

# Relating Flow Level Requirements to DiffServ Packet Level Mechanisms<sup>1</sup>

Eeva Nyberg, Samuli Aalto, Jorma Virtamo<sup>2</sup>

Helsinki University of Technology

## Abstract

We present a two level modeling approach for DiffServ mechanisms. We first sketch a flow model to capture the user requirements of differentiation and to compare differentiation to traditional best-effort TCP flow level results. We then study how the bandwidth allocations presented on the flow level can be modeled and achieved on the packet level. The DiffServ mechanism we use is the Simple Integrated Media Access (SIMA) proposal. We choose SIMA, as it employs both financial and end-to-end congestion control incentives, and is thus both a flow level and packet level mechanism. As a result of our two level modeling, we show that the differentiation achieved by SIMA is within fixed bounds predetermined by the acquired nominal bit rate of a flow. We further enhance our model to show, how the SIMA priority mechanism is also able to protect TCP flows from bandwidth exhaustion by non-TCP flows.

Keywords: Internet, TCP, best-effort, DiffServ, SIMA, pricing.

## 1 Introduction

As the history and literature on Quality of Service (QoS) shows, there is a natural tendency to perceive differentiation requirements on the flow level rather than on the packet level. However, the future Internet and its architects face the challenge of how to offer flow level quality requirements by means of packet level mechanisms. The hard QoS guarantees of Integrated Services (IntServ) and Resource Reservation Protocol (RSVP) [1] tried to bypass this question by introducing flow control and flow level scheduling. The solution was not scalable, and packet level mechanisms through the introduction of Differentiated Service (DiffServ) [2] were once again considered.

The Internet research still lacks efforts in coupling the packet level QoS mechanisms of DiffServ, e.g. Assured Forwarding (AF) [3] and Expedited Forwarding (EF) [4], to flow level analysis. On the other hand, flow level bandwidth allocation and fairness research, e.g. [5], [6], and [7], continue to assume that weighted fair bandwidth allocations between different service classes are somehow achieved and evade the question of how to do so without flow control or per flow scheduling.

Our paper attempts to bridge the gap between packet level mechanisms and flow level requirements. Of the proposed DiffServ mechanisms, we choose one that has received less attention than AF or EF, the Dynamic RT/NRT and related Simple Integrated Media Access (SIMA) proposal introduced in [8] and [9]. We choose SIMA for two reasons. First, it relies both on financial and end-to-end congestion control incentives to achieve differentiated bandwidth allocation. Both incentives are flow level mechanisms and have thus been modeled and studied on the flow level as a means to control elastic traffic. Furthermore, these incentives to control congestion, [10] and [11], are widely accepted as foundations for the future Internet.

---

<sup>1</sup>Submitted for publication

<sup>2</sup>E-mail: {eeva.nyberg, samuli.aalto, jorma.virtamo}@hut.fi

Secondly, on the packet level, SIMA relies on simple FIFO queuing and threshold mechanisms to achieve differentiation. SIMA is thus composed of both flow level and packet level mechanisms that have separately been modeled, and is a natural choice for our two level modeling.

The functionality of SIMA with TCP flows has been studied with the help of simulations in [12] and using a test network in [13], while our contribution is an analytical flow and packet model. The results obtained in our analysis and earlier studies are alike, thus also validating the use of simple analytical models in bringing insight to DiffServ performance.

We first present and compare the flow level and packet level model of SIMA for elastic flows, and then show how the SIMA mechanism is able to protect TCP flows from bandwidth exhaustion by non-TCP flows. We thus show how fair differentiation is achieved between elastic flows, and how the same differentiation mechanism differentiates TCP flows from non-TCP flows in a fair manner.

We will first, in section 2, model on the flow level how the SIMA priority levels are determined on the basis of price and rate sent and what kind of bandwidth allocation this mechanism amounts to. We will show that in a single link network, with ideal greedy TCP sources, the bandwidth allocation varies from equal allocation to allocation in proportion to price, depending on the congestion of the network and the number of priority levels used. Section 3 concerning the modeling and analysis of packet level mechanisms is divided into two parts. In section 3.1, the SIMA packet level model is first introduced for one buffer to show the similarities of the bandwidth allocation between the ideal flow model and the packet level. In section 3.2 we demonstrate how the SIMA specification with two buffers, one for each delay class, with coupled threshold mechanisms are enough to separate TCP flows from non-TCP flows. Section 4 concludes the paper. The SIMA concept is discussed in the Appendix.

## 2 Flow level bandwidth allocation: multiple TCP service classes

In this section, our main focus is to show how multiple TCP classes can be constructed. The different needs of elastic flows are taken into account by allowing equal bandwidth division inside a class, but different ratios between classes. The traditional best-effort TCP class is extended to two or three service classes and studied using a simple game theoretic approach on the flow level. The model is similar to traditional fairness models for elastic traffic, e.g. [5] and [6], in the sense that the time scale is such that the number of flows is constant with the average bit rate of a flow determined by the bandwidth allocation. In [7] bandwidth allocation was studied for a stochastic arrival model of connections. Our model differs from the fairness models as it does not assume fixed weights between different classes, but studies how these weights could be achieved without per flow scheduling or explicit rate flow control.

In flow level fairness research, TCP is ideally thought of as resulting in bandwidth allocation which is max-min fair. Loosely speaking, an allocation is max-min fair if the minimum bandwidth allocation is maximized subject to capacity constraints. In [14], it was, however, shown that TCP in a network tends to realize proportional fairness rather than max-min fairness. Proportional fairness penalizes long routes in order to improve overall throughput, whereas max-min fairness treats routes more equally.

However, in the case of a network with a single link, both fairness criteria result in bandwidth equally shared between competing TCP flows, often modeled as a processor sharing mechanism. We will, in this paper, assume a single link network, where bandwidth is divided equally inside a TCP class, and study how the weights of weighted fairness are achieved across TCP classes.

On the flow level, the DiffServ proposal that we consider, SIMA, is based on two key ideas. First, at the access node, the conditioner determines the priority levels of flows based on the ratio of the actual bit rate sent to a default sending rate purchased. The default sending rate is called the nominal bit rate (NBR) and with the NBR a price may be associated. Secondly, a flow is classified into two delay classes, the real-time (rt) and non-real time (nrt) classes, with packets handled in separate buffers for both classes. The states of both buffers together determine which priority classes should be accepted to the system through the use of a threshold function in the core node.

