SWINBURNE UNIVERSITY OF TECHNOLOGY

Literature Review Series: Delay/Rate based Congestion Avoidance in TCP

David Hayes

dahayes@swin.edu.au

Centre for Advanced Internet Architectures (CAIA)
Swinburne University of Technology

Outline

- Introduction
- Background
  - Current TCP congestion avoidance
  - Base measurements
- Quick early work overview
- Algorithm outlines
  - CARD
  - Packet pair flow control
  - TCP-LP
  - Vegas
  - FAST
  - Compound TCP
  - DUAL
  - Hamilton
  - Other
- Conclusions
- Bibliography

Introduction

- Promise low latency zero loss[1]
- Delay based intuition:
  - delay$\uparrow \equiv$ queue$\uparrow \implies$ indicates congestion
- Rate based intuition:
  - Send rate $>$ receive rate $\implies$ indicates congestion
- Basic questions:
  - How is congestion determined?
  - and if congested, how should cwnd be adjusted
- Issues:
  - Noise of measurements
  - Correlation of measurements with congestion
  - Compatibility with existing TCP algorithms

[1] congestion related

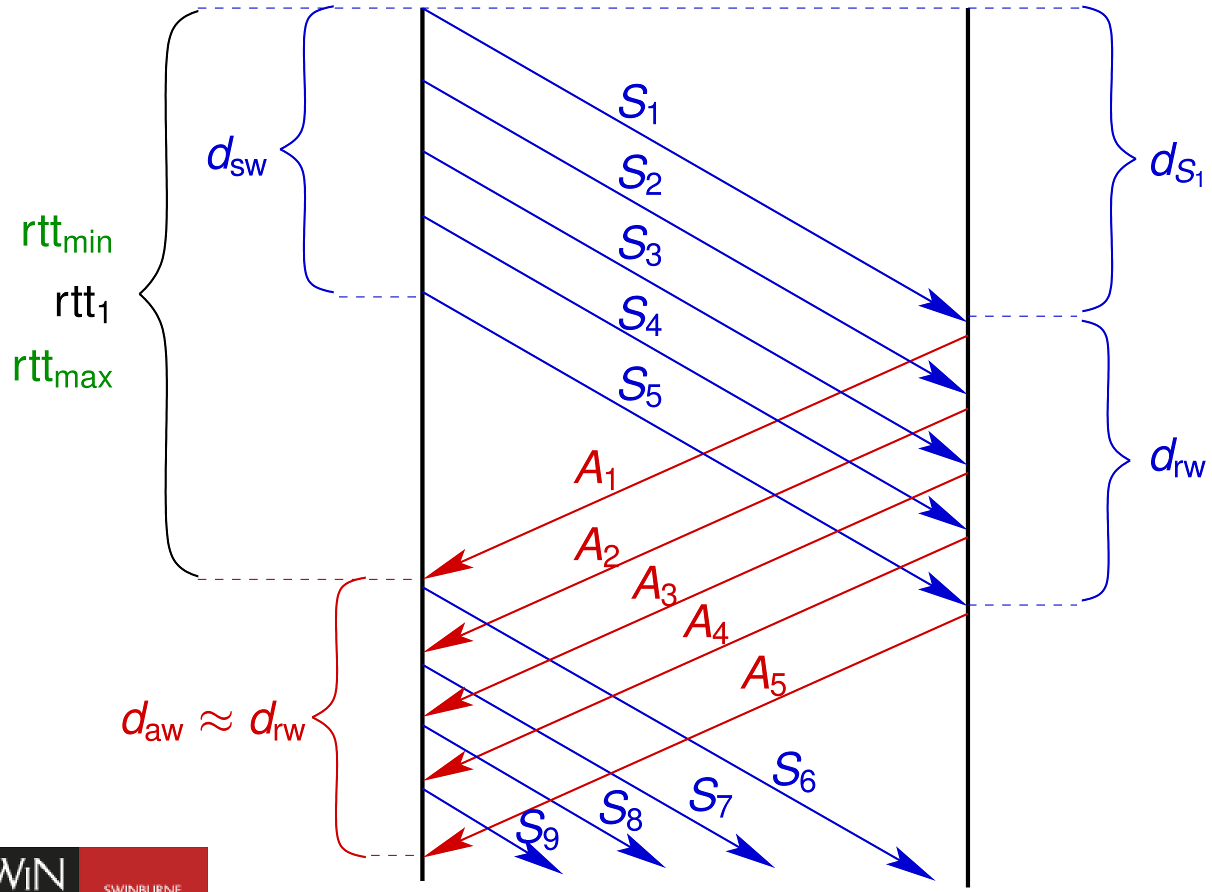
Background: TCP NewReno congestion avoidance

