Internet Congestion Control: Motivations

Steven Low
Caltech
netlab.CALTECH.edu
Outline

- Brief history of Internet congestion control
- Who/why cares about protocols
- How much can be improved
- Tech transfer

Lee Center enables impact
Control milestones

1969 - ARPANet
1974 - TCP
1988 - Tahoe
2006 - New TCP’s

Flow control:
Prevent overwhelming receiver

Congestion control – Gen 1:
Prevent overwhelming network

CC – Gen 2:
Serve new applications
theory based
Control milestones

1969
ARPANet

1974
TCP

1988
Tahoe

Flow control: Prevent overwhelming receiver

Congestion control – Gen 1: Prevent overwhelming network

Driven by infrastructure growth

2006
New TCP’s

CC – Gen 2: Serve new applications

By application growth
Backbone milestones

1969
ARPANet

1988
HTTP
Tahoe
T1, NSFNet

1991

1996
Napster
T3, NSFNet

1999
OC12
MCI

2003
OC48
vBNS

2006
OC192
Abilene

Backbone speed:
50-56kbps, ARPANet

Network is exploding
Internet at birth (1969)

1 October, SDS

1 November, IBM

2 September, SDS

December, DEC

Source: http://www.computerhistory.org/exhibits/internet_history/full_size_images/1969_4-node_map.gif
Internet growth

Hobbes' Internet Timeline Copyright ©2006 Robert H Zakon
http://www.zakon.org/robert/internet/timeline/

<table>
<thead>
<tr>
<th>DATE</th>
<th>HOSTS</th>
<th>DATE</th>
<th>HOSTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/69</td>
<td>4</td>
<td>05/82</td>
<td>235</td>
</tr>
<tr>
<td>06/70</td>
<td>9</td>
<td>08/83</td>
<td>562</td>
</tr>
<tr>
<td>10/70</td>
<td>11</td>
<td>10/84</td>
<td>1,024</td>
</tr>
<tr>
<td>12/70</td>
<td>13</td>
<td>10/85</td>
<td>1,961</td>
</tr>
<tr>
<td>04/71</td>
<td>23</td>
<td>02/86</td>
<td>2,308</td>
</tr>
<tr>
<td>10/72</td>
<td>31</td>
<td>11/86</td>
<td>5,089</td>
</tr>
<tr>
<td>01/73</td>
<td>35</td>
<td>12/87</td>
<td>28,174</td>
</tr>
<tr>
<td>06/74</td>
<td>62</td>
<td>07/88</td>
<td>33,000</td>
</tr>
<tr>
<td>03/77</td>
<td>111</td>
<td>10/88</td>
<td>56,000</td>
</tr>
<tr>
<td>12/79</td>
<td>188</td>
<td>07/89</td>
<td>130,000</td>
</tr>
<tr>
<td>08/81</td>
<td>213</td>
<td>10/89</td>
<td>159,000</td>
</tr>
</tbody>
</table>

New Survey
Old Survey
Control milestones

1969: ARPANet
1974: TCP
1983: ARPANet cut-over to TCP/IP
1988: Tahoe
Flow control: Prevent overwhelming receiver
2006:
TCP (1974)

- Window control: ack-clocking, timeout, retransmission
  - For correct delivery of packet sequence over unreliable networks
  - To prevent overwhelming receiver (through Advertised Window in TCP header)

“We envision [HOST retransmission capability] will occasionally be invoked to allow HOST accommodation to infrequent overdemands for limited buffer resources, and otherwise not used much.”

