Tuning, Tweaking and TCP

(and other things happening at the Hamilton Institute)

David Malone and Doug Leith

16 August 2010

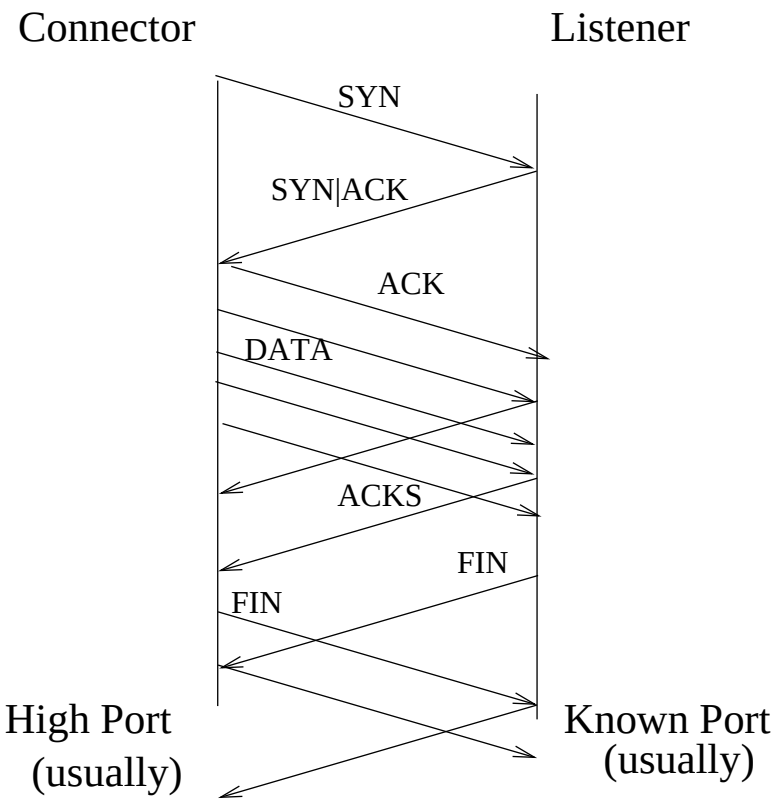
The Plan

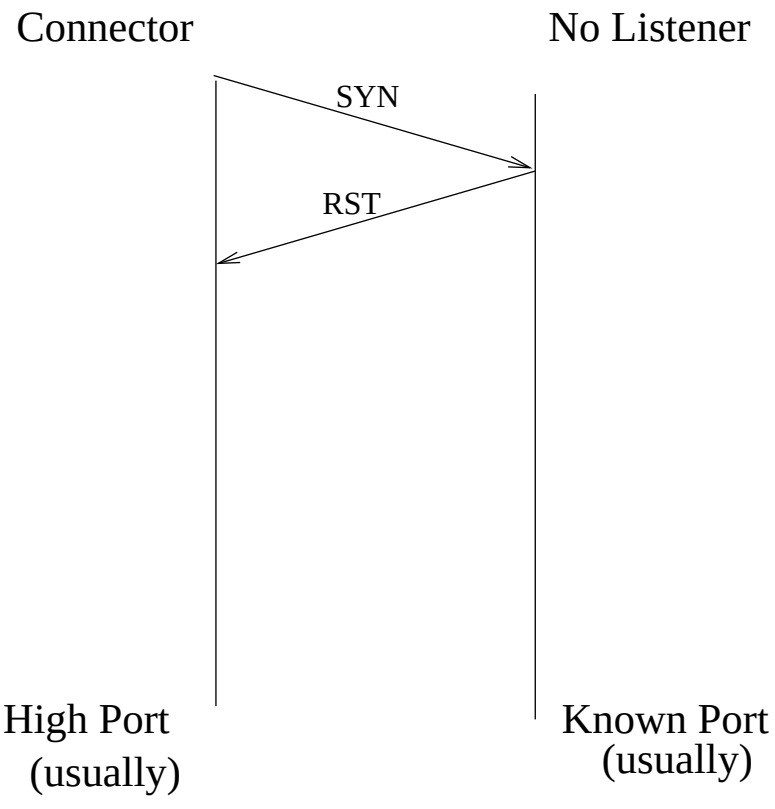
- Intro to TCP (Congestion Control).
- Standard tuning of TCP.
- Some more TCP tweaks.
- Tweaking congestion control.
- What else we do at the Hamilton Institute.