## 2.1 Flow model

Assume greedy TCP sources that tend to maximize the minimum bandwidth allocation and a network enhanced with the conditioner of SIMA, where flows are classified into priority classes depending on the ratio of their sending rate to their subscribed nominal bit rate. See appendix or [8] for more details on SIMA. The more traffic flows send, the worse the priority. In such a setting the TCP flows tend to maximize the sending rate, but the conditioner and priority level allocation may penalize an increase in sending rate by a drop in priority level. The priority mechanism then allocates less bandwidth to flows in lower priorities. We thus aim at studying in what magnitude the NBR of flows, determining the priority level, affect the bandwidth allocation.

Assume that flows are divided into groups  $l \in \mathcal{L} = \{1, \dots, L\}$  according to the NBR purchased. Each group consists of  $n_l$  identical flows. Flows are classified to priority level  $i \in \mathcal{I} = \{1, \dots, I\}$  using the SIMA conditioner. In this section, we denote the actual bit rate of a flow in group  $l$  by  $\beta_l$  and the NBR purchased by flows in group  $l$  by  $nbr_l$ .

### 2.1.1 Network with two groups of flows

Let us consider a single link network, with the link capacity scaled to one. The network is used by two NBR groups,  $L = 2$ , and they have a choice between two priority levels,  $I = 2$ . Our goal is to study how weighted fairness with weights proportional to NBRs is achieved in such a network.

Group 2 flows have acquired  $k$  times more NBR than group 1 flows

$$nbr_2 = k \cdot nbr_1.$$

Flows in group  $l = 1$  are in priority class  $i = 2$  if their actual bit rate is less than

$$b = a \cdot nbr_1.$$

Flows in group  $l = 2$  are allowed to send  $k$  times more traffic until classified to the lower priority class. Thus, the boundary rate for them is

$$kb = a \cdot nbr_2 = k \cdot a \cdot nbr_1.$$

Table 1 depicts the resulting conditions.

Table 1: Priority class assignment conditions

	$i = 1$	$i = 2$
$l = 1$	$\beta_1 \geq b$	$\beta_1 < b$
$l = 2$	$\beta_2 \geq kb$	$\beta_2 < kb$

Following the ideal TCP model, we assume that in the same priority class bandwidth is divided equally between flows. In addition, we assume that the higher priority class,  $i = 2$ , has strict priority over the lower class,  $i = 1$ . This does not cause a flow of the higher class to use up all the bandwidth and starve the flows of the lower class, as the assignment of priority classes depends on the bit rate sent by the flows. If a flow in the higher priority class increases its bit rate above the boundary  $b$  or  $kb$  it falls to the lower priority class, and divides capacity equally with the other flows in the lower class. The coupling of sending rate and assignment of priority class prevents starvation of bandwidth typical to normal priority mechanisms. The inclusion of price and NBR, on the other hand, gives the flows the right to control the boundary at which point the priority class changes.

Flows from both NBR groups may be in the same priority class or they may be in different priority classes, altogether four different states that our system may be in. When the flows are in different priority

classes, the flows in the higher priority class each receive equal amount of bandwidth and flows in the lower priority class divide equally, among themselves, the remaining bandwidth. Table 2 summarizes the above conditions, where  $\beta_{l,i}$  is the actual bit rate of flow in NBR group  $l$  and in priority class  $i$  and  $n_{l,i}$  is the number of flows in the corresponding state.

Table 2: Bandwidth allocation for NBR groups

$\beta_{l,i}$	$i = 1$	$i = 2$
$l = 1$	$\max(\frac{1-n_{12}\beta_{12}-n_{22}\beta_{22}}{n_{11}+n_{21}}, 0)$	$\min(\frac{1}{n_{12}+n_{22}}, b)$
$l = 2$	$\beta_{11}$	$\min(\max(\frac{1}{n_{12}+n_{22}}, \frac{1-n_{12}b}{n_{22}}), kb)$

Assume further that while the flows inside a NBR group are by all means identical, they also behave in the same manner, i.e. inside a group all flows belong to the same priority class. Therefore we have  $n_{l,i} = n_l$  when the flows are in priority class  $i$  and zero otherwise.

For the four different states possible for the link, the actual bit rates  $\beta_l$  received by the flows in NBR group  $l$  are shown in table 3.

Table 3: Bandwidth allocation in the four possible network states

$l = 2$	$l = 1$	
	$i = 1$	$i = 2$
$i = 1$	$\beta_1 = \max(\frac{1}{n_1+n_2}, 0)$ $\beta_2 = \max(\frac{1}{n_1+n_2}, 0)$	$\beta_1 = \min(\frac{1}{n_1}, b)$ $\beta_2 = \max(\frac{1-n_1\beta_1}{n_2}, 0)$
$i = 2$	$\beta_1 = \max(\frac{1-n_2\beta_2}{n_1}, 0)$ $\beta_2 = \min(\frac{1}{n_2}, kb)$	$\beta_1 = \min(\frac{1}{n_1+n_2}, b)$ $\beta_2 = \min(\max(\frac{1}{n_1+n_2}, \frac{1-n_1b}{n_2}), kb)$

If each group of flows individually optimizes its actual bit rate  $\beta$ , the following four scenarios can be observed as a function of the number of flows  $n_1$  and  $n_2$ .

1. In times of low load, when the following condition holds

$$n_1kb + n_2kb < 1,$$

there is no advantage in being in the higher priority class, where the bit rate would be limited. For both groups, the bit rate achieved in the lower priority class is more than the bit rate in the higher priority class. Thus all flows are in the lowest priority class sharing equally the bandwidth of the link.

2. As the number of flows increases the low load condition does not hold

$$n_1kb + n_2kb > 1,$$

and there is not enough bandwidth for all flows to send at rate  $kb$ . If, however, the condition

$$n_1b + n_2kb < 1$$

holds, then it follows that  $1/n_2 > kb$ . Thus flows in group  $l = 2$  move up to priority class  $i = 2$ , as there they always get the boundary rate  $kb$ , more than if they stayed in the lower priority class. As a result flows in group  $l = 1$  stay in priority class  $i = 1$ , as they can continue sending at rate higher than their boundary rate  $b$ . Moving up a priority class would require them to reduce their sending rate to or below  $b$ .

