Measurements of IPv6 Path MTU Discovery Behaviour

Ben Stasiewicz
Matthew Luckie

Network Research Group
Introduction

• Internet communications are most efficient when the largest possible packet size is used.

• Path MTU Discovery (PMTUD) used to find the largest packet size an Internet path can accommodate.

• Common perception that PMTUD is unreliable in IPv6.

• Implemented a PMTUD test and used it to survey a number of dual-stacked servers on the Internet.
PMTUD Recap

Router

Big packet

Packet Too Big

Smaller packet
Fragmentation

IPv4

• Intermediate routers can fragment packets.
• A packet whose size exceeds the next-hop MTU will be fragmented unless the IP-DF bit is set.
• Fragmentation has an adverse effect on performance.
• About 97% of web servers set the DF bit.

IPv6

• Intermediate routers cannot fragment IPv6 packets. Only the sending node can.
• A packet whose size exceeds the next-hop MTU will be discarded and cause an ICMPv6 PTB to be sent.
PMTUD in IPv6

• The success of PMTUD is particularly important in IPv6!

• Tunneled IPv6 connectivity is currently common.
  • These tunnels have smaller MTUs
  • Packets are more likely to be too big (and discarded)
  • Therefore PMTUD is needed more often in IPv6
Problems

• Firewalls filtering PTB messages.
• IPv6 Tunnels not sending PTB messages
• Creates PMTUD black holes
• Bewildering to the end user
  • Connection successfully establishes but then hangs.
IPv6 PMTUD Workarounds

1. Clamp MTU on IPv6 interfaces to 1280 bytes.
2. Rewrite the MSS in SYN packets to 1220 bytes.
   • Only affects TCP
   • Not ideal: reduced communication efficiency.
   • Preferable to fix the ICMP filtering problem.
     • If we hope to use larger MTUs one day.
PMTUD Test

• Test implemented in Scamper.
  • http://www.wand.net.nz/scamper/

• Tests an Internet host's ability to do PMTUD.
  • Supports PMTUD testing in IPv4 and IPv6.
  • Can test HTTP, SMTP and DNS servers.
  • Easy to add support for other application protocols.

• Runs on systems that use the IPFW firewall.
  • Mac OS X and FreeBSD
PMTUD Test - Operation

• Establish a TCP connection to the target server.
  • TCP Maximum Segment Size (MSS) = 1440 bytes

• Send a request packet
  • Specially crafted in an attempt to elicit a large response.

• Algorithm used for determining PMTUD success/failure depends on the response packet size:
  • Larger than 1280 bytes - Reduce Packet Size (RPS)
  • Less than or equal to 1280 bytes – Frag Header

• Post-test analysis used to detect additional successes and failures (not part of Scamper).
Reduce Packet Size (RPS) Algorithm

• Does the server use smaller response packets after it is sent a PTB message asking it to do so?
  • Yes – PMTUD Success
  • No – PMTUD Failure (likely due to ICMP filtering)

• Requires large response packets from the server:
  • IPv6 – Larger than 1280 (IPv6 Minimum PMTU) bytes

• Idea taken from:
  • Measuring the evolution of transport protocols in the Internet
    Alberto Medina, Mark Allman, Sally Floyd
    ACM/SIGCOMM Computer Communication Review 35 (2) 2005
Reduce Packet Size - Inferring Success

Scamper

Request

Response
Size = 1500 bytes

Server

PTB (MTU = 1280)

Response Retransmission
Size = 1280 bytes
Reduce Packet Size – Inferring Failure

Scamper

Request

Response
Size = 1500 bytes

TCP Timeout

PTB (MTU = 1280)

TCP Timeout

Response Retransmission
Size = 1500 bytes

TCP Timeout

PTB (MTU = 1280)

TCP Timeout

Response Retransmission
Size = 1500 bytes

TCP Timeout

PTB (MTU = 1280)

TCP Timeout

Response Retransmission
Size = 1500 bytes

Server
Does the server include a fragmentation header in its response packets after it is sent a PTB specifying an MTU < 1280 bytes? (See RFC 2460 Section 5)

- Yes – PMTUD Success
- No – Too Small

Can only be used to infer PTMUD success.

