Aggregate-based Congestion Control:
Addressing Congestion from Large-scale Traffic Patterns

Sally Floyd

March 26, 2001

From joint work with Steve Bellovin, John Ioannidis, Ratul Mahajan, Vern Paxson, Scott Shenker, and others...

SFI workshop on
The Internet as a Large-Scale Complex System

Topics:

- Observations on the Internet as a large-scale complex system:
  - Heterogeneity and change.
  - What we know about the current Internet.
  - End-to-end congestion control.

- Addressing Congestion from Large-scale Traffic Patterns: Controlling bullies, crowds, and mobs.
  - Controlling misbehaving or high-bandwidth flows (i.e., bullies).
  - Controlling flash crowds (i.e., crowds).
  - Controlling Denial-of-Service attacks (i.e., mobs).

Sub-themes:

- The Internet is a work in progress, with no central control or authority, many players independently making changes, and many forces of change (e.g., new technologies, new applications, new commercial forces, etc.)

- So far, the success of the Internet has rested on the IP architecture's robustness, flexibility, and ability to scale, and not on its efficiency, optimization, or fine-grained control.

- The rather decentralized and fast-changing evolution of the Internet architecture has worked reasonably well to date. There is no guarantee that it will continue to do so.

- The Internet is like the elephant, and each of us is the blind man who knows only the part closest to us.
  - The part of the Internet that I see is end-to-end congestion control.

Change and heterogeneity as conditions of the Internet:

- New link-level technologies: e.g., wireless.

- Higher bandwidth in some parts of the network, and very low bandwidth in other parts (e.g., wireless).
  - Cheaper bandwidth leads to higher connectivity between ASes.

- Changes in routers: e.g., QoS mechanisms, queue management, Explicit Congestion Notification.

- Changes to end-to-end congestion control mechanisms: e.g., in TCP, and in new transport protocols.

- Changes in infrastructure: e.g., web caching, content distribution.

- Changes in applications: e.g., telephony, streaming multimedia, peer-to-peer networking, multicast.

Invariant properties of the Internet:

- 24-hour cycles in traffic patterns.
- Log-normal connection sizes (for the main body of the distribution).
- Heavy-tailed distribution of connection sizes.
- Poisson arrivals for start times of user sessions.
- Self-similarity in traffic patterns.
- Invariants in topology?
- Heterogeneity and change!
  - [Paxson and Floyd, 1997]

Do we know the traffic dynamics and protocols in the current Internet?

- Measurements of response times and packet loss rates:
The Internet Traffic Report, the Internet Weather Report.

- Measurements of packet size distributions, protocol breakdown.

- Identification of congestion control behaviors of web servers.

- How is the traffic on a link characterized in terms of round-trip times, end-to-end congestion experienced by the packets on that link, etc.?

- We don't know much about the actual deployment of queue management mechanisms, traffic engineering, and a wide range of other issues.

  – [Web Page on Measurement Studies]

Why do we need end-to-end congestion control?

- As a tool for the application to better achieve its own goals:
E.g., minimizing loss and delay, maximizing throughput.

- To avoid congestion collapse.
  - Congestion collapse occurs when the network is increasingly busy, but little useful work is getting done.
  - E.g., congested links could be busy sending packets that will be dropped before reaching their destination.
  - Tragedy of the commons is avoided in part because the “players” are not individual users, but vendors of operating systems and other software packages.

- Fairness (in the absence of per-flow scheduling).

TCP congestion control:

- Packet drops as the indications of congestion.

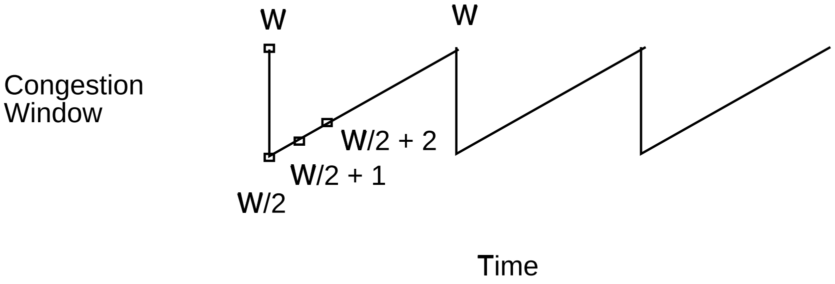
- TCP uses Additive Increase Multiplicative Decrease (AIMD) [Jacobson 1988].
  - Halve congestion window after a loss event.
  - Otherwise, increase congestion window each RTT by one packet.

