Tunnel MTU and Advisory Packet Too Big Messages

IETF92 intarea Session

Fred L. Templin
fltemplin@acm.org

draft-templin-6man-linkadapt

Tunnels Always Reduce Effective MTU

- Encapsulation header(s) consume data bytes
- Breaks “1500 Everywhere” assumption
- Exacerbated by tunnels within tunnels
- Tunnels perform link adaptation (RFCC2460) if MTU is insufficient

Upholding “1500 Everywhere” Assumption

- Tunnel ingress has three fragmentation zones:
  - (size <= 1280-ENCAPS) – send without fragmenting (no PTBs will result)
  - (size > 1500) – send without fragmenting (PTBs may result)
  - (1280-ENCAPS < size <= 1500) – send with fragmentation and determine whether fragmentation is necessary
- Probe to see if 1500’s can get through:
  - If yes, suspend fragmentation
  - If no, continue fragmenting
- If fragmentation is needed (i.e., “link adaptation”) tell the original source (?)
  - ➢ **“Advisory PTBs”**

“Advisory” PTBs

- When it has to fragment, the tunnel ingress can send PTB with a size smaller than 1280 **subject to rate limiting**. It can then:
  - Discard the payload packet (i.e., PTB as “loss” indication), or
  - Fragment the delivery packet (i.e., PTB as “advisory” indication)
- When the source gets the PTB, it “must” include a frag header in future packets but need not reduce the size of packets below 1280 (per RFC2460)
  - Not all sources do this
  - Sources that don’t do it are non-compliant
- Source could instead:
  - Reduce the size of the packets it sends to a size smaller than 1280
  - Fragment future packets that are no larger than 1500 (IPv6 minMRU) so the tunnel ingress doesn’t have to fragment

Tunnel Ingress Options

- When a source sends an “atomic fragment”, i.e. an IPv6 packet no larger than 1500 with a fragment header but (M=0; Offset=0), the tunnel ingress can:
  - Fragment the payload packet into two fragments, then encapsulate and send both fragments in separate delivery packets. **These fragments will be reassembled by the final destination, which is required to reassemble at least 1500.**
  - Perform “tunnel fragmentation” on the payload packet then encapsulate and send both fragments in separate delivery packets**. These fragments will be reassembled at the tunnel egress, but the ingress needs to know the egress can reassemble this much (AERO says 2KB minimum).**
  - Encapsulate the payload packet, then fragment the delivery packet. **These fragments will be reassembled at the tunnel egress, but the egress is only required to reassemble 1500, which might not leave enough room for encapsulation headers.**

Non-IP Encapsulations

- IP/GRE/Ethernet – Ethernet needs to see **1518**, and **there is no such thing as a PTB**
- Means that egress must be able to reassemble at least **1518+ENCAPS**, and that fragmentation cannot be avoided
- Might actually need more than 1518 for some IEEE encodings
- Ethernet-within-Ethernet encapsulations??

Documents of Interest

- draft-templin-6man-linkadapt
- draft-templin-aerolink
- draft-templin-aeromin
- draft-ietf-intarea-gre-ipv6
- draft-herbert-gue