- Congestion is indicated by packet loss
- The congestion window, cwnd, is adjusted with every ack as follows:

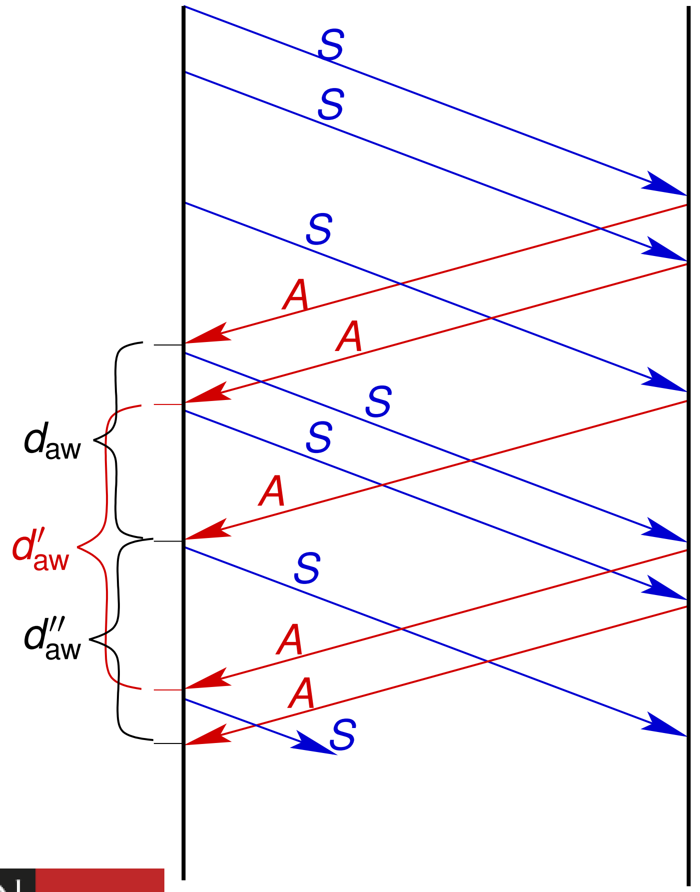
$$w_{j+1} = \begin{cases} \beta w_j & \text{packet loss} \\ w_j + 1/w_j & \text{otherwise} \end{cases}$$

  where in this case w is in packets.
- Multiplicative decrease
- Additive increase

