

Path MTU Discovery and Filtering ICMP

Marc Slemko

This document explains the details of how path MTU discovery (PMTU-D) combined with filtering ICMP messages can result in connectivity problems. If you are familiar with the terms discussed, let's start by defining what we are talking about MTU.

The maximum transmission unit is a link layer restriction on the maximum number of bytes of data in a single transmission (ie. frame, cell, packet, depending on the terminology). The below table shows some typical values for MTUs, taken from RFC-1191:

MTU	Where Commonly Used
65535	Hyperchannel
17914	16 Mbit/sec token ring
8166	Token Bus (IEEE 802.4)
4464	4 Mbit/sec token ring (IEEE 802.5)
1500	Ethernet
1500	PPP (typical; can vary widely)
576	X.25 Networks

Path MTU

The smallest MTU of any link on the current path between two hosts. This may change over time since the route between two hosts, especially on the Internet, may change over time. It is not necessarily symmetric and can even vary for different types of traffic from the same host.

Fragmentation

When a packet is too large to be sent across a link as a single unit, a router can fragment the packet. This means that it splits it into multiple parts which contain enough information for the receiver to glue them together again. Note that this is not done on a hop-by-hop basis, but once fragmented a packet will not be put back together until it reaches its destination. Fragmentation is undesirable for numerous reasons, including:

- If any one fragment from a packet is dropped, the entire packet needs to be retransmitted. This is a very significant problem.
- It imposes extra processing load on the routers that have to split the packets.
- In some configuration, simpler firewalls will block all fragments because they don't contain the header information for a higher layer protocol (eg. TCP) needed for filtering.

DF (Don't Fragment) Bit

This is a bit in the IP header that can be set to indicate that the packet should not be fragmented by routers, but instead an ICMP "can't fragment" error is returned sent to the sender and the packet is dropped.

ICMP Can't Fragment Error

This error (type 3 (destination unreachable), code 4 (fragmentation needed but don't-fragment bit set)) is returned by a router when it receives a packet that is too large for it to forward and the DF

Path MTU Discovery and Filtering ICMP

Marc Slemko

bit is set. The packet is dropped and the ICMP error is sent back to the origin host. Normally, this tells the origin host that it needs to reduce the size of its packets if it wants to get through. Recent systems also include the MTU of the next hop in the ICMP message so the source knows how big its packets can be. Note that this error is only sent if the DF bit is set; otherwise, packets are just fragmented and passed through.

MSS

The MSS is the maximum segment size. It can be announced during the establishment of a TCP connection to indicate to the other end the largest amount of data in one packet that should be sent by the remote system. Normally the packet generated will be 40 bytes larger than this; 20 bytes for the IP header and 20 for the TCP header. Most systems announce a MSS that is determined from the MTU on the interface that the traffic to the remote system passes out from the system through.

Path MTU Discovery (PMTU-D)

Now you know that Path MTUs vary. You know that fragmentation is bad. The solution? Well, one solution is Path MTU Discovery. The idea behind it is to send packets that are as large as possible while still avoiding fragmentation. A host does this by starting by sending packets that have a maximum size of the lesser of the local MTU or the MSS announced by the remote system. These packets are sent with the DF bit set. If there is some MTU between the two hosts which is too small to pass the packet successfully, then an ICMP can't fragment error will be sent back to the source. It will then know to lower the size; if the ICMP message includes the next hop MTU, it can pick the correct size for that link immediately, otherwise it has to guess.

The exact process that systems go through is somewhat more complicated to account for special circumstances. For full details, see RFC-1191.

A good indication of if a system is trying to do PMTU-D is to watch the packets it is sending with something like tcpdump or snoop and see if they have the DF bit set; if so, it is most likely trying to do PMTU-D.

Now, to the problem with ICMP filtering and PMTU-D

Now we get to the problem. Many network administrators have decided to filter ICMP at a router or firewall. There are valid (and many invalid) reasons for doing this, however it can cause problems. ICMP is an integral part of the Internet and can not be filtered without due consideration for the effects.

In this case, if the ICMP can't fragment errors can not get back to the source host due to a filter, the host will never know that the packets it is sending are too large. This means it will keep trying to send the same large packet, and it will keep being dropped--silently dropped from the view of any system on the other side of the filter. While a small handful of systems that implement PMTU-D also implement a way to detect such situations, most don't and even for those that do it has a negative impact on performance and the network.

If this is happening, typical symptoms include the ability for small packets (eg. request a very small web page) to get through, but larger ones (eg. a large web page) will simply hang. This situation

Path MTU Discovery and Filtering ICMP

Marc Slemko

can be confusing to the novice administrator because they obviously have some connectivity to the host, but it just stops working for no obvious reason on certain transfers.

There is one solution, and several workarounds, for this problem. They include:

- Fix your filters! The real problem here is filtering ICMP messages without understanding the consequences. Many packet filters will allow you to setup filters to only allow certain types of ICMP messages through. If you reconfigure them to let ICMP can't fragment (type 3, code 4) messages through, the problem should disappear. If the filter is somewhere between you and the other end, contact the administrator of that machine and try to convince them to fix the problem.
- Reduce the MTU on the machines at one end or the other. This is a workaround and should not be done unless necessary. If you reduce the MTU on the system trying to do path MTU discovery to a point where it is less than or equal to the former path MTU, it will no longer try sending packets large enough to cause problems. Similarly, if you change the MTU on the system on the other end, it will advertise a lower MSS so the sending system will only send packets with data that fits into that MSS.
- Disable PMTU-D; if you control access to the machine that is trying to do PMTU-D, and are unable to get the person administering the bogus filter to fix it, disabling PMTU-D will fix the problem for data sent by that machine. Data being received by the machine, however, can still run into the problem. With the size that HTTP requests are growing to, this could start to be a problem more and more; historically, HTTP requests have nearly always been small enough to fit through links with small MTUs in one packet. Disabling PMTU-D is simply a workaround, and should not generally be done unless necessary or you know what you are doing.

So how can using RFC 1918 addresses for router links cause problems?

On many routers, a separate IP address in the same subnet is required for each end of a point to point link. This can use address space if there are a large number of such links. Since the actual address of the links doesn't appear to impact much, many people use RFC 1918 private address space for such links. The blocks included in this are:

10.0.0.0 - 10.255.255.255 (10/8 prefix)
172.16.0.0 - 172.31.255.255 (172.16/12 prefix)
192.168.0.0 - 192.168.255.255 (192.168/16 prefix)

If you are using such addresses, then ICMP messages (including "can't fragment" errors) will normally be generated using such addresses. Since many networks filter incoming traffic from such reserved addresses, the net result is the same as if all ICMP were being filtered and can cause the same problems.