- In heavy congestion, when a retransmitted packet is itself dropped, use exponential backoff of the retransmit timer.

- Slow-start: start by doubling the congestion window every roundtrip time.

The “steady-state model” of TCP:

- The model: Fixed packet size B in bytes.
  - Fixed roundtrip time R in seconds, no queue.
  - A packet is dropped each time the window reaches W packets.
  - TCP’s congestion window: W, $\frac{W}{2}$, $\frac{W}{2}$ + 1, ..., $W - 1$, W, $\frac{W}{2}$, ...

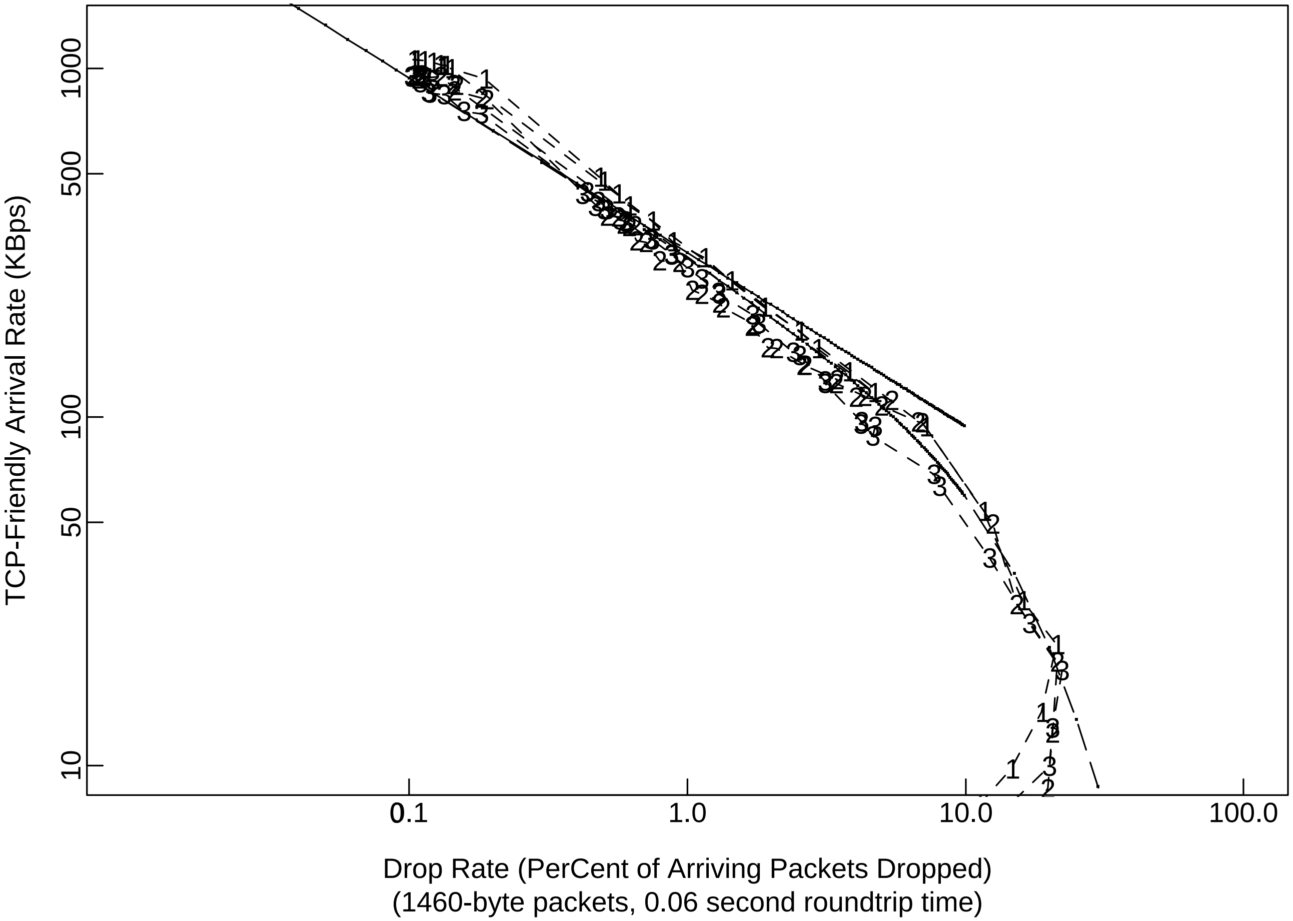


- The maximum sending rate in packets per roundtrip time: W
  - The maximum sending rate in byes per second: WB/R
  - The average sending rate T: $T = (3/4)WB/R$

