       A Widely Deployed Solution to the Generic Routing Encapsulation (GRE)
                           Fragmentation Problem

Abstract

   This memo describes how many vendors have solved the Generic Routing
   Encapsulation (GRE) fragmentation problem. The solution described
   herein is configurable. It is widely deployed on the Internet in its
   default configuration.

Copyright Notice

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 does not describe fragmentation procedures.
   Lacking guidance from the specification, vendors have developed
   implementation-specific fragmentation solutions. A GRE tunnel will
   operate correctly only if its ingress and egress nodes support
   compatible fragmentation solutions. [RFC4459] describes several
   fragmentation solutions and evaluates their relative merits.

   This memo reviews the fragmentation solutions presented in [RFC4459].
   It also describes how many vendors have solved the GRE fragmentation
   problem. The solution described herein is configurable and has been
   widely deployed in its default configuration.

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

1.1. Terminology

   The following terms are specific to GRE:

   o GRE delivery header - an IPv4 or IPv6 header whose source address
      represents the GRE ingress node and whose destination address
      represents the GRE egress node. The GRE delivery header
      encapsulates a GRE header.

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

   o 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 in GRE are described in
      [RFC2784] and [RFC2890]. Procedures for encapsulating IPv6 in GRE
      are described in [IPv6-GRE]. 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.

   o GRE delivery packet - a packet containing a GRE delivery header, a
      GRE header, and the GRE payload.

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

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

   The following terms are specific to MTU discovery:

   o 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
      LMTU in the reverse direction.

   o 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 nodes are connected through an Equal-Cost Multipath
      (ECMP), the PMTU is equal to the minimum LMTU of all links
      contributing to the multipath.

   o 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 GRE ingress and the GRE
      egress nodes minus the GRE overhead.

   o Path MTU Discovery (PMTUD) - a procedure for dynamically
      discovering the PMTU between two nodes on the Internet. PMTUD
      procedures for IPv4 are defined in [RFC1191]. PMTUD procedures
      for IPv6 are defined in [RFC1981].

   The following terms are introduced by this memo:

   o Fragmentable Packet - a packet that can be fragmented by the GRE
      ingress node before being transported over a GRE tunnel. That is,
      an IPv4 packet with the Don't Fragment (DF) bit equal to 0 and
      whose payload is larger than 64 bytes. IPv6 packets are not
      fragmentable.

   o ICMP Packet Too Big (PTB) message - an ICMPv4 [RFC792] Destination
      Unreachable message (Type = 3) with code equal to 4 (fragmentation
      needed and DF set) or an ICMPv6 [RFC4443] Packet Too Big message
      (Type = 2).

1.2. 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].

2. Solutions

2.1. RFC 4459 Solutions

   Section 3 of [RFC4459] identifies several tunnel fragmentation
   solutions. These solutions define procedures to be invoked when the
   tunnel ingress router receives a packet so large that it cannot be
   forwarded through the tunnel without fragmentation of any kind. When
   applied to GRE, these procedures are:

   1. Discard the incoming packet and send an ICMP PTB message to the
       incoming packet's source.

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

   3. Encapsulate the incoming packet in a single GRE header and GRE
       delivery header. Perform source fragmentation on the resulting
       GRE delivery packet.

   As per RFC 4459, Strategy 2 is applicable only when the incoming
   packet is fragmentable. Also as per RFC 4459, each strategy has its
   relative merits and costs.

2.2. A Widely Deployed Solution

   Many vendors have implemented a configurable GRE fragmentation
   solution. In its default configuration, the solution behaves as
   follows:

   o When the GRE ingress node receives a fragmentable packet with
      length greater than the GMTU, it fragments the incoming packet and
      encapsulates each fragment within a complete GRE header and GRE
      delivery header. Fragmentation logic is as specified by the
      payload protocol.

   o When the GRE ingress node receives a non-fragmentable packet with
      length greater than the GMTU, it discards the packet and sends an
      ICMP PTB message to the packet's source.

   o When the GRE egress node receives a GRE delivery packet fragment,
      it silently discards the fragment without attempting to reassemble
      the GRE delivery packet to which the fragment belongs.

   In non-default configurations, the GRE ingress node can execute any
   of the procedures defined in RFC 4459.

   The solution described above is widely deployed on the Internet in
   its default configuration. However, the default configuration is not
   always appropriate for GRE tunnels that carry IPv6.

   IPv6 requires that every link in the Internet have an MTU of 1280
   octets or greater. On any link that cannot convey a 1280-octet
   packet in one piece, link-specific fragmentation and reassembly must
   be provided at a layer below IPv6.

   Therefore, the default configuration is appropriate for tunnels that
   carry IPv6 only if the network is engineered so that the GMTU is
   guaranteed to be 1280 bytes or greater. In all other scenarios, a
   non-default configuration is required.

   In the non-default configuration, when the GRE ingress router
   receives a packet lager than the GMTU, the GRE ingress router
   encapsulates the entire packet in a single GRE and delivery header.
   It then fragments the delivery header and sends the resulting
   fragments to the GRE egress node, where they are reassembled.

