Congestion Control for Interactive Real-Time Applications
Sanjeev Mehrotra, Jin Li

Introduction:

Real-time communication applications are used by millions of users on a daily basis.
Examples include VoIP, video conferencing, online meetings (such as WebEx, Live Meeting, Adobe Connect), tele-presence, online games, and virtual reality. They are extremely valuable across enterprise, small business and consumer markets.

The performance of real-time communication applications depends both on throughput as well as delay and packet loss. As an example, video conferencing generates perceivable delay degradation to humans if the end-to-end delay is larger than 200ms or the uncorrectable packet loss is larger than 1%. At the same time, these applications are also bandwidth intensive, as they want to boost the audio/visual quality of the content. Thus, a rate control protocol for real-time applications needs to maximize the throughput while operating at a low delay and low packet loss state (low operating congestion level). This differs significantly from rate control for non-interactive Internet traffic, such as file downloads and video on demand, whose dominant data transfer pattern is the transfer of large data chunks in one direction, which the user consumes only after the chunks are delivered. Though often perceived as a real-time application, video-on-demand (Internet TV) is actually *non-real* time in nature as it usually has a client buffer which is at least an order of magnitude larger than the round-trip time, which absorbs fluctuations in delay and packet loss.

Existing congestion control protocols:

There is a very large collection of literature on congestion control. They can be roughly classified in these three dimensions:

- Window-based congestion control (e.g., TCP variants) vs. rate-based congestion control (e.g., TCP-friendly rate control (TFRC) vs. available bandwidth estimation (ABE) techniques.
- Sender-side congestion control vs. receiver-side congestion control.
- Algorithms that directly set the transmission rate (TFRC and ABE) vs. those that dynamically adjust the rate up or down (TCP and other utility maximization based approaches).

For congestion control of real-time bandwidth intensive applications, the existing congestion control protocols have several issues.

- The operating congestion level in terms of queuing delay or packet loss may be too high for real-time applications (for example many TCP variants and TFRC respond to loss as congestion signal, and for networks such as Cable Modem or ADSL, it may not experience loss until the congestion level, and hence the delay, is very high)

- For noisy networks, such as wireless and mobile networks, the observed packet delay or packet loss may not due to congestion. Such non-congestion induced packet delay or loss may inject excessive false congestion signal (i.e., noise in congestion signal), and may result in overreaction of congestion control algorithm which leads to link underutilization
- Available bandwidth estimation techniques may suffer from poor performance on networks where packet gaps are highly variable. For example, on FIOS and Cable Modem, it may estimate the line rate, instead of the allocated bandwidth, which leads to overestimate of the available bandwidth, while on a noisy network, it may underestimate the available bandwidth.
- ISP rate limiting policies such as PowerBoost in Comcast may improve short-lived web downloads but hurt real-time application performance as congestion signals is temporarily inhibited during the PowerBoost period.
- PowerBoost type features can also improve performance of short-lived flows, but can hamper bandwidth estimation techniques due to dynamic capacity fluctuations.
- The scheduling behavior of cellular networks such as EV-DO, HSPA, LTE as well as users dynamically entering and leaving the network can also cause changing capacity, which lead to incorrect bandwidth estimation result.

Our position:

For interactive real-time applications, both delay and packet loss affects application performance. Moreover, it is possible to obtain additional congestion signal, e.g., the Explicit Congestion Notification (ECN) flag in certain restricted network situation, one-way delay trend. It is desirable for the congestion control protocol to respond to a mix of congestion signal, that is, to respond to delay congestion signal, to loss congestion signal, to ECN congestion signal, and/or to increase in one way delay trend [4, 5, 6].

It is also desirable for the congestion control protocol to achieve a fair share of the bottleneck link whenever there are multiple flows present in the network. Theoretical analysis has shown that that utility maximization based congestion control approach can result in lower operating congestion levels while achieve fair sharing of the network bandwidth [1, 2]. Such a property is desirable.

In addition, depending on the type of application, instead of using just a window based or rate based congestion control, a hybrid window and rate based congestion control which allows an application to achieve desired operating congestion levels may sometimes be needed [3].

References:

1. M. Chen, M. Ponec, S. Sengupta, J. Li, P. A. Chou, “Utility maximization in peer-to-peer systems”, ACM SIGMETRICS 2008, Annapolis, Maryland, Jun. 2-6, 2008.

2. M. Chen, S. Sengupta, M. Ponec, P. A. Chou, J. Li, “Peer-to-peer utility maximization”, CISS’08, Princeton, NJ, Mar. 2008.

3. S. Mehrotra, J. Li, S. Sengupta, M. Jain, S. Sen “Hybrid Window and Rate Based Congestion Control for Delay Sensitive Applications”, Proc. GLOBECOM 2010, Miami, FL, Dec. 2010.

4. S. Mehrotra, T. Wynn, J. Li, S. Sengupta, “CONGESTION CONTROL FOR DELAY SENSITIVE APPLICATIONS”, Patent application, USPTO application no. 20110216648, filed April 2010.

5. N.K. Srinivas, N. Abdo, S. Mehrotra, T. Wynn, “REMOTE PRESENTATION OVER LOSSY TRANSPORT WITH FORWARD ERROR CORRECTION”, USPTO application no. 20110219287, filed March 2010.

6. M. A. Attar, M. Sridharan, P. K. Patel, B. Prabhaker, D. A. Maltz, J. D. Padhye, and A. G. Greenberg, “COMMUNICATION TRANSPORT OPTIMIZED FOR DATA CENTER ENVIRONMENT”, USPTO application no. 20110211449, filed Feb. 26, 2010.