TCP and Network Congestion Control

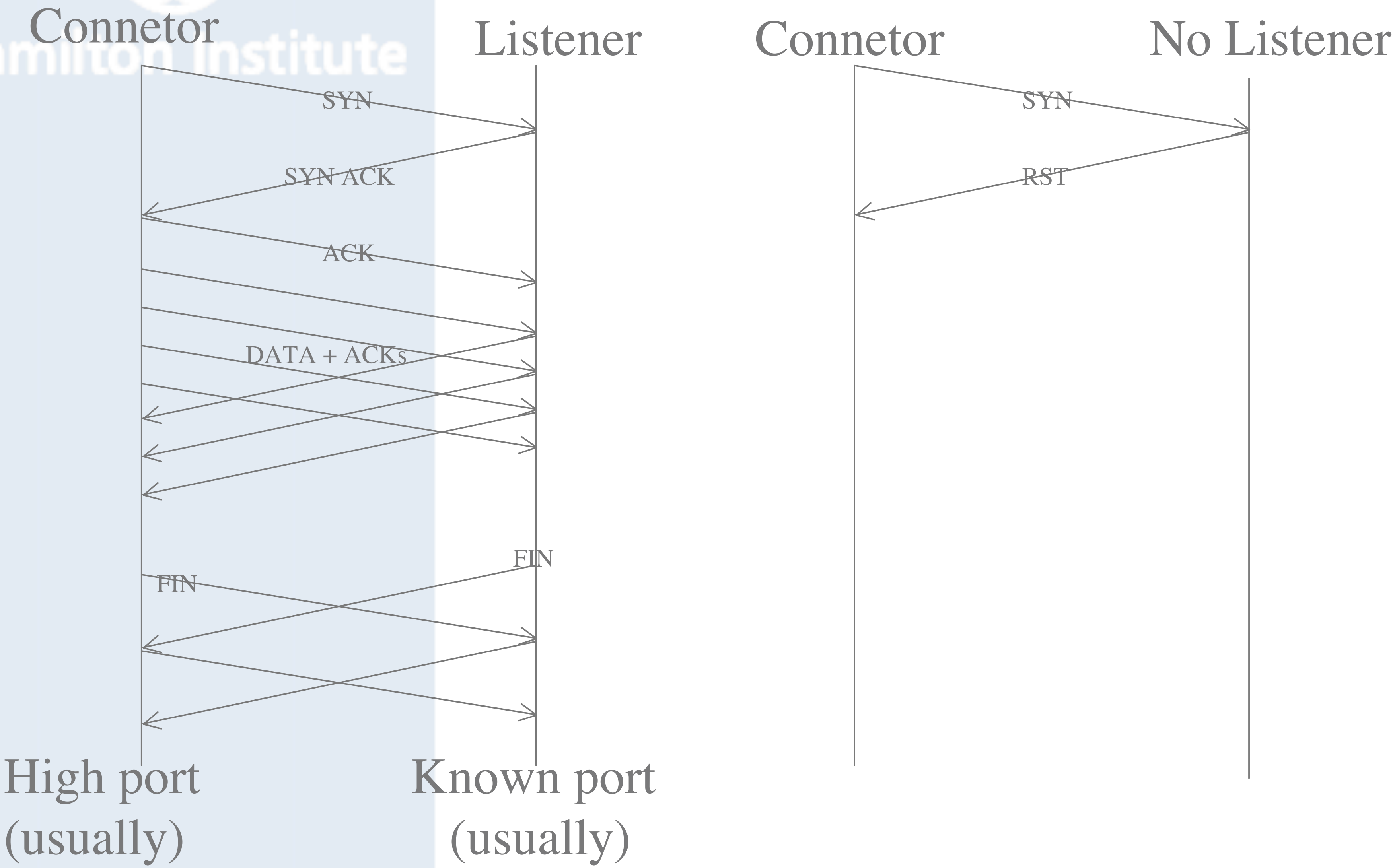
Amazon Dublin, December 2005

(David Malone and Doug Leith)

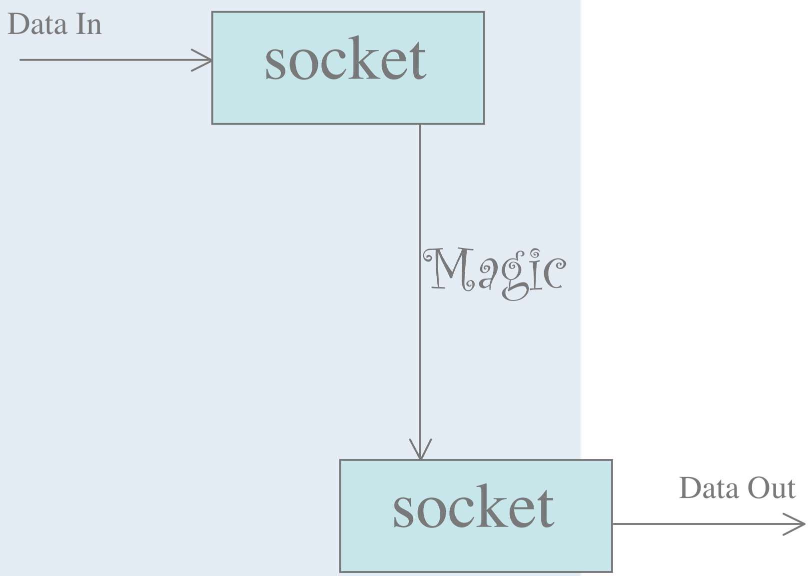
What TCP does for us

- 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.

