Transport: Advanced

G54ACC – IP and Up

Lecture 4

Recap

- Congestion control cf. flow control
- Adapt transmission window based on timers
- TCP reacts to loss by backing off
  - Loss is assumed to be caused by network congestion
- TCP reacts to successful delivery by probing for more bandwidth
  - In the simple case, generates a “saw tooth” usage

Contents

- Performance and reliability
- TCP congestion control
- ECN marking
- Multimedia

Contents

- Performance and reliability
  - Simple schemes
  - Congestion collapse
- TCP congestion control
- ECN marking
- Multimedia

Achieving Reliability

- Revisit: How do we achieve reliability?
  - Retransmit lost data: how to detect loss?
  - Enable Rx to reconstruct missing data (FEC &c) or re-request
- Detecting loss: Rx explicitly acknowledges
  - Automatic Repeat ReQuest
- Simplest possible: Stop’n’Wait
  - Tx waits for ack(N) before sending data(N+1)
  - Really poor performance in high-latency high-bandwidth networks

Reliability and Performance

- *Rate control*: never send too fast for network
  - How to estimate appropriate rate?
- *Sliding window*: allow unacknowledged data in flight
  - How to determine correct window size?
- *Retransmission TimeOut*:
  - How long to wait to decide a segment is lost?
- Require estimates of dynamic quantities
  - Actually, proxies for the network's current capacity

First Attempt (pre-1988)

- Go-Back-N ARQ
  - Timeout implies loss
  - Retransmit from lost packet onward
- Self-clocking behaviour
  - Only Tx next segment when previous segment acknowledged
- If loss occurs, a lot of throughput is wasted
  - Consider *goodput* vs. *throughput*

Congestion Collapse

- If network load gets too high, experience *congestion collapse*
  - Initially observed on NSFNet Oct/1986
  - LBNL—UCB (400yds/3 hops) dropped 32kb/s to 40b/s
- Why?
  - Router buffers fill, traffic is discarded, hosts retransmit
  - But early TCP (4.2BSD) was naive! Retransmit **more** since data was lost...
  - Boom!
- Solution: Van Jacobson's congestion control algorithm (SIGCOMM'88): Slow-start and Congestion avoidance

Contents

- Performance and reliability
- TCP congestion control
  – Avoidance and recovery
  – Tahoe, Reno, NewReno, SACK, Vegas
  – ECN marking
- Multimedia

Congestion Control

- Need to determine current parameters
  - Window size for most TCPs
  - Either estimate, or *actively probe*
- Additive-Increase, Multiplicative-Decrease
  - Maintain a *congestion window*, `cwnd`
  - `awnd` gives Rx window flow control
  - `wnd = min(awnd, cwnd)`
- First introduced as TCP Tahoe in 4.3BSD (1988)

Congestion Avoidance: TCP Tahoe

- Congestion Avoidance: given `wnd = N` segments
  - Segment acknowledged: `cwnd += 1/cwnd`
  - RTO: `cwnd := 1; ssthresh /= 2`
  - But how to initialise `cwnd`?
- Slow-Start (actually, *exponential growth*!)
  - Original behaviour was constant initial `cwnd`
  - Every segment acknowledged, `cwnd += 1`
  - Stops when loss occurs or `cwnd == ssthresh`
  - Adjust `ssthresh` up once we reach avoidance

Congestion Avoidance: TCP Tahoe

- Fast Retransmit
  - Detecting just a single loss takes a full RTO
    - 500ms or more, at 500ms granularity
  - Rx can signal a "hole" in the stream
  - Send *duplicate ACKs*, typically 3
  - Tx retransmits the relevant segment
  - `ssthresh := max(wnd/2, 2); cwnd := 1`
  - Return to slow-start

