Multi-MTU subnets

draft-van-beijnum-multi-mtu-00

IETF-69

Iljitsch van Beijnum

The Problem

	Mbps	MTU	Pkts/sec
Ethernet	10	1500	813
Fast ethernet	100	1500	8127
Gigabit ethernet	1000	1500	81274
10 gigabit ethernet	10000	1500	812744

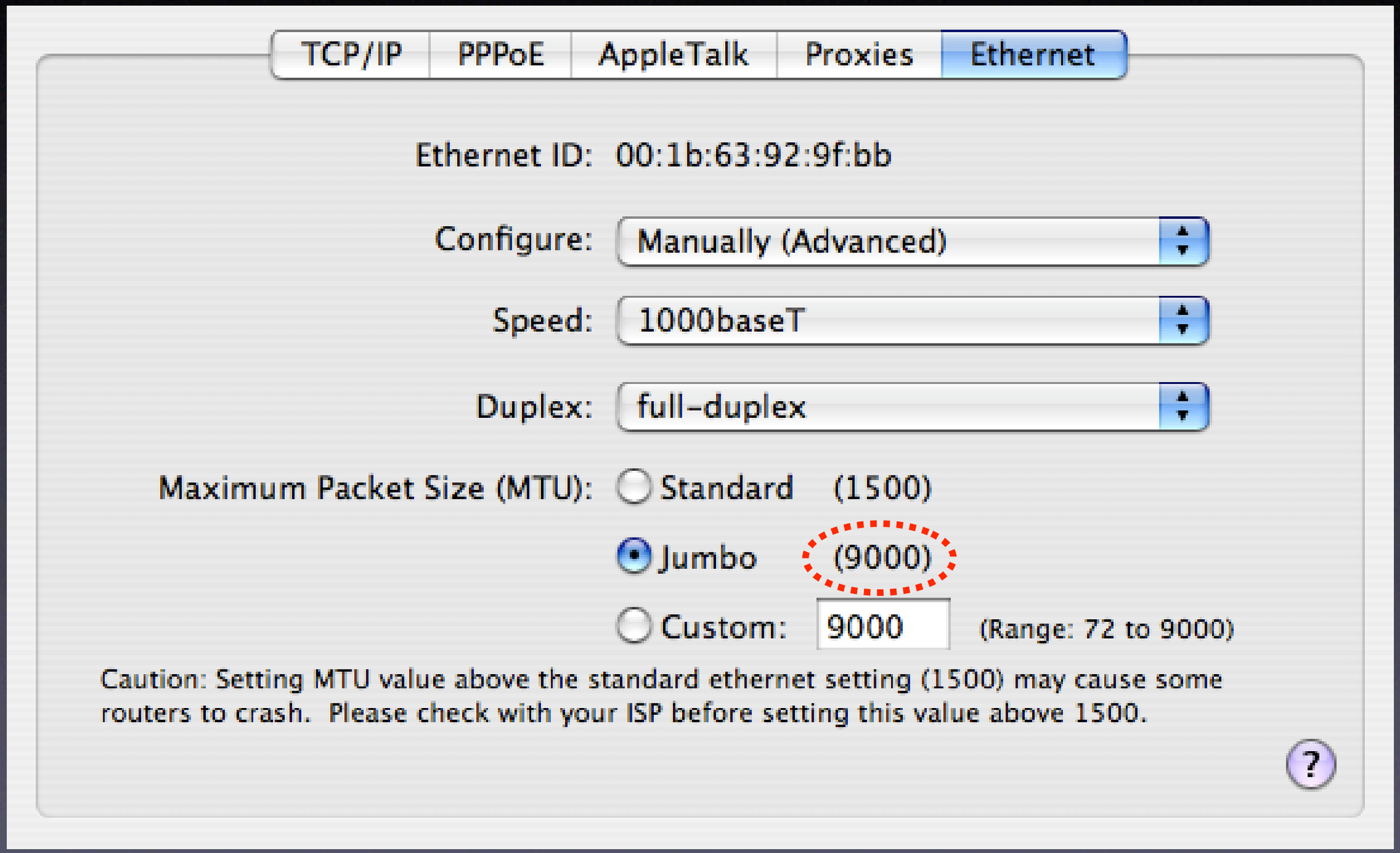
Why still 1500?