- The packet drop rate p: $p = \frac{1}{(3/8)W^2}$

- The average sending rate T in bytes/sec: $T = \frac{\sqrt{1.5}B}{R\sqrt{p}}$

Verifying the “steady-state model” of TCP:



Solid line: the simple equation characterizing TCP

Numbered lines: simulation results

The “steady-state model” of TCP: an improved version.

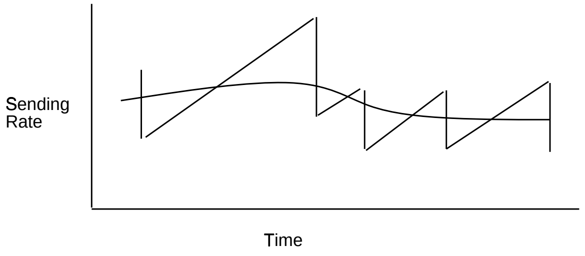
$$T = \frac{B}{RTT\sqrt{\frac{2p}{3}} + (2RTT)(3\sqrt{\frac{3p}{8}})p(1+32p^2)} \tag{1}$$

T: sending rate in bytes/sec

B: packet size in bytes

p: packet drop rate

– J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, Modeling TCP Throughput: A Simple Model and its Empirical Validation Proceedings of SIGCOMM’98



Equation-based congestion control:

- Use the TCP equation characterizing TCP’s steady-state sending rate as a function of the RTT and the packet drop rate.

- Over longer time periods, maintain a sending rate that is a function of the measured roundtrip time and packet loss rate.

- The benefit: Smoother changes in the sending rate in response to changes in congestion levels.

- The justification: It is acceptable not to reduce the sending rate in half in response to a single packet drop.

- The cost: Limited ability to make use of a sudden increase in the available bandwidth.

-

- Addressing Congestion from Large-scale Traffic Patterns.

Questions about congestion in the Internet:

- How often do routers have periods of unusually-high packet drop rates?

- Which routers? (E.g., access routers? last-mile routers? routers for transoceanic links?)

- For periods of high packet drop rates, how often is it due to:
  – A few flows not using end-to-end congestion control?
  – Legitimate flash crowds?
  – DOS attacks?
  – Network problems (e.g., routing failures)?
  – Diffuse general congestion?

Bullies (misbehaving or high-bandwidth flows):

- Flow: defined by source/destination IP addresses and port numbers.
  - Example: a single TCP connection.

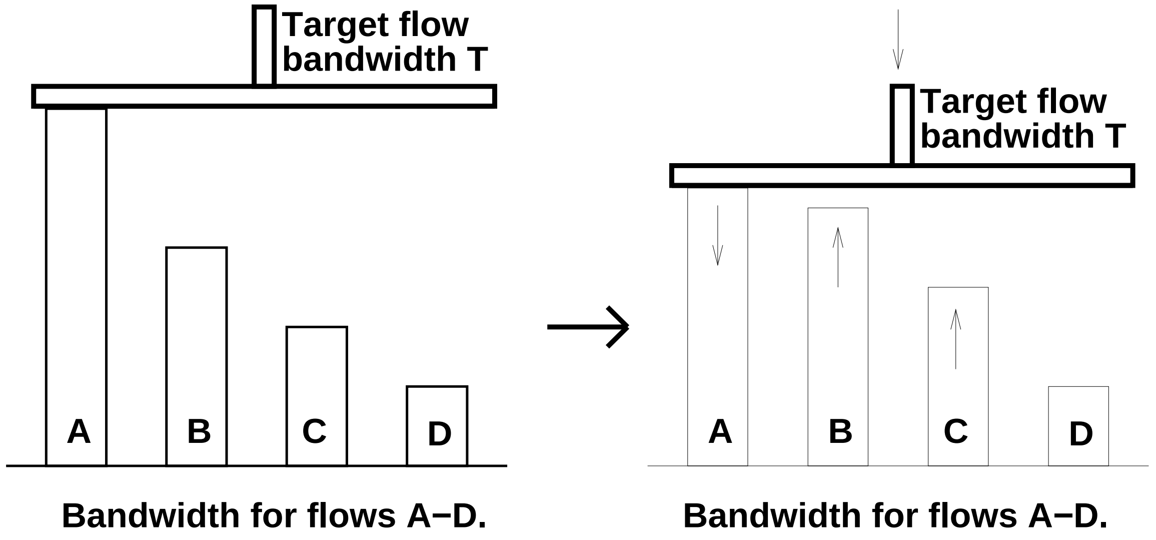
- Problem: Preventing congestion collapse from congested links carrying undelivered packets.

