Networks are Analog

but MTU is not

Matt Mathis
2024-11-05

Alarming discovery when PLPMTUD was almost done

- We (PSC) encountered a gbic with a "Jumbo Failure"
  - It ran error free at 1500B
  - 1% packet loss at 4kB
  - Device was under beta test and had to be returned
    - So we were not able to diagnosing it
- The words "Jumbo" on the brochure had/have no formal specification
  - Failing to deliver jumbograms is not contractually enforceable
  - Because there is no standard and no conformance tests

A workaround for RFC 4821

- Added an explicit requirement at the top of section 4:

  All links MUST enforce their MTU: links that might non-deterministically deliver packets that are larger than their rated MTU MUST consistently discard such packets.

- But note that implementing this is technically out of scope for the IETF
  - For Ethernet, the IEEE "owns" the error rate specification
- Furthermore the details might be very subtle

Pathological example using historical 10 Mb/s coax

- Coax has no intrinsic MTU limits
  - MTU is determined by the receivers ability to accurately decode the last bit of the packet
  - Which is mostly determined by the accuracy of the clocks at the sender and receiver
    - Traditional Ethernet only did phase acquisition during the packet preamble
    - Clocks free ran during data
    - Clocks had to have significantly less than one bit time relative drift
  - These are are all continuous valued processes
    - Analog and not digital
- Consider 10 Mb/s jumbo
  - Vendor A, increases the **accuracy** of both sender and receiver clocks
  - Vendor C, adds a **continuous** Phase Locked Loop during data reception
    - This requires logic switching time << one bit time
  - Vendor N, does **nothing** to change the design, because jumbo seems to work anyhow

Now, consider formal interoperability testing

- V-N does not reliably interoperate with itself under all test conditions
  - e.g. It is likely to fail if the sender and receiver are at different temperatures
- V-A to V-C works under all conditions, because V-C can track V-A's drift
- V-C to V-A is likely to fail under some conditions
- **V-A and V-C should work with themselves under all test conditions**
- All other combinations are likely to deliver packets under some conditions
  - e.g. if the temperatures are just right the clocks might be accurate enough
  - **Thus all FAIL RFC 4821**
- However, there is no jumbo specification for 10 Mb/s Ethernet
  - and both V-A and V-C implement plausible extensions to the standard

My assertions, without proof

- IEEE specifies waveforms, thresholds, tolerances and testing methodology
  - It does not specify implementations
- Updating Ethernet by replacing 1500B with some other constant fails as a spec
  - Does not yield a self consistent spec (If nothing else, the error rate calculations depend on MTU)
  - Different vendors might make different assumptions about else what needs to be changed
  - No easy way to do conformance testing w/o multi-vendor or interoperability testing
  - Absolutely no guarantee that different devices interoperate under all conditions
- As a community we have mostly been using ad hoc interoperability testing
  - "it seems to work"
  - May be prone to PLPMTUD false pass failures, where PLPMTUD is overly optimistic

Underlying problem: there are two different MTU limits

- Digital L2 Limit, as determined by a register and enforced by a counter
  - Accessed by drivers, software and protocols
  - This is the MTU limit that we, the IETF, understand and use
- Parametric L1 limits as determined by the underlying HW design
  - Analog or statistical properties such as clock stability, error rates, etc
  - Implicit in confirming that the HW can meet a specified error bounds
  - This is the MTU that the IEEE understands
- IEEE 802.3 owns the MTU alignment between the layers
  - But has refused many requests from the IETF and others to adopt a "jumbo" work item
    - e.g. draft-kaplan-isis-ext-eth-02 [1999&2000]

What are some of the IEEE's blockers?

- IEEE has repeatedly rejected requests to standardize jumbo
- Zeroconf plug-n-play is Ethernet's biggest driver
  - (Nearly) zero engineering cost most of the time
  - We can't ask to break this – so deploying jumbo probably requires engineering anyhow
- Need new inband signalling (or other options)
- Due to "false pass" risks heterogeneous networks will be problematic
  - We (IETF) probably need to add MTU signaling to dhcp, etc
  - Will still need a lot of ad-hoc engineering and testing
- Changing the MTU is likely to change error rate specifications
  - They need to be reviewed and reengineered

My wish

- An ID to sketch "Internet optimized Ethernet"
  - Larger MTU (assume multiple values)
  - Relax some of the error rate specifications
    - Internet has multiple layers of error recovery: retransmissions at L4 and L7
    - Multiple layers of integrity protection: checksums, HMAC, app data fingerprints
    - 1E-14 error rate is overkill, probably by several orders of magnitude
  - New signaling (or something) to protect Ethernet plug&play
  - Review impact to other subsystems, such as switch buffer carving
- Use the IETF process to sketch an IEEE work item
  - Don't officially introduce it to the IEEE before we have allies on the inside

My Needs

- ID editor or at least a coauthor
- Ethernet expert(s)
  - Experience with IEEE 802 process and politics
  - Sufficient knowledge of the silicon to anticipate technical issues
    - Need an accurate sense of what might be easy or hard to implement
  - Preferably somebody who has experience on both sides of the issues
    - e.g. has also built data center scale CLOS networks
- A WG venue for the work

Questions?