feedback from the network (or the receiver) to control the rate from such sources [1]. This feedback can be achieved by the use of ECN (Explicit Congestion Notification) [2] or RTCP reports [3]. 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 [4], or with use of scalable video encoding [5]. 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 [6]. 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 [6]. 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. In [7] a steady state analysis of the TCP congestion avoidance is carried out. This analysis is extended in [8] where TCP's fast retransmit mechanism and the effect of timeouts are taken into account. Another approach is taken in [9] where a fluid-based analysis is used to describe the behaviour of long-lived TCP flows over networks of IP routers adopting a RED (Random Early Detection) [10] AQM (Active Queue Management) scheme with one bottleneck link. In [11] the authors extend their model to consider multi-bottleneck links, while in [12] both short-lived and long-lived TCP flows are represented using a similar fluid-based model.

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 [13-18].

Our analysis is based on some simple rate control policies for streaming sources based on feedback from a congested buffer, first studied in [19]. These studies are based on fluid-flow modeling and discrete time Markov chains (DTMC). The DTMC approach makes it possible to take the delay in the feedback loop into account, but are only applicable to simple systems. More complex systems will lead to state explosion. In this paper we focus on a fluid-flow model of such systems and are interested in studying these schemes more deeply by investigating how realistic these schemes are in the sense of amount of signalling required and how this rate control will affect the quality at the receiving end.

In part two we present the model used in [19] to analyse the control policies based on feedback from the network. In part three a level crossing analysis is carried out to study how this control scheme will affect the dynamics in the queue. Some results from these analyses are presented in part four, and part five concludes the paper.

For a system without any rate control the sending rate will only depend on the number of active sources and the behaviour of the queue can be described by the following set of differential equations [20]:

$$(i \cdot h - c) \cdot \frac{d}{dx} F_i(x) = F_{i-1}(x) \cdot p(i-1, i) + F_{i+1}(x) \cdot p(i+1, i) - F_i(x) \cdot p^*(i) \quad (1)$$

$$, 0 \le i \le n, 0 < x < m$$

where $F_i(x) = P(X(t) \le x, I(t) = i)$ as $t \to \infty$.

These equations can be rewritten in matrix notation $\underset{\sim}{D} \frac{d}{dx} \underset{\sim}{F}(x) = \underset{\sim}{M} \underset{\sim}{F}(x)$ which gives the solution:

$$\underset{\sim}{F}(x) = \sum_k a_k \cdot e^{z_k \cdot x} \cdot \underset{\sim}{\varphi}_k \quad (2)$$

where φ_k and z_k denote the eigenvectors and eigenvalues of $\underset{\sim}{D}^{-1} \underset{\sim}{M}$. The constants a_k are determined by looking at the boundary conditions:

- For all states where $i \cdot h > c$ the queue is always increasing, so the queue length cannot be zero $\Rightarrow F_i(0) = 0$
- For all states where $i \cdot h < c$ the queue is always decreasing, so the queue length cannot be on its limit $\Rightarrow F_i(m^-) = P(I = i)$

A system with a step wise rate reduction (figure 2a) can be solved as for the general model, but here we have to look at each rate level separately. For each level we then get a system of differential equations describing the system in the same form as in (1), with h replaced by $r(x) = h \cdot k(x)$. If we look at level j, the solution is found as in the general model by:

$$\underset{\sim}{F}_j(x) = \sum_k a_{k,j} \cdot e^{z_{k,j} \cdot x} \cdot \underset{\sim}{\varphi}_{k,j} \quad (3)$$

which is only valid in the interval where *k(x)* is constant. For each level we have to determine the constants $a_{k,j}$. These are found by looking at the following conditions:

- The state probabilities are continuous and we then have:

  $F_{i,j}(q_j) = F_{i,j+1}(q_j) \quad , j = 1, 2, \ldots j_{max} - 1$
- In level 1: for all states where $i \cdot r(x) > c$ the queue is always increasing, so the queue length cannot be zero $\Rightarrow F_{i,1}(0) = 0$
- In the last level: for all states where $i \cdot r(x) < c$ the queue is always decreasing, so the queue length cannot be on its limit $\Rightarrow F_{i,j_{max}}(m^-) = P(I = i)$

Based on the results for $\underset{\sim}{F}(x)$ in (2) and (3), we can now find the cumulative queue size distribution, and the loss probability in the buffer as the amount of data arriving when it is full:

If we look at the queue size distribution at a small area we can find the total crossing rate $\eta(x)$ at x by using Little's formula:

$$\eta(x) = \eta_0(x) + \eta_1(x) = \frac{E(\text{numbers in } [x, x+dx])}{E(\text{time in } [x, x+dx])} = \frac{\sum_i f_i(x) \cdot dx}{\sum_i \frac{P(I=i) \cdot dx}{|c - r_i|}} = \frac{\sum_i f_i(x)}{\sum_i \frac{P(I=i)}{|c - r_i|}} \quad (6)$$

where $f_i(x) = \frac{dF_i(x)}{dx}$.