- 1500 has been the (IEEE) law for 30 years
- old equipment handles >1500 badly
- Higher speed ethernet segments must interconnect with older ones
- Can't fragment or negotiate neighbor properties at ethernet level

Big Packet Advantages

- More room for additional headers without path MTU discovery breakage
- Lower overhead, especially with large headers
- Less per packet work in hosts = faster
- Less per packet work in routers = possible power/heat savings
- Better TCP performance

But...


TCP/IP
PPPoE
AppleTalk
Proxies
Ethernet
Ethernet ID: 00:1b:63:92:9f:bb
Configure:
Manually (Advanced)
Speed:
1000baseT
Duplex:
full-duplex
Maximum Packet Size (MTU):
Standard (1500)
Jumbo (9000)
Custom:
9000
(Range: 72 to 9000)
Caution: Setting MTU value above the standard ethernet setting (1500) may cause some routers to crash. Please check with your ISP before setting this value above 1500.
?

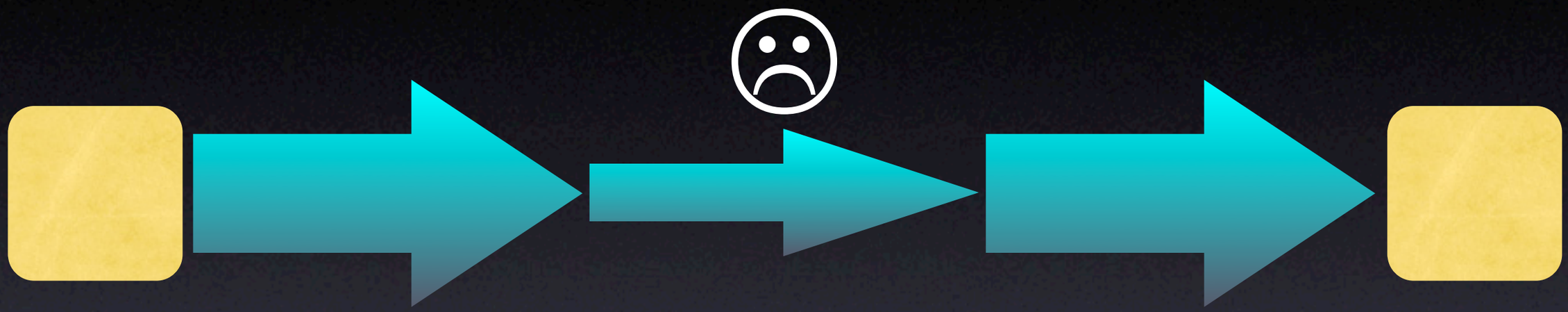
Jumboframes

- Lots of gigabit ethernet equipment supports larger packets: "jumboframes"
- Common value: ±9000 bytes
  - but no standard non-standard size
- "mini jumbos" of ±2000 bytes common in lower-speed switches

Disadvantages (1)

- More delay and jitter
  - so only do 1500+ at 1000 Mbps or faster
- Depend more on path MTU discovery
  - see the problem if ***you*** break PMTUD
  - can always reduce MTU (not increase...)
  - few problems with large MTU in middle

PMTUD



Disadvantages (2)

- More packet loss from bit errors
  - ideal pkt size = √(overhead bytes / BER)
- More undetected bit errors (?)
  - naive: more errors/packet, but fewer packets = no difference
  - complex: hamming distance makes CRC32 much stronger than expected
  - use stronger FCS for jumboframes?

The Solution

- Remove limitation that all nodes on subnet must use the same MTU
  - use standard MTU as default
    - negotiate per-neighbor MTU (and test)
- Hardware vendors must implement reasonable hardware MTUs
- Administrators may override at any point

The Protocol

- Learn IPv6 neighbor MTU from neighbor discovery option
- Send test packet
- Now ignore TCP MSS option and subnet MTU and use neighbor MTU
- IPv4: same thing but slightly different

Be Careful

- Router advertisement option:
  - MTUs for different link speeds
    - off-link MTU (for TCP MSS option)
- New "switch advertisement"
  - let switches advertise supported MTU

Questions

- What do you think?
  - stick to 1500 bytes until the end of time?
  - experimental?
  - standards track?
  - go to IEEE in asbestos suit?
- Feedback: iljitsch@muada.com