3. As the number of flows continues to increase and the link is further congested the previous condition does not hold, and

$$n_1b + n_2kb > 1.$$

However, the load is still such that

$$n_1b + n_2b < 1.$$

Now flows in group 1 move up to priority  $i = 2$  as otherwise their sending rate would be reduced to less than  $b$ . The flows in group 2 are, however, still sending more than  $b$ , as there is still enough bandwidth in the link. Therefore, though the two groups of flows are in the same priority class, the bandwidth is not divided equally between the two classes.

4. As the congestion deepens, and more flows are introduced to the link, the flows in group 2 have to also reduce their sending rate, and when condition

$$n_1b + n_2b > 1$$

holds, all flows are in the highest priority class with equal bit rates of less than  $b$  units.

The actual bit rates of each flow as a function of the number of flows is therefore equal when

$$\begin{aligned} n_1kb + n_2kb &< 1, \text{ or} \\ n_1b + n_2b &> 1. \end{aligned}$$

These correspond to times of low load, when there is no need to differentiate between flows, as there is enough bandwidth for everyone, and times of very high load, which as a condition should be very rare. In all other cases there is a difference in the bandwidth received by the flows in group  $l = 1$  and flows in group  $l = 2$ , the ratio of bandwidth being at most equal to the nominal bit rate ratio  $k$ . The differences occur in the middle area, in cases of moderate congestion.

### 2.1.2 Numerical results

The bandwidth allocation, as a function of flows in two NBR classes, in a system with two priority classes, with  $(nbr_1, nbr_2) = (0.04, 0.08)$  and  $(nbr_1, nbr_2) = (0.02, 0.08)$ , i.e. NBR ratio  $k = 2$  and  $k = 4$  respectively, is depicted in figure 1. The number of flows of group 1 and 2 are on the  $x$ - and  $y$ -axis respectively. The black areas correspond to equal bandwidth allocation, while the white areas correspond to bandwidth allocation equal to  $k$ . The figure clearly shows that the middle area, where differentiation is achieved, increases as  $k$  increases.

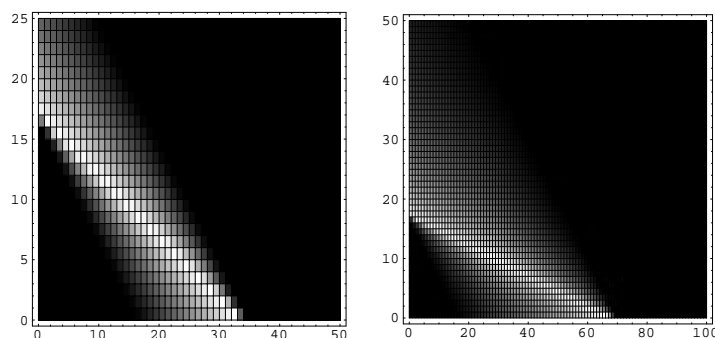


Figure 1: Bandwidth allocation for flows with  $k = 2$  and  $k = 4$  as function of number of flows,  $n_1$  ( $x$ -axis) and  $n_2$  ( $y$ -axis).

We have also studied the model for three priority classes. As the number of priority classes i.e. differentiation classes is increased the middle area becomes larger and only rarely do all flows receive the same

bandwidth. The middle area grows also as the ratio of NBR,  $k$ , grows. Thus with many priority classes and large differences in NBRs, differentiation is at its best.

We have also extended the simple flow model when one of the NBR groups is composed of non-TCP flows. We will, however, not consider the results here, but present them in section 3.2 in terms of the packet model.

The key point of the model and numerical examples of this section is to demonstrate how differentiation should be done among classes of TCP traffic. As a result the model shows how fair differentiation respecting the cooperative nature of the Internet can be brought to the network with simple priority mechanisms. If the TCP friendliness advocated in [10] becomes a reality in the Internet, all flows will behave in the way sketched above and non-TCP flows need not even be taken into account.

We have demonstrated how, in a single link network, the bandwidth share received by competing TCP flows depends on the NBR purchased by the flows. We assumed a fixed number of flows and studied how the weighted bandwidth allocation depended on the congestion of the link. The SIMA mechanism did not allocate bandwidth in proportion to fixed weights, rather the weights depended on the state of the network. The share of bandwidth was, however, at least equal and at most in proportion to the fixed weights determined by the NBR or price paid. Any allocation in proportion to weights could be achieved by per flow scheduling. The challenge, however, is to design such packet level mechanisms that achieve the required bandwidth allocation without per flow scheduling. This is the topic of the next section.

### 3 Zooming in on the packet level mechanisms

To achieve the above flow level performance by packet level mechanisms requires a conditioner, which marks the packets of a flow to the priority level according to the bit rate sent and the nominal bit rate purchased. A scheduling mechanism, where flows in higher priority classes are treated superior to flows in lower classes, is also needed. The FIFO queue with thresholds is such a queuing discipline, where the threshold tightens as the priority level decreases. As the number of priority classes increases, it is more reasonable to increase the number of thresholds in the buffer than to introduce more queues, one for each priority class.

Assured Forwarding (AF) would have the required buffer mechanism, with the drop precedence class being the priority classes, but it lacks a clear conditioner mechanism at the flow level relating the price and the bit rate sent of a flow to the priority class.

Pricing mechanisms would be another natural choice. In the Smart Market pricing [11] and other usage based pricing schemes, e.g. congestion pricing, price is determined by a real-time Vickery auction. In essence, pricing mechanisms attempt to couple the congestion of the network with a price, giving the users and flows the incentive to adjust their bandwidth accordingly. This is, however, not the same as the priority mechanism used in the flow model.

Recall that in the flow model the ratio between the nominal bit rate groups was constant, and thus the price differentiation between the groups remained fixed throughout the scenario. This kind of capacity based pricing [15] is more than the flat rate pricing of the best-effort Internet, but less complicated than the continuous usage based pricing scheme of congestion pricing. Introducing continuous usage based pricing is not a restoration of the traditional Internet, based on flat rate pricing and simple control, but an attempt to renew congestion control to rely heavily on financial incentives and continuous control.

