

Transport Protocols: Bulk File Transfer

Stanislav Shalunov <shalunov@internet2.edu>

Google, 2005-06-02

Advanced Networking: Why Care?

- Advanced networking = fat pipes + new protocols
- Fast streams (fast, in 2005, starts at 100 Mb/s and goes to 10 Gb/s)
- Usually, not something encountered during web crawling or serving web pages
- Easily encountered during dataset replication, sharing data, etc.

Bulk Transfers

- Very important for research and education networks
- The killer application for high-performance networks so far
 - What else do we need fat pipes for?
- Several flavors:
 - straightforward huge file transfer
 - interactive high-throughput applications
 - instrument data transfer

Problem Exists Below Application

- Remains unsolved even in its most simple form (file transfer)
 - best current practice: open n TCP streams, send data
 - typical current practice: $n = 1$ (FTP, HTTP, etc.)
 - worst current practice: $n = 1$, application broken (SCP)
- Expected performance on Abilene (links are not congested):
~100 Mb/s
- Typical performance: less than 3Mb/s (Abilene)
- Source: Internet2 NetFlow Weekly Reports,
<http://netflow.internet2.edu/weekly/>
- The *wizard gap* gets wider

Wizard Gap

(With apologies to Matt Mathis for abusing his term.)

- wizard gap = $\log_{10} \frac{\text{performance}_{\text{wizard}}}{\text{performance}_{\text{regular user}}}$
- In 1993, the gap was 0.5.
- In 1997, the gap was 1.5.
- In 2001, the gap was 2.5.
- In 2005, the gap is 3.5.
- This is a bad law!
- (You also want to be a wizard.)

Top Reasons of Poor Performance (maybe 80% of cases)

- **Bad transport protocols** (layer 4)
- Ethernet duplex mismatch (layer 2)
- Bad last-hop cables (layer 1)

Bad last-hop cables (layer 1)

- Too twisted pair (stop rolling the chair)
- Dirty fiber
- Air gaps

Ethernet duplex mismatch (layer 2)

- One side thinks duplex is full
- The other side thinks it's half
- Problem most easily created while trying to cure or prevent it ("let me set that to full duplex")
- Not the topic; if interested, see "Detecting Duplex Mismatch on Ethernet," <http://www.pam2005.org/PDF/34310138.pdf>

Conventional TCP: Bad Transport

- Fundamental problems:
 - Unstable for high-speed networks (because it ignores most information given by the network: uses 0.00000000347 bits/packet if loss is 10^{-10})
 - Too sensitive to non-congestive packet loss (even after minor fixes): need 0.000000007% loss—less than 10^{-10} —to get 10Gb/s with 1500-B MTU and 100-ms RTT
 - Before a loss happens, buffers need to fill: delay is at least doubled
- Implementation problems
 - Buffers are laughably small:
 - * Normal default buffer sizes: 8 kB, 16 kB, 32 kB, 64 kB
 - * Even 64kB over 70ms limits throughput to 7.5Mb/s
 - * Good default would be: 8 MB with autotuning
 - No provisions for automatic buffer increases

Remedies for TCP's maladies

(In increasing order of invasiveness.)

1. Tuning: buffers, window scaling, timestamps, SACK
2. Use multiple streams
3. **Something else**
4. Replace the kernel and use a different congestion control
5. Replace all routers and kernels

Internet2 Bulk Transport Working Group

- <http://transport.internet2.edu/>
- A group of congestion control researchers and high-end users
- Started in late October 2004
- Goal: do better than conventional TCP
- Most immediate deliverable: a design space survey (almost done)

Transport tool

- High performance
- Completely end-to-end: no router modifications
- Suitable for both bulk file transfer and interactive multimedia
- Portable, easy to install and use (no kernel modifications)
- Advanced congestion control using existing research
- Tolerance for minor non-congestive packet loss
- Security

Design Space for the Tool

- Current version: transport-design-space-10.pdf
- Available from <http://transport.internet2.edu/>
- Specify requirements
- Document independent design questions
- Converge on a design

Design Space Dimensions

- Explicit signaling
- Implicit signaling
- Kernel- vs user-space
- Protocol features
- Window- vs rate-based
- TCP-compatible vs TCP-friendly
- State at sender vs receiver
- Single vs multiple streams
- UI and API

Current ideas about the design

- TCP-friendly, not TCP-compatible
- Security nonces
- Implicit congestion signaling
- Delay-based, with fallback to loss-based
- User-space tool with UDP
- State at receiver where possible
- API and a file transfer/distribution application

Open questions

- Telling a full queue from an empty one
- Noise filtering
 - for current delay
 - * exponentially decaying weighted average?
 - * median of last n ?
 - for minimum
 - * route changes
 - * 1st percentile?
 - * none?
- Controller selection

What's your take?

- Your feedback counts
- Mailing list: transport@internet2.edu
- Teleconferences: about every other Friday, 9AM Pacific Time

More...

- <http://transport.internet2.edu/>
- Join the mailing list: transport@internet2.edu
- Help us with the work

Contributors

Lawrence D. Dunn (Cisco), Yunhong Gu (UIC),
Steven Low (Caltech), Injong Rhee (NCSU),
Steven Senger (UW–La Cross),
Bartek Wydrowski (Caltech), Lisong Xu (UNL)

Questions?