Fluid-flow modeling of rate control policies for streaming sources

Erling Austreim   Peder J. Emstad
austreim@q2s.ntnu.no   peder@q2s.ntnu.no

Centre of Quantifiable Quality of Service in Communiation Systems (Q2S)
Department of Telematics, Norwegian University of Science and Technology (NTNU)
Trondheim, Norway

Extended Abstract

Up to now the traffic volume in the Internet has mostly been TCP traffic due to browsing applications and file transfer. Streaming applications require bounds on delay and loss, and especially high bandwidths. Now when this bandwidth becomes available, such applications have become more popular. The use of TCP as transport protocol is inappropriate for this real time traffic since the protocol combines error control and congestion control. Most of these applications therefore use RTP on top of UDP as their transport protocol, which do not have any built-in mechanisms for congestion control, in contrast to the traditional TCP protocols.

In the case of congestion, the TCP sources will typically reduce their sending rates due to implicit feedback from the network. The streaming sources on the other hand will send with the same high rate, leading to a situation where sources which react to congestion in the network may get less than their fair share of the bandwidth. Generally the exact behaviour of an admitted session like a video stream is unknown, so rate control policies are needed in addition to admission control.

To ensure fairer bandwidth sharing and control with the quality of the streaming applications, one therefore may use mechanisms at higher layers which reduce the sending rates from the streaming sources in the case of network congestion. One way to do this is to use explicit feedback from the network (or the receiver) to control the rate from such sources. This feedback can be achieved by the use of ECN (Explicit Congestion Notification) or RTCP reports. ECN lets a congested node in the network signal back to the sources when congestion is detected while RTCP reports makes it possible for the receiver to transmit an estimate of available bandwidth to the sources.

Streaming audio and video are examples of popular streaming applications, which can make use of such explicit feedback. The rate can here be controlled directly by the encoder by dynamically adjusting the quantization parameters, or with use of scalable video encoding. Dynamically adjusting the quantization parameters makes it possible for the sources to adapt to a specified rate, if such rate feedback signals are available. A drawback with such direct rate control though is the lack of responsiveness. With the use of scalable encoding the media stream consists of a base stream and a number of enhancement streams, where the reception of the enhancement streams will increase the quality at the receiving end. The rate from the sources can here be shaped by regulating the number of enhancement layers sent. Another rate shaping approach used is server selective frame discard, where frames are dropped preemptively in an intelligent manner to minimize the distortion due to packet loss and make best possible use of the network resources. All these rate control mechanisms will reduce the decoded quality of the media at the receiver, but allows the sources to cooperate and adapt in the case of network congestion.

To avoid the effects the increasing amount of unresponsive traffic can have on fairness and congestion in the networks, rate control schemes for these types of traffic have to be analysed. Former literature have thoroughly analysed TCP congestion control mechanisms. Rate control schemes for non-TCP traffic have also attracted attention recently. Most of these allow for the rate from the sources to be TCP-friendly and are based on encoders capable of adapting to a specified rate.

We analyse some simple rate control policies for streaming sources based on feedback from a congested buffer. The analysis is based on a fluid-flow approximation of the traffic where the sources are modeled as simple ON-OFF sources. The rate control schemes applied use set points in the buffer of a congested node. Further, we study what information can be used to decide how often and which feedback information is sent back to the sources. This information can be based on the instantaneous queue size, or preferably some averaged observations of the queue size crossing the set points in the buffer. By using these averages some negative effects by applying rate control for streaming traffic can be reduced. Feedback decision based on the instantaneous queue size will easily lead to very much signaling which again will require the sources to change their target rate very often. This is hard to achieve and also undesirable because of the effects this will have on the receiving end. Rapid changes in the sources target rate will mean rapid quality fluctuations for the receiver of the streaming media.

The fluid-flow analysis carried out give some insight into what can be achieved by letting the sources sending rate be dependent on the load of one congested node by the use of some simple rate control policies. The loss probability, amount of signaling required and how these rate control policies affect the quality at the receiving end is studied.