The main advantage of the SIMA proposal is that bandwidth allocation is based on both financial incentives through the NBR as well as on network incentives through end-to-end congestion control.

The next section introduces our packet level model of SIMA which relies on TCP and packet loss feedback for congestion control, enhances it by introducing discrete pricing groups and takes into account the different delay requirements of traffic in the future Internet. Furthermore, we show how the SIMA packet level mechanism achieves similar weighted fairness in bandwidth allocation than our ideal flow model of section 2.

#### 3.1 Model of a simple SIMA packet mechanism

In this section, we present our packet level model of SIMA. The threshold mechanisms and the scheduling unit are included in the buffer model, while the rate adjustment is modeled as a TCP mechanism. These

two sub models are related through a feedback equation and the conditioner assigning priority levels. The setting is the same as in section 2. We have a single link network with  $L$  NBR groups, a fixed number of flows, and  $I$  priority classes. We then study, using the packet model, to what bandwidth allocation the system converges. As in section 2 we study the interdependency between the NBR and the allocation in terms of the state of the network, i.e.  $(n_l; l \in \mathcal{L})$ , where  $n_l$  refers to the number of flows in NBR group  $l$ .

First we present the buffer model of the scheduling unit. We start by studying only the non-real time buffer in order to compare the packet model results to the flow model. Inclusion of the real-time buffer will follow in section 3.2.

### 3.1.1 Buffer model for one delay class

In SIMA the scheduling algorithm discards a packet based on the occupancy level of both the real-time (rt) and the non-real-time (nrt) buffers. The arrival process of admitted packets should therefore be modeled as a state dependent arrival process. In order to preserve analytical tractability we assume that the arrival process within each priority class is Poissonian. Using this model we can calculate the loss probability and throughput for each priority class and compare the actual level of service received by the different classes.

Let us first concentrate on one buffer, say the nrt buffer. Denote the acceptance thresholds of priority class  $i$  by  $K_i$ ,  $i \in \mathcal{I}$ , with  $K_I = K$ , the size of the buffer. The packet transmission time is assumed to be exponentially distributed with mean  $1/\mu$ , and let  $\lambda(i)$  denote the packet arrival rate of priority class  $i$ . Define the cumulative sum of arrival intensities of those priority classes accepted into the system as  $\lambda_i = \sum_{k=i}^I \lambda(k)$ . The buffer can then be modeled as an  $M/M/1/K$  queue, with state dependent arrival intensities as depicted in figure 2.

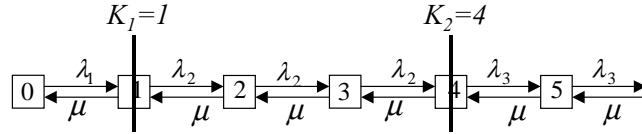


Figure 2: State transition diagram for one buffer modeled as an  $M/M/1/K$  queue.

The stationary distribution of the system is then

$$\pi_m = \left( \sum_{k=1}^I \frac{\lambda(k)}{\mu} \right)^m \pi_0 = \left( \frac{\lambda_1}{\mu} \right)^m \pi_0,$$

for  $m = 1, \dots, K_1 - 1$ . Correspondingly,

$$\pi_{K_i+m} = \left( \prod_{j=1}^i \left( \frac{\lambda_j}{\mu} \right)^{K_j - K_{j-1}} \right) \cdot \left( \frac{\lambda_{i+1}}{\mu} \right)^m \pi_0, \quad (1)$$

for  $i = 1, \dots, I - 1$  and  $m = 0, \dots, K_{i+1} - K_i - 1$ . Finally, the probability that the buffer is full is

$$\pi_K = \left( \prod_{j=1}^I \left( \frac{\lambda_j}{\mu} \right)^{K_j - K_{j-1}} \right) \pi_0.$$

The probability that the buffer is empty  $\pi_0$  is determined from the normalization condition  $\sum_{i=0}^K \pi_i$ .

Now, the probability  $p(i)$  that packets belonging to priority level  $i$  will be lost is simply

$$\begin{aligned} p(i) &= \sum_{j=K_i}^K \pi_j = p(i+1) + \sum_{j=K_i}^{K_{i+1}-1} \pi_j \\ &= p(i+1) + \sum_{m=0}^{K_{i+1}-K_i-1} \pi_{K_i+m}, \end{aligned} \quad (2)$$

for  $i = 1, \dots, I - 1$ , with  $\pi_{K_i+m}$  defined in equation (1). For the highest priority level, the loss probability is

$$p(I) = \pi_K = \left( \prod_{j=1}^I \left( \frac{\lambda_j}{\mu} \right)^{K_j - K_{j-1}} \right) \pi_0. \quad (3)$$

It is clear that, as the priority level increases, the probability that a packet will be discarded decreases. Furthermore, from the above equations one notices that the rate at which  $p(i)$  decreases with increasing  $i$  also decreases. The discarding probability is therefore a decreasing and convex function of the priority level.

The throughput of each priority level is defined as the net rate at which packets leave the system, i.e.

$$\lambda_{\text{eff}}(i) = \lambda(i)(1 - p(i)).$$

### 3.1.2 The TCP feedback mechanism

After the user has chosen the NBR and a priority level is assigned to the flow, the TCP mechanism adjusts the sending rate according to the feedback signal from the buffer system, represented by the loss probability. Both the NBR purchased as well as the state of the network thus determine the priority level. If the network is congested the loss probabilities will be high for the lower priority classes, the TCP will then drastically decrease the sending rate and the flow will move up a priority class, until the sending rate stabilizes in accordance with the total load of the network.

Here we use a basic TCP model, where we assume that the TCP mechanism is in congestion avoidance mode, resulting in the differential equations discussed in [16] for aggregates of flows. Furthermore, we assume that the dynamics of the buffer is faster than that of TCP resulting in the equilibrium relation between arrival intensity, round trip time (RTT) and loss probability of the TCP mechanism, namely the stable point [16]

$$\lambda = \frac{1}{RTT} \sqrt{2 \frac{1-p}{p}}. \quad (4)$$

In the case of many NBR groups and priority levels, and thus differing loss probabilities, equation (4) needs to be formulated for each NBR group separately.

