Intarea Working Group
Internet-Draft
Intended status: Best Current Practice
Expires: January 3, 2015

R. Bonica
Juniper Networks
C. Pignataro
Cisco Systems
J. Touch
USC/ISI
July 2, 2014

A Fragmentation Strategy for Generic Routing Encapsulation (GRE) draft-bonica-intarea-gre-mtu-05

Abstract

This memo specifies a default GRE tunnel fragmentation strategy, which has been implemented by many vendors and widely deployed on the Internet.

This memo also specifies requirements for GRE implementations. Having satisfied these requirements, a GRE implementation will execute the default GRE tunnel fragmentation strategy, specified herein, with minimal configuration. However, with additional configuration, the GRE implementation can execute any of the tunnel fragmentation strategies defined in RFC 4459.

Requirements Language

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 RFC 2119 [RFC2119].

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 January 3, 2015.

Table of Contents

1. Introduction

Generic Routing Encapsulation (GRE) [RFC2784] [RFC2890] can be used to carry any network layer protocol over any network layer protocol. GRE has been implemented by many vendors and is widely deployed in the Internet.

The GRE specification, by design, does not describe procedures to address fragmentation. Lacking guidance from the specification, vendors have developed implementation-specific fragmentation strategies. Because fragmentation procedures are local to the GRE ingress router, devices implementing one fragmentation strategy can interoperate with devices that implement another fragmentation strategy. Operational experience has demonstrated the relative merits of each strategy. [RFC4459] describes several fragmentation strategies and evaluates the relative merits of each.

This memo reviews the fragmentation strategies presented in [RFC4459]. It also specifies a default GRE tunnel fragmentation strategy, which has been implemented by many vendors and widely deployed on the Internet.

Finally, this memo specifies requirements for GRE implementations. Having satisfied these requirements, a GRE implementation will execute the default GRE tunnel fragmentation strategy, specified herein, with minimal configuration. However, with additional configuration, the GRE implementation can execute any of the strategies defined in[RFC4459].

This memo specifies requirements beyond those stated in [RFC2784]. However, it does not update [RFC2784]. Therefore, a GRE implementation can comply with [RFC2784] without satisfying the requirements of this memo.

This memo addresses point-to-point unicast GRE tunnels that carry IPv4, IPv6 or MPLS payloads. All other tunnel types are beyond the scope of this document.

1.1. Terminology

The following terms are specific to GRE and are taken from [RFC2784]:

- GRE delivery header - an IPv4 or IPv6 header whose source address is that of the GRE ingress and whose destination address is that of the GRE egress. The GRE delivery header encapsulates a GRE header.

- GRE header - the GRE protocol header. The GRE header is encapsulated in the GRE delivery header and encapsulates GRE payload.

- GRE payload - a network layer packet that is encapsulated by the GRE header. The GRE payload can be IPv4, IPv6 or MPLS. Procedures for encapsulating IPv4 and IPv6 in GRE are described in [RFC2784] and [RFC2890]. Procedures for encapsulating MPLS in GRE

are described in [RFC4023]. While other protocols may be delivered over GRE, they are beyond the scope of this document.

- GRE delivery packet - A packet containing a GRE delivery header, a GRE header, and GRE payload.

- GRE payload header - the IPv4, IPv6 or MPLS header of the GRE payload

- GRE overhead - the combined size of the GRE delivery header and the GRE header, measured in octets

The following terms are specific MTU discovery:

- link MTU (LMTU) - the maximum transmission unit, i.e., maximum packet size in octets, that can be conveyed over a link. LMTU is a unidirectional metric. A bidirectional link may be characterized by one LMTU in the forward direction and another MTU in the reverse direction.

- path MTU (PMTU) - the minimum LMTU of all the links in a path between a source node and a destination node. If the source and destination node are connected through an equal cost multipath (ECMP), the PMTU is equal to the minimum LMTU of all links contributing to the multipath.

- GRE MTU (GMTU) - the maximum transmission unit, i.e., maximum packet size in octets, that can be conveyed over a GRE tunnel without fragmentation of any kind. The GMTU is equal to the PMTU associated with the path between the tunnel ingress and the tunnel egress, minus the GRE overhead

- Path MTU Discovery (PMTUD) - A procedure for dynamically discovering the PMTU between two nodes on the Internet. PMTUD procedures rely on a router's ability to deliver ICMP [RFC0792] [RFC4443] feedback to the host that originated a packet. PMTUD procedures for IPv4 are defined in [RFC1191]. PMTUD procedures for IPv6 are defined in [RFC1981].

- Packetization Layer PMTU Discovery (PLPMTUD) - An alternative to PMTUD that is designed to operate correctly in the absence of ICMP feedback from a router to the host that originated a packet. PLPMTUD procedures are defined in [RFC4821].

