Intarea Working Group
Internet-Draft
Intended status: Best Current Practice
Expires: November 30, 2013

R. Bonica
Juniper Networks

May 29, 2013

Generic Routing Encapsulation (GRE) Fragmentation Strategy
draft-bonica-intarea-gre-mtu-00

Abstract

This memo documents a GRE fragmentation strategy upon which many vendors have converged. Specifically, it defines procedures to be executed by GRE ingress routers. It is published so that those building new implementations will be aware of best common practice. It is also published so that those building applications over GRE will understand how GRE works.

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 will expire on November 30, 2013.

Copyright Notice

Table of Contents

1. Problem Statement

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 interoperate with devices that implement another fragmentation strategy.

However, implementors and network operators have discovered that some fragmentation strategies work better than others. A poorly chosen

fragmentation strategy can cause operational issues, including black-holing, packet reassembly on GRE egress routers and unexpected interactions with Path MTU Discovery [RFC1191] [RFC1981].

This memo documents a GRE fragmentation strategy upon which many vendors have converged. Specifically, it defines procedures to be executed by GRE ingress routers. It is published so that those building new implementations will be aware of best common practice. It is also published so that those building applications over GRE will understand how GRE works.

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 enumerates design goals. Section 3 defines procedures that all GRE ingress routers must execute.

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 tunneled over IPv4[RFC0791] and Section 4.2 is applicable when GRE is tunneled 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 tunnel ingress and whose destination address is that of the GRE tunnel 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].

- 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 without fragmentation

- path MTU (PMTU) - the minimum LMTU of all the links in a path between a source node and a destination node

- 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

2. Design Goals

The following is an ordered list of design goals for this specification:

1. Avoid black-holing

2. Avoid fragmentation

3. If fragmentation cannot be avoided, avoid fragmentation procedures that require reassemby on the GRE egress router.

As an alternative to fragmentation, the procedures described herein rely on PMTU Discovery at the payload source. Therefore, the procedures described herein cause the GRE ingress router to provide the payload source with all ICMP feedback required for PMTU Discovery.

3. Common Procedures

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) Discovery

Implementations MUST maintain a local data structure that reflects the TMTU of each GRE tunnel that originates on the node. The TMTU MUST be equal to the PMTU associated with the path between the tunnel ingress and the tunnel egress, minus the GRE overhead.

By default, implementations MUST discover the PMTU associated with the path between the tunnel ingress and the tunnel egress. PMTU discovery procedures defined in [RFC1191] and [RFC1981] and will never permit the PMTU to exceed the LMTU associated with the first IP hop in the path to the tunnel egress.

However, implementations MUST include a configuration option that disables PMTU Discovery for GRE tunnels. This configuration option may be required to mitigate certain denial of service attacks (see Section 7). When PMTU discovery for GRE tunnels is disabled, the TMTU for a tunnel MUST default to the LMTU associated with the first IP hop in the path to the tunnel egress, minus the GRE overhead. However, implementations MAY include a configuration option through which the TMTU can be set to another value, which is likely to be lower.

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

When the GRE ingress router tunnels an IPv4 payload over IPv4, and the DF Bit in the payload header is set to 1 (Don't Fragment), the GRE ingress router MUST set the DF bit in the delivery header to 1.

When the GRE ingress router tunnels an IPv4 payload over IPv4, and the DF Bit in the payload header is set to 0 (May Fragment), by default, the GRE ingress router MUST set the DF bit in the delivery header to 1. However, implementations MAY include a configuration option that allows the DF bit to be copied from the payload header to the delivery header.

When the GRE ingress router tunnels an IPv6 payload over IPv4, the GRE ingress router MUST set the DF bit in the delivery header to 1.

The GRE ingress router MUST NOT emit a delivery header in which the MF bit is set to 1 (More Fragments).

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 DF bit in the payload header is set to 1 (Don't Fragment), the GRE ingress router MUST discard the packet and sent an ICMPv4 [RFC0792] 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 DF bit in the payload header is set to 0 (May Fragment), 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

7.1. VPN Considerations

[RFC4364] introduces the concept of a Virtual Routing and Forwarding Table (VRF). When a GRE ingress router forwards an ICMP message to the payload source, it MUST forward that message using the appropriate VRF. Failure to do so would a) cause information to leak between VRFs and b) prevent the ICMP message from reaching its intended destination.

Specifically, the GRE ingress router MUST forward the ICMP message using the VRF that is associated with the interface upon which the payload arrived.

7.2. Attacks Against PMTU Discovery

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.

8. Acknowledgements

The authors would like to thank John Scudder, Jeff Haas and Jagadish Grandhi for their constructive comments.

9. 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.E. and R.M. 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.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[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.

Author's Address

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

Email: rbonica@juniper.net