What does TCP do for us?

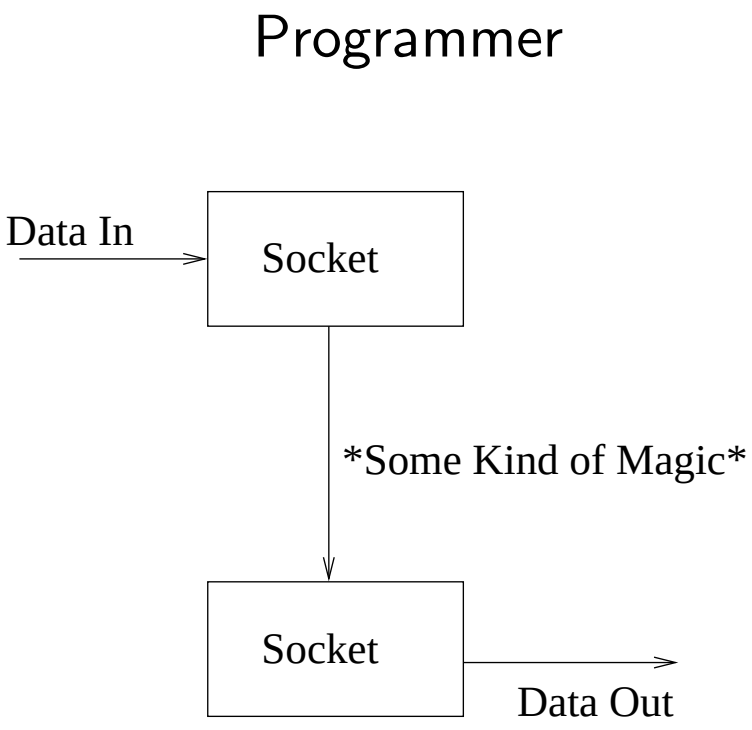
- Demuxes applications (using port numbers).
- Makes sure lost data is retransmitted.
- Delivers data to application in order.
- Engages in congestion control.
- Allows a little out-of-band data.
- Some weird stuff in TCP options.