- Testing to 688 IPv6-enabled web servers found that less that half of them exhibited this behaviour.
- Using it to infer failure would result in many false positives

Does not require large response packets.
Frag Header – Inferring Success

Scamper

Request

Response
Size = 1100 bytes
Frag Hdr = no

Server

PTB (MTU = 1000)

Response Retransmission
Size = 1108 bytes
Frag Hdr = yes
Post-test Analysis – Inferring Success

• Through successful PMTUD a server can learn of a smaller MTU in the path between it and Scamper.

• Scamper was not involved and is unaware of this
  • It only sees the end result – a smaller response packet.

• The following criteria is used to infer when a server learns of a 1280 byte tunnel (PMTUD Success):
  • Server MSS > 1220
  • Received a 1280 byte response packet from the server.
  • Another data packet followed it.
Post-test Analysis – Inferring Failure

• PMTUD Failure can mean that Scamper does not receive a server’s response packet.
  • These are real-world failures that cause connections to hang.
  • Test result = No Data.

• Repeat test but with smaller MSS of 1220 bytes
  • All server response packets can make it to Scamper without being discarded for being too big (IPv6 Min PMTU = 1280)

• If this time the response packet is received:
  • No Data → PMTUD Failure
HTTP - Eliciting Large Packets

• Prior to testing a web server a script finds a URL to a large object that it serves.

• An HTTP GET request for the object should result in a large response packet from the web server.

• This is done separately for IPv4 and IPv6.
SMTP - Eliciting Large Packets

Different MTAs require different methods:

• Sendmail
  • Send the commands “HELP EHLO\r\nHELP\r\n”.

• Exim
  • Specify a really long domain name in the EHLO.

• Postfix
  • Send multiple EHLOs in the same packet.

• All three techniques were implemented but in the end we only tested Sendmail. The techniques for Exim and Postfix might be considered a breach of mail server etiquette. Would like to hear your opinions on this.
DNS - Eliciting Large Packets

• Long TXT record configured for tbit.staz.net.nz

• A recursive query for this should result in a large packet.

• Can therefore use this to test recursive name servers.

Batch Test - Address Collection

• To qualify for testing a server must be:
  • Dual-stacked
  • Have global unicast IPv4 and IPv6 addresses.
  • Be reachable on both of these addresses.

• Started with the Alexa Top 1 Million Websites List.
  • 987,891 unique domains

• Web Servers – www.$domain

• Mail Servers – Query each domain for a MX record.

• DNS Servers – Query each domain for a NS record.
### Batch Test - Vantage Points

<table>
<thead>
<tr>
<th>Vantage Point</th>
<th>Location</th>
<th>IPv6 Connectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ1</td>
<td>New Zealand</td>
<td>Tunneled (6to4)</td>
</tr>
<tr>
<td>NZ2</td>
<td>New Zealand</td>
<td>Native</td>
</tr>
<tr>
<td>US1</td>
<td>United States</td>
<td>Native</td>
</tr>
<tr>
<td>NL1</td>
<td>Netherlands</td>
<td>Native</td>
</tr>
<tr>
<td>IE1</td>
<td>Ireland</td>
<td>Native</td>
</tr>
</tbody>
</table>

Vantage point has a significant effect on the results

- NZ1 is behind a transparent web proxy.
  - All HTTP PMTUD tests went to the same host.
- IE1 has a 1280 byte tunnel configured on the next hop.
  - Server response packets limited to 1280 bytes
Batch Test

• Test Population
  • 825 dual-stacked web servers.
  • 643 dual-stacked mail servers.
  • 1504 dual-stacked name servers.