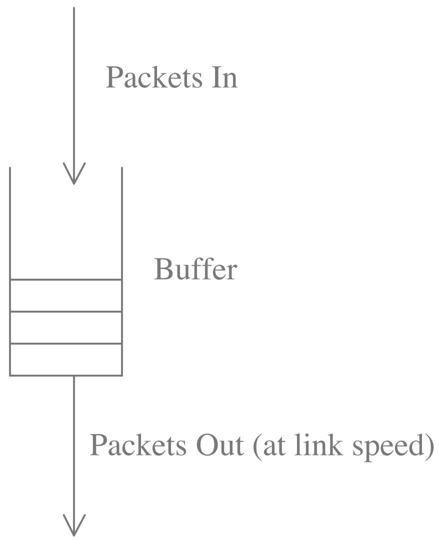
Standard Picture of TCP



Other views of TCP



Programmer's View

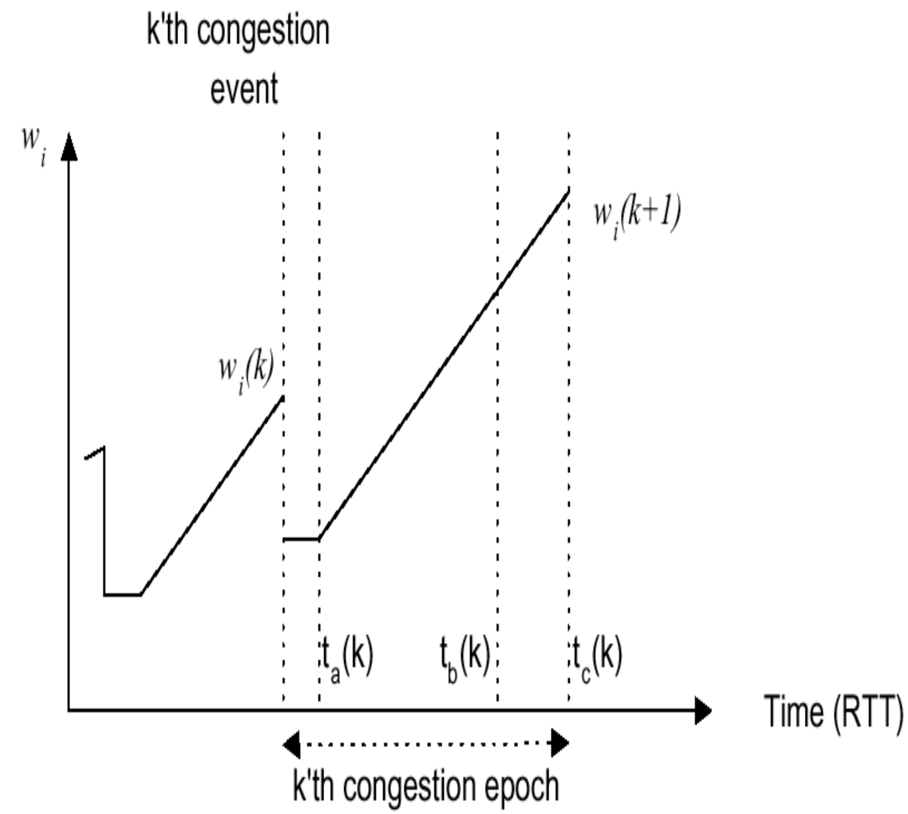


Network View

Congestion Control

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

The Congestion Window



- Additive increase, multiplicative decrease (AIMD).
- To fill link need to reach BW x Delay.
- Backoff by 1/2 => buffer at bottleneck link should be BW x Delay.
- Fairness (responsiveness, stability, …)

Basic TCP tuning

- Network stack has to buffer in-flight data.
- Need BW x delay sized sockbuf!
- /proc/net/core/{r,w}mem_max store something like sockbuf sizes.
- /proc/net/ipv4/tcp_{r,w,}mem store min/def/max for tcp windows.
- (or sockopt SO_SNDBUF/SO_RCVBUF).
- For large transfers, crank up to few MB.
- Memory will be wired, so need to use balance.