- The answer: Either the use of end-to-end congestion control, or a guarantee that packets that enter the network will be delivered to the receiver.

- The concrete incentive to users: Provide mechanisms in routers that, in times of high congestion, police high-bandwidth flows contributing to that congestion.

Controlling High-Bandwidth Flows at the Congested Router

- Max-min fairness is an acceptable policy for flows.
  - – Per-flow scheduling gives max-min fairness.



- Implementation issues:
  - – detecting high-bandwidth flows;
  - – deciding the bandwidth limit for rate-limiting those flows.

Mechanisms for Controlling High-Bandwidth Flows

- Use the packet drop history at the router to detect high-bandwidth flows.

- The target bandwidth in pkts/sec from the TCP throughput equation is $\frac{\sqrt{1.5}}{R\sqrt{p}}$, for:

  R: a configured round-trip time

  p: the current packet drop rate

.

- Monitored flows are rate-limited before the output queue.

- Monitored flows could be misbehaving flows (e.g., not using end-to-end congestion control) or conformant flows with small round-trip times.

- Identifying which monitored flows are *misbehaving* would be a separate step.

  – [Mahajan and Floyd, 2000]

Crowds (flash crowds):

- Example: The Starr Report, September 11, 1998:

“Nothing in recent times has caused a spike quite like that: not the Olympics (Nagano or Atlanta); not the beginning or end of the World Cup.”

- Example: The Victoria’s Secret Internet fashion show, May 18, 2000.

- Example: The Slashdot Effect:

  – “The spontaneous high hit rate upon a web server due to an announcement on a high volume news web site.”

- Problem: Protecting other traffic on congested links.

Mobs (Denial of Service Attacks):

- Example: Denial of Service attacks, February 7 and 8, 2000:

  – Attacks on a large number of web sites across the U.S.

  – “It’s completely clear that the entire Internet had higher packet loss and far lower reachability for several hours.” - John Quarterman.

- Problem: Limiting the damage to the legitimate traffic at the site.

- Problem: Protecting the rest of the Internet.

The Mechanisms of Aggregate-based Congestion Control:

- Detect sustained congestion, as characterized by a persistent, high packet drop rate.

- Look at the packet drop history:
  - See if some aggregate is heavily represented in the packet drop history.
  - An aggregate is defined by destination address prefix, source address prefix, etc.

- If an aggregate is found:
  - Preferentially drop packets from the aggregate before they are put in the output queue, to rate-limit aggregate to some specified bandwidth limit.

– [Mahajan et al, 2001]

Traffic Aggregates are Different from Flows:

- Similarities between the mechanisms for controlling aggregates and flows:
  - Both use the packet drop history for identification.
  - Both use rate-limiting before the output queue.

- Differences:
  - Per-flow scheduling does not control aggregates.
  - There is no simple fairness goal for aggregates, as for flows.
  - Control of aggregates is heavily affected by policy, customer relationships, differentiated services, etc.
  - A single flow could be in several different aggregates:
    - E.g., destination 192.0.0.0/12, or source www.victoriasecret.com.
  - Aggregate-based congestion control (ACC) should only be invoked for extreme congestion.

A Thought Experiment of Aggregate-based Congestion Control (ACC):

- Under normal conditions, with no flash crowd:
  - N aggregates A_1-A_n share link with background traffic.
  - Packet drop rate p (e.g., $p = 0.01$).

- During flash crowd i from aggregate A_i, with no ACC at the router:
  - The drop rate is p_i (e.g., $p_i = 0.2$).
  - The throughput for A_j, for $j \neq i$, is roughly $\frac{1}{\sqrt{p_i/p}}$ of its value without the flash crowd (e.g., 1/5-th of its old value).

- During flash crowd i, with ACC at the router:
  - Assume that during the flash crowd, A_i is restricted to at most half the link bandwidth:
  - A_i's throughput is at worst halved, compared to the flash crowd with no ACC.
  - All other traffic has its throughput at worst halved, compared to times with no flash crowd (and its packet drop rate at most quadrupled).

