SWINBURNE UNIVERSITY OF TECHNOLOGY

News from CAIA's NewTCP Project

Delay-based TCP and improved instrumentation of FreeBSD's TCP stack

Presented by Michael Welzl on behalf of:

David Hayes, Lawrence Stewart

{dahayes,lastewart}@swin.edu.au

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

FreeBSD As A Research Platform

- Modular congestion control
  - In svn project branch, coming to a FreeBSD release soon
  - Available as a stand alone patch on the NewTCP website
  - BSD licenced NewReno, HTCP, CUBIC, Vegas, HD & CHD implementations available
  - New v0.10.0 release contains many improvements and paves way for shared CC between multiple transports e.g. TCP and SCTP
  - Supported by Cisco Systems
- KHELP and Enhanced RTT
  - Kernel Helper (KHELP) framework makes modularising “stuff” easy
  - Enhanced RTT (ERTT) KHELP module hooks TCP stack to maintain an RTT estimate appropriate for CC use
    - Used by Vegas, HD and CHD CC modules
    - ERTT supported by Cisco Systems

FreeBSD As A Research Platform

- Statistical Information for TCP Research (SIFTR)
  - FreeBSD kernel module to gather TCP connection data as CSV
  - Some similarity to Web100 but event driven and more variables
  - v1.2.3 has been integrated into FreeBSD and will appear in 8.2+
  - Supported by Cisco Systems and the FreeBSD Foundation
- Deterministic Packet Discard (DPD)
  - Adds 'pls' (packet loss set) option for dummynet pipes
  - e.g. ipfw pipe 1 config pls 1,5-10,30 would drop packets 1, 5-10 inclusive and 30
- Dummynet Forensic logging support
  - Log pipe/queue state on each packet event as CSV
- TCP stack improvements including RFC 3465 & reassembly queue autotuning
  - Supported by the FreeBSD Foundation

Delay-based Congestion Control

- Implementation of the algorithm proposed by Budzisz et al. [1] (we call it HD)
  - Probabilistic backoff based on inferred path queueing delay

