Network Working Group
Internet-Draft
Intended status: Informational
Expires: September 29, 2013

F. L. Templin, Ed.
Boeing Research & Technology

March 28, 2013

Operational Considerations for Tunnel Fragmentation and Reassembly draft-generic-v6ops-tunmtu-13.txt

Abstract

The Maximum Transmission Unit (MTU) for popular IP-in-IP tunnels is currently recommended to be set to 1500 (or less) minus the length of the encapsulation headers when static MTU determination is used. This requires the tunnel ingress to either fragment any IP packet larger than the MTU or drop the packet and return an ICMP Packet Too Big (PTB) message. Concerns for operational issues with Path MTU Discovery (PMTUD) point to the possibility of MTU-related black holes when a packet is dropped due to an MTU restriction. The current "Internet cell size" is effectively 1500 bytes (i.e., the minimum MTU configured by the vast majority of links in the Internet) and should therefore also be the minimum MTU assigned to tunnels, but this has proven to be problematic in common operational practice. This document therefore discusses operational considerations for tunnel fragmentation and reassembly necessary to accommodate this Internet cell size.

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 http://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 September 29, 2013.

Table of Contents

1. Introduction

The Maximum Transmission Unit (MTU) for popular IP-in-IP tunnels is currently recommended to be set to 1500 (or less) minus the length of the encapsulation headers when static MTU determination is used. This requires the tunnel ingress to either fragment any IP packet larger than the MTU or drop the packet and return an ICMP Packet Too Big (PTB) message [RFC0791][RFC2460]. Concerns for operational issues with Path MTU Discovery (PMTUD) [RFC1191][RFC1981] point to the possibility of MTU-related black holes when a packet is dropped due to an MTU restriction. The current "Internet cell size" is effectively 1500 bytes (i.e., the minimum MTU configured by the vast majority of links in the Internet) and should therefore also be the minimum MTU assigned to tunnels, but this has proven to be problematic in common operational practice.

[RFC4459] discusses "MTU and Fragmentation Issues with In-the-Network Tunneling" and provides a comprehensive study of the various techniques that could be applied to alleviate the issues, including:

1. Fragmenting all too big encapsulated packets to fit in the paths, and reassembling them at the tunnel endpoints.

2. Signal to all the sources whose traffic must be encapsulated, and is larger than fits, to send smaller packets, e.g., using PMTUD.

3. Ensure that in the specific environment, the encapsulated packets will fit in all the paths in the network, e.g., by using MTU bigger than 1500 in the backbone used for encapsulation.

4. Fragmenting the original too big packets so that their fragments will fit, even encapsulated, in the paths, and reassembling them at the destination nodes. Note that this approach is only available for IPv4 under certain assumptions.

After considerable effort by many individuals since the publication of [RFC4459], these four alternatives continue to cover the domain of potential solutions - all of which have drawbacks and/or impracticalities. In this document, we discuss further considerations within the framework of the only solution alternative that can be applied generically - namely, fragmentation and reassembly at the tunnel endpoints.

2. Tunnel Fragmentation and Reassembly

Pushing the tunnel MTU to 1500 bytes or beyond is met with the challenge that the addition of encapsulation headers would cause an inner IP packet that is 1500 bytes (or slightly smaller) to appear as a slightly larger than 1500 byte outer IP packet on the wire, where it may be too large to traverse the path in one piece. When an IP tunnel configures an MTU smaller than 1500 bytes, packets that are small enough to traverse earlier links in the path toward the final destination may be dropped at the tunnel ingress which then returns a PTB message to the original source. However, operational experience has shown that the PTB messages can be lost in the network [RFC2923], in which case the source does not receive notification of the loss.

It is therefore highly desirable that the tunnel configure an MTU of at least 1500 bytes even though encapsulation would cause some tunneled packets to be slightly larger than 1500 bytes. In that case, the tunnel ingress would need to make special adaptations to deliver packets that are no larger than 1500 bytes yet larger than can be accommodated in a single piece.

One possibility is to use IP fragmentation of the inner IP layer protocol before encapsulation so that inner packet fragments can be delivered via the tunnel without loss due to a size restriction and then reassembled at the final destination. This option removes the burden from the tunnel endpoints, but is only available for IPv4 packets (since IPv6 deprecates router fragmentation [RFC2460]), and is further only available when the IPv4 header sets the Don't Fragment (DF) bit in the IPv4 header to 0.

A second possibility is to use IP fragmentation of the outer IP layer protocol following encapsulation so that the outer packet fragments can be delivered via the tunnel without loss due to a size restriction and then reassembled at the tunnel egress. This option is available for tunnels over both IPv4 and IPv6, and indeed the tunnel ingress is permitted to use IPv6 fragmentation since it is acting as a "host" (i.e., and not a router) for the encapsulated packets it produces. While IPv6 fragmentation is assumed to be "safe at all speeds", IPv4 fragmentation can be dangerous at high data rates due to the possibility of Identification field wrapping while reassemblies are still active [RFC4963][RFC6864]. Also, if outer IP fragmentation were used the tunnel ingress has no assurance that the egress can reassemble packets larger than 1500 bytes, since the Minimum Reassembly Unit (MRU) is 1500 bytes for IPv6 [RFC2460] and only 576 bytes for IPv4 [RFC1122]. Finally, recent studies have shown that IPv6 fragments are sometimes dropped in the network due to middlebox misconfigurations [I-D.taylor-v6ops-fragdrop].