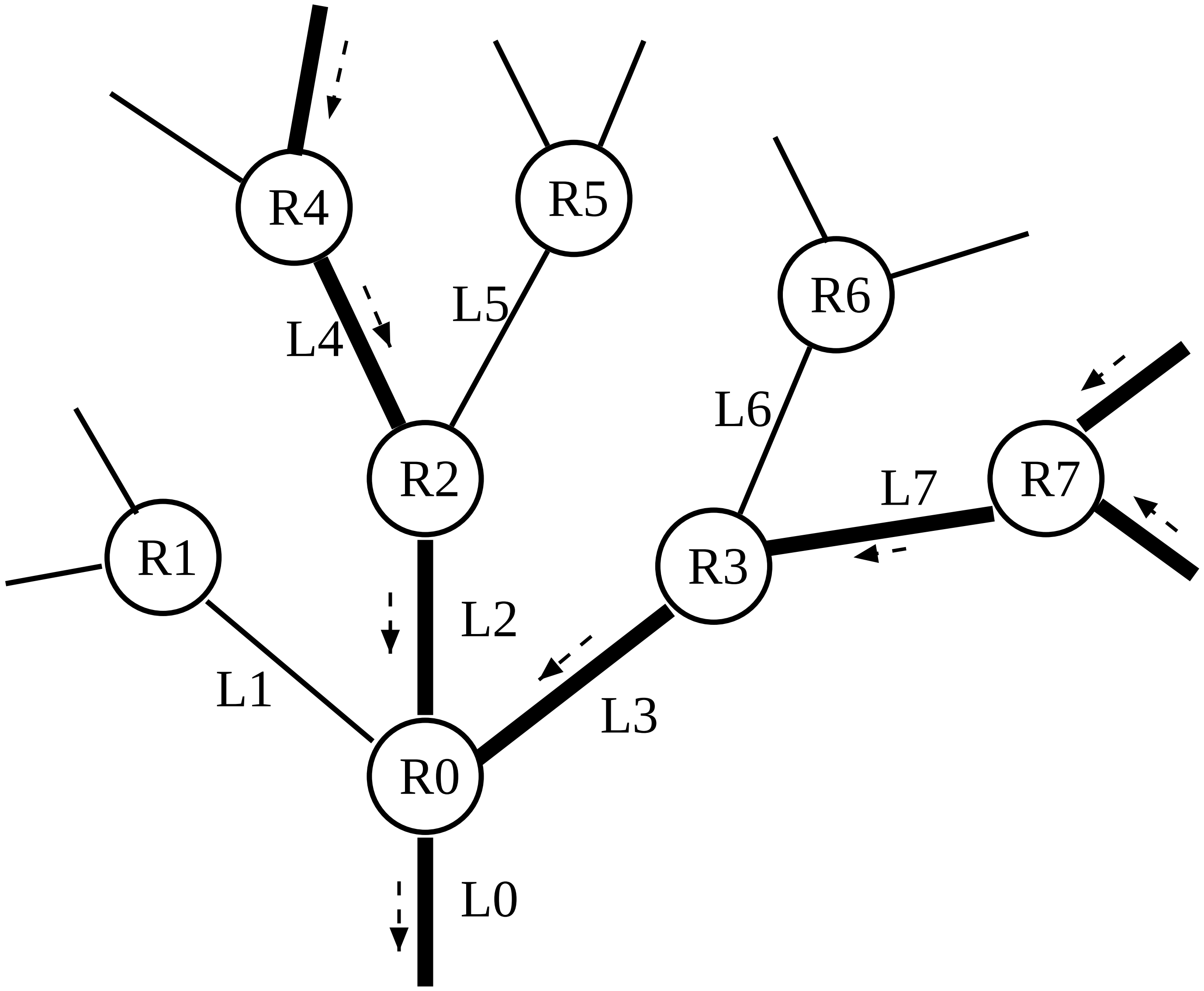
Now consider a Denial of Service (DOS) Attack:

- If an aggregate causing congestion is from a DOS attack, then the aggregate will contain both malicious traffic and legitimate, “good” traffic.

- We can not necessarily trust the IP source addresses.

- “Pushing-back” some of the rate-limiting of the aggregate to neighboring, upstream routers:
  - Limits the damage from the DoS attack, reducing wasted bandwidth upstream.
  - In some cases, allows rate-limiting to be concentrated more on the malicious traffic, and less on the good traffic within the aggregate.
  - Does not assume valid IP source addresses.