The following terms are introduced by this memo:

- fragmentable packet - An IPv4 packet with DF-bit equal to 0 and whose payload is larger than 64 bytes

- o ICMP Packet Too Big (PTB) message - an ICMPv4 [RFC0792] Destination Unreachable message with code equal to 4 (fragmentation needed and DF set) or an ICMPv6 [RFC4443] Packet Too Big message

2. Strategic Overview

2.1. Candidate Strategies

Section 3 of [RFC4459] identifies several strategies that a tunnel ingress router can execute in order to prevent payload packets with size greater than the GMTU from being black-holed inside of a tunnel. When applied to GRE, these actions are:

1. Discard the payload packet and send an ICMP PTB message to the payload source. The ICMP PTB message specifies the GMTU associated with the GRE tunnel. Upon receipt of the ICMP PTB message, the payload source revises its estimate of the PMTU associated with the payload destination. As a result, the payload source refrains from sending packets to that destination with size greater than the GMTU.

2. Fragment the payload packet and encapsulate each fragment within a complete GRE header and GRE delivery header.

3. Encapsulate the payload packet in a single GRE header and GRE delivery header. If the GRE payload is fragmentable and the GRE delivery header is IPv4, set the DF-bit on the GRE delivery header to 0, allowing the GRE delivery packet to be fragmented downstream. Also, if the delivery packet is IPv4 or IPv6 and the GRE delivery packet size exceeds the GMTU, fragment the GRE delivery packet.

In Strategies 1) and 2) the GRE payload packet is fragmented, and the task of reassembly is assigned to the payload destination. By contrast, in Strategy 3) the GRE delivery packet is fragmented, and the task of reassembly is assigned to the GRE egress router. In scenarios where the GRE egress router is not known to have sufficient compute and memory resources to support reassembly, Strategies 1) and 2) are preferable to Strategy 3).

However, Strategy 1) is effective only if the payload source executes PMTUD procedures and the GRE ingress router can deliver ICMP PTB messages to the payload source. In scenarios where the payload source does not execute PMTUD procedures or the GRE ingress router cannot deliver ICMP PTB messages to the payload source , Strategies 2) and 3) are preferable to Strategy 1).

Strategy 2) is applicable only when the GRE payload is fragmentable. In all other cases, Strategies 1) or 3) are required.

Finally, Strategies 1) and 2) are effective only if the GRE ingress router maintains a sufficiently conservative estimate of the GMTU. Likewise, Strategy 3) is effective only if the GRE ingress router maintains a sufficiently conservative estimate of the GMTU or the GRE delivery packet is IPv4. Therefore, Strategy 3) is preferable to Strategies 1) and 2) when the GRE ingress router does not maintain a sufficiently conservative estimate of the GMTU and the GRE delivery header is IPv4.

2.2. Default Strategy

This section describes a default GRE fragmentation strategy that has been implemented by many vendors and has been widely deployed on the Internet.

When the GRE ingress router receives a non-fragmentable payload packet with length greater than the GMTU, the GRE ingress router discards the packet and sends an ICMP PTB message to the payload source. Upon receipt of the ICMP PTB message, the payload source revises its estimate of the PMTU associated with the payload destination. As a result, the payload source refrains from sending packets to that destination with size greater than the GMTU. See Strategy 1), above.

When the GRE ingress router receives a fragmentable packet with length greater than the GMTU, if fragments the payload packet and encapsulates each fragment within a complete GRE header and GRE delivery header. See Strategy 2), above.

3. Generic Requirements for GRE Ingress Routers

3.1. General

GRE ingress routers MUST satisfy all of the requirements stated in [RFC2784].

3.2. GRE MTU (GMTU) Estimation and Discovery

GRE ingress routers MUST support a configuration option through which a PMTU estimate can be associated with a GRE tunnel. The PMTU estimate reflects an estimate of the PMTU that the GRE ingress router associates with the GRE egress router. The default value of this configuration item MUST be less than or equal to the LMTU of the next-hop to the GRE egress router. However, GRE ingress routers MUST

permit network operators to explicitly configure this value to be greater or less than its default.

GRE ingress routers SHOULD execute either PMTUD or PLPMTUD procedures to further refine their PMTU estimate. However, if an implementation supports PMTUD or PLPMTUD for GRE tunnels, it MUST include a configuration option that disables those procedures. This configuration option may be required to mitigate certain denial of service attacks (see Section 8).

GRE ingress routers MUST set the GMTU estimate to a value that is less than or equal to the PMTU estimate minus the GRE overhead. The ingress router's GMTU estimate will not always reflect the actual GMTU. It is only an estimate. When the GMTU associated with a tunnel changes, the tunnel ingress router will not discover that change immediately. Likewise, if the ingress router performs PMTUD procedures and tunnel interior routers cannot deliver ICMP feedback to the tunnel ingress, GMTU estimates may be inaccurate.