### 3.1.3 Packet model: SIMA and TCP

Figure 3 presents the dynamics of SIMA and the TCP feedback mechanism. The flows are assigned priorities

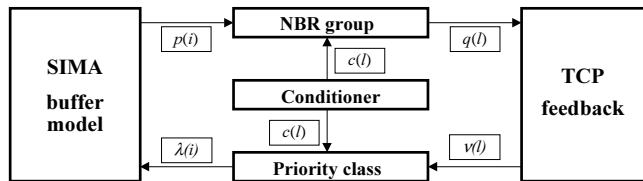


Figure 3: SIMA using TCP, an iteration view

$c(l)$  at the conditioner with the following priority level assignment to take into account varying number of priority levels,

$$c(l) = \max \left[ \min \left[ \left[ \lceil I/2 \rceil + 0.5 - \frac{\ln \frac{\nu(l)}{nbr(l)}}{\ln 2} \right], I \right], 1 \right]. \quad (5)$$

The packet arrival intensity  $\nu(l)$  corresponds to a long term average of the MBRs of individual flows. Figure 4 depicts the relationship between priority level and arrival intensity.

The aggregate arrival intensity  $\lambda(i)$  of priority class  $i$  is then

$$\lambda(i) = \sum_{l:c(l)=i} n_l \nu(l). \quad (6)$$

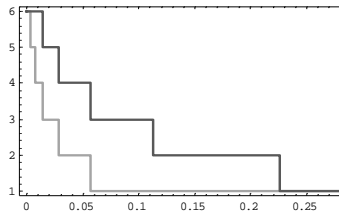


Figure 4: Evolution of priority levels of two NBR groups as function of intensity, with ratio of NBRs = 4, in a system with 6 priority levels.

From the buffer model we get the loss probabilities  $p(i)$  for each priority class  $i$ , see equations (2) and (3). The packet loss probability,  $q(l)$ , for a flow belonging to NBR group  $l$  is then

$$q(l) = p(c(l)) \quad (7)$$

Finally, equation (4) for each NBR group is

$$\nu(l) = \frac{1}{RTT} \sqrt{2 \frac{1 - q(l)}{q(l)}}. \quad (8)$$

Solving the equilibrium bandwidth allocation  $\nu(l)$  for each group  $l \in \mathcal{L}$  amounts to solving the set of equations (2), (3), (5)-(8) for the unknowns  $p(i)$ ,  $\lambda(i)$ ,  $c(l)$ ,  $q(l)$ , and  $\nu(l)$ , where  $i \in \mathcal{I}$  and  $l \in \mathcal{L}$ . We wish to study how bandwidth is divided among flow groups and how the division of bandwidth depends on the state of the network, i.e. the number of flows in the network. From the resulting equilibrium solution, we can also deduce how flows are divided into priority classes depending on the state of the link.

Notice that the discrete priority levels make equation 6 piecewise continuous as shown in figure 4 and from equations (2) and (3) we notice that the loss probability  $q(l)$  is also a piecewise continuous function of  $\nu(l)$ . The set of equations can, however, for computational purposes, be made continuous through parameterization. Due to lack of space we will not present the specifics of this simple technique.

### 3.1.4 Numerical Results

We solved the set of equations (2), (3), (5)-(8) under various scenarios. The parallel situation of figure 1 in section 2 for the packet model was solved for two flow groups,  $L = 2$ , and two priority classes,  $I = 2$ . The other parameters were set to  $\mu = 1$ ,  $RTT = 1000$ ,  $K_1 = 27$  and  $K_2 = 39$ .

Figure 5 shows the ratio,  $\nu(2)/\nu(1)$ , of bandwidth allocations for flows with  $(nbr(1), nbr(2)) = (0.04, 0.08)$  and  $(nbr(1), nbr(2)) = (0.02, 0.08)$ , i.e.  $k = 2$  and  $k = 4$ , respectively, as a function of number of flows  $n_1$  and  $n_2$  in group  $l = 1$  and  $l = 2$ . As in figure 1, the black areas correspond to equal bandwidth allocation, while the white areas correspond to bandwidth allocation equal to  $k$ .

Figures 1 and 5 are alike for both  $k = 2$  and  $k = 4$ , thus showing that the flow level setting of section 2 can be achieved using packet level mechanisms.

Figure 6 shows how increasing the number of priority classes to  $I = 3$  differentiates further the bandwidth allocation. Only the situation  $(nbr(1), nbr(2)) = (0.02, 0.08)$  is considered.

Figure 7 shows the same effect of increasing priority levels for  $(nbr(1), nbr(2)) = (0.04, 0.08)$  with  $RTT = 1000$  and  $K_I = 78$  as a two-dimensional cross-section of the diagonal, i.e. bandwidth allocation at points where the both NBR groups have the same number of flows. The trajectories of  $\nu(1)$  and  $\nu(2)$  are black and gray, respectively. Compared are  $I = 3$  and  $I = 6$  priority levels. The total number of flows is on the  $x$ -axis. The figure shows that more than three priority levels are needed, as with over 6 priority classes only rarely, and under normal circumstances never, do all packets and flows share bandwidth in equal shares.

## 3.2 Packet model for two delay classes

The previous section showed how bandwidth is divided between TCP flows inside the nrt queue. In this section, we extend the model to include all parts of the SIMA specification, namely how real time traffic

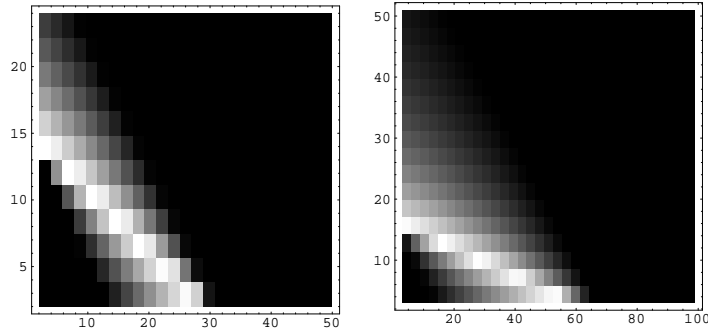


Figure 5: Bandwidth allocation for flows with  $k = 2$  (left) and  $k = 4$  (right) as function of number of flows,  $n_1$  ( $x$ -axis) and  $n_2$  ( $y$ -axis).