Existing TCP mods

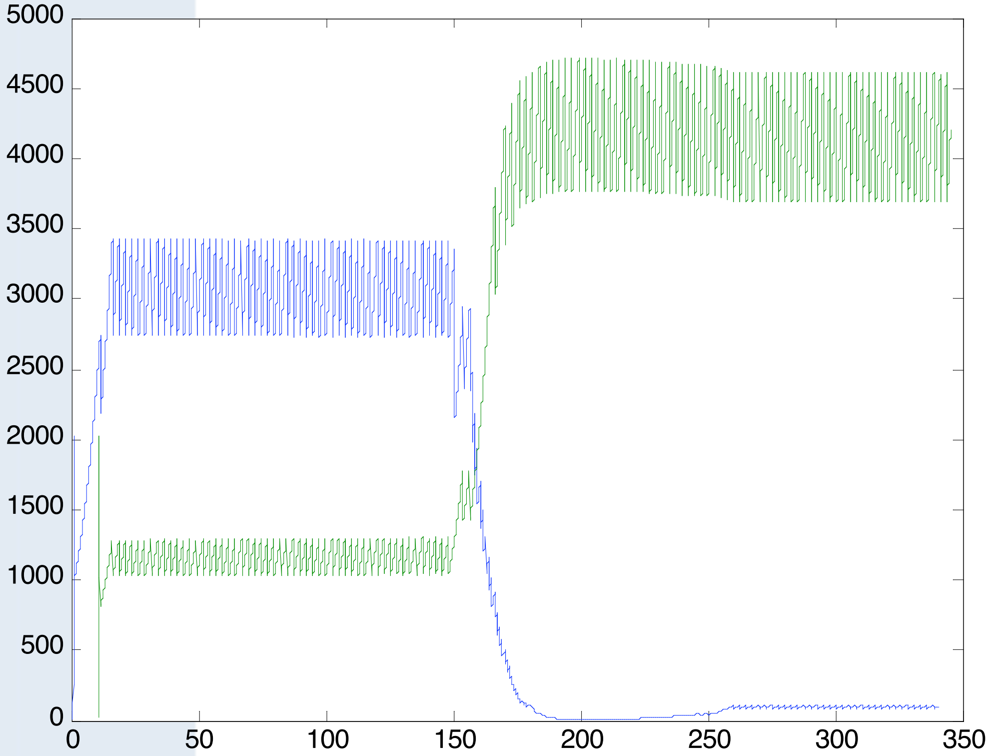
Hamilton Institute

- Window scaling.
- TCP traditionally ACKs last contiguous byte. SACK - transmits information about gaps.
- TCP usually uses drops as feedback signal. ECN allows use of few bits in IP header.
- Timestamps: more accurate RTT estimates.
- MD5 checksums.

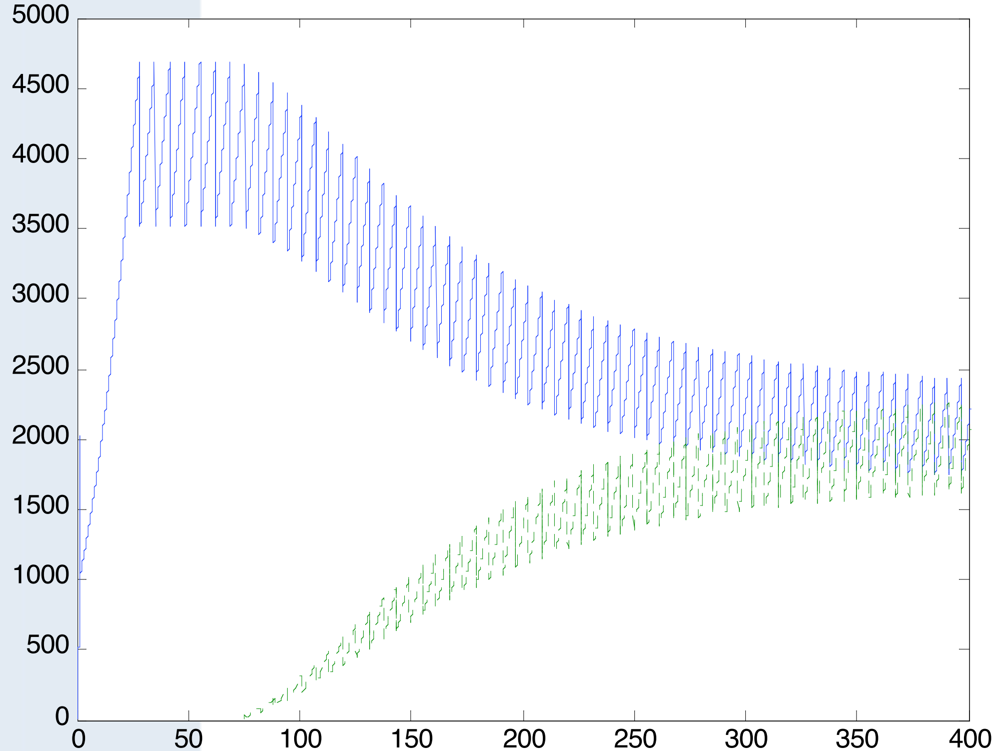
Problems for Congestion Control

- Packet loss caused by other factors.
- Filling a big link at one-packet-per-round-trip.
- The combination is bad for high speed long distance links.
- Problem was flagged up: various solutions being studied (BIC, Scalable TCP, High-Speed TCP, FAST TCP, H-TCP, …)

