The Case for Informed Transport Protocols

Stefan Savage
Neal Cardwell    Tom Anderson

University of Washington
Our position

Wide-area network performance:
- is important
- is limited by inefficient congestion control
- can be improved by sharing information among hosts
The desktop bandwidth gap
(or, “why systems people should care about wide-area networking”)
Why is it so bad?

- Not enough capacity
  - Maybe - exchanges or shared access links
- Poor network load balancing
  - Maybe - routing inefficiency common
- Poor use of available bandwidth
  - Probably - TCP-style congestion control wasn’t designed for today’s Internet
A student’s view of the Internet then...

Long flows, thin pipes, infrequent losses

Typical late-1980’s grad students

ftp BSD4.2

Internet (56kbps)

ftp rogue

ucbvax.cs.berkeley.edu
… and the Internet now

Short flows, fat pipes, frequent losses

Typical 1999 grad students

5% drop rate

Internet (>155Mbps)

www.etrade.com
The congestion control problem

- Mismatch between:
  - TCP-style end-to-end congestion control
    - Assumes long flows, low bandwidth, few losses
  - Internet applications & infrastructure
    - Has short flows, high bandwidth, high loss rate

- Key problems:
  - Hosts are ignorant about the network
  - They take too long to learn
TCP connection establishment

- *Don’t know* RTT
- Set initial timeout to 3-6 secs
- Result: many connections are delayed
**TCP slow start**

- *Don’t know* bottleneck BW
- Increase sending rate until packet is lost
- Result: poor utilization of bandwidth; excess burstiness

\[
BW \leq \frac{B}{RTT \cdot \left[ \log_{1.5} \left( \frac{B}{2 \cdot MSS} + 1 \right) \right]}
\]

*BW*: bandwidth  
*RTT*: round-trip time  
*MSS*: Maximum segment size  
*B*: data transfer size
TCP congestion avoidance

- Don’t know available bandwidth
- Half sending rate when a packet is lost
- Result: underutilization of bandwidth

\[
BW \leq \left( \frac{MSS}{RTT} \right) \sqrt{\frac{3}{2}} \sqrt{\frac{1}{p}}
\]

*BW*: Bandwidth    
*RTT*: round-trip time
*MSS*: Maximum segment size    
*p*: packet loss rate
TCP stability

- **Assumptions**
  - Traffic is elastic
  - A packet loss will reduce the sending rate

- **Reality**
  - Traffic is inelastic
  - Most flows too short to see packet losses
  - Offered load driven by mouse clicks, not TCP congestion control
Informed Congestion Control

- **Key idea:**
  - Share network information among hosts
  - Make many small flows behave more like one large flow

- **What you do with it?**
  - Set initial RTT estimate
  - Set initial sending rate
  - Aggregate management of sending rate
  - Aggregate detection of congestion
Sample architecture

Congestion control gateway

<table>
<thead>
<tr>
<th>Network</th>
<th>RTT</th>
<th>Bttleneck capacity</th>
<th>Flows</th>
<th>Drop rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>washington.edu</td>
<td>50ms</td>
<td>4.0Mbps</td>
<td>7</td>
<td>0.02</td>
</tr>
<tr>
<td>mit.edu</td>
<td>30ms</td>
<td>8.6Mbps</td>
<td>5</td>
<td>0.00</td>
</tr>
<tr>
<td>whitehouse.gov</td>
<td>2ms</td>
<td>40Mbps</td>
<td>35</td>
<td>0.10</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ways to share information

- Hosts requests info (SPAND)
- Gateway piggybacks info on ACK (ECN)
- Gateway sends info out-of-band (IP SQ)
- Proxy TCP connections (I-TCP)
Potential for sharing information

- How frequently do flows share a network bottleneck?
- Conservative estimates
  - Destination host locality
  - Destination network locality
- Traces taken from UW and Harvard
Destination locality

Cumulative fraction vs Time (secs)

- UW: Networks
- UW: Hosts
- Harvard: Hosts
Adjusted destination locality
Conclusion

- TCP-style congestion control works poorly with short flows
  - End host starts knowing nothing
  - When host learns something, its all over
- There are many opportunities to share network performance information
## Changes in assumptions

<table>
<thead>
<tr>
<th>The old Internet</th>
<th>Today’s Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Few long-lived flows</td>
<td>Many short-lived flows</td>
</tr>
<tr>
<td>Low bandwidth*delay</td>
<td>High bandwidth*delay</td>
</tr>
<tr>
<td>Low drop rates, &lt; 1%</td>
<td>High drop rates, &gt; 5%</td>
</tr>
<tr>
<td>Homogeneous &amp; well-behaved traffic</td>
<td>Heterogeneous &amp; inelastic traffic</td>
</tr>
</tbody>
</table>
Network measurement stability

How predictive is a measurement?
- Balakrishnan et al, SIGMETRICS 97
  - Throughput has temporal & spatial stability
- Seshan et al, SPAND
  - Predicts throughput within factor of 2 (~70%)
- Paxson, SIGCOMM 97, PhD Thesis
  - Loss rate measurement stable for mins/hours
  - Throughput estimate good for mins/hours