4. Procedures Affecting The GRE Deliver Header

4.1. Tunneling GRE Over IPv4

By default, the GRE ingress router MUST set the DF-bit in the delivery header to 1 (Don't Fragment). However, the GRE ingress router MUST support a configuration option that invokes the following behavior:

- when the GRE payload is IPv6, the DF-bit on the delivery header is set to 0 (Fragments Allowed)

- when the GRE payload is IPv4, the DF-bit value is copied from the payload header to the delivery header

When the DF-bit on the delivery header is set to 0, the GRE delivery packet may be fragmented by any router between the GRE ingress and egress and the GRE delivery packet will be reassembled by the GRE egress.

By default, the GRE ingress router MUST NOT emit a delivery header with MF-bit equal to 1 (More Fragments) or Offset greater than 0. However, the GRE ingress router MAY include a configuration option that allows this.

4.2. Tunneling GRE Over IPv6

By default, the GRE ingress router MUST NOT emit a delivery header containing a fragment header. However, the GRE ingress router MAY include a configuration option that allows this.

5. Procedures Affecting the GRE Payoad

This section defines procedures that GRE ingress routers execute when they receive a packet a) whose next-hop is a GRE tunnel and b) whose size is greater than the GMTU associated with that tunnel.

5.1. IPv4 Payloads

If the payload is non-fragmentable, the GRE ingress router MUST discard the packet and send an ICMPv4 Destination Unreachable message to the payload source, with code equal to 4 (fragmentation needed and DF set). The ICMP Destination Unreachable message MUST contain an Next-hop MTU (as specified by [RFC1191]) and the next-hop MTU MUST be equal to the GMTU associated with the tunnel.

If the payload is fragmentable, the GRE ingress router MUST fragment the payload and submit each fragment to GRE tunnel. Therefore, the GRE egress router will receive complete, non-fragmented packets, containing fragmented payloads. The GRE egress router will forward the payload fragments to their ultimate destination where they will be reassembled.

5.2. IPv6 Payloads

The GRE ingress router MUST discard the packet and send an ICMPv6 [RFC4443] Packet Too Big message to the payload source. The MTU specified in the Packet Too Big message MUST be equal to the GMTU associated with the tunnel.

5.3. MPLS Payloads

The GRE ingress router MUST discard the packet. As it is impossible to reliably identify the payload source, the GRE ingress router MUST NOT attempt to send an ICMPv4 Destination Unreachable message or an ICMPv6 Packet Too Big message to the payload source.

6. GRE Egress Router Procedures

This section defines procedures that all GRE egress routers must execute.

If the GRE egress router reassembles packets carrying GRE payloads, it MUST process the Explicit Congestion Notification (ECN) bits as described in Section 5.3 of [RFC3168].

7. IANA Considerations

This document makes no request of IANA.

8. Security Considerations

PMTU Discovery is vulnerable to two denial of service attacks (see Section 8 of [RFC1191] for details). Both attacks are based upon on a malicious party sending forged ICMPv4 Destination Unreachable or ICMPv6 Packet Too Big messages to a host. In the first attack, the forged message indicates an inordinately small PMTU. In the second attack, the forged message indicates an inordinately large MTU. In both cases, throughput is adversely affected. On order to mitigate such attacks, GRE implementations MUST include a configuration option to disable PMTU discovery on GRE tunnels. Also, they MAY include a configuration option that conditions the behavior of PMTUD to establish a minimum PMTU.

9. Acknowledgements

The authors would like to thank Fred Baker, Fred Detienne, Jagadish Grandhi, Jeff Haas, Vanitha Neelamegam, John Scudder, Mike Sullenberger and Wen Zhang for their constructive comments. The authors also express their gratitude to an anonymous donor, without whom this document would not have been written.

10. References

10.1. Normative References

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

[RFC0792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, September 1981.

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

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

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

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

[RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.

[RFC2890] Dommety, G., "Key and Sequence Number Extensions to GRE", RFC 2890, September 2000.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

[RFC4023] Worster, T., Rekhter, Y., and E. Rosen, "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.

[RFC4443] Conta, A., Deering, S., and M. Gupta, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", RFC 4443, March 2006.

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

10.2. Informative References

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

Authors' Addresses

Ron Bonica
Juniper Networks
2251 Corporate Park Drive Herndon
Herndon, Virginia 20170
USA

Email: rbonica@juniper.net

Carlos Pignataro
Cisco Systems
7200-12 Kit Creek Road
Research Triangle Park, North Carolina 27709
USA

Email: cpignata@cisco.com

Joe Touch
USC/ISI
4676 Admiralty Way
Marina del Rey, California 90292-6695
USA

Phone: +1 (310) 448-9151
Email: touch@isi.edu
URI: http://www.isi.edu/touch