Path MTU discovery in the presence of security gateways

Michael Richardson mcr@sandelman.ottawa.on.ca
v1.1, 30 June 1997

This document describes the problem of getting accurate Path MTU information in the presence of untrusted routers. Typical Path MTU discovery is done by sending packets with the don't fragment bit set, and listening for ICMP messages from routers that want to fragment the packets. Unfortunately, these messages could be forged, and IPsec based security system(s) can not pass make direct use of these messages. An alternate, backwards compatible algorithm is suggested.

Contents

1 Introduction

1.1 Definition of terminology

Here is a network of two security gateways, a client node and a server node.

```
C---{G1}--{R1}--{R2}...{R3}--{R4}...{Rn}--{G2}---S
```

```
C is the TCP initiator.
G1/G1 are security gateways.
Rx are routers.
.. is a link with a restricted MTU.
S is the TCP listener.
```

There are both TCP endpoints and security association end points, they will be distinguished with the following terms:

C

is the transport layer originator. `TLO`

S

is the transport layer target. `TLT`

C/G1

is a network layer originator/target pair. `NLO/NLT/`

G1/G2

is a network layer originator/target pair.

G2/S

is a network layer originator/target pair.

2 Introduction to the problem

RFC1191 describes a mechanism for finding the maximum transmission unit of an arbitrary internet path. It says:

> The basic idea is that a source host initially assumes that the PMTU of a path is the (known) MTU of its first hop, and sends all datagrams on that path with the DF bit set. If any of the datagrams are too large to be forwarded without fragmentation by some router along the path, that router will discard them and return ICMP Destination Unreachable messages with a code meaning "fragmentation needed and DF set" 7. Upon receipt of such a message (henceforth called a "Datagram Too Big" message), the source host reduces its assumed PMTU for the path.

The are several problems:

1. the ICMP "Datagram Too Big" messages are sent from a intermiate router (Rx in the diagram) to the gateway machine. They are not authenticated in anyway, nor does it appear that there is any reasonable way for the routers to prove they are legitimate members of the routing path.

   An attacker could influence the MTU used, possibly reducing the MTU of the route to an unacceptably low value. This may consistute unacceptably bad service. This is an issue to the Internet Metrics WG.

   A too high an MTU would result in excessive fragmentation, which on a loosy link, may result in very high retransmission rates. IPsec tunnels do not retransmit encrypted packets, rather they depend on the TLO node to do a retransmit, so retransmitted packets result in higher encryption loads as well. A gateway with limited CPU may start discarding more datagram fragments as it spends more time encrypting.

2. the PMTU information in the ICMP messages is difficult to relay back to the TCP/UDP (or other) stacks of the sending node. So, nodes `C` and `S` continue to send using whatever MTU they started with. This defeats the point of doing PMTU in conjunction with IPsec.

3. it would be preferable to IPsec gateways for TLO nodes to have PMTU available. This allows the IPsec gateway to ask the TLO node to reduce its PMTU by the amount of overhead the ESP takes. Otherwise, the resulting ESP datagram has to be fragmented.

There are two path MTUs:

1. the TLO/TLT PMTU

2. the NLO/NLT PMTU

The ideal transport layer PMTU is the NLx PMTU minus the overhead of the ESP header and transform. For rfc1829 ESP this number is 36 bytes, for the KSM draft ESP $rfcXXXX$ this is 52 bytes (for DES, DES/HMAC-MD5-96).

2.1 Requirement for PMTU information

The information must be authenticated. This implies that none of the routers Rx may provide this information. It must come from either nodes/routers on the trusted side, or from the gateways themselves.

Only the two gateway nodes know the effective number of bytes of overhead.

Only the decrypting node can observe the fragmentation resulting from the sequence of routers, R1..Rn.

IPv6 does not allow for intermediate routers to fragment packets. Only the originating node may do so. Intermediate routers MUST send ICMP Datagram Too Big messages, and drop the packet. It should be noted, again, that there are two originators: C and G1.

3 Authenticated PMTU information

Both proposal one and two must be adapted slightly for IPv6. This is discussed later.

3.1 Proposal one

Gateway G1 MUST drop all non-local ICMP Host Unreachable datagrams (including "Datagram too bid") which arrive on its unprotected interface. The gateway MAY accept ICMP packets that are addressed to itself.

ICMP datagrams arriving via an authenticated (whether encrypted or not, depending only policy) at G1 SHOULD be passed to their destination node as normal.

Gateway G2 upon receiving an ESP or AH packet that needs to be reassembled, MUST take note of the size largest fragment received. This value is compared to the previous largest fragment size. If this size has changed by more than 10%, or more than 2*MSL time (i.e. 2 minutes) has passed since the previous ICMP message, then an ICMP Datagram Too Big message is generated. The largest fragment size is initialized to 576 bytes.

