1500 bytes is not a virtue

Denton Gentry, Dominet Systems
denny@dominet systems.com

Ted Seely, Sprint
tseely@sprint.net
Why 1500?

1980
- CPUs moving to 32 bit
- 1 MHz == Really Fast
- Pages 512 bytes

2001
- CPUs moving to 64 bit
- 1 GHz = ho hum
- Pages 8 Kbytes
Benefits of larger frames

• Improves bulk throughput
  – 9000 byte MTU performance improves 400%

• Why?
  – Reduce interrupts 75%
  – Reduce context switches 50%
  – Larger copies to user space
  – MTU > page size == no copy

• Not just benchmarks: backups, database replication, SANs
Other Benefits

• Carriers want to tunnel customer traffic
  – prepend one or more encapsulating IP headers.
  – 20 bytes each, 40 for IPv6
• New markets benefit from larger frames
  – SAN: disk blocks are 4K or larger
• Lots of (non-standard) jumbo frame LANs
  – Explain to your customer why RPR cannot...
Drawbacks of larger frames

• Increases jitter from transit packet
  – At 1 Gbps, increase is 60 usec
  – At 10 Gbps, increase is 6 usec

• Increases buffer size if store and forward
Internetworking
with big packets
MTU mismatch

• Ethernet will remain 1500 bytes
• How to interconnect with RPR?
  – Not a new problem: FDDI, Token Ring
  – Well-defined, proven mechanisms
• Will concentrate on TCP/IP
Mechanism #1: TCP mss

- TCP option sent in SYN & SYN+ACK
- Not a negotiation. Minimum mss wins.
Mechanism #2: Path MTU Discovery

- mss handles leaf networks
- Discover interior networks
  - Set DF (Don’t Fragment) in IP header
  - Listen for ICMP error
Mechanism #3: send small packets

- On subnet, use MTU size
- Off subnet, use 576 bytes
- TCP now uses Path MTU Discovery
Mechanism #4: Fragmentation

- Intermediate router breaks into fragments
- Mainly for UDP
Ethernet -> RPR, routed
Handled: mss
Ethernet-> RPR-> Ethernet, routed
Handled: mss
Ethernet-> RPR-> RPR, routed
Handled: mss
RPR-> Ethernet-> RPR, routed
Handled: path MTU discovery
Ethernet -> RPR, bridged
Handled: mss
Ethernet-> RPR-> Ethernet, bridged
Handled: mss
Ethernet-> RPR-> RPR, bridged
Handled: mss
RPR-> Ethernet-> RPR, bridged
Not handled, unless bridge sends ICMP
Why 1500 bytes?

• 1500 is not a virtue
• 1500 is not a magic number
• 1500 is not what the software wants
• 1500 is not mandated by IEEE 802
• 1500 is neither a dessert topping nor a floor wax