A third possibility for accommodating inner packets that are slightly too large is the use of "tunnel fragmentation" based on a mid-layer encapsulation that is inserted between the inner and outer IP headers. Tunnel fragmentation requires separate packet Identification and segmentation control bits in the mid-layer encapsulation that are distinct from those that appear in the inner and/or outer headers. As for outer fragmentation, the tunnel egress is responsible for reassembly. Tunnel fragmentation can be particularly useful for tunnels over IPv4, since the mid-layer encapsulation can include an extended Identification field that avoids the identification wrapping issue discussed above. However, tunnel fragmentation is not used in common widely-deployed tunneling mechanisms at the time of this writing. An example of tunnel fragmentation appears in SEAL [I-D.templin-intarea-seal].

Following any inner, tunnel or outer fragmentation, the ingress must allow the encapsulated packets or fragments to be further fragmented by a router on the path that configures a link with a too-small MTU. These fragments would be reassembled by the tunnel egress the same as if the fragmentation occurred within the tunnel ingress. This final form of fragmentation is undesirable and should be avoided if at all

possible through the application of fragmentation at the tunnel ingress. However, common widely-deployed tunneling mechanisms at the time of this writing make no such provisions.

3. Jumbo Packet Accommodation

In addition to failure to accommodate packets up to 1500 bytes in length, current tunneling solutions typically do not make provisions for delivering packets that are larger than 1500 bytes. As long as they are no larger than the underlying link used for tunneling, the tunnel ingress should admit such "jumbo" packets into the tunnel and allow them to either be delivered to the egress in one piece or be dropped with the possibility of a PTB message being returned. The original host will then be able to determine the correct packet sizes whether or not PTB messages are delivered if it is using [RFC4821]. However, this approach is not used in common widely-deployed tunneling mechanisms at the time of this writing.

4. Common Tunneling Mechanisms

The operational issues discussed in this document apply to existing IPv6-in-IPv4 transition mechanisms, including configured tunnels [RFC4213], 6to4 [RFC3056], Teredo [RFC4380], ISATAP [RFC5214], DSMIP [RFC5555], 6rd [RFC5969], etc.

The issues further apply to existing IP-in-IP tunneling mechanisms of all varieties, including GRE [RFC1701], IPv4-in-IPv4 [RFC2003], IPv6 -in-IPv6 [RFC2473], IPv4-in-IPv6 [RFC6333], IPsec [RFC4301], etc.

5. IANA Considerations

There are no IANA considerations for this document.

6. Security Considerations

The security considerations for the various tunneling mechanisms apply also to this document.

7. Acknowledgments

This method was inspired through discussion on the IETF v6ops and NANOG mailing lists in the May/June 2012 timeframe.

8. References

8.1. Normative References

[RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

[RFC2460] Deering, S.E. and R.M. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC4459] Savola, P., "MTU and Fragmentation Issues with In-the-Network Tunneling", RFC 4459, April 2006.

8.2. Informative References

[I-D.taylor-v6ops-fragdrop] Jaeggli, J., Colitti, L., Kumari, W., Vyncke, E., Kaeo, M., and T. Taylor, "Why Operators Filter Fragments and What It Implies", draft-taylor-v6ops-fragdrop-00 (work in progress), October 2012.

[I-D.templin-intarea-seal] Templin, F., "The Subnetwork Encapsulation and Adaptation Layer (SEAL)", draft-templin-intarea-seal-52 (work in progress), March 2013.

[RFC1122] Braden, R., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, October 1989.

[RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, November 1990.

[RFC1701] Hanks, S., Li, T., Farinacci, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 1701, October 1994.

[RFC1981] McCann, J., Deering, S., and J. Mogul, "Path MTU Discovery for IP version 6", RFC 1981, August 1996.

[RFC2003] Perkins, C., "IP Encapsulation within IP", RFC 2003, October 1996.

[RFC2473] Conta, A. and S. Deering, "Generic Packet Tunneling in IPv6 Specification", RFC 2473, December 1998.

[RFC2923] Lahey, K., "TCP Problems with Path MTU Discovery", RFC 2923, September 2000.

[RFC3056] Carpenter, B. and K. Moore, "Connection of IPv6 Domains via IPv4 Clouds", RFC 3056, February 2001.

[RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", RFC 4213, October 2005.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

[RFC4380] Huitema, C., "Teredo: Tunneling IPv6 over UDP through Network Address Translations (NATs)", RFC 4380, February 2006.

[RFC4821] Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", RFC 4821, March 2007.

[RFC4963] Heffner, J., Mathis, M., and B. Chandler, "IPv4 Reassembly Errors at High Data Rates", RFC 4963, July 2007.

[RFC5214] Templin, F., Gleeson, T., and D. Thaler, "Intra-Site Automatic Tunnel Addressing Protocol (ISATAP)", RFC 5214, March 2008.

[RFC5555] Soliman, H., "Mobile IPv6 Support for Dual Stack Hosts and Routers", RFC 5555, June 2009.

[RFC5969] Townsley, W. and O. Troan, "IPv6 Rapid Deployment on IPv4 Infrastructures (6rd) -- Protocol Specification", RFC 5969, August 2010.

[RFC6333] Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion", RFC 6333, August 2011.

[RFC6864] Touch, J., "Updated Specification of the IPv4 ID Field", RFC 6864, February 2013.

Author's Address

Fred L. Templin (editor)
Boeing Research & Technology
P.O. Box 3707
Seattle, WA 98124
USA

Email: fltemplin@acm.org