

# Increasing MTU in Research Networks

Simon Leinen, SWITCH <[simon@switch.ch](mailto:simon@switch.ch)>

# MTU

---

- Maximum Transmission Unit
- Parameter of a
  - (transmission link)
  - network link (L3 subnet)
  - network path

# History

---

- pre-1990: MTU-576 bytes for off-net hosts
  - Hosts had to assume the Internet minimum MTU for all destinations other than those directly connected.
- post-1990: Path MTU discovery
  - Source sends packets with DF (Don't Fragment) bit set
  - When next-hop MTU too small, routers drop packet and send an ICMP "destination unreachable/fragmentation needed and DF bit set" to source
- By ca. 2000
  - Most traffic on the Internet uses PMTUD
  - Most paths have a PMTU of 1500 bytes, with the bottleneck at the ends.

# MTU and Performance

---

- People have observed that
  - larger MTUs -> better throughput
  - (or lower CPU load at equal rate)
  - in particular with Gigabit speeds on current hosts
- Host issues:
  - interrupt load
    - ▷ can be alleviated with coalescing on interface
  - virtual memory interactions in zero-copy systems
    - ▷ 4096 bytes is the most common VM page size
- Router issues:
  - If everybody uses larger packets, routers have fewer packets to switch -> less work to do.
  - But... most routers aren't pps-limited, except under (D)DoS, and DoS traffic won't use large packets.

# Larger-MTU support

---

- Most backbones today run at 4470 byte MTU (Cisco default for POS and ATM).
- Most Gigabit Ethernet adapters and most non-low-end GigE switches support MTUs larger than 1500 bytes.
- Let's start using  $\geq 4470$  bytes on GigE hosts!

# Oops...

---

- ❑ MTU must be consistent over L3 subnet!
  - Can be worked around with separate (VLAN) subnets.
- ❑ Path MTU discovery isn't all that reliable.
  - But it can be made reliable.
- ❑ Overhead for nodes with many correspondents
  - where most (paths to) correspondents have smaller MTUs
- ❑ Other obstacles
  - Ethernet Exchange Points (separate VLAN solution)
  - Middleboxes
  - Differing hardware limitations for MTUs

# Internet2 Activities

---

- Move towards 9000 byte MTU from Abilene out
  - Most boxes that support >>1500 MTUs can do 9000.
- Document issues
- Help with work to make PMTUD more robust

# Current State in Europe

---

- Most NRENs and GEANT have MTU=4470B.
  - There are a few exceptions (4466 GEANT->DFN).
  - What about new GigE-based backbones?
    - ▷ Ours runs at 4470 bytes, could be increased to 9000.
- About 99.999% of all first-hop networks have MTU=1500B.

# TF-NGN

---

- Should follow these discussions
- Could help coordinate MTUs
  - between NRENs
  - between GEANT and other I2 backbones
- Could run tests between large-MTU servers
  - E.g.: {ez,ce,ls,ix}mp{1,2}-eth{0,1,2}.switch.ch
- Other possible experiments:
  - Different PMTUD methods
  
- But is this really worth the pain?
- Or should we just accept MTU=1500 forever?

# Further Reading

---

- Raising the Internet MTU:
  - <http://www.psc.edu/~mathis/MTU/>
  
- Jumbo Frames on Abilene:
  - <http://www.abilene.iu.edu/JumboMTU.html>