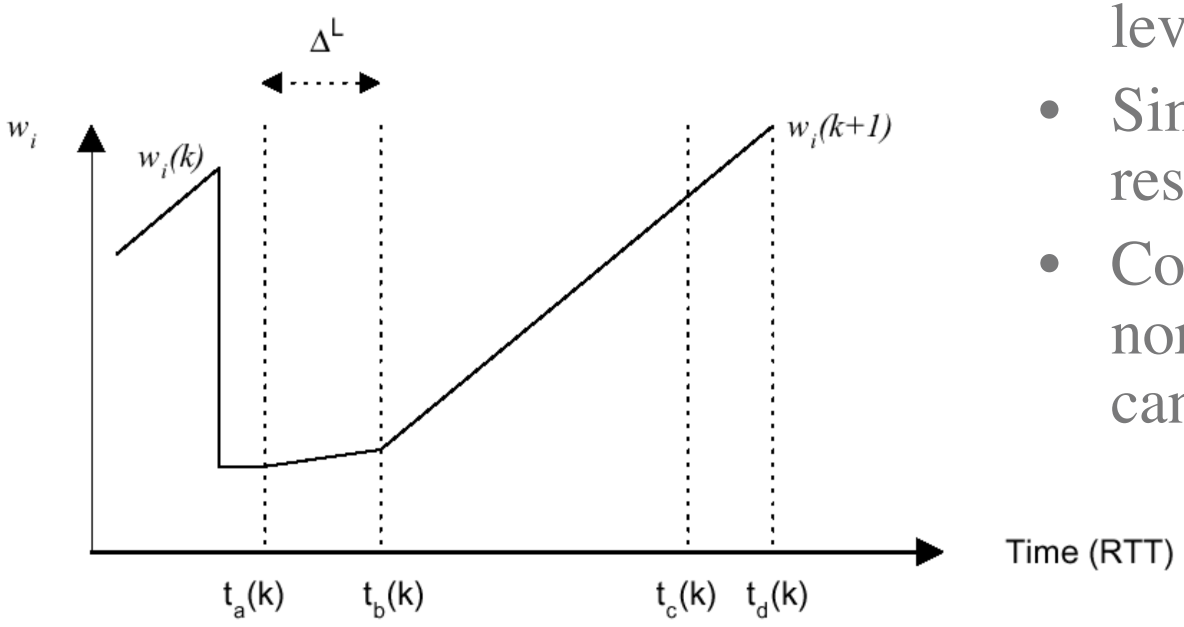
Stability issues



Convergence Issues

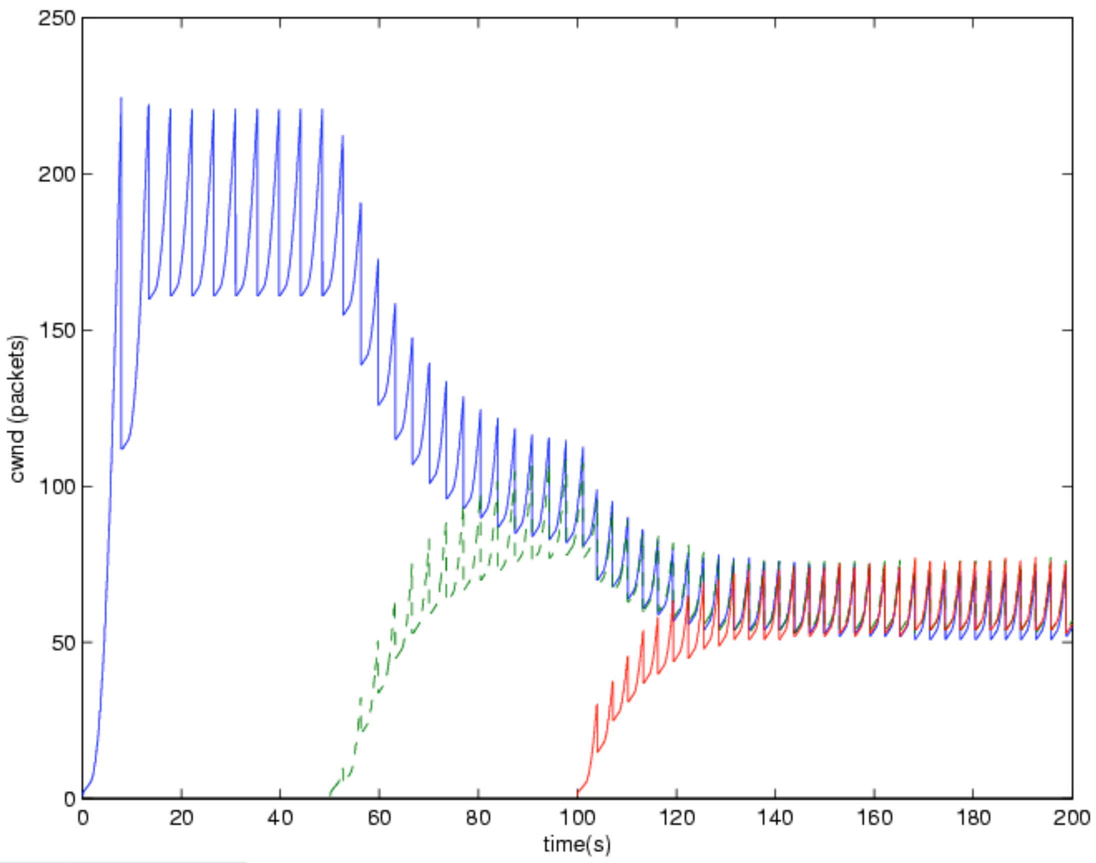


H-TCP



- Aim to make small changes we can analyse.
- Rate of increase depends on how long since last backoff.
- New flows compete on level playing field.
- Similar fairness & responsiveness.
- Competes fairly with normal TCP where it can compete.

