Intarea Working Group
Internet-Draft
Intended status: Best Current Practice
Expires: December 24, 2013

R. Bonica
Juniper Networks
C. Pignataro
Cisco Systems
June 22, 2013

A Fragmentation Strategy for Generic Routing Encapsulation (GRE)
draft-bonica-intarea-gre-mtu-02

Abstract

This memo documents a GRE fragmentation strategy that has been implemented by many vendors and deployed in many networks. It was written so that a) implementors will be aware of best common practice and b) those who rely on GRE will understand how implementations work. The scope of this memo is limited to point-to-point GRE tunnels. All other tunnel types are beyond the scope of this memo.

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 December 24, 2013.

Copyright Notice

Table of Contents

1. Introduction

Generic Routing Encapsulation (GRE) [RFC2784] 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 on the Internet.

[RFC2784], by design, does not describe procedures that affect fragmentation. Lacking guidance from the specification, vendors have developed implementation-specific fragmentation strategies. For the most part, devices implementing one fragmentation strategy can interoperate with devices that implement another fragmentation

strategy. Operational experience has demonstrated the relative merits of each strategy. Section 3 of [RFC4459] describes four fragmentation strategies and evaluates the relative merits of each.

This memo documents a GRE fragmentation strategy that has been implemented by many vendors and deployed in many networks. It was written so that a) implementors will be aware of best common practice and b) those who rely on GRE will understand how implementations work. The scope of this memo is limited to point-to-point GRE tunnels. All other tunnel types are beyond the scope of this memo.

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

1.1. How To Use This Document

This memo is presented in sections. Section 2 reviews four fragmentation strategies presented in [RFC4459] and provides an overview the strategy described herein.

Section 3 defines generic requirements for GRE ingress routers. These include compliance with the specifications of [RFC2784] and Tunnel MTU Estimation and Discovery.

Section 4 defines procedures affecting generation of the GRE delivery header. It is divided into two subsections. Section 4.1 is applicable when GRE is delivered over IPv4 [RFC0791] and Section 4.2 is applicable when GRE is delivered over IPv6 [RFC2460].

Section 5 defines procedures for handling payloads that are so large that they cannot be forwarded through the GRE tunnel without fragmentation. Section 5.1 is applicable when the payload is IPv4, Section 5.2 is applicable when the payload is IPv6 and Section 5.3 is applicable with the payload is MPLS.

Section 6 discusses IANA considerations and Section 7 discusses security considerations.

1.2. 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]. 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 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.

- tunnel MTU (TMTU) - the maximum transmission unit, i.e., maximum packet size in octets, that can be conveyed over a GRE tunnel without fragmentation. The TMTU 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 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 MTU Discovery (PLMTUD) - An extension of PMTUD that is designed to operate correctly in the absence of ICMP feedback from a router to the host that originated a packet. PLMTUD procedures are defined in [RFC4821]

The following terms are introduced by this memo:

- o fragmentable packet - all IPv4 packets with DF-bit equal to 0
- o non-fragmentable packet - all IPv4 packets with DF-bit equal to 1. Also, for the purposes of this document, all IPv6 packets are considered to be non-fragmentable.

2. Candidate Strategies and Strategic Overview

2.1. Candidate Strategies

Section 3 of [RFC4459] identifies the following tunnel fragmentation strategies:

1. Fragmentation and Reassembly by the Tunnel Endpoints
2. Signalling the Lower MTU to the Sources
3. Encapsulate Only When There is Free MTU
4. Fragmentation of the Inner Packet

In Strategy 1, the tunnel ingress router encapsulates the entire payload, without fragmentation, into a single GRE-delivery packet. It then fowards the GRE-delivery packet in the direction of the tunnel egress. If the GRE-delivery packet exceeds the LMTU of any link along the path to the tunnel egress, the router directly upstream of that link fragments it. The tunnel egress router reassembles the GRE-delivery packet, de-encapsulates its payload, and processes the payload appropriately.

In Strategy 2, the tunnel ingress router performs PMTUD procedures or some variant thereof (e.g., PLMTUD). When the tunnel ingress router receives a non-fragmentable IPv4 packet so large that it cannot be forwarded through the tunnel, it discards the packet and sends an ICMPv4 [RFC0792] Destination Unreachable message to the packet source, with type equal to 4 (fragmentation needed and DF set). The ICMP Destination Unreachable message contains a Next-hop MTU (as specified by [RFC1191]) and the next-hop MTU is equal to the TMTU associated with the tunnel. If the ICMPv4 message reaches the packet source, and if the packet source executes PMTUD procedures, the packet source adjusts its PMTU for the packet destination and emits subsequent packets with size less than the TMTU.

In Strategy 3, the network is engineered so that all network ingress links have LMTU less than the TMTU of any tunnel contained by the network. In this case, all packets entering the network are small

enough to be forwarded through any tunnel contained by the network, without fragmentation. The entire issue is thus avoided.

