Congestion control: stochastic models for the Internet protocol TCP

References:

- A. Bose, I. Kaj, *Meanfield approximation for multiple TCP sources interacting in a RED packet loss environment,* In preparation (C.f. Chapter 6.4.5)
- I. Kaj, J. Olsén, *Throughput modeling and simulation for single connection TCP-Tahoe, Proceedings of ITC'01, Dec 2001*
- I. Kaj, J. Olsén, *Stochastic equilibrium modeling of the TCP dynamics in various AQM environments,* Preprint Uppsala University, Sept 2001, Submitted
- A. Kumar, *Comparative performance analysis of versions of TCP in a local network with a lossy link,* IEEE/ACM Trans. on Networking August 1998
- J. Padhye, V. Firoiu, D. Towsley, J. Kurose, *Modeling TCP Throughput: a simple model and its empirical validation,* Computer Communication Review 28 (1998) 303-314

Outline of talk

- TCP dynamics; ACK-based adaptive rate control
- Stochastic models for single connection TCP
- Renewal-reward analysis of throughput
- Towards interaction models, the RED environment
- Mean-field approximation

Brief introduction to TCP

- Users are allowed to increase sending rate (window size W packets per round-trip-time) during loss-free transmission ($b = 1, 2, 4$ packets cleared per ACK packet)
- Users consent to reduce rate upon packet loss indication from congested network.

Slow-start mode:

$$W' = (1 + 1/b)W \qquad \text{exponential growth}$$

Congestion avoidance mode:

$$W' = W + \underbrace{\frac{1}{W} + \frac{1}{W + \frac{1}{W}} + \dots}_{[W/b] \text{ terms}} \qquad \text{linear growth}$$

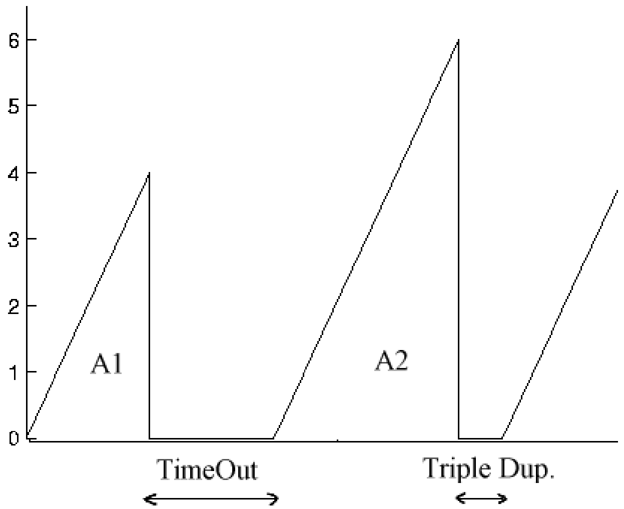
Loss indications: Triple-duplicate acknowledgement loss (TD) or time-out loss (TO)

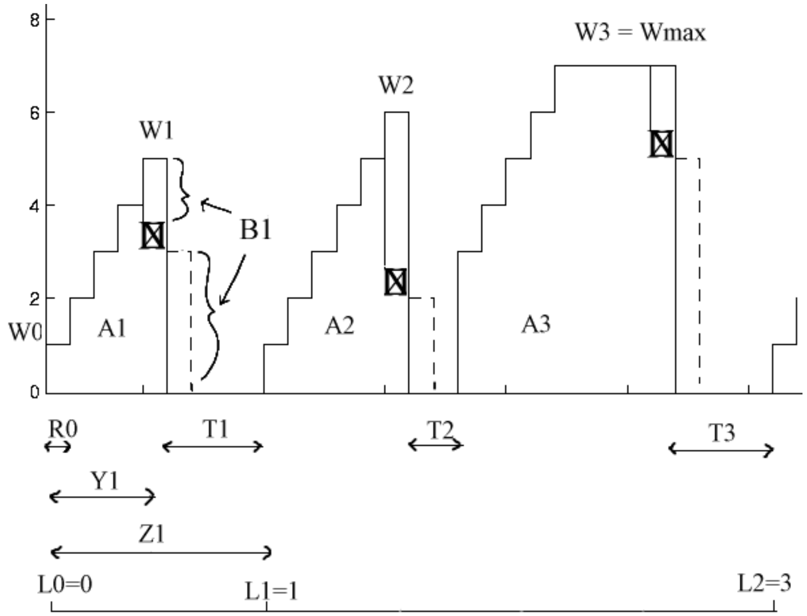
TCP-Tahoe: Responds to loss indication by resetting window to minimal value, $W \to 1$

TCP-Reno (Fast retransmit): Upon TD loss reset window to half of previous value, $W \to W/2$

Stochastic model for simple special case

Find throughput (packets per round-trip-time) for single connection persistent web-transfer under TCP-Tahoe without slow-start as a function of packet-loss probability p.