Quicker Convergence



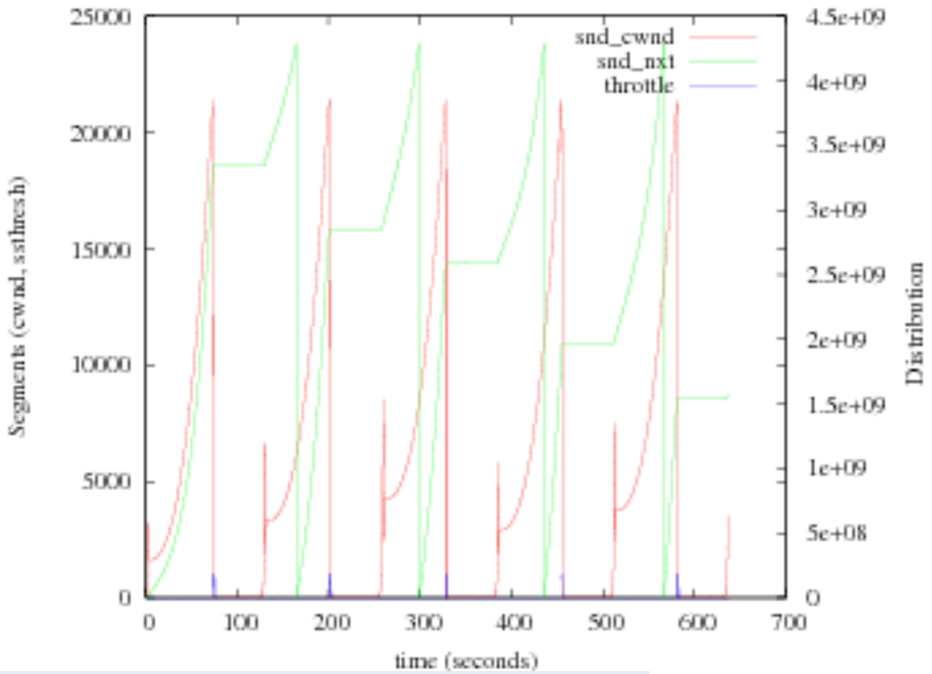
More Linux tuning

- As of 2.6.13 Linux allows you to choose the congestion control technique.
- Hidden behind TCP_CONG_ADVANCED.
- Can use /proc/sys/net/ipv4/tcp_congestion_control
- Older versions can disable/enable with /proc/sys/net/ipv4/tcp_{bic,vegas_cong_avoid,westwood}
- Some bugs fixed recently, so new kernels useful.

