PMTUD Over Vxlan (draft-saum-nvo3-pmtud-over-vxlan)

Saumya Dikshit
Vengada Prasad Govindan

Requirement(s)

- Operational Requirements
  - Vxlan Overlay in a typical datacenter deployment leverages an underlay L3 network for establishing tunnels (VTEPs)
  - The end-user device (VMs in hypervisor hosting applications, Blade servers) connects over Vxlan core (tunnels) in a operationally disparate network, which is transparent from end-user perspective.
  - The ICMP errors generated in underlay are not translated or percolated to end-user devices rather black-holed in the Vxlan Gateways (tunnel end-points)
    - ICMP Packet Too Big (and corresponding ipv4) also meet the same fate.
  - Potential overhead to Fragment and Reassemble packets in the core network.
    - Manageability Headache: Explicit Configuration required to ensure fragmentation and reassembly to be avoided

Requirement(s)

- Business Requirements
  - Optimal usage of bandwidth in core
    - More fragments lead to more metadata and more bandwidth usage in the core.
  - More usage of computational (cpu) resource for fragmentation and reassembly.
  - With millions of VMs connected to cloud in MSDC deployment with WebOTT based applications leading the graph with 90% of intra-datacenter traffic, the above two bullets can potentially lead to havoc.
  - Considering IPv6 only datacenter as what future beholds
    - IPv6 PMTUD becomes mandatory for end-devices
  - Dual stack devices (applications) connecting to legacy IPv4 deployments surely need a solution too.

Problem Description

- RFC1981 states IPV6 Path MTU discovery is based on the "Packet too big" icmpv6 error code, generated by the networking device(s) which is/are capable of generating such messages on encountering packet paths which go over link with MTU size smaller than packet size.
- Current standard based implementation(s) may lead to black-holing in case of data center deployments with Vxlan as core.
  - Vxlan Gateway(or TOR) MAY not set the DF bit in the outer IP header encapsulation.
  - Vxlan Gateway(or TOR) is incapable of relaying icmp error "Fragmentation Needed and Don't Fragment was Set", generated by IPv4 enabled network device (underlay network), to IPv6 enabled end-point host/vm/server(source of the original packet).
  - Same problems are expected with all permutations and combinations (ipv4 and ipv6) of underlay and overlay network.

Potential Solution

- Vxlan Gateway encapsulates the Vxlan header onto the client packet at ingress overlay tunnel and also decapsulates the overlay header for packets egressing out
  - Solution becomes more apt to be installed on devices playing such role (Top of Rack device).
- Vxlan gateways should set the DF-bit in Outer IP header encapsulation for client packets that are wrapped with vxlan, thus ensuring, ICMP error packet is generated for packet size exceeding the link MTU in (ipv4) underlay network.
- Vxlan gateway devices should translate the ICMP error "Destination Unreachable" with code 'Fragmentation Needed and Don't Fragment was Set', into a ICMPv6 error 'Packet too big' packet and vice versa.
- This mandates that original packet in icmp error message MUST carry information about the inner payload(original packet). This aids in relaying the errors to end-point devices.