$A_n = \#$ of packets transmitted in cycle n before first loss, $\{A_n\}_{n>0}$ independent
$B_n = \#$ of packets transmitted in cycle n after first loss
$W_n =$ window size at end of cycle n
$Y_n = \#$ of effective transmission rounds in cycle n including *loss round*

Renewal reward analysis

Asymptotic throughput obtained as renewal cycle average

$$\text{Throughput}_{\text{Tahoe}} = \frac{E(A) + E(B)}{E(Y) + 1 \cdot (1 - Q) + T_0 \cdot Q}$$

where

$$Q = P(\text{loss indication causes time-out}).$$

In RED (Random Early Discard) environment it is reasonable to assume independent packet loss,

$$A_n \in \text{Ge}(p) \qquad B_n \in \text{Bin}(W_n - 1, 1 - p) \qquad p = \lim_{n \to \infty} \frac{n}{\sum_{k=1}^{n} (A_k + 1)} = \frac{1}{E(A) + 1}$$

and obtain

$$\text{Throughput}_{\text{Tahoe}} = \frac{(1 - p)(1 - p + pE(W_\infty))}{pE(Y_\infty) + p(1 + (T_0 - 1)Q)}$$

Continuous approximation

Let $(V_n)_{n\geq 1}$ denote sequence of i.i.d. exponentially distributed random variables with mean one. Using $A_n \approx V_n/p$, based on the convergence in distribution

$$p\,A_n \Rightarrow V_n \quad p \to 0$$

we may apply the continuous approximation

$$W_n = \sqrt{2A_n/b} \approx \sqrt{2V_n/bp} \qquad Y_n = \sqrt{2bA_n} \approx \sqrt{2bV_n/p}.$$

As we have $E(\sqrt{V_1}) = \Gamma(3/2) = \sqrt{\pi/2}$ this yields

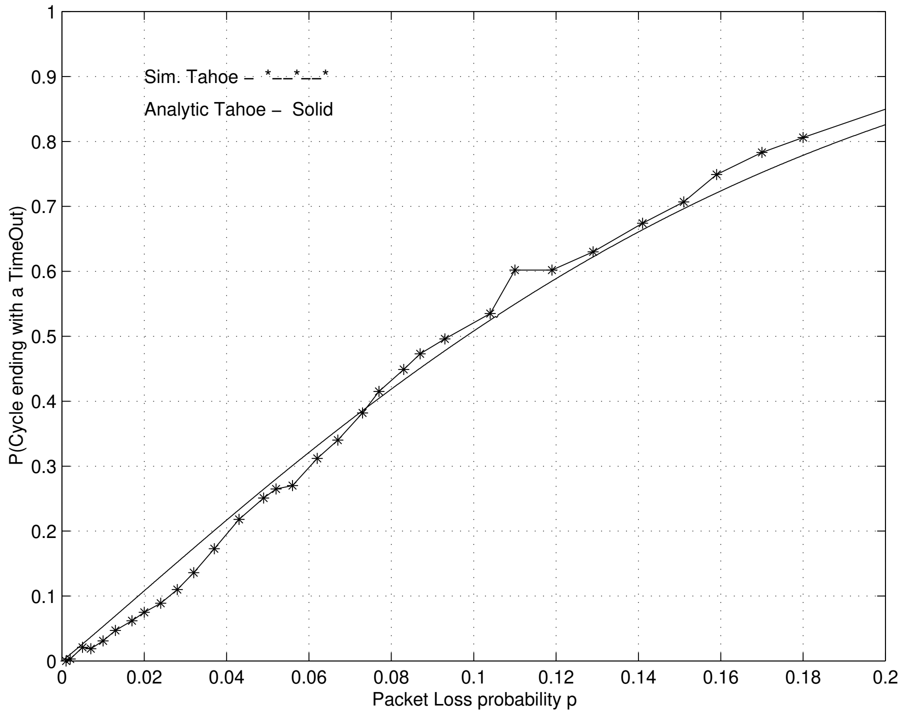
$$\text{Throughput}_{\text{Tahoe}} = \frac{(1-p)(1-p+pE(W_\infty))}{pE(Y_\infty)+p(1+(T_0-1)Q)} = \frac{(1-p)(1-p+\sqrt{p\pi/2b})}{\sqrt{b\pi p/2}+p(1+(T_0-1)Q)}$$