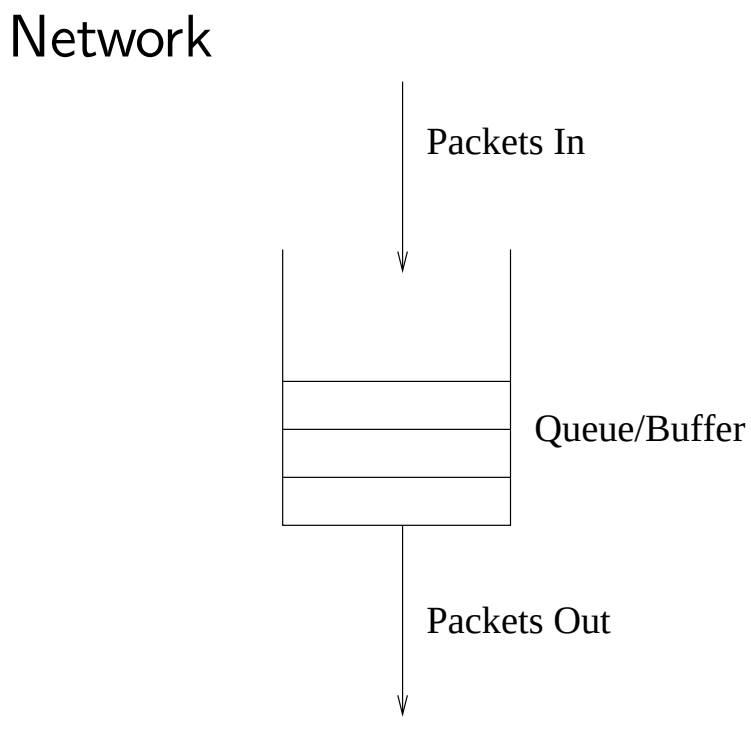
Standard TCP View





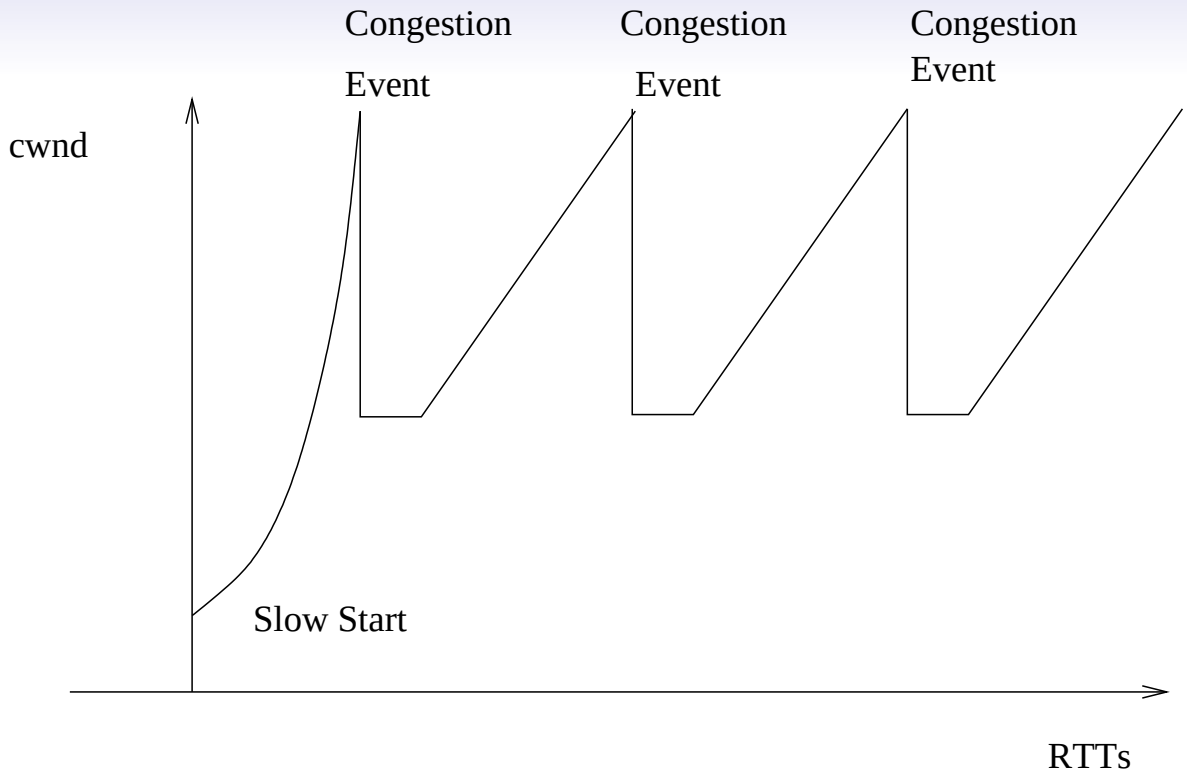
Other Views


Programmer
Data In
Socket
Some Kind of Magic
Socket
Data Out


Network
Packets In
Queue/Buffer
Packets Out

TCP Congestion Control

- TCP controls the number of packets in the network.
- Packets are acknowledged, so reverse flow of ACKs.
- Receiver advertises window to avoid overflow.
- Congestion window (cwnd) tries to adapt to network.
- Slow start mechanism to find rough link capacity.
- Congestion avoidance to gradually adapt.
- Timeouts for emergencies!



