TCP

Kai Shen

TCP: Overview

- Connection-oriented:
  - handshaking (exchange of control msgs) to initialize sender, receiver state before data exchange
- Reliable data transfer:
  - guaranteed arrival, no error, in order
- Flow controlled:
  - sender does not overwhelm receiver
- Congestion controlled:
  - sender does not overwhelm the network
- No delay or bandwidth guarantee.

TCP Segment Structure

- source port #
- dest port #
- sequence number
- acknowledgement number
- Urg data pnter
- Options (variable length)
- application data (variable length)
- Receive window
- Urg data pnter
- checksum
- not used
- path MSS
- RST, SYN, FIN: connection estab (setup, teardown commands)
- PSH: push data now (generally not used)
- ACK: ACK # valid
- URG: urgent data (generally not used)

Maximum Segment Size (MSS)

- MSS is the maximum TCP segment that’d fit into the link layer frame
- Local MSS
- Path MSS
  - Try and error probing

- MSS
- header
- not used
- seq, ack, win, checksum
- source port, dest port
- urgent pointer
- options (variable length)
- application data (variable length)
- receive window
- urgent data pointer
- checksum
- header length
- not used
- source port, dest port
- sequence number, acknowledgement number
- options (variable length)
- application data (variable length)
- receive window
- urgent data pointer
- checksum
- header length
- not used
**TCP Reliable Data Transfer**

- TCP provides reliable data transfer service on top of IP’s unreliable service
  - Pipelined transmissions
  - Cumulative ACKs
  - When the receiver receives out-of-order segments, it buffers them and re-ACKs the last in-order data
  - Retransmit a single segment at each timeout
  - The sender retransmits at timeout or when receiving duplicate ACKs
- Somewhere between Go-back-N and Selective Repeat, closer to which?

**TCP Timeout**

Q: principles for setting transmission timeout value?

- too short: premature timeout and unnecessary retransmissions
- too long: slow reaction to segment loss
- longer than normal RTT (round trip time)
  - but RTT varies

![RTT distribution]

**Expected Round Trip Time**

Derive expected RTT from past RTT measurement.

- Basic measurement: measured time from segment transmission until ACK receipt
- Stability: RTT fluctuates, want to avoid instability (pre-mature reaction to short-term spikes)
  - average several recent measurements, not just current RTT
- Agility: in case things do change, we want to adjust accordingly
  - give more recent measurements higher weight

**EWMA – Exponentially Weighted Moving Average**

- influence of past samples decreases exponentially fast

\[
\text{ExpectedRTT} = \frac{\text{SampleRTT}_1 + \alpha \times \text{SampleRTT}_2 + \alpha^2 \times \text{SampleRTT}_3 + \ldots}{1 + \alpha + \alpha^2 + \ldots}
\]

\[
\text{SampleRTT}_1 \text{ is RTT for the most recent data segment, } \text{SampleRTT}_2 \text{ is RTT for the next recent data segment, etc.}
\]

\[
\Rightarrow \text{ExpectedRTT} = \alpha \times \text{ExpectedRTT}_{\text{last}} + (1-\alpha) \times \text{SampleRTT}_1
\]

- typical value in TCP: \( \alpha = 0.875 \)
Example RTT Estimation

TCP Timeout

Setting the timeout:
- **ExpectedRTT** plus “safety margin”
- larger variation in RTT → larger safety margin
- **TimeoutInterval** = **ExpectedRTT** + 4***DevRTT**
- **DevRTT** also estimated using EWMA of past measurements

TCP Sender Events and Processing

Data ready to send:
- create segment with seq #
- seq # is byte-stream number of first data byte in segment
- start timer
- timeout value: we just decided it!!

Timeout:
- retransmit segment that caused timeout
- restart timer

ACK rcvd:
- slide sender window if acknowledges previously unacked segments
- retransmit if 3 duplicate ACKs

TCP byte-oriented seq. #’s and ACKs

Seq. #’s:
- byte stream “number” of first byte in segment’s data

ACKs:
- seq # of next byte expected from other side
- cumulative ACK
- piggybacked in data segments in the reverse direction

Host A

User types 'ls
'

Seq=42, ACK=79, data = 'ls
'

host ACKs receipt of 'ls
', echoes back 'a.c
'

Host B

Seq=79, ACK=45, data = 'a.c
'

host ACKs receipt of echoed 'a.c
'
TCP in Action: Cumulative ACK

- Host A
- Host B
- Time
- Cumulative ACK scenario

- Slide send window base to 120
- Cumulative ACK scenario

Fast Retransmission

- Time-out period often relatively long:
  - long delay before resending lost packet
- When receiver receives out-of-order segments, it re-ACKs the last in-order byte
- If sender receives 3 ACKs for the same data, it supposes that segment after ACKed data was lost:
  - fast retransmission: resend segment before timer expires, restart timer

TCP in Action: Duplicate ACKs and Fast Retransmission

- Host A
- Host B
- Time
- Cumulative ACK scenario

- 3 duplicate ACKs resend 92

Outline

- Segment structure
- Reliable data transfer
- Flow control
- Connection management
**TCP Flow Control**

- Receive side of TCP connection has a receive buffer:
- App process may be slow at reading from buffer.

**Flow Control:**
- Receiver advertises spare room by including value of RcvWindow in segments
- Sender limits unACKed data to RcvWindow, therefore guarantees receive buffer doesn’t overflow

---

**TCP Connection Management**

**Establishment:**
- TCP sender, receiver establish “connection” before exchanging data segments
- Initialize TCP variables: starting seq. #s, MSS, buffers, flow control info (e.g. RcvWindow)

**Teardown:**
- Free up resources after mutually close

---

**TCP Connection Establishment**

**Three-way handshake:**

**Step 1:** client (active open) sends TCP SYN segment to server
- specifies initial seq #
- no data

**Step 2:** server (passive open) host receives SYN, replies with SYNACK segment
- server allocates buffers
- specifies server initial seq. #

**Step 3:** client receives SYNACK, replies with ACK segment, which may contain data

---

**TCP Connection Teardown**

**Closing a connection:**
- close socket: close(sockfd);

**Step 1:** A (active closing host) sends TCP FIN control segment.

**Step 2:** B (passive closing host) receives FIN, replies with ACK. Closes connection, sends FIN.

**Step 3:** A receives FIN, replies with ACK. Enters “timed wait” – resend ACK in case it is lost.

**Step 4:** B receives ACK. Connection closed.
TCP State Transition Diagram

1. CLOSED
2. LISTEN
3. SYN_RCVD
4. SYN_SENT
5. ESTABLISHED
6. CLOSE_WAIT
7. LAST_ACK
8. FIN_WAIT_2
9. FIN_WAIT_1
10. CLOSE
11. CLOSE_WAIT
12. FIN_WAIT_2
13. FIN_WAIT_1
14. ACTIVE
15. PASSIVE

Disclaimers:
- Parts of the lecture slides contain original work of James Kurose, Larry Peterson, and Keith Ross. The slides are intended for the sole purpose of instruction of computer networks at the University of Rochester. All copyrighted materials belong to their original owner(s).