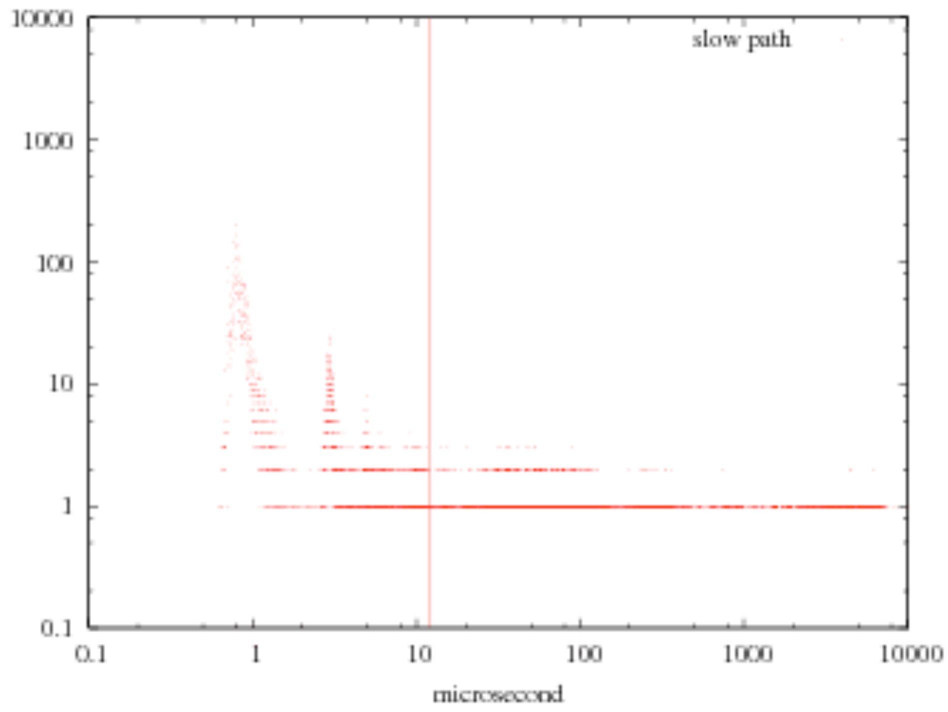
Practical Issues

- Congestion control isn't the only issue.
- Implementation is important.
- Testing is important: land speed records.
- Web 100 project to instrument Linux.
- Important stack tuning to be done.
- http://www.psc.edu/networking/projects/pathdiag/
- http://www.csm.ornl.gov/~dunigan/netperf/web100.html
- http://www.psc.edu/networking/projects/hpn-ssh/