- Reno: Additive increase, multiplicative decrease (AIMD).
- To fill link need to reach BW $\times$ Delay.
- E.g. 1Gbps Ethernet Dublin to California
  $80000 \times 0.2 \approx 16000$ 1500B packets.
- Backoff $1/2 \Rightarrow$ buffer at bottleneck should be BW $\times$ Delay.
- Fairness (responsiveness, stability, ...)

Basic TCP Tuning

- Network stack has to buffer in-flight data.
- Need BW $\times$ delay sized sockbuf!
- `/proc/net/core/{r,w}mem_max` $\leftarrow$ sockbuf size limits.
- `/proc/net/ipv4/tcp_{r,w}mem` $\leftarrow$ min/def/max for TCP wnd.
- (or sockopt `SO_SNDBUF`/`SO_RCVBUF`).
- For large transfers, crank up to few MB.
- Traditionally kernel non-pageable memory, so need to balance.

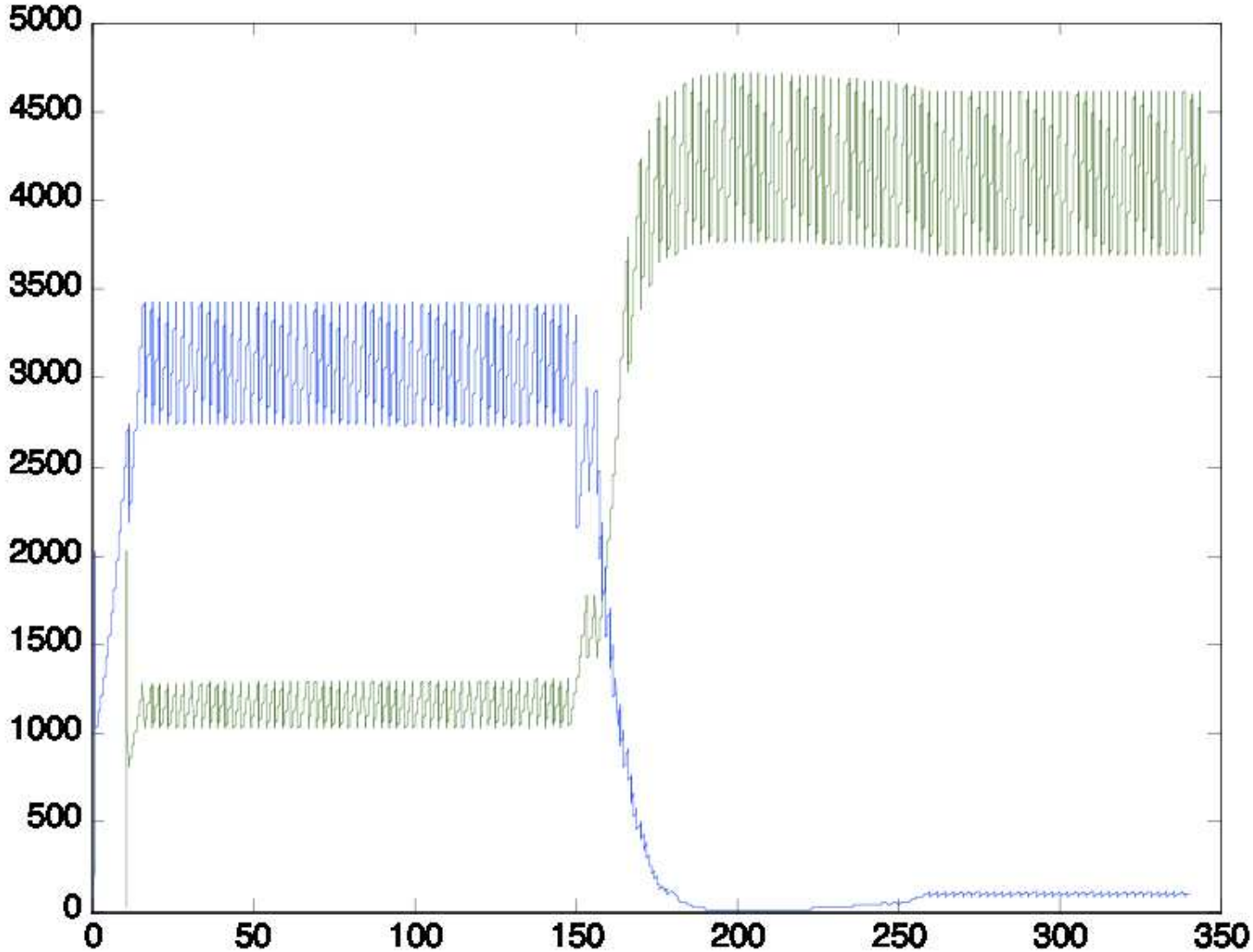
Common TCP Extensions

- Window scaling.
- Timestamps: better RTT estimates and duplicate detection.
- SACK lets receiver do more than ACK last contiguous byte.
- ECN lets receiver find out about congestion without drops.
- MD5 checksums.
- ABC.
- Now IETF work on new CA schemes.
- Also Google work on initial cwnd.

