HIGH-SPEED VARIANTS

Scalable TCP

The Sluggishness of TCP

- 10 Gbps network with 100 ms round-trip time
  - Desired $c_{win} \approx 83,000$ packets

- Initial bandwidth discovery:
  - $SSThresh$ usually set to no more than 32-64 segments
  - Would take hours to achieve a sending rate of 10 Gbps

- Bandwidth rediscovery after timeout:
  - $C_{win}$ reset to 1
  - Additive increase would still take hours to recover 10 Gbps throughput
The Sluggishness of TCP

- Steady-state Congestion Avoidance behavior:
  - If congestion events occur frequently, average throughput will be less than $C$
  - To achieve 10 Gbps with TCP, only 1 in $(2 \times 10^{10})$ packets should be dropped
    - This is past the limits of achievable fiber error rates
    - Packet loss rate of $10^{-7}$ is reasonable to expect

Considerations for Scalable TCP Design

- Scalability for utilization of high available bandwidth (AB)
- TCP-friendliness:
  - Impact on (coexistence with) legacy traffic
    - Where legacy traffic occupies different fractions of capacity
- Flow rate variance
  - If AB is not changing, neither should the data sending rate
- Convergence
  - If AB changes, how quickly does the protocol converge to the new network state
- Stability:
  - Is the protocol operation stable around an equilibrium?
  - Or does it oscillate or diverge?
Scalable TCP: Basic Idea

- Multiplicative increase:
  - Increase window more aggressively
  - Standard TCP: \( c_{\text{win}} = c_{\text{win}} + 1 \)
  - Scalable TCP: \( c_{\text{win}} = (1 + a)c_{\text{win}} \)

- Multiplicative decrease:
  - Decrease window less aggressively
  - \( c_{\text{win}} = b.c_{\text{win}} \), where \( b > 0.5 \)

Time to recover previous sending rate after loss is independent of \( C \)
(depends on RTT, \( a \), \( b \))

If capacity is stable, average link utilization achieved is independent of \( C \)

Scalable TCP: Response Curve

- Response curve:
  - Relation between \( c_{\text{win}} \) and end-to-end signaling rate
  - Only signal that TCP or Scalable respond to: packet drops

- For small end-to-end drop rates,
  - \( \text{avg Scalable cwin} \approx \frac{d1}{PP} \)
  - \( \text{avg TCP cwin} \approx \sqrt{\frac{1.5}{PP}} \)

Different values of avg cwin for different loss rates
Scalable cwin could be lower or higher!
Achieving TCP Friendliness

- How to ensure coexistence with conventional, low-speed TCP traffic?
  - Define LowThresh
  - When \( cwin < \text{LowThresh} \)
    - Adopt TCP \( cwin \) behavior
  - When \( cwin > \text{LowThresh} \)
    - Adopt high-speed growth behavior

Selecting \( a, b \)

- For a \( \text{LowThresh} \) of 16 MSS,
  - From the response curve, when \( cwin \approx \text{LowThresh} \)
    - \( \frac{a}{b} = \sqrt{1.5}\, P \)
  - Given \( b \), this determines value of \( a \)

- Setting \( b \):
  - [13] studies flow-rate variation for scalable TCP
    - Covariance in sending rate is proportional to \( \sqrt{b} \)
    - Smaller the value of \( b \), smaller is the flow-rate variance
  - Scalability to high speeds requires a large value of \( a \)
    - \( \frac{\log(1-b)}{\log(1+a)} \)
    - Also requires a large value of \( b \)
  - In above tradeoff, author found \( b = 1/8 \) to be reasonable
    - \( b = 1/8; \ a = 0.01 \)
Other protocols

- Protocols where window-growth is a function of $c_{\text{win}}$:
  - High-speed (motivation for Scalable TCP)
    - $a$ and $b$ depend on value of $c_{\text{win}}$ (and expected loss rate)
  - BIC/CUBIC (cubic growth function, centered around expected bandwidth)

- Delay-based protocols (do not create losses to detect incipient congestion)
  - FAST (touted as high-speed version of Vegas)
  - Compound
    - How to make a delay-based protocol co-exists competitively with loss-based protocols?

- Protocols that depend on explicit router support:
  - DCTCP
    - Relies on ECN to reduce queue sizes/losses in data centers
  - XCP
    - Routers compute and convey new value of $c_{\text{win}}$
  - RCP
    - Routers explicitly guide flow-rate (to minimize flow-completion time)