Background: Base timing measurements



Background: Base timing measurements



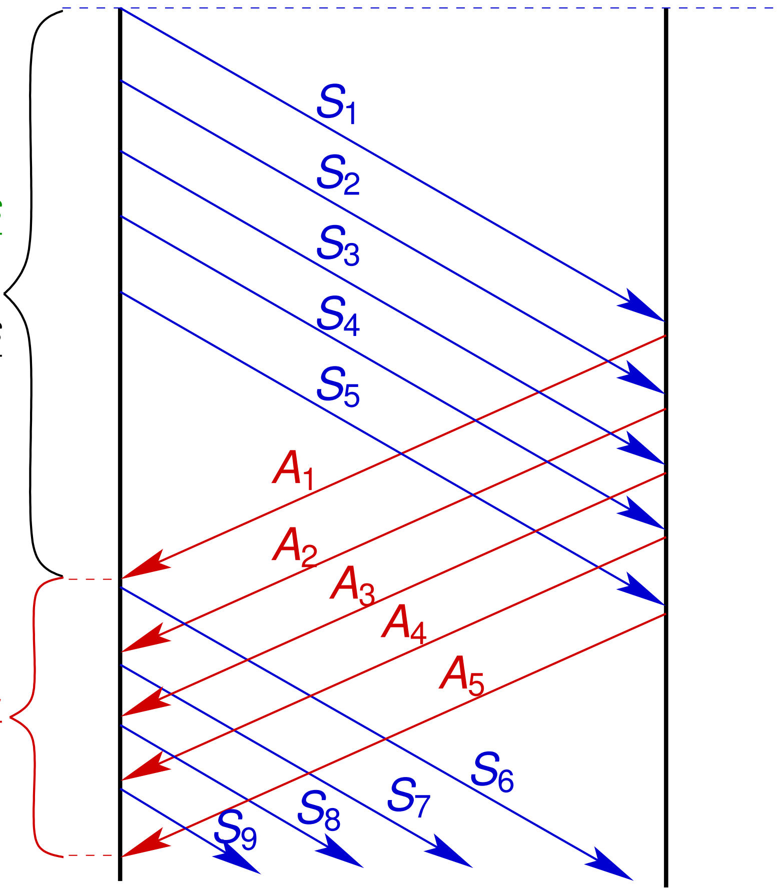
- Note: Queueing at FIFO network nodes can increase **or** decrease the interpacket times

Background: Base rate measurements

$$T_{max} = \frac{\sum^{w} s}{\text{rtt}_{min}}$$

$$T_1 = \frac{\sum^{w} s}{\text{rtt}_1}$$

$$R_a = \frac{\sum^{w-a_1} A_i}{d_{aw}}$$



Quick early work overview

- [Clark et al., 1985]&[Clark et al., 1987] NETBLT RFCs 996&998
- [Jacobson, 1988][a] – footnote on connectionless rate based AIMD.
- [Jain, 1989][b] normalised delay gradient.
- [Wang and Crowcroft, 1992][c] DUAL algorithm.
- [Brakmo and Peterson, 1995][d] TCP Vegas.

[a] V. Jacobson, “Congestion avoidance and control,” in *SIGCOMM ’88: Symposium proceedings on Communications architectures and protocols*. New York, NY, USA: ACM, 1988, pp. 314–329

[b] R. Jain, “A delay-based approach for congestion avoidance in interconnected heterogeneous computer networks,” *SIGCOMM Comput. Commun. Rev.*, vol. 19, no. 5, pp. 56–71, 1989

[c] Z. Wang and J. Crowcroft, “Eliminating periodic packet losses in the 4.3-Tahoe BSD TCP congestion control algorithm,” *SIGCOMM Comput. Commun. Rev.*, vol. 22, no. 2, pp. 9–16, Apr. 1992

[d] L. S. Brakmo and L. L. Peterson, “TCP Vegas: end to end congestion avoidance on a global internet,” *IEEE J. Sel. Areas Commun.*, vol. 13, no. 8, pp. 1465–1480, Oct. 1995

Algorithms: CARD [Jain, 1989]

- CARD - Congestion Avoidance using RTT Delay
- Uses queueing theory to determine knee of throughput graph
- Delay gradient, $\frac{drtt}{dw}$
- Conditional increase/decrease of window based on Normalised Delay Gradient:

$$\text{NDG} = \left(\frac{\text{rtt}_j - \text{rtt}_{j-1}}{\text{rtt}_j + \text{rtt}_{j-1}}\right)\left(\frac{w_j + w_{j-1}}{w_j - w_{j-1}}\right)$$

and

$$w_{j+1} = \begin{cases} \beta_j w_j & \text{NDG} > 0 \\ w_j + \alpha & \text{otherwise} \end{cases}$$