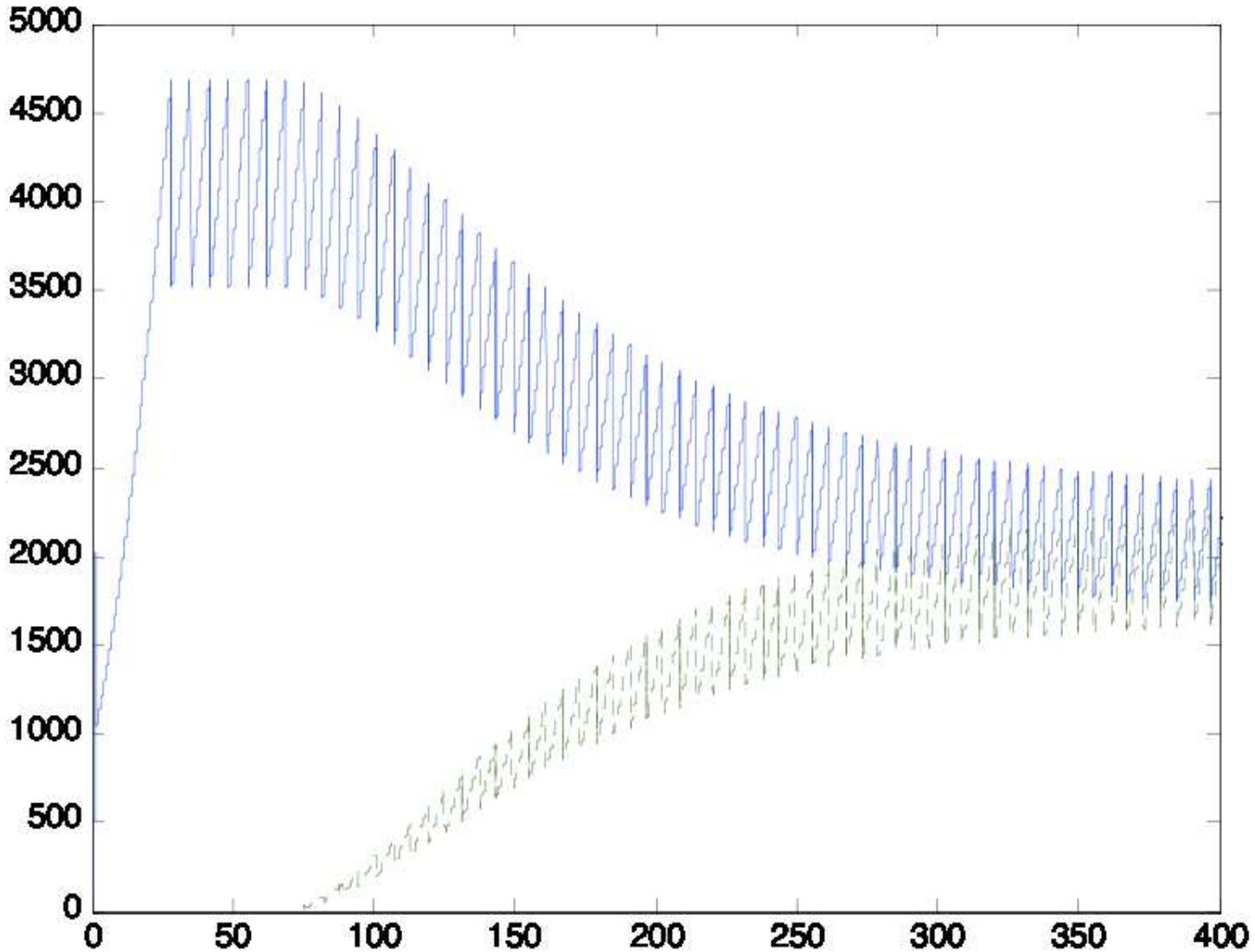
Problems for Congestion Control

- Packet loss caused by other factors.
- Filling a big link at one-packet-per-round-trip.
- Combined, Reno bad for high speed long distance links.
- Problem was flagged up: various solutions considered (Scalable TCP, HS TCP, FAST TCP, BIC, ...)
- In practice, you'll see Reno, Cubic and Compound.

Stability


5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300
350

Convergence



H-TCP

- Our CA scheme.
- Available in Linux.
- Aim to make small changes that could be analysed.
- Rate of increase depends on how long since last backoff.
- New flows compete on level playing field.
- Can use some nice ideas like adaptive backoff.

Defaults

Reno What most OSes are still using. Textbook: increase by 1 per RTT, backoff by one half.

Cubic Default in Linux for a long time (previously BIC and Reno). Increases as a cubic function of time since backoff. Friendlier than BIC.

Compound MS's congestion control. Uses two cwnds — one based on loss and one on delay. Available in Vista/2008/Windows 7.

IETF Drafts in TCP-M.

More Linux Tuning

- Linux allows you to choose the congestion control technique.
- Hidden behind `TCP_CONG_ADVANCED`.
- Can use `/proc/sys/net/ipv4/tcp_congestion_control`