The ICMP datagram is addressed from gateway G2 to the originating node C, and gives a size that is based on the maximum fragment size (above), minus the IPsec overhead. The ICMP datagram is sent via the tunnel on which the IPsec packet was a member. I.e. the ICMP is encapsulated.

A packet arriving at G1 with the DF bit set, does not cause the DF bit to be set on the encapsulating datagram.

3.2 Proposal two

Gateway G1 MUST drop all non-local ICMP Host Unreachable datagrams (including "Datagram too bid") which arrive on its unprotected interface. The gateway MAY accept ICMP packets that are addressed to itself.

ICMP datagrams arriving via an authenticated (whether encrypted or not, depending only policy) at G1 SHOULD be passed to their destination node as normal.

Gateway G1 MUST maintain a PMTU value with its SPI/Security Association state. Packets arriving from node C with the DF bit set, and that are bigger than the PMTU value, MUST be discarded, and an ICMP Datagram Too Big message sent. In other words, the security gateway acts as a router would if the IPsec tunnel were in fact a physical interface. The PMTU value is initialized to either to the MTU of the interface on which outgoing ESP packets would travel, minus the ESP overhead.

Gateway G2 upon receiving an ESP or AH packet that needs to be reassembled, MUST take note of the size largest fragment received. This value is compared to the previous largest fragment size. If this size has changed by more than 10%, or more than 2*MSL time (i.e. 2 minutes) has passed since the previous ICMP message, then an ICMP Datagram Too Big message is generated. The largest fragment size is initialize to 576.

The ICMP datagram is addressed from gateway G2 to gateway G1, and gives a size that is based on the maximum fragment size (above), minus the IPsec overhead. The ICMP datagram is sent via the tunnel on which the IPsec packet was a member. I.e. the ICMP is encapsulated and encrypted.

A packet arriving at G1 with the DF bit set (but fitting in the MTU of the SA), does not cause the DF bit to be set on the encapsulating datagram. If the DF bit was copied, and a routing change reduced the PTMU, the datagram to be dropped, and never reach G2, so news of the PMTU change would not be relayed.

3.3 Differences

This section is still under construction. Input is requested:

1. the ICMP is generated by the near router in proposal two.
2. the ICMP in the tunnel potentially carries addresses which would not satisfy filtering rules.

4 Limits to this solution: IPv6

The major problem in the IPv6 case is that the far end gateway G2 will not see no packets if the PMTU estimate is too big. An ICMP will only be received by G1 if the PMTU estimate is small enough to transit all routers.

In order to grow the PMTU, either initially, or to take advantage of a routing change, the gateway G1 must therefore send probe packets of a larger size, knowing that the packet will be lost if the probe is too big. There are other reasons why the packet, or the response may be lost, so the probe must be done again anyway.

Further, the path may suddendly experience a drop in PMTU due to a routing change. In that case, no packets will be received at G2, so G1 must also occasionally send probes of a smaller size if it hasn't received an ICMP message in 2*MSL time. (note, this number is probably too big)

Making smaller packets is easy: the gateway can use the fragmentation facilities of IPv6 to split up an encrypted packet. A larger packet can be produced by adding more padding before encryption.

5 Security Considerations:

This entire document discusses a security protocol.

6 References:

RFC − 1825

R. Atkinson, "Security Architecture for the Internet Protocol", RFC-1825, August 1995.

RFC − 1191

J. Mogul, S. Deering, "Path MTU Discovery", RFC-1191, November 1990.

KSM − AH

New AH draft.

metrics

I. M. ISP, "How fast can it go?", draft-ietf-metrics-00.txt, work in progress: Jan. 20, 1997

*Gupta*97 − 1

V. Gupta, S. Glass, "Firewall Traversal for Mobile IP: Goals and Requirements", draft-ietf-mobileip-ft-req-00.txt, work in progress: Jan. 20, 1997

*Gupta*97 − 2

V. Gupta, S. Glass, "Firewall Traversal for Mobile IP: Guidelines for Firewalls and Mobile IP entities", draft-ietf-mobileip-firewall-trav-00.txt, work in progress: March 17, 1997

6.1 Author's Address

```
Michael C. Richardson
Sandelman Software Works Corp.
152 Rochester Street
Ottawa, ON K1R 7M4
Canada

Telephone:   +1 613 233-6809
EMail:       mcr@sandelman.ottawa.on.ca
```

6.2 Expiration and File Name

This draft expires January 9, 1997

Its file name is draft-richardson-ipsec-pmtu-disocovery-00.txt