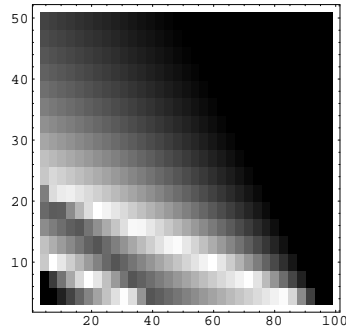


Figure 6: Bandwidth allocation for flows with  $k = 4$  and  $I = 3$  as function of number of flows,  $n_1$  ( $x$ -axis) and  $n_2$  ( $y$ -axis).

with stringent delay requirements is serviced in the model and how this can be done without starving the nrt queue of bandwidth. Studying both rt and nrt buffers gives us insight into evaluating the performance of SIMA in protecting the nrt-TCP flows from bandwidth exhaustion by the rt-non-TCP flows. We are thus interested in how the SIMA mechanism divides bandwidth between TCP and non-TCP flows.

Remembering that the packets in the rt buffer have strict priority over the packets in the nrt buffer, and that the non-TCP flows do not change their sending rate in accordance with the loss probability feedback, it is up to the threshold mechanism to drop excess non-TCP flows and to the conditioner to give lower priority to flows sending over the purchased rate.

### 3.2.1 Buffer model

In the case of two buffers, the discarding levels depend on both buffer contents. The two buffer case is modeled as two dependent  $M/M/1/K$  queues, with the other queue (rt buffer) having strict priority over the other. As was the case for the single buffer, the arrival intensities are state dependent, according to some threshold function. In this paper, we use for computations a threshold function derived from equation (9) in appendix, also used in [17]. For arrival intensities  $\lambda^{rt}(i)$  and  $\lambda^{nrt}(i)$  of priority class  $i$  with the notation  $\lambda_i^{rt} = \sum_{k=i}^I \lambda^{rt}(k)$  and  $\lambda_i^{nrt} = \sum_{k=i}^I \lambda^{nrt}(k)$  we have the transition diagram depicted in figure 8. The two buffer case is quite similar to the single buffer setting. The only difference is the equation for calculating  $p(i)$ , which can only be solved numerically.

Note that the discarding of an nrt packet depends on both the buffer content of the nrt buffer as well as that of the rt buffer. The above mechanism is designed to guarantee that priority levels are fixed across delay classes and hence to avoid starvation of the nrt packet queue. An alternative mechanism is to assign a fixed (minimum) weight to the nrt buffer, along the lines of the AF specification. However, assigning

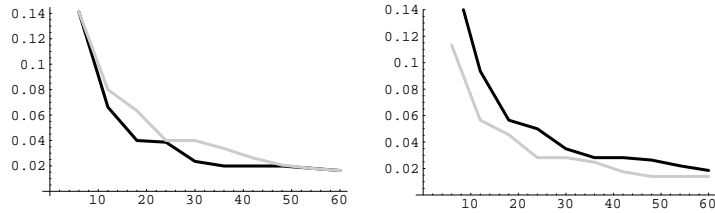


Figure 7: Bandwidth allocation for flows when  $I = 3$  (left) and  $I = 6$  (right) as function of total number of flows.

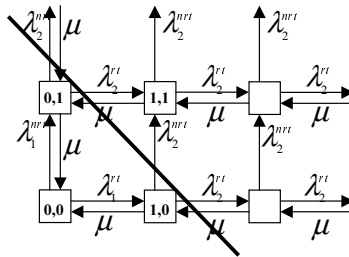


Figure 8: An example state transition diagram with fixed threshold (two priority levels) for a two buffer priority queue.

weights dynamically according to the change in load level is not trivial, and wrongly assigned weights may result in unfair service [18].

### 3.2.2 Numerical examples

Solving the equilibrium intensities is done by assuming that the non-TCP flows in the *rt* class send traffic at a constant intensity while the TCP flows in the *nrt* class adjust their sending rate as described earlier. The equilibrium sending rate  $\nu(l)$  is thus only solved for the TCP flows, and the non-TCP flows are modeled as non-responsive and constant background traffic. Notice however, that the priority class of the *rt* flows is determined by the SIMA conditioner, thus a flow with a high sending rate has a low priority.

We have the following scenario in terms of the free parameters:  $RTT = 1000$ ,  $K_{rt} = 13$ ,  $K_{nrt} = 39$  and  $I = 3$ . We have two NBR groups,  $L = 2$ , one group of *nrt*-TCP flows with  $nbr(1) = 0.04$  and one group of *rt*-non-TCP flows with  $nbr(2) = 0.08$ . Thus, the ratio between NBRs is  $k = 2$ . The set of equations (2), (3), (5-8) are only solved for  $l = 1$ . The non-TCP flows have a fixed sending rate  $\nu(2)$  of 0.039, 0.079, and 0.16 chosen so that the flows, when  $I = 3$ , are assigned priorities  $c(2) = 3$ ,  $c(2) = 2$ , and  $c(2) = 1$ , respectively. In the figures to follow the trajectories are solid, gray, and dashed for  $\nu(2) = 0.039$ ,  $\nu(2) = 0.079$ , and  $\nu(2) = 0.16$ .

Figure 9 shows how the SIMA conditioner and threshold mechanism are able to protect TCP flows from non-TCP flows when 33% of the traffic is TCP traffic serviced by the *nrt* buffer. Note that the condition  $I = 1$  is the traditional best-effort TCP, while the differentiation achieved with SIMA is represented by  $I = 3$  case.

Finally, in figure 10, 66% of the traffic is TCP traffic serviced by the *nrt* buffer.

It is realistic to assume that 10% of Internet traffic is real-time and non-TCP traffic. The figures show that the SIMA conditioner and threshold mechanism are able to protect TCP flows from non-TCP flows under such circumstances. When 66% of the traffic originates from non-TCP flows and all are assigned highest priority, TCP flows may receive only a small fraction of the bandwidth. Even when as many as one third of flows are non-TCP flows, the SIMA mechanisms are able to ensure that the TCP flows get a reasonable share of bandwidth.