3. Implementation Details

   This section describes how many vendors have implemented the solution
   described in Section 2.2.

3.1. General

   The GRE ingress nodes satisfy all of the requirements stated in
   [RFC2784].

3.2. GRE MTU (GMTU) Estimation and Discovery

   GRE ingress nodes support a configuration option that associates a
   GMTU with a GRE tunnel. By default, GMTU is equal to the MTU
   associated with the next hop toward the GRE egress node minus the GRE
   overhead.

   Typically, GRE ingress nodes further refine their GMTU estimate by
   executing PMTUD procedures. However, if an implementation supports
   PMTUD for GRE tunnels, it also includes a configuration option that

   disables PMTUD. This configuration option is required to mitigate
   certain denial-of-service attacks (see Section 4).

   The GRE ingress node's estimate of the GMTU will not always be
   accurate. It is only an estimate. When the GMTU changes, the GRE
   ingress node will not discover that change immediately. Likewise, if
   the GRE ingress node performs PMTUD procedures and interior nodes
   cannot deliver ICMP feedback to the GRE ingress node, GMTU estimates
   may be inaccurate.

3.3. GRE Ingress Node Procedures

   This section defines procedures that GRE ingress nodes execute when
   they receive a packet whose size is greater than the relevant GMTU.

3.3.1. Procedures Affecting the GRE Payload

3.3.1.1. IPv4 Payloads

   By default, if the payload is fragmentable, the GRE ingress node
   fragments the incoming packet and encapsulates each fragment within a
   complete GRE header and GRE delivery header. Therefore, the GRE
   egress node receives several complete, non-fragmented delivery
   packets. Each delivery packet contains a fragment of the GRE
   payload. The GRE egress node forwards the payload fragments to their
   ultimate destination where they are reassembled.

   Also by default, if the payload is not fragmentable, the GRE ingress
   node discards the packet and sends an ICMPv4 Destination Unreachable
   message to the packet's source. The ICMPv4 Destination Unreachable
   message code equals 4 (fragmentation needed and DF set). The ICMPv4
   Destination Unreachable message also contains a next-hop MTU (as
   specified by [RFC1191]), and the next-hop MTU is equal to the GMTU
   associated with the tunnel.

   The GRE ingress node supports a non-default configuration option that
   invokes an alternative behavior. If that option is configured, the
   GRE ingress node fragments the delivery packet. See Section 3.3.2
   for details.

3.3.1.2. IPv6 Payloads

   By default, the GRE ingress node discards the packet and sends an
   ICMPv6 [RFC4443] Packet Too Big message to the payload source. The
   MTU specified in the Packet Too Big message is equal to the GMTU
   associated with the tunnel.

   The GRE ingress node supports a non-default configuration option that
   invokes an alternative behavior. If that option is configured, the
   GRE ingress node fragments the delivery packet. See Section 3.3.2
   for details.

3.3.1.3. MPLS Payloads

   By default, the GRE ingress node discards the packet. As it is
   impossible to reliably identify the payload source, the GRE ingress
   node does not attempt to send an ICMP PTB message to the payload
   source.

   The GRE ingress node supports a non-default configuration option that
   invokes an alternative behavior. If that option is configured, the
   GRE ingress node fragments the delivery packet. See Section 3.3.2
   for details.

3.3.2. Procedures Affecting the GRE Deliver Header

3.3.2.1. Tunneling GRE over IPv4

   By default, the GRE ingress node does not fragment delivery packets.
   However, the GRE ingress node includes a configuration option that
   allows delivery packet fragmentation.

   By default, the GRE ingress node sets the DF bit in the delivery
   header to 1 (Don't Fragment). However, the GRE ingress node also
   supports a configuration option that invokes the following behavior:

   o When the GRE payload is IPv6, the DF bit on the delivery header is
      set to 0 (Fragments Allowed).

   o When the GRE payload is IPv4, the DF bit is copied from the
      payload header to the delivery header.

   When the DF bit on an IPv4 delivery header is set to 0, the GRE
   delivery packet can be fragmented by any router between the GRE
   ingress and egress nodes.

   If the GRE egress node is configured to support reassembly, it will
   reassemble fragmented delivery packets. Otherwise, the GRE egress
   node will discard delivery packet fragments.

3.3.2.2. Tunneling GRE over IPv6

   By default, the GRE ingress node does not fragment delivery packets.
   However, the GRE ingress node includes a configuration option that
   allows this.

   If the GRE egress node is configured to support reassembly, it will
   reassemble fragmented delivery packets. Otherwise, the GRE egress
   node will discard delivery packet fragments.

3.4. GRE Egress Node Procedures

   By default, the GRE egress node silently discards GRE delivery packet
   fragments without attempting to reassemble the GRE delivery packets
   to which the fragments belongs.

   However, the GRE egress node supports a configuration option that
   allows it to reassemble GRE delivery packets.