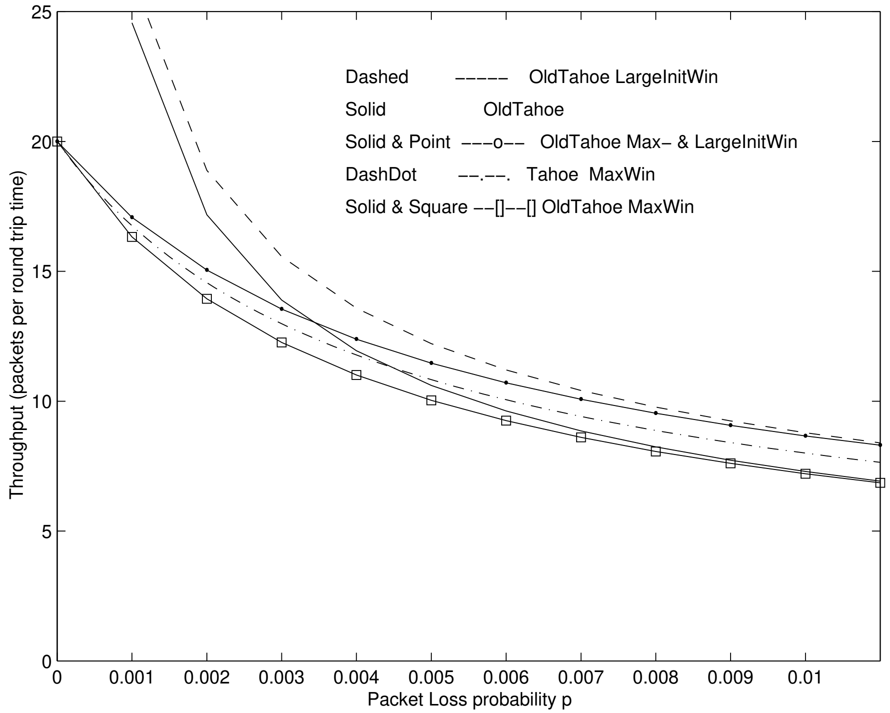
Throughput TCP-Tahoe, continued

Remains to find time-out probability

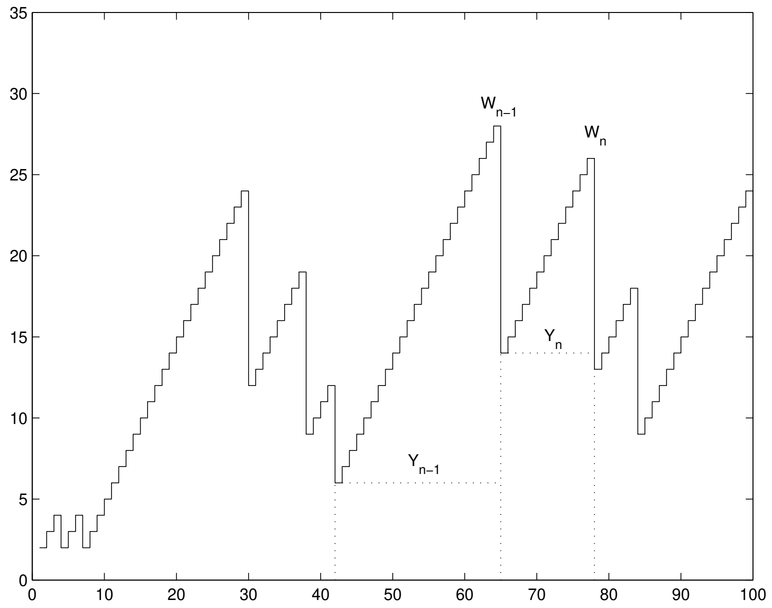
$$Q(p) = P(B_\infty < 3) \approx \int_0^3 f_W(v)\,dv + \frac{1}{2}\int_3^\infty (v-1)(v-2)p^{v-3} f_W(v)\,dv$$

where $f_W(v) = bpv\, e^{-bpv^2/2}$, $v \geq 0$.