Before

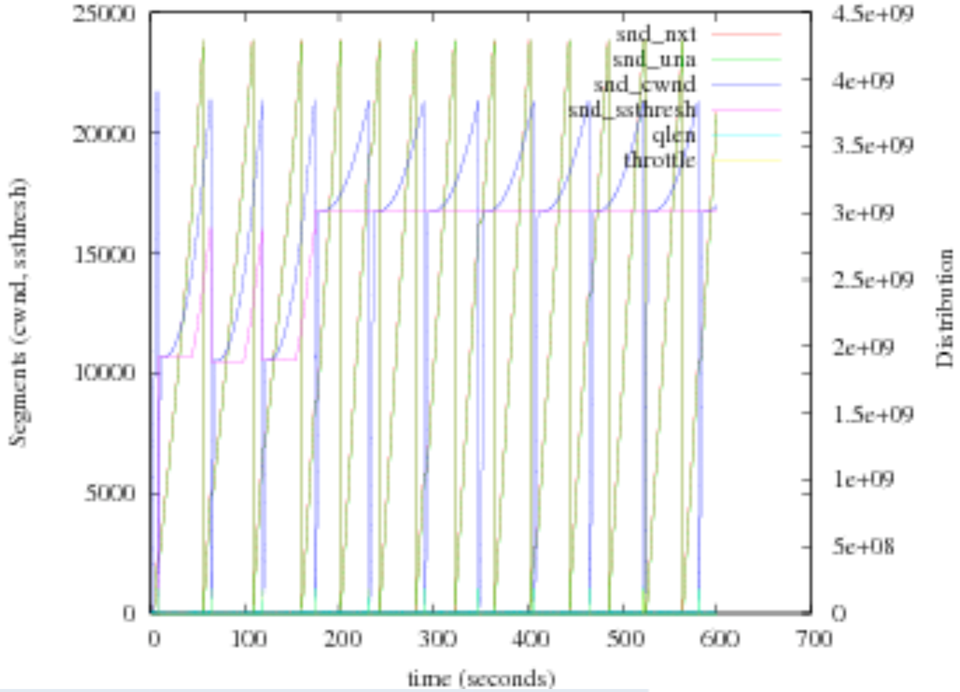


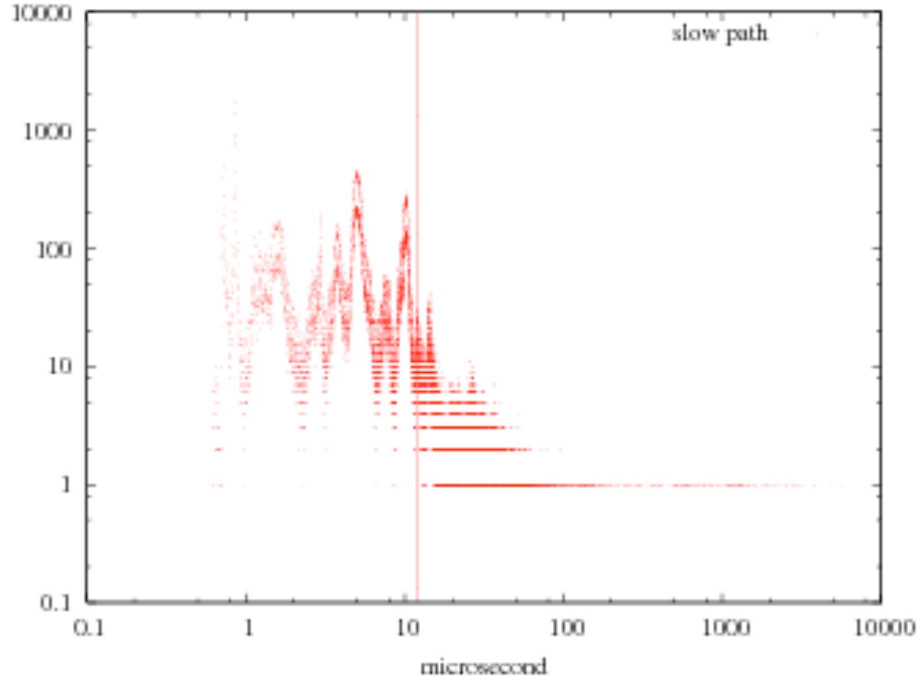
snd_cwnd
snd_nxt
throttle
Segments (cwnd, ssthresh)
Distribution
time (seconds)
0
100
200
300
400
500
600
700
25000
20000
15000
10000
5000
4.5e+09
4e+09
3.5e+09
3e+09
2.5e+09
2e+09
1.5e+09
1e+09
5e+08



slow path
microsecond
0.1
1
10
100
1000
10000

After


snd_nxt
snd_una
snd_cwnd
snd_ssthresh
qlen
throttle
Segments (cwnd, ssthresh)
Distribution
time (seconds)
0
100
200
300
400
500
600
700
25000
20000
15000
10000
5000
4.5e+09
4e+09
3.5e+09
3e+09
2.5e+09
2e+09
1.5e+09
1e+09
5e+08


slow path
microsecond
0.1
1
10
100
1000
10000

Other issues

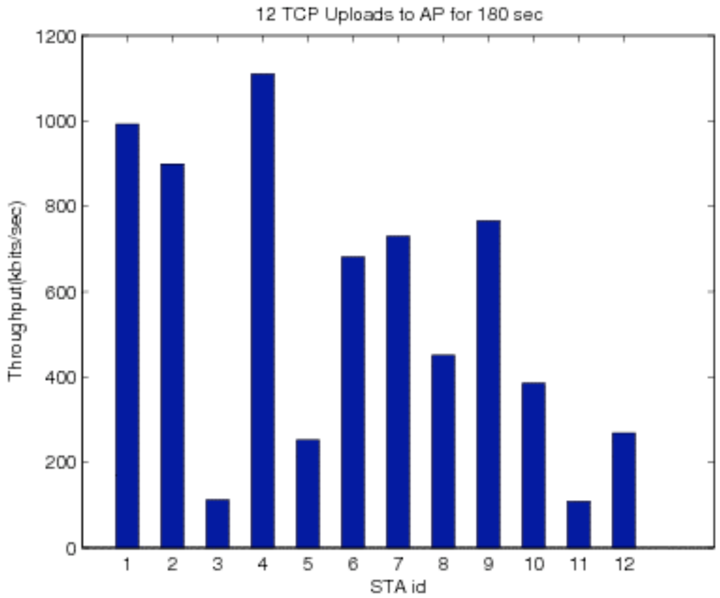
- High speed is important (packet switched vs. circuit switched), but not for everyone.
- Sizing router buffers important to everyone (cost, QoS, optics).
- Wireless interesting - random losses.
- Other interesting wireless issues too.
- Many flows don’t leave slow start => Quick start.
- For really small flows handshake is too much: T/TCP2.

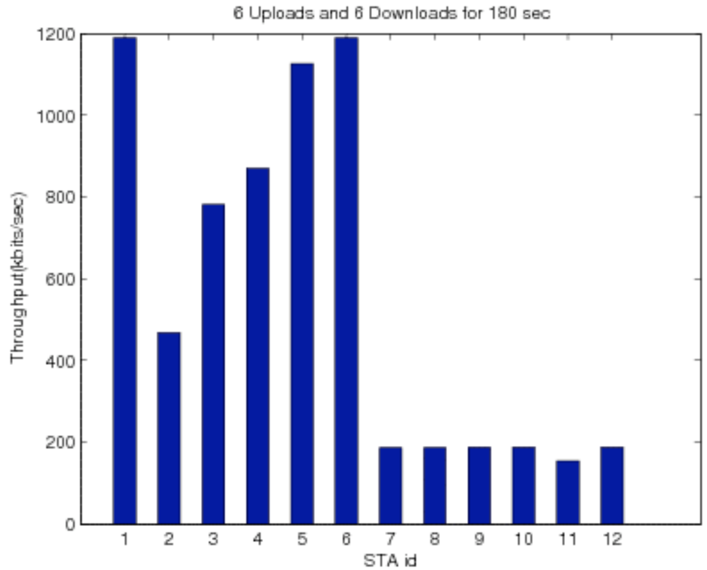
Hamilton Institute

Thanks!