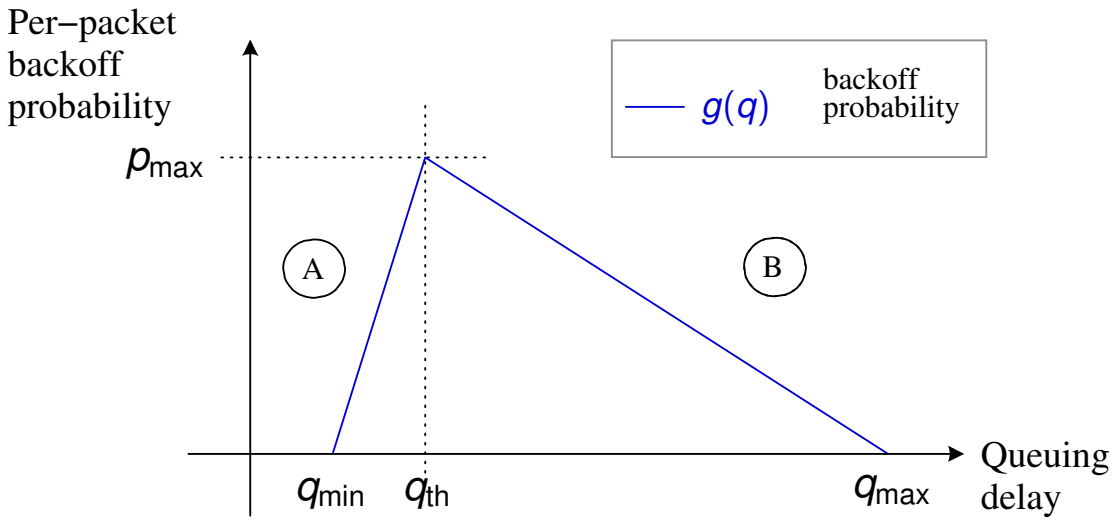
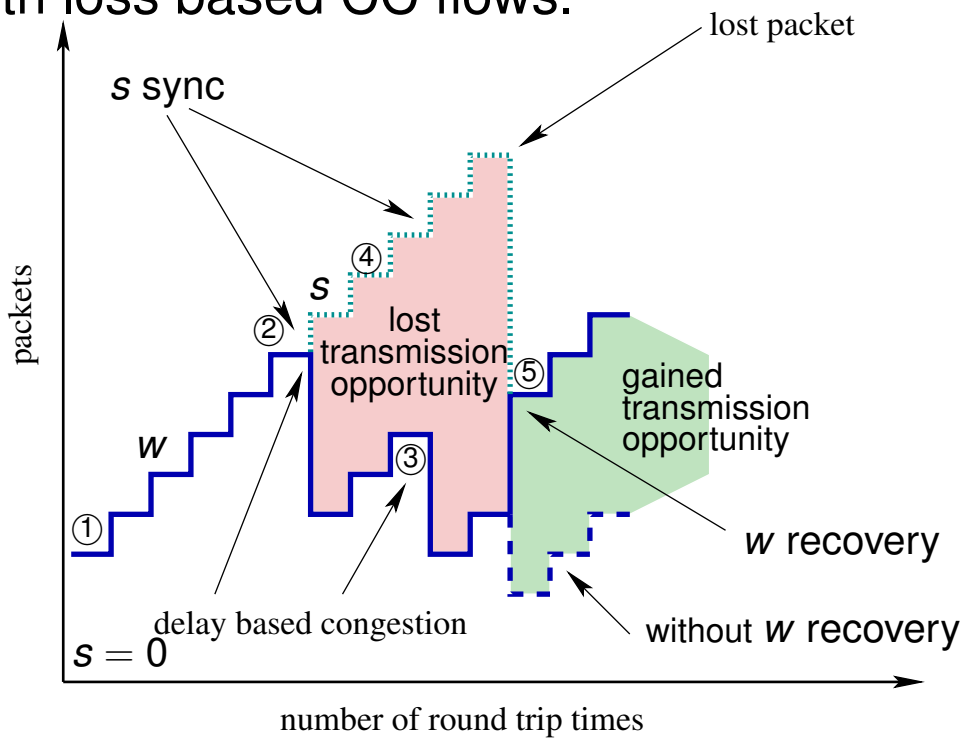


Figure: Per-packet backoff probability as a function of estimated queueing delay[1]

Delay-based Congestion Control continued

- "CHD": Enhanced HD (Hayes and Armitage [2])
  - Per RTT backoff decisions (for scalability and fairness)
  - Tolerance of non-congestion related packet loss
  - Improved coexistence with loss based algorithms in lightly multiplexed environments.

Figure: Interaction of the shadow window (s) and the congestion window (w) when competing with loss based CC flows.