Note that, if all flows were TCP friendly the ratio  $\nu(2)/\nu(1)$  would at most be  $k = 2$ . In figure 9 we observe the bound  $\nu(2)/\nu(1) < 4$ , when *rt* traffic is not in the highest priority class. Because only the TCP

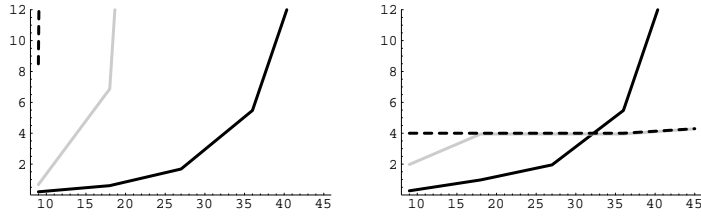


Figure 9: Ratio  $\nu(2)/\nu(1)$  between bandwidth allocations for flows with  $k = 2$  and  $n_2/n_1 = 2$ , when  $I = 1$  (left) and  $I = 3$  (right) as function of total number of flows.

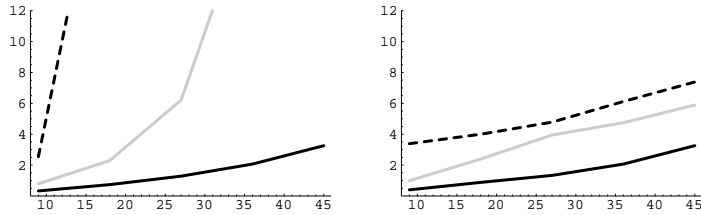


Figure 10: Ratio  $\nu(2)/\nu(1)$  between bandwidth allocations for flows with  $k = 2$  and  $n_2/n_1 = 1/2$ , when  $I = 1$  (left) and  $I = 3$  (right) as function of total number of flows.

flows adjust their sending rate, they drop their sending rate just enough to be in a priority class one higher than the rt flows, resulting in a ratio at most equal to four. In figure 10, the bandwidth ratio worsens as congestion deepens, as many TCP flows compete against each other. However, the bandwidth ratio is not as unfavorable, as in the traditional best-effort case.

The functionality of the SIMA threshold mechanism in protecting TCP flows against aggressive non-TCP flows is further demonstrated when studying the throughput ratio  $\frac{\nu(2)(1-q(2))}{\nu(1)(1-q(1))}$ . Figure 11 depicts how the non-TCP flows, which do not adjust their sending rate and do thus not move up in priority, have a high loss probability and thus a low throughput.

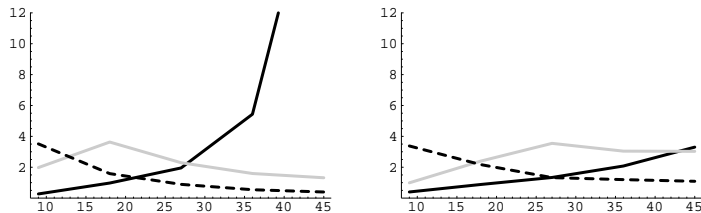


Figure 11: Ratio  $\frac{\nu(2)(1-q(2))}{\nu(1)(1-q(1))}$  between throughputs for flows with  $k = 2$ ,  $I = 3$  as function of total number of flows,  $n_2/n_1 = 2$  (left) and  $n_2/n_1 = 1/2$  (right).

The above figures show that a non-TCP flow with highest priority is not affected by the SIMA mechanism, compared to the best-effort case. Note, however, that for a flow to be in the highest priority class  $I = 3$ , it must send less than half of its NBR. Such events where many non-TCP flows are content with such little bandwidth occur rarely. Furthermore as the number of priority levels is increased to, e.g.  $I = 6$ , a flow must further cut down its sending rate in order to be in the highest priority class.

## 4 Conclusions

We studied both on the flow level and the packet level how weighted fairness could be achieved in a DiffServ network without per flow scheduling. The SIMA mechanism under study did not result in fixed weights, determined by the NBR or price paid, between bandwidth allocation of different classes. However,

our models showed that it did achieve differentiation between classes, where the weights varied between equal allocation to allocation in proportion to NBR purchased. With many priority classes the packet level mechanisms of SIMA approximate bandwidth division in fixed weights. With AF, for example, it is not clear that the buffer mechanism, using weighted fair queuing, is able to achieve weighted fairness among flows. Our models show how DiffServ fairness objectives can be quantified and taken into account in designing DiffServ packet level mechanisms.

Furthermore, the applicability of SIMA as a future Internet service model was demonstrated, as the SIMA conditioner, assigning priority levels as a function of both purchased nominal bit rate and bit rate sent by a flow, is able to protect TCP flows from bandwidth exhaustion by non-TCP flows. The two key ideas of SIMA, coupling of price paid and sending rate in determining priority levels and fixing threshold levels across delay classes, ensure that bandwidth is divided fairly inside classes of TCP flows and between non-TCP flows and TCP flows.

The models presented in the paper were kept simple to demonstrate the key ideas of modeling the future Internet based on DiffServ. We intend to broaden the two level, i.e. flow and packet level, study to include simulation results. Furthermore, we intend to deepen our study by including in the flow level model stochastic arrival model for connections. Enhancing the TCP model used in the packet level modeling is also a subject for further research. We then intend to apply our approach to modeling the AF DiffServ proposal to further demonstrate the need of quantifying DiffServ mechanisms on both the flow and the packet level. An ultimate goal is to design a DiffServ mechanism that realizes fixed weighted fairness as well as possible.

## References

- [1] J. Wroclawski, The Use of RSVP with IETF Integrated Services, Sept. 1997, RFC 2210.
- [2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, An Architecture for Differentiated Service, Dec. 1998, RFC 2475.
- [3] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, Assured Forwarding PHB Group, June 1999, RFC 2597.
- [4] V. Jacobson, K. Nichols, and K. Poduri, An Expedited Forwarding PHB, June 1999, RFC 2598.
- [5] F. Kelly, "Charging and rate control for elastic traffic," *Eur. Trans. Telecommun.*, vol. 8, pp. 33–37, 1997.
- [6] L. Massoulié and J. Roberts, "Bandwidth sharing: Objectives and algorithms," in *Proceedings of IEEE INFOCOM*, 1999, pp. 1395–1403.
- [7] G. de Veciana, T.-J. Lee, and T. Kontantopoulos, "Stability and performance analysis of networks supporting elastic services," *IEEE/ACM Transactions on Networking*, vol. 9, no. 1, pp. 2–14, Feb. 2001.
- [8] K. Kilkki, "Simple Integrated Media Access," available at <http://www-nrc.nokia.com/sima>, 1997.
- [9] K. Kilkki, *Differentiated Services for the Internet*, MacMillan Technical Publishing, 1999.
- [10] S. Floyd and K. Fall, "Promoting the use of end-to-end congestion control in the Internet," *IEEE/ACM Transactions on Networking*, vol. 7, no. 4, pp. 458–472, Aug. 1999.
- [11] J. MacKie-Mason and H. Varian, "Pricing the Internet," in *Public access to the Internet*, B. Kahin and J. Keller, Eds. Prentice-Hall, 1995.
- [12] K. Kilkki and J. Ruutu, "Simple Integrated Media Access (SIMA) with TCP," in the 4th INFORMS Telecommunications conference Boca Raton, FL, USA, Mar. 1998.