Questions?

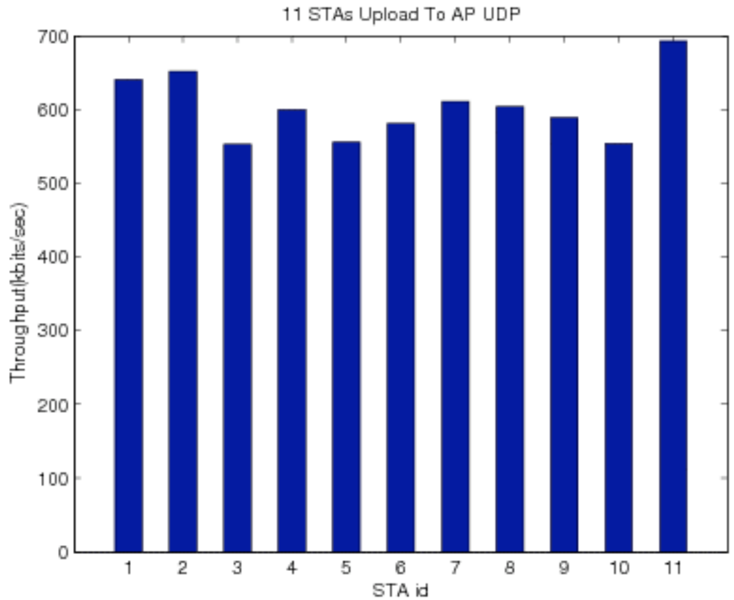
802.11 AP Before



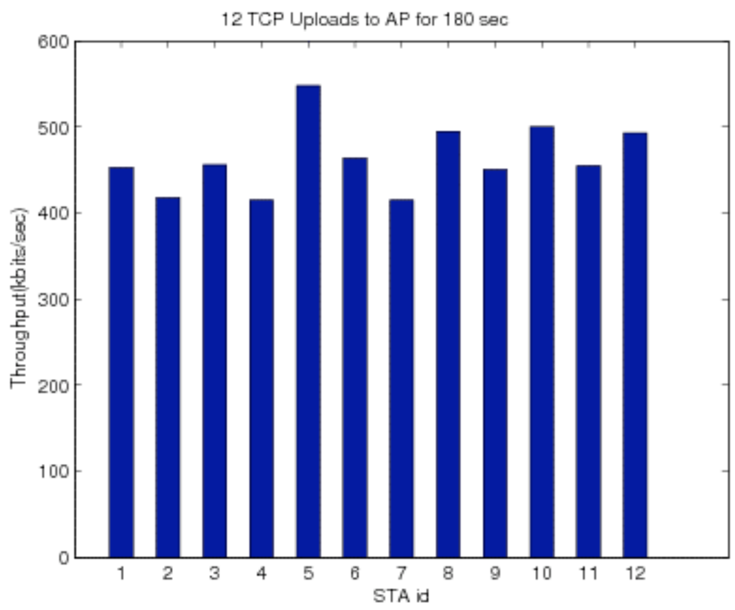


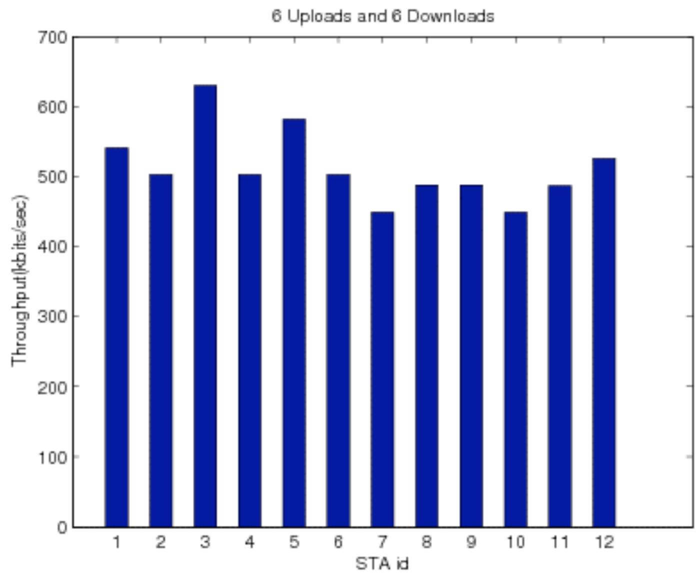
Hamilton Institute

Baseline



Adjusting 802.11e Parameters





More Questions?