Evolution: Reno, NewReno, SACK

- Reno (4.4BSD, 1990) adds *fast recovery*
  - Dup ACKs trigger retransmit
  - But now set `cwnd /= 2`
  - Then inflate `wnd` by number of dup ACKs as self-clocking means segments have left the pipe
- NewReno (4.4BSD, 1993) modifies fast recovery
  - Reno has poor performance for multiple losses
  - ACK for part of window no longer exits recovery
- <u>S</u>elective <u>ACK</u>nowledgement
  - Maintain more state to retransmit *only* lost packets

Revolution: Vegas

- Alternative approach to congestion avoidance
  - Source uses RTT to estimate #packets in pipe
  - End-to-end queuing delay as congestion measure
- for every RTT:

  $RTT_{min} = \min(RTT, RTT_{min})$

  if $wnd/RTT_{min} - wnd/RTT < \alpha$ then wnd++

  if $wnd/RTT_{min} - wnd/RTT > \beta$ then wnd--
- Never widely deployed

Marking

- Many *many* other variants
  - FACK, VENO, BIC, CUBIC, HYBLA, Compound, Westwood, Fusion, ...
  - Cf. doi:10.1109/SURV.2010.042710.00114
- Basically: loss signals congestion, then react
  - But waiting for loss takes a full RTO
- Explicit Congestion Notification, RFC3168
  - Signal congestion *early* via TOS byte and TCP rsvd byte
  - Requires router support: marking, queuing

Contents

- Performance and reliability
- TCP congestion control
- Multimedia
  - RT*P
  - TCP friendliness
  - SST

Multimedia

- TCP is not always useful
  - Messages/frames preferably to bytestream
  - Reliability can cause untimely delivery
  - Canonical example: Real-time media
- Audio/video codecs usually produce frames
  - Not a continuous bytestream
- Losing a frame is better than delaying all subsequent data
  - We can skip or interpolate missing frames

Real Time Protocol

- Encapsulates media over UDP
  - Sequencing, timestamping, delivery monitoring
  - No reliability, no QoS
- Adds a control channel (RTCP)
  - Backchannel to report statistics, participants, &c
- Transport only
  - Leaves encodings, floor control, &c to application
- Extensible

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|V=2|P|X|  CC   |M|     PT      |       sequence number         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           timestamp                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           synchronization source (SSRC) identifier            |
+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
|            contributing source (CSRC) identifiers             |
|                             ....                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      defined by profile       |           length              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        header extension                     ...
```

```
        0                   1                   2                   3
        0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
header |V=2|P|    RC   |   PT=RR=201   |             length            |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                     SSRC of packet sender                     |
       +=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
report |                 SSRC_1 (SSRC of first source)                 |
block  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  1    | fraction lost |       cumulative number of packets lost       |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |           extended highest sequence number received           |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                      interarrival jitter                      |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                         last SR (LSR)                         |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                   delay since last SR (DLSR)                  |
       +=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
report |                 SSRC_2 (SSRC of second source)                |
block  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  2    :                               ...                             :
       +=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
       |                  profile-specific extensions                  |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

TCP Friendliness

- Real-time media being sent at a given rate
  - May cause TCP to experience loss and backoff
  - Media will eventually win
- Need to adapt real-time media
  - Equation based congestion control
  - T = avg(wnd)/RTT
  - p(drop) = $2/(3*w^2)$ => w = sqrt(2/3p)
  - Thus T(p) ~= (1/RTT)*sqrt(3/2p)
- Send drop rate to Tx and it uses T(p) directly

Structured Stream Transport

- Presented in SIGCOMM'07
- Single TCP stream imposes total order
  - Cf. HTTP/1.1 later
- Alternative: enable "substreams" to be created
  - More flexible scheduling by both ends
  - Treat datagram transmission as ephemeral substream
  - Multiplex substreams onto congestion controlled channels

Summary

- Achieving both reliability and performance is *hard*!
- There're many subtle interactions particularly if your scheme is to be general
  - Even trickier if you mix in non-responsive traffic
- But try not to let TCP become a strait-jacket...