- Includes implementations: BIC, CUBIC, HS, H-TCP, HYBLA, Illinois, LP, Scalable, Vegas, Veno, Westwood, YeAH.
- Also note tcpprobe — debugging TCP congestion control.

Practical Issues

- Congestion control isn't the only issue.
- SACK processing time.
- Implementation is important.
- Testing is important: land speed records.
- `http://www.web100.org/`
- `http://www.psc.edu/networking/projects/hpn-ssh/`

Other Networking at Hamilton Institute

- Delay based TCP to keep small delays.
- TCP over WiFi.
- WiFi: modeling, 11e, channel measurement, rate control, channel allocation, no collisions, buffering, ...
- Mesh: fairness, efficient low-delay multipath, multipath with network coding, ...
- Distributed load balancing.
- Device monitoring.
- Cheat detection.
- IPv6 and Internet Measurement.

Other Work at Hamilton Institute

- Systems biology, Parkinsons, HIV, mathematical immunology, pharmacokenetics, ...
- Control systems, switched systems, control tradeoffs, automotive, robotics, ...
- Linear algebra, matrix theory, graph theory, probability, ...
- Network maths courses.

Wrap Up

- Quick review of TCP Congestion Control.
- Some tips on where to start tuning/tweaking.
- Haven't covered everything: statistical buffering, incast, ...

Thanks! Any Questions.