-- Cerf and Kahn, 1974
Congestion collapse

- October 1986, Internet had its first congestion collapse
- Link LBL to UC Berkeley
  - 400 yards, 3 hops, 32 Kbps
  - throughput dropped to 40 bps
  - factor of \(~1000\) drop!
- 1988, Van Jacobson proposed TCP congestion control

![Graph showing throughput vs. load](image-url)
Why the 1986 collapse

- 5,089 hosts on Internet (Nov 1986)
- Backbone speed: 50 – 56 kbps
- Control mechanism focused only on receiver congestion, not network congestion

- Large number of hosts sharing a slow (and small) network
  - Network became the bottleneck, as opposed to receivers
  - But TCP flow control only prevents overwhelming receivers

Jacobson introduced feedback control to deal with network congestion in 1988
Tahoe and its variants (1988)

- Jacobson, Sigcomm 1988
- + Avoid overwhelming network
- + Window control mechanisms
  - Dynamically adjust sender window based on congestion (as well as receiver window)
  - Loss-based AIMD

“… important considering that TCP spans a range from 800 Mbps Cray channels to 1200 bps packet radio links”

-- Jacobson, 1988
Control milestones

1969: ARPANet
1974: TCP
1988: Tahoe
2006: New TCP’s

Flow control: Prevent overwhelming receiver
Congestion control – Gen 1: Prevent overwhelming network
Application milestones

1969 1972
- ARPANet
- Network Mail
- Telnet
- File Transfer

1971 1973
- 1971: ARPANet
- 1972: Network Mail
- 1973: Telnet

1969 1972
- 1969: File Transfer

1969 1972
- 1972: Telnet

1971 1973
- 1971: ARPANet
- 1972: Network Mail

1988
- 1988: Tahoe
- 1988: Internet Talk Radio

1993
- 1993: Internet Phone

1995
- 1995: Napster music

1990
- 1990: Whitehouse online

2004
- 2004: AT&T VoIP

2005
- 2005: YouTube

Simple applications

Diverse & demanding applications
The first network email was sent by Ray Tomlinson between these two computers at BBN that are connected by the ARPANet.
Internet applications (2006)

- Telephony
- Music
- TV & home theatre
- Finding your way
- Mail
- Games
- Library at your finger tip
- Network centric warfare
- Software As A Service
Control milestones

- 1969: ARPANet
- 1974: TCP
- 1988: Tahoe
- 2006: New TCP’s

**Flow control**: Prevent overwhelming receiver

**Congestion control – Gen 1**: Prevent overwhelming network

**CC – Gen 2**: Serve new applications
Summary

- First 25 years of Internet (1969 – 94)
  - Networks are several orders of magnitude smaller in size, speed, and heterogeneity
  - Applications are much simpler

- Last 14 years of Internet (1994 – 08)
  - Networks are much bigger, faster, and heterogeneous
  - Applications are much more diverse and demanding

- Our control mechanisms have shown remarkable robustness, but also essentially frozen in the last 20 years
Outline

- Brief history of Internet congestion control
- Who/why cares about protocols
- How much can be improved
- Tech transfer

Lee Center enables impact
## Protocol problems

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Example problem &amp; Solution</th>
</tr>
</thead>
</table>
| Application       | P: Chattiness, slow  
|                   | S: Application acceleration                                    |
| Transport         | P: TCP congestion control  
|                   | S: TCP acceleration                                            |
| Network           | P: Path failure & congestion  
|                   | S: Overlay routing optimization                                |
| Link & Physical   |                                                                 |
Trend 1

- Exploding need to deliver large content
  - Video, software, games, business info
Trend 1: Online content is exploding

- CAGR 2005 – 2011: 36%
- Google worldwide: 46PB/month
- Library of Congress: 0.136PB
Trend 1: Video is exploding

Source: Deutsche Bank, IDC  Jan 2007
Trend 1: Multi-year growth

- Only ~10% of digital content is currently online

*Source: Deutsche Bank, CBS, ABC, NBC, FOX  Jan 2007*
Trend 2

- Exploding need to deliver large content
  - Video, software, games, business info
- Infrastructure growth to support delivery need
Trend 2: Broadband penetration

- Global broadband penetration is accelerating
  - 11 million new subscribers/month globally
  - 83% of US home Internet access is broadband by June 2007
**Trend 3**

