AERO Tunnel MTU

IETF93 - July 20, 2015

Fred L. Templin

Fred.l.Templin@boeing.com

AERO Tunnel MTU Mitigation

- RFC4459 Solutions:
  - Fragmentation and Reassembly by the Tunnel Endpoints
  - Signaling the Lower MTU to the Sources
  - Encapsulate Only When There is Free MTU
  - Fragmentation of the Inner Packet
- AERO observes that aspects of each approach are applied according to the specific situation – there is no one-size-fits-all
- https://datatracker.ietf.org/doc/draft-templin-aerolink/
- https://datatracker.ietf.org/doc/draft-herbert-gue-fragmentation/

Path MTU Discovery

- When the tunnel ingress, tunnel egress and original source are **all within the same well-managed administrative domain**, use standard Path MTU Discovery. Reasons:
  - Tunnel ingress will receive authentic Packet Too Big (PTB) messages from a router on the path to the egress w/o loss due to filtering middleboxes or spoofing from an attacker that can spoof the source address
  - Original source will receive authentic PTB messages from the tunnel ingress if the tunnel MTU is insufficient

Fragmentation and Reassembly by Tunnel Endpoints

- When the original source and/or tunnel egress are in different administrative domains than the tunnel ingress, the tunnel ingress treats each packet to be tunneled as follows:
  - If packet is <= (1280-HLEN), encapsulate and send
  - If packet is > (1280-HLEN) and <= 1500, encapsulate and fragment using **TUNNEL FRAGMENTATION** as opposed to outer or inner IP fragmentation (reason: avoids filtering middleboxes and IP ID wraparound)
  - If packet is > 1500 bytes, encapsulate and send if packet fits in first hop MTU. **Original sources that send packets larger than 1500 SHOULD use RFC4821.**
- Tunnel fragmentation uses Generic UDP Encapsulation (GUE)
- When tunnel fragmentation is used, reassembly occurs at an egress near the destination; not somewhere in the middle of the network
- Means reassembly does not impact performance-intensive nodes

Fragmentation of the Inner Packet

- If the inner packet is IPv4 with DF=0, and inner packet is larger than the smaller of 1500 and the path MTU (if known), fragment inner packet into 1024 byte fragments then encapsulate each fragment
- Reason:
  - Sources that send IPv4 packets with DF=0 must have some way of knowing that the destination is able to reassemble if necessary
  - Tunnel should let destination do the reassembly if necessary
  - **Tunnel fragmentation still applies if packet is no larger than 1500 and the tunnel path MTU is unknown**

Encapsulate Only When There is Free MTU

- More and more, links in the middle of the network between the ingress and egress configure MTUs that are larger than the size required to pass a 1500 byte tunneled packet
- Question is how tunnel ingress can tell when this is the case?
- Possible answer – probe the forward path with 1500 byte probe packets
- Problem – no way of knowing whether the probe packets will follow the same path as data packets (e.g., due to ECMP, LAG, etc.)
- Resolution
  - Operational assurance that probes follow same path as data allows optimization
  - Else, use PMTUD when possible
  - Else, tunnel fragmentation always works

Summary

- Aspects of all RFC4459 solutions are employed according to the specific situation
- No one-size-fits-all solution – a systems approach is needed
- Take advantage of known larger Path MTUs when possible
- Else, use standard Path MTU discovery when possible
- Else, use **tunnel fragmentation** instead of IP fragmentation when fragmentation is necessary
- Make sure than any necessary reassembly occurs at a tunnel egress near the edge of the network and not near the middle of the network