TCP-Reno (No time-outs)



$$A_n = \frac{1}{2} W_{n-1} Y_n + Y_n^2/2b \qquad W_n = \frac{1}{2} W_{n-1} + Y_n/b.$$

Solve using $A_n \approx V_n/p$ and iterate to obtain

$$W_\infty \stackrel{d}{=} \sqrt{\frac{2}{bp} \sum_{k=0}^{\infty} (1/2)^{2k} V_k} \qquad \text{with} \quad E(W_\infty) = \sqrt{\frac{2}{bp}} \prod_{k=1}^{\infty} \frac{1-(1/2)^{2k+1}}{1-(1/2)^{2k}} \frac{\sqrt{\pi}}{2} = 1.0797 \cdot \sqrt{\pi/2bp}$$

Network models

- Transmit m TCP connections; Congestion in bottleneck router
- DropTail (buffer overflow loss) or RED (preventive discard)
- Problem: Dimension DropTail-buffers or tune RED parameters to optimize throughput, minimize download times, etc
- Details of algorithm include: moving average estimate of queue size

$$\hat{q}_n = (1 - \beta)\hat{q}_{n-1} + \beta q_n, \quad \text{if } q_n > 0$$

and similar dynamic estimates of round-trip-times.

Modeling approaches

- (Sharma, Virtamo, Lassila) Put $z^{\beta}(t) = \hat{q}_{[t/\beta]}$, let $\beta \to 0$:

$$\frac{dz(t)}{dt} = E_{\pi}[q(z(t)] - z(t)$$

  where $E_{\pi}[q(z)]$ is mean stationary queue length conditional on $\hat{q}_n = z$.

- (Towsley et al) Numerical schemes using throughput formulas for single connection TCP. Control theoretic tools.

- Diffusion approximation for mean field interaction model

Meanfield approximation

- m separate TCP connections, bulk transfer over common node;
- Constant and equal round-trip-times;
- Node (infinite) buffer actively managed by random early drop mechanism;
- Capacity Cm packets per RTT;
- Sending rate of each connection varies according to TCP-OldTahoe, which updates window sizes $(W_k^i)_{1\le i\le m}$ once every round;
- Average congestion window during round k is $\bar{W}_k = \frac{1}{m}\sum_{i=1}^m W_k^i$;

Approximate the buffer queue length (B_k) by

$$B_k = \left(B_{k-1} + \sum_{i=1}^m W_k^i - Cm\right)^+,$$

hence the normalized queue $Q_k = B_k/m$ (packets per user) satisfies

$$Q_k = \left(Q_{k-1} + \bar{W}_k - C\right)^+.$$

The RED mechanism exercises congestion control based on the current time-averaged buffer size defined by the moving average

$$\hat{Q}_k = (1-\beta)\hat{Q}_{k-1} + \beta Q_k, \quad \hat{Q}_0 = 0$$

where $0 \leq \beta \leq 1$ is a parameter. (Floyd et al 2001: $\beta = 1 - e^{-1/C}$.) On arrival at the link buffer node, packets are subject to the RED mechanism which discards independently each packet conditionally on $\hat{Q}_k = \hat{q}$ with probability

$$p_{\text{loss}}(\hat{q}) = \begin{cases} 0 & q \leq T_{\min} \\ \dfrac{p_{max}(\hat{q} - T_{min})}{T_{max} - T_{min}} & T_{\min} < q \leq T_{\max} \\ 1 & \hat{q} > T_{\max} \end{cases}$$

where T_{min}, T_{max} and $0 \leq p_{max} \leq 1$ are parameters set by the RED protocol.

If all W_k^i packets from user i in round k pass without losses increase the sending rate to $W_k^i + 1$. If losses occur reset the window either to 1 (TCP-Tahoe). Hence

$$E[W_{k+1}^i - W_k^i | W_k = w, \hat{Q}_k = \hat{q}] = 1 - \left(1 - (1 - p_{\text{loss}}(\hat{q}))^{w^i}\right) w^i.$$

As a crude approximation

$$E[W_{k+1}^i - W_k^i | W_k = w, \hat{Q}_k = \hat{q}] = 1 - \left(1 - (1 - p_{\text{loss}}(\hat{q}))^w\right) w^i, \quad w = \frac{1}{m} \sum_{i=1}^{m} w^i.$$

This gives for the average congestion window

$$E[\bar{W}_{k+1} - \bar{W}_k | \bar{W}_k = w, \hat{Q}_k = \hat{q}] = 1 - w + (1 - p_{\text{loss}}(\hat{q}))^w w \approx 1 - p_{\text{loss}}(\hat{q}))w^2.$$