In general the set points for rate decrease and increase can be different, where the latter is the smallest of the two. This has not been studied explicitly in this paper. However, we can in this scheme distinguish between the two types of information sent back to the sources, that is ECI (Explicit Congestion Indication: reduce the rate when the queue size crosses a set point for decrease upwards) and ECR (Explicit Congestion Reduced: increase the rate when the queue size crosses a set point for increase downwards). The rate of ECIs originating from a set point at y can then be found as:

$$\eta_1(y) = \eta(y) \cdot \frac{\sum_{i, r_i > c} f_i(y)}{\sum_i f_i(y)} = \frac{\sum_{i, r_i > c} f_i(y)}{\sum_i \frac{P(I=i)}{|c - r_i|}} \quad (7)$$

Similar, the rate of ECRs originating from a set point at x can be found as:

$$\eta_0(x) = \eta(x) \cdot \frac{\sum_{i, r_i < c} f_i(x)}{\sum_i f_i(x)} = \frac{\sum_{i, r_i < c} f_i(x)}{\sum_i \frac{P(I=i)}{|c - r_i|}} \quad (8)$$

Given that the set points are set at x_i for increase and y_i for decrease, the mean time between signalling messages is given by:

$$E(T_{signal}) = \frac{1}{\sum_{x_i} \eta_0(x_i) + \sum_{y_i} \eta_1(y_i)} \quad (9)$$

Needless to say, each source need to be notified. This time will be equal to the mean time between rate changes at the sources. In the following we assume there is just one set point (the same set point for rate decrease and increase).

3.2. Measure on Quality at Receiving End

The frequency of rate changes at the sources, given by the inverse of $E(T_{signal})$ found above, will clearly affect the quality of the streaming media. To study how the scheme with step wise rate reduction will affect the receiving end we are also interested in the time the queue will be

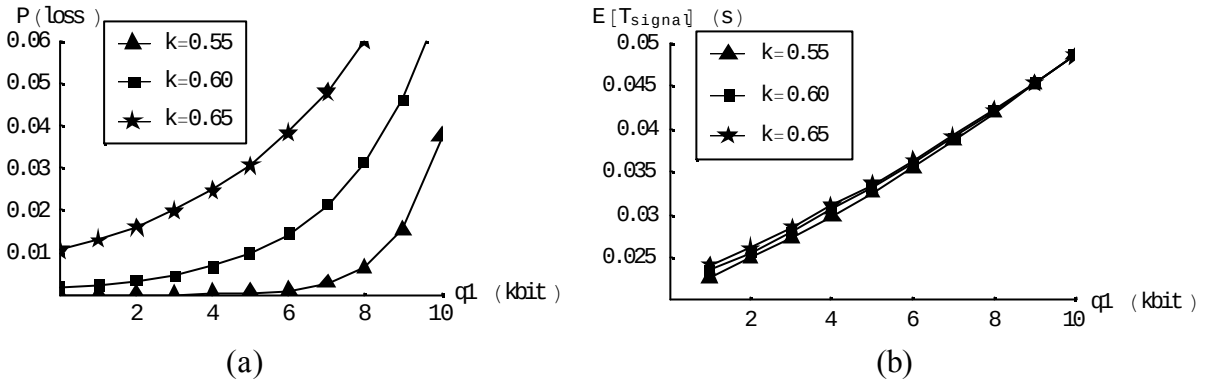

Figure 5: Loss probability (a) and mean time between signalling messages (b) as a function of the set point q_1 where the rate is changed for different rate reductions k.

queue size distribution for a case without rate control, with the distribution for a step wise rate reduction similar to the one shown in figure 2a, but with two rate levels. The analysis was carried out with traffic from one source, and the parameters as shown in table 1.

By using this step wise rate reduction with two levels the sources only need to adapt to two rate levels, and ECN can be used for signalling. This will require the sources to use their lower rate when congestion is detected. The two rate levels can be realized by a base layer and one enhancement layer in the case of scalable encoding. The rate is then reduced by cutting off the enhancement layer. Given the two rate levels the sources should adapt to, and the requirements on the loss probability, an appropriate value of the set point q_1 where the rate is changed has to be chosen. With the parameters listed in table 1 we have looked at how the loss probability is affected by the value of q_1 and by how much the rate is reduced (figure 5a). From this curve the value of q_1 can be found when the acceptable loss level is given. This makes it possible to give statistical guarantees for the loss in the node.

In the step wise shceme a signalling message is generated each time the buffer content crosses the set point. We have studied how the rate of signalling meassages is affected by the choice of this set point q_1 (see figure 5b). This gives better understanding of how the set point can be chosen such that an acceptable loss level is met with a reasonable amount of signalling. The mean time between signalling messages also gives a good indication on how large the delay in

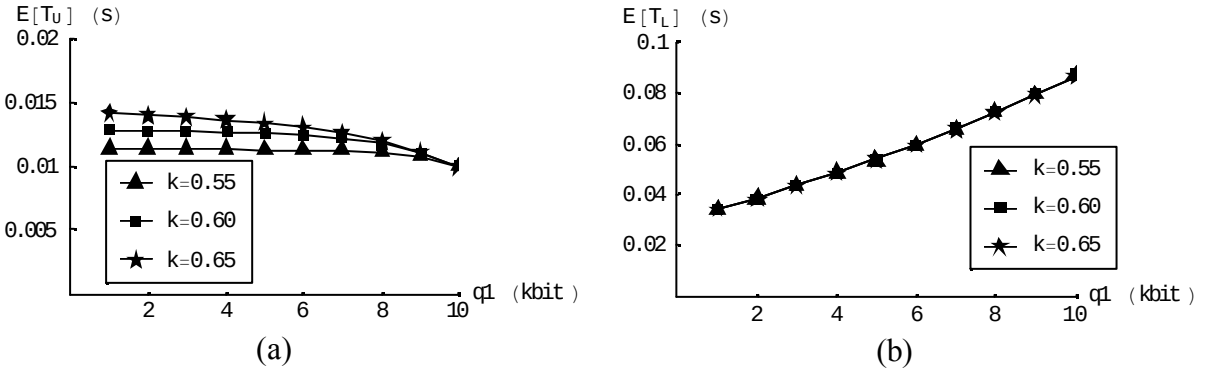

Figure 6: Mean times in upper (a) and lower (b) parts of the queue as a function of the set point q_1 for different rate reductions k.