Increasing the MTU of the Internet

Tom Scholl, AT&T Labs
NANOG 42 – Feb 19 2008
A Quick Review of the Acronyms

- **MTU** = Maximum Transmission Unit
  - The largest size packet or frame possible on a link.
- **PMTUD** = Path MTU Discovery
  - A host-based discovery mechanism for handling MTU mismatches without requiring packet fragmentation
So what is the MTU of “The Internet”? 

- Every media technology has its own MTU
  - Many technologies offer relatively large MTUs
    - HSSI/FDDI, Packet over SONET, ATM, all supported 4470 MTU
  - Ethernet
    - Official IEEE standardized MTU for Ethernet is 1500 bytes

- But Ethernet sets the de facto standard at 1500
  - Ethernet is by far the most common edge technology
    - Thus rendering larger frame support in the middle useless
  - Path MTU Discovery is very easy to break
    - Just ask anyone who has ever tried to run PPPoE or other tunnels over Ethernet without lowering their host MTU/MSS.
Benefits of Bigger Packets

• Simplistic View: Bigger packets = Fewer Packets/sec
  • Fewer routing lookups, fewer copies, fewer interrupts, etc
  • All true, and has linear benefits, but misses the big picture

• Big Picture View
  • 1500 bytes is horrifically inefficient for high speed transfers
  • Hosts perform memory transactions in page sized chunks
    • Typically 4096 or 8192 bytes
  • The ultimate goal is packets carrying page-sized payloads
    • Allows for zero-copy optimization technique called page flipping.
    • Avoids copying the packet from NIC to kernel to user-land to kernel to disk by simply remapping the memory pages.
Did somebody say Jumbo Frame?

- What is a Jumbo Frame?
  - No standard – Anything larger than 1500 bytes really.
- The official 1500 byte MTU value for Ethernet has remained unchanged since the original in 1980.
  - IEEE has repeatedly failed to standardize anything larger.
  - However, almost every modern Ethernet port produced in the past few years supports some kind of Jumbo Frame
    - Purely due to customer demand, not part of any a standard.
- Unfortunately, there is no standardization on the frame size above 1500.
Path MTU Discovery

- Fragmenting and reassembling packets is hard
  - Slow path for routers, impacts performance on hosts too
- PMTUD detects lower MTU to avoid fragmentation
  - Host sends packets with Don’t Fragment (DF) bit set
  - If path MTU is too small, router sends ICMP NeedFrag
  - Host receives ICMP and lowers packet sizes accordingly
- But PMTUD is remarkably easy to break
  - If the ICMP NeedFrag packet is blocked, PMTUD breaks
  - If any router pairs have mismatched MTUs, ICMP breaks
  - If PMTUD breaks, traffic is blackholed, potentially forever
Why is Path MTU Mismatch so fatal?

Router A cannot transmit ICMP NeedFrag messages back to its Source (Host A) since it does not know that Router B cannot handle the large frame.
Inter-Provider Jumbo Frame Support

- Can be accomplished via Point to Point links
  - Just make sure both sides agree on the MTU
- Significantly harder via Multipoint links (IX VLAN)
  - No mechanism exists to negotiate MTU per IP/next-hop
  - Even though many routers support per-next-hop MTUs

- How can this be fixed?
  - For Ethernet, the proper fix would be via ARP
    - But good luck getting THAT implemented.
  - Alternate hacks include negotiation via BGP.
  - Or just picking a number and hoping everyone supports it.
• Some Internet Exchanges have gone the route of separate VLANs for jumbos (NetNOD)
  • We have IPv4, IPv6, Multicast and VoIP peering VLANs, do we need a jumbo too?
What is a good target MTU value?

- A common pitch is “somewhere around” 9kB
  - Design goal is 8192 data payload + some room for headers
  - Headers like TCP, IPv4 or IPv6, IPSec, PPPoE, L2TP, etc.
    - AKA “don’t try to send the largest possible packet every time”.
    - And then tunneling through the Internet might actually work!
  - For the most part, this is a good improvement over 1500.

- In the long term, much larger values may make sense
  - IPv4/Ethernet support up to 65535 length packets/frames.
  - IPv6 supports 32 bit values (4 billion bytes) for length.
Implementation Caveats

- Not every vendor talks about MTU in the same way
  - Do they mean 1500 the frame payload?
  - Do they mean 1514 the frame + headers?
  - Do they mean 1518 the frame + headers + FCS?
  - Depends on the vendor and where you’re configuring it.
  - Oh and does that include 802.1q overhead or not?
  - Hope you’ve got your calculator!

- What OS’s expect you to include overhead?
  - Juniper JUNOS
  - Cisco IOS-XR
  - Alcatel TimOS
Implementation Caveats (cont’d)

- Not all pieces of equipment can support 9kB:
  - Older Cisco Gear
    - Fast Ethernet Port Adaptors
    - Engine 1 1xGigE, Engine 2 3xGigE Trident
  - Older Juniper PICs
    - M160 4xGigE PICs (4500), 8/12/48xFE PICs (1536)
- Enabling jumbos may be production impacting
  - Re-carving buffers, etc, on some routers, hosts, or NICs.
Implementation Caveats (cont’d)

It’s may be unrealistic to expect jumbo support all the way to the home user for some time

- OS or NIC may not support jumbos
- Home router may not support jumbos
- DSL Modem (ATL-R), Cable modem may not support jumbos
- Broadband aggregation networks may not support large frame sizes
  - Example DSLAM, LAC in L2TP Model
- BRAS (first L3 hop) may not support jumbo frames to the customer or to the backbone
- Providers network may lack jumbo support
- Not all public peers on an Internet Exchange support jumbos
- Transil & Peers may not support jumbos within their own network
Actions Required

• Have the IEEE standardize on a MTU value?
• Need a negotiation method to discover neighbor MTU
• Need a less breakable replacement for PMTUD
Resources

- http://darkwing.uoregon.edu/~joe/jumbo-clean-gear.html
Send questions, comments, complaints to:

Tom Scholl, AT&T Labs
tom.scholl@att.com