The buffer recursion is

$$Q_k - Q_{k-1} = \big(Q_{k-1} + \bar{W}_k - C\big)^+ - Q_{k-1} = \bar{W}_k - (Q_{k-1} + \bar{W}_k) \wedge C$$

so that

$$E(Q_k - Q_{k-1} | \bar{W}_k = w, Q_{k-1} = q) = w - (q + w) \wedge C.$$

The time-averaged buffer satisfies

$$\hat{Q}_k - \hat{Q}_{k-1} = -\beta \hat{Q}_{k-1} + \beta Q_k$$

hence

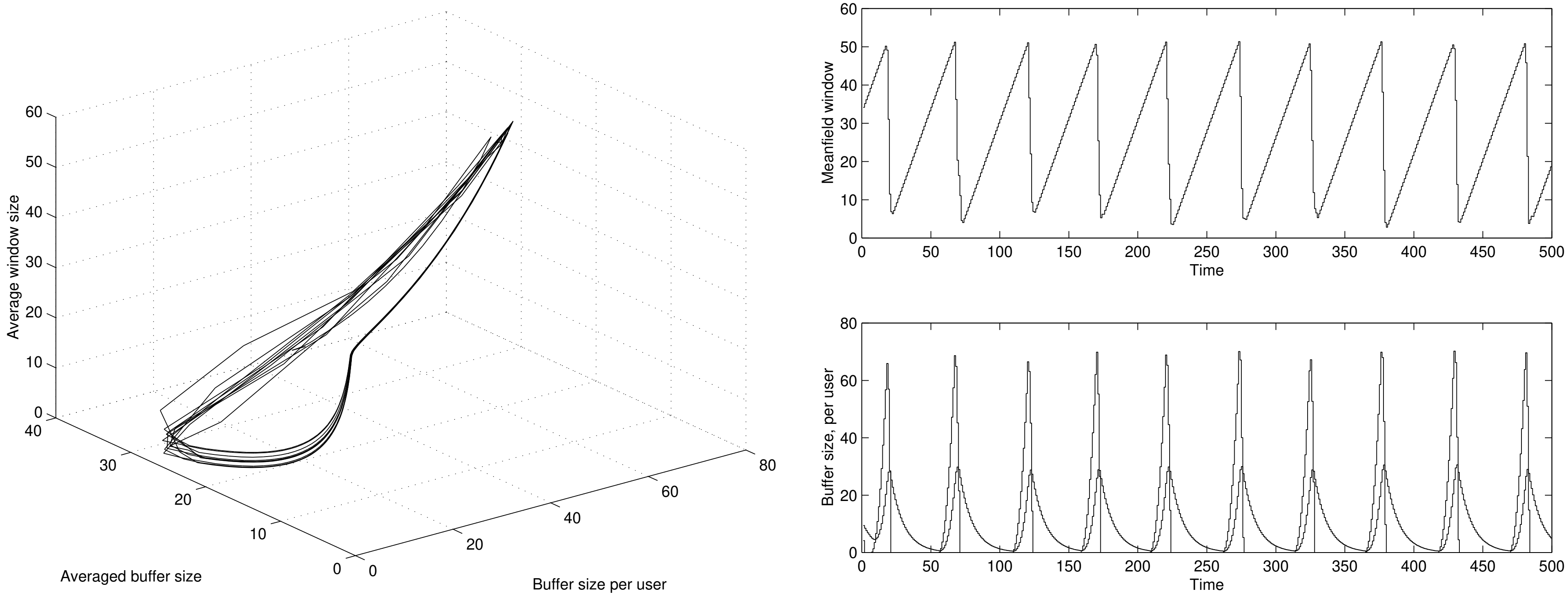
$$E(\hat{Q}_k - \hat{Q}_{k-1} | \bar{W}_k = w, Q_{k-1} = q, \hat{Q}_{k-1} = \hat{q}) = \beta(q + w - \hat{q} - (q + w) \wedge C).$$

We are led to study the recursive system

$$\begin{cases} \bar{w}_{k+1} = \bar{w}_k + 1 - p_{\text{loss}}(\hat{q}_k)\, \bar{w}_k^2 \\ q_{k+1} = q_k + \bar{w}_k - (q_k + \bar{w}_k) \wedge C \\ \hat{q}_{k+1} = \hat{q}_k + \beta(q_k + w_k - \hat{q}_k - (q_k + w_k) \wedge C). \end{cases}$$

The corresponding continuous time version is the system of coupled non-linear ordinary differential equations

$$\begin{cases} \dot{w}_t = 1 - p_{\text{loss}}(\hat{q}_t) w_t^2 \\ \dot{q}_t = w_t - ((q_t + w_t) \wedge C)) \\ \dot{q}_t = \beta(q_t + w_t - \hat{q}_t - (q_t + w_t) \wedge C). \end{cases}$$


Average window size
Averaged buffer size
Buffer size per user
Meanfield window
Time
Buffer size, per user
Time