- Exploding need to deliver large content
  - Video, software, games, business info
- Infrastructure growth to support delivery need
- **Centralize IT** to reduce costs of management, space, power, cooling
  - Exacerbated by virtualization of infrastructure and personalization of content
Trend 3: Centralization

- Power & cooling cost is escalating, fueling centralization
  - CAGR (2005-10): power & cooling 11.2%, new server spend 2.7%
  - In 2005, 1,000 servers cost $3.8M to power & cool in 4yrs; 2% increase in electricity cost raises cost by $200K

Source: IDC Sept 2006

Worldwide Expense
(US: $4.5B in 2006; EPA)
Trend 3: Centralization

![Graph showing spending trends](image)

Source: IDC, VIRTUALIZATION 2.0: THE NEXT PHASE IN CUSTOMER ADOPTION
Protocols becoming bottleneck

- Exploding need to deliver large content
  - Video, software, games, business info
- Infrastructure growth to support delivery need
- Centralize IT to reduce costs of management, space, power, cooling
  - Exacerbated by virtualization of infrastructure and personalization of content

- Implication
  - More large contents over longer distance
  - Served from centralized data centers

“You could only get that sustained rate if you are delivering within 100 miles, due to the way current Internet protocols work.”
Tom Leighton, MIT/Akamai, Oct 2007

Solution Approaches

- Protocol problems degrade performance of long-distance transfers
- Current approach: reduce distance
- FastSoft approach: fix protocol problems
Outline

- Brief history of Internet congestion control
- Who/why cares about protocols
- How much can be improved
  - Equilibrium
  - Dynamics
  - Heterogeneous protocols
- Tech transfer
Duality model of TCP/AQM

- TCP/AQM
  \[ x^* = F(R^T p^*, x^*) \]
  \[ p^* = G(p^*, Rx^*) \]

- Equilibrium \((x^*, p^*)\) primal-dual optimal:
  \[
  \max_{x \geq 0} \sum U_i(x_i) \quad \text{subject to} \quad Rx \leq c
  \]

- \(F\) determines utility function \(U\)
- \(G\) guarantees complementary slackness
- \(p^*\) are Lagrange multipliers
Duality model of TCP/AQM

- Equilibrium \((x^*, p^*)\) primal-dual optimal:

\[
\max_{x \geq 0} \sum U_i(x_i) \quad \text{subject to} \quad Rx \leq c
\]

Mo & Walrand 2000:

\[
U_i(x_i) = \begin{cases} 
\log x_i & \text{if } \alpha = 1 \\
(1-\alpha)^{-1} x_i^{1-\alpha} & \text{if } \alpha \neq 1
\end{cases}
\]

- \(\alpha = 1\) : Vegas, FAST, STCP
- \(\alpha = 1.2\) : HSTCP
- \(\alpha = 2\) : Reno
- \(\alpha = \infty\) : XCP (single link only)
Some implications

- Equilibrium
  - Always exists, unique if $R$ is full rank
  - Bandwidth allocation independent of AQM or arrival
  - Can predict macroscopic behavior of large scale networks

- Counter-intuitive throughput behavior
  - Fair allocation is not always inefficient
  - Increasing link capacities do not always raise aggregate throughput
Traditional: integrator model

Aggregate FAST rate

\[ \dot{p} = \frac{1}{c} \left( \sum_{i} x_i(t) - c \right) \]

Basic assumption

\[ x_i(t) = \frac{w_i(t - \tau_f^i)}{d_i + p(t)} \]
Traditional: integrator model

![Graph showing queue size vs. time with different models: NS-2, Static model, Integrator model, and Joint model. The graph indicates that the integrator model lags true link dynamics.]
New: integrate micro effects

aggregate FAST rate

\[ \dot{p} = \frac{1}{c} \left( \sum_i x_i(t) - c \right) \]

\[ \int_t^{t+\tau(t)} x(s) \, ds = w(t + \tau(t)) \]

Can recover all previous link models
New model tracks dynamics

![Graph showing comparison of models with Testbed, Static link model, Joint link model, Integrator link model, and Proposed model.