4. Security Considerations

   In the GRE fragmentation solution described above, either the GRE
   payload or the GRE delivery packet can be fragmented. If the GRE
   payload is fragmented, it is typically reassembled at its ultimate
   destination. If the GRE delivery packet is fragmented, it is
   typically reassembled at the GRE egress node.

   The packet reassembly process is resource intensive and vulnerable to
   several denial-of-service attacks. In the simplest attack, the
   attacker sends fragmented packets more quickly than the victim can
   reassemble them. In a variation on that attack, the first fragment
   of each packet is missing so that no packet can ever be reassembled.

   Given that the packet reassembly process is resource intensive and
   vulnerable to denial-of-service attacks, operators should decide
   where the reassembly process is best performed. Having made that
   decision, they should decide whether to fragment the GRE payload or
   GRE delivery packet accordingly.

   Some IP implementations are vulnerable to the Overlapping Fragment
   Attack [RFC1858]. This vulnerability is not specific to GRE and
   needs to be considered in all environments where IP fragmentation is
   present. [RFC3128] describes a procedure by which IPv4
   implementations can partially mitigate the vulnerability. [RFC5722]
   mandates a procedure by which IPv6-compliant implementations are
   required to mitigate the vulnerability. The procedure described in

RFC 5722 completely mitigates the vulnerability. Operators SHOULD
ensure that the vulnerability is mitigated to their satisfaction on
equipment that they deploy.

PMTUD 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. In order to mitigate
such attacks, GRE implementations include a configuration option to
disable PMTUD on GRE tunnels. Also, they can include a configuration
option that conditions the behavior of PMTUD to establish a minimum
PMTU.

5. References

5.1. Normative References

   [RFC792]   Postel, J., "Internet Control Message Protocol", STD 5,
               RFC 792, DOI 10.17487/RFC0792, September 1981,
               <http://www.rfc-editor.org/info/rfc792>.

   [RFC1191]   Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191,
               DOI 10.17487/RFC1191, November 1990,
               <http://www.rfc-editor.org/info/rfc1191>.

   [RFC1858]   Ziemba, G., Reed, D., and P. Traina, "Security
               Considerations for IP Fragment Filtering", RFC 1858,
               DOI 10.17487/RFC1858, October 1995,
               <http://www.rfc-editor.org/info/rfc1858>.

   [RFC1981]   McCann, J., Deering, S., and J. Mogul, "Path MTU Discovery
               for IP version 6", RFC 1981, DOI 10.17487/RFC1981, August
               1996, <http://www.rfc-editor.org/info/rfc1981>.

   [RFC2119]   Bradner, S., "Key words for use in RFCs to Indicate
               Requirement Levels", BCP 14, RFC 2119,
               DOI 10.17487/RFC2119, March 1997,
               <http://www.rfc-editor.org/info/rfc2119>.

   [RFC2784]   Farinacci, D., Li, T., Hanks, S., Meyer, D., and P.
               Traina, "Generic Routing Encapsulation (GRE)", RFC 2784,
               DOI 10.17487/RFC2784, March 2000,
               <http://www.rfc-editor.org/info/rfc2784>.

   [RFC2890] Dommety, G., "Key and Sequence Number Extensions to GRE",
              RFC 2890, DOI 10.17487/RFC2890, September 2000,
              <http://www.rfc-editor.org/info/rfc2890>.

   [RFC3128] Miller, I., "Protection Against a Variant of the Tiny
              Fragment Attack (RFC 1858)", RFC 3128,
              DOI 10.17487/RFC3128, June 2001,
              <http://www.rfc-editor.org/info/rfc3128>.

   [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed.,
              "Encapsulating MPLS in IP or Generic Routing Encapsulation
              (GRE)", RFC 4023, DOI 10.17487/RFC4023, March 2005,
              <http://www.rfc-editor.org/info/rfc4023>.

   [RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet
              Control Message Protocol (ICMPv6) for the Internet
              Protocol Version 6 (IPv6) Specification", RFC 4443,
              DOI 10.17487/RFC4443, March 2006,
              <http://www.rfc-editor.org/info/rfc4443>.

   [RFC5722] Krishnan, S., "Handling of Overlapping IPv6 Fragments",
              RFC 5722, DOI 10.17487/RFC5722, December 2009,
              <http://www.rfc-editor.org/info/rfc5722>.

5.2. Informative References

   [IPv6-GRE] Pignataro, C., Bonica, R., and S. Krishnan, "IPv6 Support
              for Generic Routing Encapsulation (GRE)", Work in
              Progress, draft-ietf-intarea-gre-ipv6-10, June 2015.

   [RFC4459] Savola, P., "MTU and Fragmentation Issues with In-the-
              Network Tunneling", RFC 4459, DOI 10.17487/RFC4459, April
              2006, <http://www.rfc-editor.org/info/rfc4459>.

Authors' Addresses

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

   Email: rbonica@juniper.net

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

   Email: cpignata@cisco.com

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

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