Illustration of pushback.



Questions about Aggregate-based Congestion Control?

- ACC helps traffic not in the aggregate, but why should we restrict the bandwidth given to a single aggregate in the first place?

- When does ACC with Pushback help an attacker to deny service to legitimate traffic within the aggregate?

-

-

Pushback, Traceback, and Source Filtering:

- With Pushback, a router rate-limiting packets from aggregate A might ask upstream routers to rate-limit that aggregate on the upstream link.

- Pushback is orthogonal to ”traceback”, which tries to trace back an attack to the source.
  - Traceback allows legal steps to be taken against the attacker.
  - Traceback by itself does not protect the other traffic in the network.

- Pushback is orthogonal to source filtering, which limits the ability to spoof IP source addresses.
  - Source filtering is important in any case.
  - Pushback can be useful even when source addresses can be trusted.

References

Web pages:

[Web Page on Measurement Studies] Measurement Studies of End-to-End Congestion Control in the Internet, ”http://www.aciri.org/floyd/ccmeasure.html”.

Papers:

[Floyd and Fall, 1999] Floyd, S., and Fall, K., “Promoting the Use of End-to-End Congestion Control in the Internet”, IEEE/ACM Transactions on Networking, August 1999.

[Jacobson, 1988] Jacobson, V., Congestion Avoidance and Control. Proceedings of SIGCOMM ’88 (Palo Alto, CA, Aug. 1988) URL “http://www-nrg.ee.lbl.gov/nrg-papers.html”.

[Mahajan et al, 2001] Ratul Mahajan, Steven M. Bellovin, Sally Floyd, John Ioannidis, Vern Paxson, and Scott Shenker, “Controlling High Bandwidth Aggregates in the Network”, draft, February 2001.

[Mahajan and Floyd, 2000] Mahajan, R., and Floyd, S., “Controlling High-Bandwidth Flows at the Congested Router”, draft, November 2000.

[Paxson and Floyd, 1997] Paxson, V., and Floyd, S., “Why We Don’t Know How To Simulate The Internet”, 1997 Winter Simulation Conference, December 1997.

Additional References:

[Clark88] D. D. Clark, The Design Philosophy of the DARPA Internet Protocols, SIGCOMM 88, August 1988.

[Floyd and Fall, 1999] Floyd, S., and Fall, K., Promoting the Use of End-to-End Congestion Control in the Internet, IEEE/ACM Transactions on Networking, August 1999. URL “http://www.aciri.org/floyd/papers.html”.

[FJ94] Floyd, S., and Jacobson, V., The Synchronization of Periodic Routing Messages. IEEE/ACM Transactions on Networking, V.2 N.2, p. 122-136, April 1994. URL “http://www.aciri.org/floyd/papers.html”.

[RFC 2481] Ramakrishnan, K.K., and Floyd, S., A Proposal to add Explicit Congestion Notification (ECN) to IP. RFC 2481, Experimental, January 1999. URL “http://www.aciri.org/floyd/papers.html”.

[S99] Vernon Schryver, Email Message-Id: 199910191533.JAA22180@calcite.rhyolite.com to the end2end-interest mailing list.

[TCP-Friendly Web Page] The TCP-Friendly Web Page, URL “http://www.psc.edu/networking/tcp_friendly.html”.

The future of congestion control in the Internet: several possible views:

- View #1: No congestion, infinite bandwidth, no problems.
- View #2: The “co-operative”, end-to-end congestion control view.
- View #3: The game theory view.
- View #4: The congestion-based pricing view.
- View #5: The virtual circuit view.
- The darker views: Congestion collapse and beyond.

Global traffic dynamics:

- Synchronized routing messages [FJ94].

- Undesired synchronization or emergent behavior for other network traffic?

  – Possible feedback loop: The TCP feedback loop of a data packet followed by an acknowledgement packet followed by another data packet.

  – Possible feedback loop: Feedback loops in the network of connections A, B, and C, with a loop where A and B share a congested link, B and C share a congested link, and C and A share a congested link.

“What simulations and measurements of prototype implementations do you have that show that it is better than alternatives? What objective concrete evidence do you have that it is worth the trouble of changing many 1,000,000s of hosts and many 100,000 routers?”

- [S99], Email to the end2end-interest mailing list.