but small packets whose destination IP falls in the SSM group range. This should be a reserved IP and should not be used for any other regular multicast stream.

2. The Probe packets are different from the actual packets that the Source is sending. This algorithm runs on the Routers and not on the actual Source sending the Stream.

3. The receiver Routers will also run periodic probing to the Source(s). As part of the probe the receiving Routers will run Path MTU protocol to the Source Device. The PMTU will run only for Active Sources when they receive the Probe packets. This is the reason, the Sender device needs to send periodic probe packets.

4. This will be performed at all the Receiver Routers (Designated Router). All these Receiver Routers would also use the same Source which would be specifically reserved

for PMTU computation. This is the PIM SSM source for
the specified Group.

5. There are two options, one is the receiver Router (Host Connected DR) themselves sending a PIM Join for these Groups to the sources or optionally it can act on this by receiving an IGMP v3 join. In the latter case , the Host device need to send IGMP v3 joins to the Sources for Computing Path MTU

6. The Receiver DR (Host Connected) would compute PMTU to the Source by sending Probe packets of different sizes.

7. Once the receiver Router has computed the PMTU to the Source connected DR, the PMTU will be sent to the Source Router via a new option in PIM Join packet or a new type of PIM packet. A new ICMP packet is not chosen for this as this algorithm is supposed to run inside the PIM Application.

8. Once the Source Connected Designated Router receives the PMTU for all the connected paths, it would compute the minimum MTU and send it back to the Source device. This takes away all the computation headache from the Source Device. The Source device will get the periodic MTU update from all the Routers and should never send any packets with a MTU higher than this. The assumption is that TCP/IP stack with ICMP packets is implemented in all the Sources, so internally it can handle the ICMP packets.

9. The probing packets sent by the sender device can be of reduced frequency to prevent congestion

10. The receiver can keep sending the probe packets as long as it has an intended Host.

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

Utkarsh Srivastava
Hewlett Packard India Software Operations Pvt. Ltd.
Survey # 192, Whitefield Road,
Mahadevapura Post, Bangalore 560048. India
Phone: (+91) 8948794936
EMail: usrivastava@hpe.com