Packetization Layer path MTU discovery
draft-ietf-pmtud-method-07

Matt Mathis <mathis@psc.edu>

John Heffner <jheffner@psc.edu>

IETF 66

July 2006

Montréal

Outline

- Draft status
  - Document history and status
    - (extended) WGLC ending July 24
  - Review core algorithms
  - Future opportunities (1 slide)
- Background and Vision (2 slides)
  - One step of a longer process

Document evolution

- Was Intended to replace RFC1191 and RFC1981
- Now an extension to RFC1191 and RFC1981
  - Faster convergence when ICMP processing works
  - Guarantees strictly more robust
    - PLPMTUD (“MTU probing”) for ICMP black hole recovery
  - Does not change the algorithm except initial conditions
  - Optimized for easy deployment
- Removed language:
  - Tighten ICMP processing rules (out of scope?)
  - Required complexity to address some obscure failure cases
    - Now mentioned only as a suggestion

The Algorithm

- MTU discovery in TCP or other protocol
  - Responsible for choosing packet boundaries
- Probe with progressively larger packets
    - Choose from range search_low to search_high
  - Successful
    - Raise effective MTU and search_low to probe size
  - Unsuccessful (only the probe was lost)
    - Reduce search_high to probe size (or ICMP PTB size)
    - Suppress Congestion Control (next slide)
  - Inconclusive (lost more than just the probe)
    - Assume congestion or other unrelated loss
    - Pause and probe again

Suppressing Congestion Control

- No window reduction on lost probes if:
  - No other losses
  - Probes are further apart than normal CC loss rate
    - At least as many RTTs as the window size in packets
    - Probing will not cause more loss than the TCP friendly rate
  - This is specified in RFC2119 language
    - **Only** this part that affects interoperation with other standards
- Note that MTU probing parallels CC
  - Use losses to judge the appropriateness of window and/or MTU adjustments

A question of starting values

- Can start with large effective MTU
  - Fully mimic classical ICMP based PMTUD
    - Depend on ICMP PTB to bring the MTU down
    - On full stop timeout, restart w/ small effective MTU
  - No probing unless there are problems
  - Now the default (change from our initial vision)
- Can start with small effective MTU
  - Start out by probing
  - Does not need ICMP at all
- Can use a heuristic to pick initial eff_MTU

Supported protocols

- Full descriptions for TCP and SCTP
- Discussion of IP fragmentation as a PLP
  - Can continue to use RFC1191
  - Requires an adjunct protocol to probe
- Discussion of datagram/application probing
  - E.g. NFS
- Easy to extrapolate to other protocols
  - Support for a “echo” and “pad” makes it easier
- Pick initial eff_MTU per protocol

Open Issues

- Problems with probing
  - Striping (inverse MUX) across dissimilar MTU
  - Devices that ignore DF
  - Raising MTU causing parametric problems
    - Over/underflow rate adaptation FIFO between clock domains
  - It is not known if these are real worries
  - The goal is to be strictly more robust than 1191
    - Preserving ICMP processing does this without extra complexity
- MIB ?
  - Not enough of the algorithm is specified as a standard
  - No suitable table index

Opportunities

- Can make tradeoffs on ICMP paranoia
  - Application/Vendor/Site can discard suspect ICMP
  - Spurious discards may cause full stop timeouts
  - **But will not cause unrecoverable hangs**
- Heuristic choice of initial eff_MTU
  - Tradeoff search time vs risk of persistent timeouts
- Long term - reduce the reliance on ICMP
  - Future make all ICMP advisory(?)
- Unlock 1500 Byte “Internet Cell Size”

Background and Vision

- The 1500 Byte “Internet Cell Size” is too small
  - 1500B@10 Gb/s is 1.2 µS
  - 53B@OC-3 is 2.7µS
  - See http://www.psc.edu/~mathis/MTU
- Jumbograms are really hard to deploy
  - RFC1191 forces “4 party” reconfiguration
    - Both end-systems
    - All switches and routers on the path
    - All other systems sharing broadcast domains
- Nearly impossible in production networks

The Vision

- PLPMTUD will be deployed
  - Solves current problems with tunnels, etc
  - Does not require uniform MTU per broadcast domain
- Then, for selected device types
  - Set default eff_MTU to 1500
  - Unilaterally set interface MTU large
- Then “easy” to create E2E “jumbo” paths
- Will unlock 1500 Bytes local optima