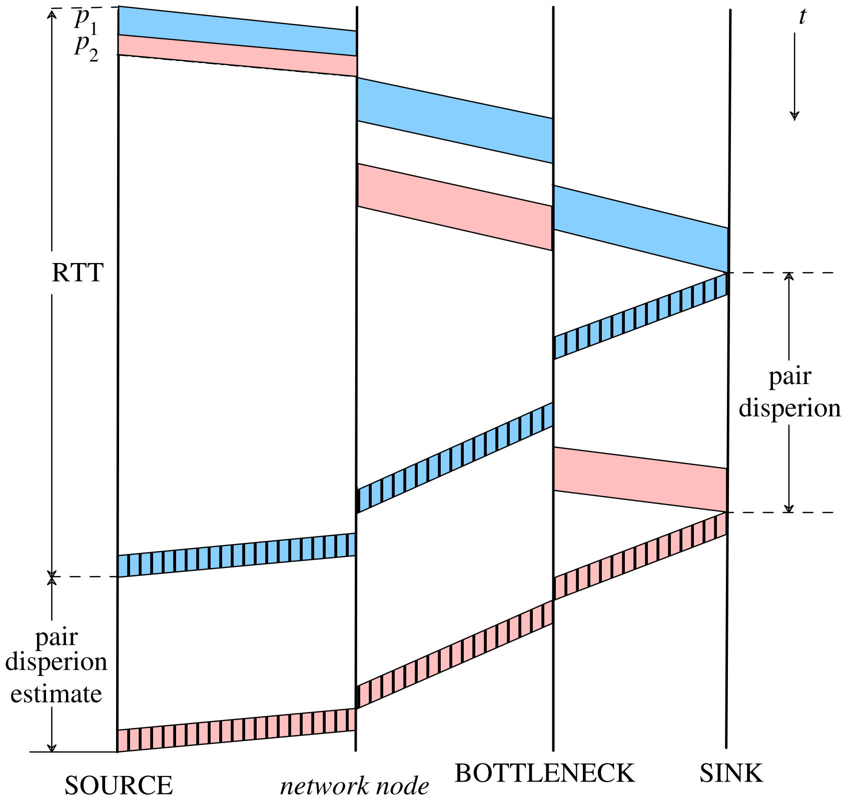
- Algorithm derived using D/D/1 queues
- Use in stochastic networks require enhancements

Algorithms: Packet pair flow control [Keshav, 1994]

- Full transport protocol proposal and analysis
- All data is sent as back-to-back pairs
- Available send rate is:

$$T = \frac{\text{size}(p_2)}{\text{pair dispersion}}$$

- Presumes routers use round robin scheduling



Algorithms: TCP-LP [Kuzmanovic and Knightly, 2006]

- Low Priority TCP
- Based on *relative* one way delay: $d_i = \text{ts}_{\text{rx},i} - \text{ts}_{\text{tx},i}$
  - Send and receive clocks do not need to be synchronised.
  - They do need to be the same frequency.
- Congestion: $c_i = \begin{cases} 1 & \overline{d}_i > d_{\text{min}} + \delta(d_{\text{max}} - d_{\text{min}}) \\ 0 & \text{otherwise} \end{cases}$

  where $\delta \in (0, 1)$
- Cwnd adjustment —

$$w_i = \begin{cases} \frac{w_{i-1}}{2} & (c_i = 1) \wedge (\text{itti} = 0) \\ 1 & (c_i = 1) \wedge (\text{itti} = 1) \\ w_{i-1} + \frac{1}{w_{i-1}} & (c_i = 0) \wedge (\text{itti} = 0) \end{cases}$$

  itti – interference timeout timer indication (debounce)
- Requires feedback of delay measurement
- Requires accurate estimates of $d_{\text{max}} - d_{\text{min}}$

Algorithms: Vegas [Brakmo and Peterson, 1995]

- Iconic rate based TCP
- Defines two rates:

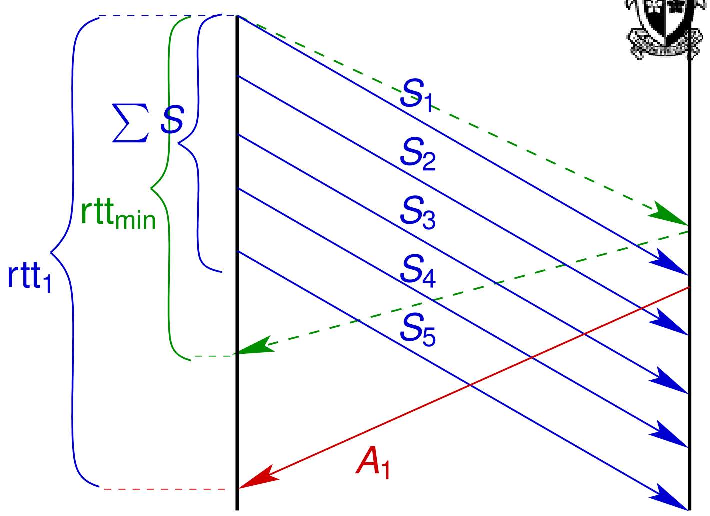
$$\text{actual} = \frac{\sum S}{\text{rtt}}$$

$$\text{expected} = \frac{w}{\text{rtt}_{\text{min}}}$$

and $\text{diff} = \text{expected} - \text{actual}$

- window adjustment:

$$w \leftarrow \begin{cases} w - 1 & \text{diff} > \beta \\ w + 1 & \text{diff} < \alpha \\ w & \text{otherwise} \end{cases}$$



- Usually $w = \sum S$
- Then $\tau_{\text{diff}} = \text{rtt} - \text{rtt}_{\text{min}}$
- where $\tau_{\text{diff}} = \text{diff}\left(\frac{\text{rtt}+\text{rtt}_{\text{min}}}{w}\right)$
- requires accurate estimate of rtt_{min}
- AIAD

Algorithms: FAST [Wei et al., 2006]

- Enhanced Vegas type algorithm
- MIMD — AIMD to slow for high BDP networks
- Uses delay as a rich (non binary) congestion indicator
- Cwnd is updated at regular time intervals (Δt):

$$w_{t+\Delta t} = \min\left\{2w_t,\ \gamma\left(\frac{\mathrm{rtt}_{\min,i}}{\overline{\mathrm{rtt}}_i} w_t + \alpha\right) + (1-\gamma)w_t\right\}$$

- For MIMD, $\alpha(w_t, q_i)$
  - increase is proportional to the size of cwnd and the network queueing delay.

Algorithms: Compound TCP [Tan et al., 2006]

- In high speed high BDP networks aims to increase:
  - efficiency
  - RTT fairness and TCP fairness
- In MSW Vista and 7
- Uses Vegas' rates: diff $=$ (expected $-$ actual)$\mathrm{rtt}_{\min}$
- Provides NewReno+ performance throughput
  - The send window, win_j, is calculated as: $\mathrm{win}_j = \min(w_j + \mathrm{dwnd}_j, \mathrm{awnd}_j)$
  - where w_j is NewReno's cwnd
  - and dwnd_j is the delay based window.
  - and awnd_j is the receivers advertised window.

Algorithms: Compound TCP *continued*

- The delay window is calculated as follows:

$$\text{dwnd}_{j+1} = \begin{cases} \text{dwnd}_j + \alpha \left((\text{win}_j)^k - 1\right)^+ & \text{diff} < \gamma \\ \left(\text{dwnd}_j - \zeta \text{diff}\right)^+ & \text{diff} \geq \gamma \\ \left(\text{win}_j(1-\beta) - \frac{\text{cwnd}}{2}\right)^+ & \text{on loss} \end{cases}$$