In Strategy 4, the tunnel ingress router performs PMTUD procedures or some variant thereof (e.g., PLMTUD). When the tunnel ingress router receives a fragmentable IPv4 packet so large that it cannot be forwarded through the tunnel without fragmentation, it fragments the payload and encapsulates each payload fragment in to a complete, separate GRE-delivery packet. It forwards those complete packets to the tunnel egress router which de-encapsulates them and forwards each payload fragment, individually and without re-assembly, to the payload destination. The payload destination reassembles packet.

Strategy 3 is attractive because it avoids fragmentation. However, networks cannot always be designed to meet the requirements of Strategy 3. When this is the case, Strategies 1, 2 and 4 become applicable.

Strategy 2 is also attractive, because it avoids fragmentation. However, Strategy 2 requires the payload source and the tunnel egress to execute PMTUD procedures. PMTUD procedures require ICMP feedback from downstream routers and fail when the network blocks required ICMP messages. Therefore, Strategy 2 can cause blackholing in networks that block ICMP.

Strategy 1 is an attractive alternative to Strategy 1, because it does not rely on PMTUD. However, Strategy 1 may not be feasible in many operational environments because it assigns the task of reassembly to the tunnel egress router. When the tunnel supports high data rates, reassembly at the tunnel egress is not cost-effective.

Strategy 4 moves the task of packet reassembly from the tunnel egress to the payload destination. However, it is applicable only when the payload is fragmentable. Furthermore, it requires the tunnel ingress router to perform PMTUD procedures and fails when the network blocks ICMP messages from tunnel interior to the tunnel ingress.

2.2. Strategic Overview

The fragmentation strategy described herein, has two modes of operation. The default mode resembles Strategies 2 and 4, above. When a GRE ingress router runs in the default mode, and it receives a non-fragmentable packet that is too large to forward through the tunnel, it behaves as described in Strategy 2, above. When the it receives a fragmentable packet that is too large to forward through the tunnel, it behaves as described in Strategy 4, above. In neither case will the GRE ingress router fragment the GRE-delivery packet.

When GRE is delivered over IPv4, the DF-bit on the delivery header is always set to 1 (Don't Fragment).

Default mode operation is desirable with the following conditions are true:

- the payload source supports PMTUD procedures
- the tunnel ingress supports PMTUD procedures
- the network does not block ICMP messages required by PMTUD

Realizing that some devices do not support PMTUD and that some networks indiscriminately block ICMP messages, the fragmentation strategy described herein includes a non-default mode, which incorporates some characteristics of Strategy 1, above.

When a GRE ingress router runs in the non-default mode, and it receives a non-fragmentable packet that is too large to forward through the tunnel, it behaves as described in Strategy 2, above. When the it receives a fragmentable packet that is too large to forward through the tunnel, it behaves as described in Strategy 4, above. In neither case will the GRE ingress router fragment the GRE-delivery packet. In this respect, the default and non-default modes are identical to one another.

However, if the ingress router delivers fragmentable payload over IPv4, it copies the DF-bit value from the payload header to the delivery header. Therefore, the GRE delivery packet may be fragmented by any router between the GRE ingress and egress. When this occurs, the GRE delivery packet is reassembled by the GRE egress.

The non-default mode of operation is desirable in some scenarios where networks block ICMP messages required by PMTUD.

3. Generic Requirements for GRE Ingress Routers

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

3.1. General

Implementations MUST satisfy all of the requirements stated in [RFC2784].

3.2. Tunnel MTU (TMTU) Estimation and Discovery

Implementations MUST maintain a running TMTU estimate. The TMTU associated with a tunnel MUST NOT, at any time, be greater than the LMTU associated with the next-hop towards the tunnel egress minus the GRE overhead.

Implementations SHOULD execute either PMTUD or PLMTUD procedures to further refine their TMTU estimate. If they do so, they MUST set the TMTU to a value that is less than or equal to the discovered PMTU minus the GRE overhead.

However, if an implementation supports PMTUD or PLMTUD 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 7). When PMTUD is disabled, the TMTU MUST be set to a value that is less than or equal to the LMTU associated with the next-hop towards tunnel egress, minus the GRE overhead.

The ingress router's TMTU estimate will not always reflect the actual TMTU. It is only an estimate. When the TMTU 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, TMTU estimates may be inaccurate.

4. Procedures Affecting The GRE Deliver Header

This section defines procedures that GRE ingress routers execute while generating the GRE delivery 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). Also, 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 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.

4.2. Tunneling GRE Over IPv6

The GRE ingress router MUST NOT emit a delivery header containing a fragment header.

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 TMTU 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 type 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 TMTU 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 TMTU 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. IANA Considerations

This document makes no request of IANA.

7. 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.

8. Acknowledgements

The authors would like to thank Jagadish Grandhi, Jeff Haas, 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.

9. References

9.1. 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.

[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.

9.2. 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