Research on the QoS Control Technology Based on Terminal

Jing AN*, Lei ZHANG, An-qi LI and Fang DONG

No 23, Taiping Road, Beijing, China

*Corresponding author

Keywords: Congestion control, QoS.

Abstract. This paper studies the QoS control technology based on terminal, congestion control. Through the analysis and compare of the characteristics of various control method, it is emphatically introduced the control scheme of hybrid rate control which can promote the quality of multimedia data.

Introduction

The current QoS control technology can be roughly divided into two categories, based on network and terminal. Due to the QoS control technology based on network (such as routers and switches) is in need of network nodes, its technical implementation involves to the updated and expanded of network infrastructure, so it is limited in the practical application. And QoS control technology based on the terminal does not require the participation of network nodes, it is controlled by the application layer design for multimedia transmission, which can be compatible with the existing Internet, to achieve the required QoS requirements. So this paper mainly studies the QoS control technology based on terminal. The QoS control technology based on terminal mainly includes congestion control and error control. In the progress of the multimedia information transmission on network, each network node is choosed by the router, and packet must carry on the line. So there is packet transmission delay in the process, when the delay over a certain period of time, it will take the lost packet processing, after all of these affect the quality of multimedia data processing. So it is used congestion control to reduce latency and eliminate the packet loss, and it is used error control to improve quality in the case of packet loss recovery in the QoS control based on terminal. These two kinds of control technology was introduced in this paper.

Congestion Control: The QoS Control Technology Based on Terminal

According to statistics, in 1999 about 90% of the traffic on the Internet is through TCP transmission, the continuous improvement of the TCP protocol maintains the stability of the Internet. But with the increase of user access bandwidth and the development of VoIP, network audio and video applications, the UDP protocol transmission flow greatly increases, this caused great impact to the TCP traffic. Because of the network transmission is time-varying, if these UDP applications without the use of end-to-end congestion control or use non-standard rate control, it will produce the bad influence of two aspects: on the one hand, lead to network congestion, make the business quality fell sharply, even lead to collapse of the network; On the other hand, the application of using TCP protocol transmission constitutes unfair competition, it means bandwidth is stole from the TCP application to meet their own needs. Therefore, end-to-end congestion control become a inevitable requirement of multimedia information real-time transmission, that is, the application of multimedia real-time transmission must be able to change along with the network and dynamically adjust the transmission rate, and has the TCP friendliness.

At present there are two methods of congestion control. One is based on the window, it is an unit with the number of packets, by slowly increasing the congestion window and network bandwidth available for matching. When the network congestion is detected, the congestion window is quickly reduced to reduce and avoid network congestion, the TCP protocol is adopted the method. Another
is based on the rate, it uses the number of bits per second for a unit, estimates available bandwidth of the network at first, and then adjusts the data sent rate.

It as far as possible makes the transmission bandwidth is matched with the available bandwidth of the connection link, to reduce or avoid the occurrence of network congestion. There are some drawbacks in congestion control method based on window: it is easily to cause the packet of sudden, a window of the loss is not easy to recover, the retransmission delay is too long, etc., it does not apply to real-time multimedia information transmission. Therefore, the multimedia information real-time transmission uses congestion control method based on the rate. There are three kinds of solution in congestion control based on rate: based on the sender, based on the receiving terminal and based on the hybrid control.

The Rate Control Based on Sending Terminal

The rate control based on sending terminal adapts to the network by adjusting the sending rate of the sending terminal. If the sending rate and network bandwidth are matched, the packet loss rate can be greatly reduced. A feedback channel is usually needed, the network state information will be feedback to the sender that detected from the receiving terminal, the sending rate is adjusted by the sender according to network state information, to reduce the transmission of the packet loss rate. For unicast, according to the difference of the send rate sending strategy, there are two rate control methods, based on the test and based on the formula.

The Method Based on Test. In this method, a threshold of packet loss rate Pth is set at first, the packet loss rate P is detected in the receiver, and feedback to the sender through feedback channel, the corresponding strategy is used by the sender to adjust the sending rate. The sending rate is constantly adjusted by sender, to make the packet loss rate P that shows in feedback information is always less than the set threshold. There are two ways to adjust the sending rate: AIMD and MIMD.