• Data collected for each test
  • Result of the PMTUD test
  • Server MSS
  • All packets sent and received during the test
PMTUD Test Results – HTTP IPv6

n = 825

Success/Failure
- Post-test Analysis
- Fragmentation Header
- Reduce Packet Size

Other
- No Data
- TCP Reset
- No Connection
- Too Small

Failure Rate: 2.6%

© THE UNIVERSITY OF WAIKATO • TE WHARE WANANGA O WAIKATO
4 May 2010
PMTUD Test Results – SMTP IPv6

n = 122

Success/Failure
- Post-test Analysis
- Fragmentation Header
- Reduce Packet Size

Failure Rate: 4.4%

Success

Failure

Other

- No Data
- No Connection
- Too Small

© THE UNIVERSITY OF WAIKATO • TE WHARE WANANGA O WAIKATO

4 May 2010
PMTUD Test Results – DNS IPv6

Success/Failure
- Post-test Analysis
- Fragmentation Header
- Reduce Packet Size

Other
- No Data
- TCP Reset
- No Connection
- Too Small

n = 1504

Failure Rate: 1.1%
Server MSS – HTTP IPv6

- Other: 75.4%
- 1440: 8.7%
- 1220: 5.2%
- 1420: 2.7%
- 1380: 0.7%
- Other: 7.3%
PMTUD Test Web Interface

Email: 

URL: 

IPv4  IPv6

IPv4 Address: 

IPv6 Address: 

Submit

Before running PMTUD tests you must first register your email. Click here to do so.

http://www.staz.net.nz/pmtud.php
Conclusion

• Results suggest that PMTUD failure in IPv6 is not as prevalent as widely believed.
  • Combined failure rate (HTTP, SMTP and DNS) is 1.9%

What you can do to help:

• Run the PMTUD test to a host on your network.
  • using scamper yourself
  • using the web interface

• Read and implement RFC 4890
  • ICMPv6 Filtering Recommendations
Allow PTB Messages

**ipfw**

ipfw add <num> allow icmp from <src> to <dst> icmptypes 3
ipfw add <num> allow ipv6-icmp from <src> to <dst> icmp6types 2

**iptables**

iptables -A <chain> -s <src> -d <dst> -p icmp –icmp-type fragmentation-needed -j ACCEPT
ip6tables -A <chain> -s <src> -d <dst> -p ipv6-icmp –icmpv6-type packet-too-big -j ACCEPT

**IOS**

access-list <id> permit icmp <src> <dst> packet-too-big
ipv6 access-list <id> permit icmp6 <src> <dst> packet-too-big

**JUNOS**

[edit firewall family inet filter <name>]
set term <name> from protocol icmp
set term <name> from icmp-type unreachable
set term <name> from icmp-code fragmentation-needed
set term <name> then accept

[edit firewall family inet6 filter <name>]
set term <name> from next-header icmp6
set term <name> from icmp-type packet-too-big
set term <name> then accept
Acknowledgements

Those who provided test machines for my use:
Dan Wing (Cisco) and Ken Key
Bill Walker (Snap Internet)

Those who ran PMTUD tests on my behalf:
Emile Aben (RIPE)
David Malone (National University of Ireland)

A big thank you to RIPE for giving me the opportunity to present at this conference!
Links

<table>
<thead>
<tr>
<th>Service</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAND</td>
<td><a href="http://www.wand.net.nz/">http://www.wand.net.nz/</a></td>
</tr>
<tr>
<td>Scamper</td>
<td><a href="http://www.wand.net.nz/scamper/">http://www.wand.net.nz/scamper/</a></td>
</tr>
<tr>
<td>Web Interface</td>
<td><a href="http://www.staz.net.nz/pmtud.php">http://www.staz.net.nz/pmtud.php</a></td>
</tr>
<tr>
<td>RFC 4890</td>
<td><a href="http://www.ietf.org/rfc/rfc4890.txt">http://www.ietf.org/rfc/rfc4890.txt</a></td>
</tr>
</tbody>
</table>
Any Questions?

Ben Stasiewicz <ben@wand.net.nz>
Matthew Luckie <mjl@wand.net.nz>