- Increase rule, where $\alpha = \frac{1}{8}$ is the multiplicative increase factor relative to window size ($k = 0.75$)
- Delay decrease rule, relative to diff (the queued data)
- Loss decrease rule, $\beta = 0.5$
- requires accurate estimate of rtt_{min}

note: $\text{win}_j = \min(w_j + \text{dwnd}_j, \text{awnd}_j)$

Algorithms: DUAL [Wang and Crowcroft, 1992]

- Designed to supplement loss based congestion control
- Delay based measurements provide "slow tuning" of cwnd every 2^{nd} RTT

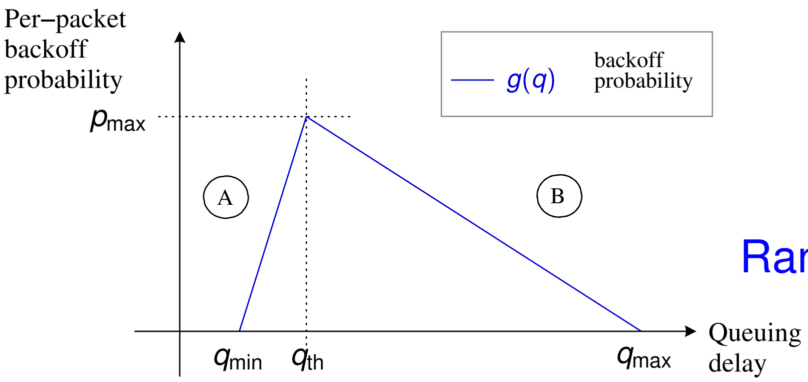
$$w \leftarrow \begin{cases} \beta w & \text{rtt} > \frac{(\text{rtt}_{\text{min}} + \text{rtt}_{\text{max}})}{2} \\ w & \text{otherwise} \end{cases}$$

where $\beta = \frac{7}{8}$

- Attempts to keep network buffers half full
- Smaller multiplicative decrease
- Relies on accurate estimates of rtt_{min} and rtt_{max}

Algorithms: Hamilton [Budzisz et al., 2009]

- Designed for coexistence with loss based TCP
- Inspired by Active Queueing techniques (as was PERT [Kotla and Reddy, 2008])



$$w_{i+1} = \begin{cases} \frac{w_i}{2} & X < g(q_i) \\ w_i + \frac{1}{w_i} & \text{otherwise} \end{cases}$$

Random multiplicative decrease

- Region **B** stable when queueing delay is high
- Region **A** stable when queueing delay is low
- AIMD matches NewReno
- Relies on accurate estimates of rtt_{min} and rtt_{max}

Algorithms: Others of Interest

- [King et al., 2005] — TCP-Africa
  - Two modes: Fast delay based, and slow NewReno based.
  - Compound TCP is based on some of Africa's ideas
- [Baiocchi et al., 2007] — YeAH-TCP
  - Yet Another Highspeed TCP
  - Two modes like Africa
  - Provides performance improvements on lossy paths.
- A number of schemes propose traffic shaping TCP's send rate
  - [Karandikar et al., 2000] – ABR like
  - [Wu et al., 2002] – leaky bucket
  - [Abendroth et al., 2002] – improved leaky bucket for network burstiness.

Conclusions

- Delay can provide an earlier indication of congestion than loss
- As such it will become important in high BDP networks:
  - Even aggressive loss based protocols have very long cwnd oscillations and cannot use the available bandwidth.
- Issues:
  - Compatibility with existing TCPs
  - Inaccurate estimates of rtt_{min} and rtt_{max}
- Send and receive rates are hard to measure (except in FQing networks)
  - Rate based flow control?
- CAIA's work in the next seminar

Bibliography I

[Clark et al., 1985] D. Clark, M. Lambert, and L. Zhang, "NETBLT: A bulk data transfer protocol," RFC 969, Dec. 1985, obsoleted by RFC 998. [Online]. Available: http://www.ietf.org/rfc/rfc969.txt

[Clark et al., 1987] D. Clark, M. Lambert, and L. Zhang, "NETBLT: A bulk data transfer protocol," RFC 998 (Experimental), Mar. 1987. [Online]. Available: http://www.ietf.org/rfc/rfc998.txt