Rate control algorithm based on AIMD can be described as follows:

\[
\text{if } (p \leq P_{th}) \quad r = \min\{(r + \text{AIR}, \text{MaxR}) \quad \text{else} \quad r = \max\{\alpha \times r, \text{MinR}\}
\]

(1)

P is the packet loss rate; P_{th} is the threshold value of congestion packet loss rate; r is sending the sending rate; AIR is the rate of increase factor; MaxR and MinR represent the minimum and maximum sending rate transmission speed; \(\alpha\) is the multiplication factor, and \(0 < \alpha < 1\).

To achieve the friendliness of the multimedia streaming and TCP flows, the sending rate set by sender should not be significantly more than the sending rate of TCP flow.

In the absence of congestion, multimedia streaming transmission speed is increased to achieve the ideal maximum rate. When congestion is occurred, the transmission speed is reduced quickly. At the same time, the continuous monitoring of packet loss and delay. This AIMD algorithm based on rate can achieve fair bandwidth sharing between different flow, as well as lower packet loss rate and higher bandwidth utilization.

The Method Based on Formula. In rate control based on the formula, to achieve the friendliness of the multimedia streaming and TCP flows, it is needed to limit the bandwidth of the multimedia stream. TCP flow rate formulas are directly used by a lot of rate control. A calculation formula has been put forward by Sally Floyd, in this formula, using TCP response function (see 2) as the sending rate adjusted by the sender.

\[
T = \frac{S}{R \sqrt{\frac{2p}{3} + t_{\text{retransmission}} \left(\frac{3p}{8} p(1 + 32p^2)\right)}}
\]

(2)

T is TCP transmission speed limit of the stable state, S is the size of the packet, R is the round-trip time, P is packet loss rate of the stable state, \(t_{\text{retransmission}}\) is TCP retransmission timeout.

In the rate control based on sending terminal, it is needed to receive the feedback information and then adjust the sending rate according to the feedback information. Feedback information will
reflect the time delay, jitter, packet loss, and so on directly or indirectly. It is mainly used for unicast, also be used for multicast.

**The Rate Control Based on Receiving Terminal**

In the rate control based on receiving terminal, the receiver adjust the received audio bit rate by increasing or decreasing channel. Normally, it is used for layered coding transmission of multicast. There are several layers in the media file. In the real-time transmission, a layer corresponds to a channel. For example, the basic layer corresponds to a multicast address, number of enhancement layer combination also corresponding to different multicast address. The receiver adjusts the number of receiving layers and sending rate indirect by selecting different multicast address. So the advantages of multicast and unicast can be better combined in this method.

Similar to rate control based on sending terminal, there are two kinds of method in receiving terminal, one is based on the test the other is based on formula. Based on the test method, usually has two situations: first, there is no congestion. In this situation, the receiver use available bandwidth to add a channel (layers), increasing the client receiving rate, then if congestion situation is not happened, it will continue increasing, otherwise, it will give up increased channel. Second, there is congestion. In this situation, reducing the rate by reducing a channel (layers). In the method based on formula, the rate control processing can also use the formula mentioned ahead, just like the sending terminal based on the formula method.

**Hybrid Rate Control**

In the hybrid rate control method, the sending rate is adjusted according to the feedback channel information by the sending terminal, at the same time, the receiver channel (layers) is increased or decreased by the receiving terminal.

**Principle**

In hybrid method, there is a two-way feedback channel between the sender and the receiver. The sender can use feedback channel to collect the receiver bandwidth demand information, and through the optimization algorithm to dynamically adjust the video coding rate of layers and different levels, at the same time, it can also through the feedforward channel to inform the current adjutive results to the receiver in broadcast way. As the adjustive basis, the receiving terminal can decide to increase or decrease the keeping video layers, according to its own effective receive bandwidth after receiving the message from the sender. Compared with the method based on the sender or the receiver, the multichannel is used by hybrid method, and each channel transmission rate is not fixed, so it can be adjusted according to the condition of network congestion.