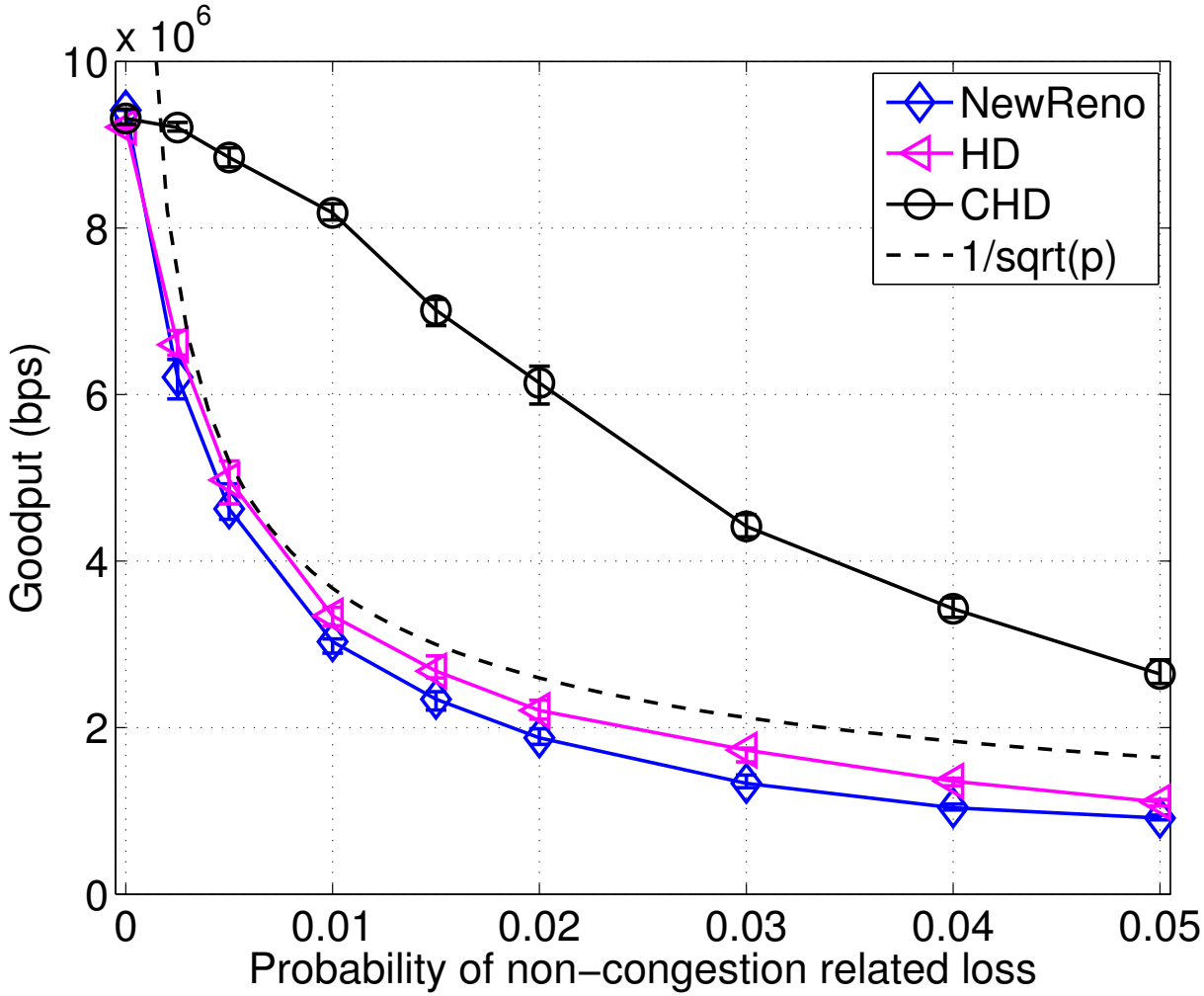


Figure: Comparison of the goodput of NewReno, HD, and CHD when there are non-congestion related losses

(10 Mbps bottleneck, 40 ms baseRTT, 100 ms queue at the bottleneck)

Delay-based Congestion Control continued

- Issues with inferred queueing delay CC signals:
  - Unfairness when BaseRTT estimate is wrong
  - Setting queueing delay thresholds (depends on network path)
- We have been revisiting the idea of delay gradient as a congestion signal (CDG – Hayes and Armitage [3]).
- Why?
  - Does not require an accurate estimate of baseRTT
  - Thresholds are less dependent on network path
- Hybrid
  - Combining the strengths of a threshold system based on inferred queueing delay, with the strengths of a delay-gradient approach may provide a more robust mechanism.

Further Information

- Contact
  - David Hayes <dahayes@swin.edu.au>
  - Lawrence Stewart <lastewart@swin.edu.au>
  - Grenville Armitage <garmitage@swin.edu.au>
- Links
  - `http://caia.swin.edu.au/urp/newtcp/`
  - `http://caia.swin.edu.au/freebsd/etcp09/`

Acknowledgements

- Cisco Systems

- The FreeBSD Foundation

References

[1] 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.*, 13(7):555–557, July 2009.

[2] David A. Hayes and Grenville Armitage. Improved coexistence and loss tolerance for delay based TCP congestion control. In *35th Annual IEEE Conference on Local Computer Networks (LCN 2010)*, Denver, Colorado, USA, October 2010. (to be presented).

[3] David A. Hayes and Grenville Armitage. Revisiting TCP congestion control using delay gradients. 2010. (submitted to ACM CoNEXT).