[Jacobson, 1988] V. Jacobson, "Congestion avoidance and control," in *SIGCOMM '88: Symposium proceedings on Communications architectures and protocols*. New York, NY, USA: ACM, 1988, pp. 314–329

Bibliography II

[Jain, 1989] R. Jain, “A delay-based approach for congestion avoidance in interconnected heterogeneous computer networks,” *SIGCOMM Comput. Commun. Rev.*, vol. 19, no. 5, pp. 56–71, 1989

[Wang and Crowcroft, 1992] Z. Wang and J. Crowcroft, “Eliminating periodic packet losses in the 4.3-Tahoe BSD TCP congestion control algorithm,” *SIGCOMM Comput. Commun. Rev.*, vol. 22, no. 2, pp. 9–16, Apr. 1992

[Keshav, 1994] S. Keshav, “Packet-pair flow control,” Only available on web http://www.cs.cornell.edu/skeshav/doc/94/2-17.ps, 1994

Bibliography III

[Brakmo and Peterson, 1995] L. S. Brakmo and L. L. Peterson, “TCP Vegas: end to end congestion avoidance on a global internet,” *IEEE J. Sel. Areas Commun.*, vol. 13, no. 8, pp. 1465–1480, Oct. 1995

[Wei et al., 2006] D. X. Wei, C. Jin, S. H. Low, and S. Hegde, “FAST TCP: Motivation, architecture, algorithms, performance,” *IEEE/ACM Trans. Netw.*, vol. 14, no. 6, pp. 1246–1259, Dec. 2006

[Kuzmanovic and Knightly, 2006] A. Kuzmanovic and E. Knightly, “TCP-LP: low-priority service via end-point congestion control,” *IEEE/ACM Trans. Netw.*, vol. 14, no. 4, pp. 739–752, Aug. 2006

Bibliography IV

[Tan et al., 2006] K. Tan, J. Song, Q. Zhang, and M. Sridharan, "A compound TCP approach for high-speed and long distance networks," in *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, Apr. 2006, pp. 1–12

[Budzisz et al., 2009] L. Budzisz, R. Stanojevic, R. Shorten, and F. Baker, "A strategy for fair coexistence of loss and delay-based congestion control algorithms," *IEEE Commun. Lett.*, vol. 13, no. 7, pp. 555–557, Jul. 2009

[Kotla and Reddy, 2008] K. Kotla and A. Reddy, "Making a delay-based protocol adaptive to heterogeneous environments,"

Bibliography V

in *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, Jun. 2008, pp. 100–109

[King et al., 2005] R. King, R. Baraniuk, and R. Riedi, "TCP-africa: An adaptive and fair rapid increase rule for scalable TCP," in *IEEE INFOCOM 2005*, 2005, pp. 1838–1848

[Baiocchi et al., 2007] A. Baiocchi, A. P. Castellani, and F. Vacirca, "YeAH-TCP: Yet another highspeed TCP," in *PFLDnet 2007*, Feb. 2007. [Online]. Available: http://infocom.uniroma1.it/~vacirca/yeah/yeah.pdf

[Karandikar et al., 2000] S. Karandikar, S. Kalyanaraman, P. Bagal, and B. Packer, "TCP rate control," *SIGCOMM Comput. Commun. Rev.*, vol. 30, no. 1, pp. 45–58, Jan. 2000

[Wu et al., 2002] C.-S. Wu, M.-H. Hsu, and K.-J. Chen, "Traffic shaping for tcp networks: Tcp leaky bucket," in *TENCON '02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering*, vol. 2, Oct. 2002, pp. 809–812

[Abendroth et al., 2002] D. Abendroth, K. Below, and U. Killat, "The interaction between TCP and traffic shapers - clever alternatives to the leaky bucket," in *Global Telecommunications Conference, 2002. GLOBECOM '02. IEEE*, vol. 2, Nov. 2002, pp. 1507–1511