**The Algorithm of Sending Terminal**

In order to describe, constraint conditions as follows,

1. The cumulative layered method is used by the sender, the maximum level classification constraints is L. The first layer as the basic layer, next is the enhancement layer in turn.
2. \( b_i \) correspond to each bit rate, in which \( i = 1, 2, ..., L \). Bit rate constraint of every level is \( b_i \subseteq \beta_{\text{max}},...,\beta_{\text{max}} \), they are a limited number of discrete values. If \( c_i = \sum_{j=1}^{i} b_j \), \( i = 1, 2, ..., i \leq L \), \( c_i \) is the sum of bit rate from the first layer to the i layer. \( \beta_i = c_i, ..., c_i \) is defined as cumulative rate vector, \( i \leq L \), in which \( c_i = b_i \) is the lower limit of basic layer bit rate.
3. If a number of users to join multicast session is N, the receiving rate collection is \( R_i, R_{i1}, ..., R_{iN} \), fair index definition of receiving terminal is \( \beta_{\text{r}} \), \( \beta = \Psi_{\text{r}} \), \( i \leq N \), in which \( \Psi_{\text{r}}, \beta = \max c_i : c \leq R_i, c \in c_i, c_i, ..., c_i \), it is the current effective receiving bit rate for the receiver i. Obviously, \( J_i, \beta_i \leq J_1 \). Set \( J_i, \beta_i = \sum_{J_i} \), \( \beta_i \) is general receive fair index of N receivers.

According to the above setting, QoS control of sending terminal can be defined as: in the constraint of condition (1) and (2), how to determine the current maximum layers \( J_i \) and layered accumulation rate vector \( \beta_i \), make \( J_i, \beta_i \) to obtain the maximum value.
Algorithm description as follows:

(1) Collecting the feedback message of RR, the current collection \( r_1, r_2, r_3, \ldots, r_N \) is obtained by the receiver, and the general receive fair index is calculated under the the current layers \( l \) and layered accumulation rate vector \( \beta_l = c_1, c_2, \ldots, c_l \).

(2) According to the constraint conditions (1) and (2), it is combined all of ri that more than \( \Sigma L_i = b_{\max} \) into one number that equals to \( \Sigma L_i = b_{\max} \), and deleted all of ri that less than \( b_{\min} \), that are not included in the calculation of the average equity index. In this way, the number of effective rate collection (duplicates) by the receiver actual remaining \( N' = N - m - n + 1 \), in which \( m \) is the item number of ri that more than \( \Sigma L_i = b_{\max} \), \( n \) is the number of ri that less than \( b_{\min} \). The rate collection expect to receive of the receiving terminal after preprocessing is \( r_1', r_2', r_3', \ldots, r_N' \).

(3) Formula \( J_r, \beta_l = \Sigma J_{r_i}, \beta_l \) was revised to \( J_r', \beta_l' = \Sigma J_{r_i}', \beta_l' \), calculation is done by data inserting, to search for \( l' \) and \( \beta_l' \) that get maximum of the overall equity index. Define tolerance adjustment \( \epsilon \), when \( J_r', \beta_l' - J_r, \beta_l > \epsilon \), the corresponding optimization adjustment is done by encoder according to the parameters \( l' \) and \( \beta_l' \), on the other hand, the current status is maintained to ensure the relative stability of the system.

(4) Return to the first step. In the implementation of above process, the sender send a SR message that about the current number of video layered and a cumulative rate vector to the receiver in every TSR seconds by multicast. In which TSR= TAdj/k, k is integer that greater than 1, TAdj is cycle of the algorithms carrying out for the sender, also known as the sending terminal adjustment cycle. The choose of TAdj is not too small, otherwise it will increase the computing burden of sender and lead to an adjustment of oscillation. It is not conducive to transmission and receiving of real-time video data, at the same time, it also can take difficult to get a correct collection and statistic rate report. Analysis shows that TAdj values in 10 ~ 15s is reasonable.

The Algorithm of Receiving Terminal

(1) To estimate the parameters \( p, t_{RTT} \) and \( t_{max} \).

(2) If the message SR is received, to the next step, otherwise returns to the previous step.

(3) Get current \( \beta_l \) of sending terminal by analysing the SR message, (1) is used to calculate the expecitive receive rate \( r \) of receiver.

(4) According to \( \beta_l \) and the expecitive receive rate \( r \) to adjust the number of booking video layer.

(5) Return to the first step.

In the implementation of the above steps, the receiver sends a RR message that about the expecitive receiving rate in every \( T_{re} \) second to the sender. At the same time, the message is also used as a request to estimate \( t_{RTT} \).

Regardless of the control strategy is based on the screen or the rate, and the speed adjustment is based on the sending terminal or receiving terminal, the adjusted rate is finally decided by the bandwidth, processing, display and other factors of the receiver. If these factors suggest that part of the send data by sender is redundant and useless, then the sender should send less useless data to avoid the waste of resources. The coding support is needed by congestion control strategy. For multi-rate encoding, the control strategy is realized by the dynamic choosing of different video stream rate. For scalable encoding, the realization of the control strategy is completed by real-time adjusting of layers. If divisional services and integrated services are combined, the effect of the congestion control strategy is more ideal.

References