- **Testbed**
- **Static link model**
- **Joint link model**
- **Integrator link model**
- **Proposed model**

*FastSoft, Inc. - Confidential*
Some implications

- Interaction of paced and un-paced flows

- FAST stability depends on heterogeneity of RTT’s, not their particular values

- Important in applications where dynamics is persistent
Multiple equilibria

Tang, Wang, Hegde, Low, Telecom Systems, 2005

FastSoft, Inc. - Confidential
Multiple equilibria

Tang, Wang, Hegde, Low, Telecom Systems, 2005
Heterogeneous protocols

- Equilibrium: \( p \) that satisfies

\[
x_i^j(p) = f_i^j \left( \sum_l R_{li} m_l^j (p_l) \right)
\]

\[
y_l(p) := \sum_{i,j} R_{li}^j x_i^j(p) \begin{cases} 
\leq c_l \\
= c_l \quad \text{if} \quad p_l > 0
\end{cases}
\]

Duality model no longer applies!

\( p_l \) can no longer serve as Lagrange multiplier
Global uniqueness

Theorem
- If price heterogeneity is small, then equilibrium is globally unique

Corollary
- If price mapping functions $m^j_i$ are linear and link-independent, then equilibrium is globally unique

\[ m^j_i \in [a, 2^{1/L} a] \text{ for any } a_i > 0 \]
\[ \hat{m}^j_i \in [a^j, 2^{1/L} a^j] \text{ for any } a^j > 0 \]

e.g. a network of RED routers with slope inversely proportional to link capacity almost always has
Local stability: `uniqueness' $\xrightarrow{}$ stability

**Theorem**
- If *price heterogeneity* is *small*, then the unique equilibrium $p$ is locally stable

Linearized dual algorithm: $\delta \dot{p} = \gamma J(p^*) \delta p(t)$

Equilibrium $p$ is *locally stable* if

$$\text{Re} \lambda(J(p)) < 0$$
Local stability: `converse’

**Theorem**
- If all equilibria $p$ are locally stable, then it is globally unique

**Proof idea:**
- For all equilibrium $p$: $I(p) = (-1)^L$
- Index theorem: $\sum_{eq \; p} I(p) = (-1)^L$
Deployment: FAST in a box

FAST inside

Internet

Throuput: LA → Tokyo

<table>
<thead>
<tr>
<th>File size (MB)</th>
<th>with Aria</th>
<th>without Aria</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.8x</td>
<td>6.3x</td>
</tr>
<tr>
<td>0.5</td>
<td>3.8x</td>
<td>21.8x</td>
</tr>
<tr>
<td>1</td>
<td>17.6x</td>
<td>32.5x</td>
</tr>
<tr>
<td>5</td>
<td>28.1x</td>
<td>28.1x</td>
</tr>
<tr>
<td>10</td>
<td>233Mbps</td>
<td>233Mbps</td>
</tr>
</tbody>
</table>

Throuput: San Fran → MIT

Reno avg: 35Mbps

FAST avg: 233Mbps
Extreme loss resilience

Heavy packet loss in Sprint network: Aria increased throughput by 120x!
Outline

- Brief history of Internet congestion control
- Who/why cares about protocols
- How much can be improved
- Tech transfer

Lee Center enables impact
High Tech Startup

Industry

Academia & Govt

VC

Market

Admin

Operations

Marketing

Technical Support

Human Resource

Engineering

Finance
Remarks

- Integration of theory, prototyping, experiment, infrastructure
  - Industrial partnership like ACCESS is key

- Idea ≠ Technology ≠ Product ≠ Company
  - Entrepreneurship takes a lot more

- Closing the loop achieves more

Diagram:

Theory → Experiment
Remarks

- Integration of theory, prototyping, experiment, infrastructure
  - Industrial partnership like ACCESS is key

- Idea $\neq$ Technology $\neq$ Product $\neq$ Company
  - Entrepreneurship takes a lot more

- Closing the loop achieves more