- [13] J. Harju, Y. Koucheryavy, J. Laine, S. Saaristo, K. Kilkki, J. Ruutu, H. Waris, J. Forsten, and J. Oinonen, "Performance measurements and analysis of TCP flows in a differentiated services WAN," in Proceedings of the Local Computer Networks conference (LCN 2000), Tampa, Florida, USA, Nov. 2000, pp. 1 – 10.
- [14] F. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 15, no. 49, pp. 237–255, 1998.
- [15] N. Semret, R. Liao, A. Campbell, and A. Lazar, "Pricing, provisioning and peering: Dynamic markets for differentiated Internet services and implications for network interconnections," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 12, pp. 2499–2513, Dec. 2000.
- [16] F. Kelly, "Mathematical modelling of the Internet," in *Proc. of Fourth International Congress on Industrial and Applied Mathematics*, 1999, pp. 105–116.
- [17] J. Laine, S. Saaristo, J. Lemponen, and J. Harju, "Implementation and measurements of simple integrated media access (SIMA) network nodes," in *Proceedings for IEEE ICC 2000*, June 2000, pp. 796–800.
- [18] Constantinos Dovrolis, Dimitrios Stiliadis, and Parameswaran Ramanathan, "Proportional differentiated services: Delay differentiation and packet scheduling," in *Proceedings of ACM SIGCOMM*, 1999, pp. 109–120.
- [19] G. Schultz, "A simulation study of the SIMA priority scheme," Unpublished manuscript, 1999.

## A Simple Integrated Media Access

In order to be self-contained, we include in the appendix a summary of the Simple Integrated Media Access (SIMA) specification [8]. Differentiated Services mechanisms aim at determining how network capacity should be divided during overload situations. SIMA presents a proposal for this, and furthermore includes a basis for charging. The entity that influences both the charging and division of bandwidth is the nominal bit rate introduced earlier. The user chooses the NBR, with the amount charged from the user being a function of NBR. However, the NBR does not alone affect the service acquired by the user. The received service is a function of the network load through the congestion feedback of TCP determining the sending rate of a user. Furthermore, the priority class of a flow is determined by the ratio of MBR to NBR, where MBR refers to the momentary bit rate of the flow. The user can also make a distinction in service according to application or delay requirement by choosing to label the traffic as real-time (rt) or non-real-time (nrt) traffic. Figure [8] depicts the functional entities of SIMA.

The NBR is permanent, it has a charge associated with it and is related to an organization (e.g. network interface), a user (e.g. IP address), or a flow (e.g. IP address and port number). The simplest approach is to assign a NBR to each interface, while the most useful approach in terms of performance, is to have a NBR associated with each flow. In the following, we assume that the NBR entity is the flow.

In addition to purchasing a certain value of NBR, the user or application labels the flows sent in one of the two delay classes. The real-time class is designated for flows requiring low delay and jitter. This is achieved by having small real-time buffers serviced prior to non-real-time buffers and favoring smooth traffic with small, i.e. less than 0.1 ms, traffic variations.

Once NBR is purchased, the delay class is chosen, and traffic is sent to the network, a priority level is associated with the packets sent. The conditioner at the access node assigns priorities per flow. In order to assign the priority level, also called drop preference, the traffic rate of the flow has to be measured. The measurement of the momentary bit rate is done by averaging the traffic sent by the flow. A proposed measuring principle is the exponential moving average, with different parameters for the rt and nrt traffic. Non-real time applications that can have variations in time scales of over 10 ms would not benefit from marking flows as real-time, as the bit rate measurements for real-time class traffic is more sensitive to traffic variations, giving thus worse priorities during peak rates.

With the momentary bit rate determined, the conditioner at the access node assigns the priority level at the arrival of the  $j$ :th packet of  $l$ :th flow to

$$PL(l, j) = \max \left[ \min \left[ \left\lceil \left[ I/2 \right] + 0.5 - \frac{\ln \frac{MBR(l, j)}{NBR(l)}}{\ln 2} \right\rceil, I \right], 1 \right]$$

where  $I$  is the total number of priority levels. The above equation is derived so that when the ratio of MBR to NBR is one, the flow is given medium priority. As the MBR of a flow doubles (halves) the priority decreases (increases) by one unit.

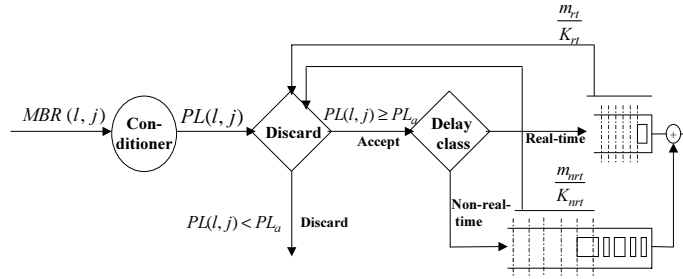


Figure 12: The SIMA specification

Once the momentary bit rate of the flow is measured and the flow is assigned a priority level at the access node, the packets are marked with the priority level. In the core nodes, each packet is handled based on the DiffServ code point (DSCP) information storing the priority level and rt/nrt classification information. The delay class of the packet determines which of the two queues the packet is directed to in the scheduling unit.

Before this, however, there is a cell discarding system, which determines if the packet can enter the scheduling unit. The discarding is solely based on the priority level of the packet, the delay class does not affect the decision whatsoever. At the discarding system there is an accepted level of priority,  $PL_a$ , calculated from the current buffer contents of both the scheduling units queues. Some of the possible equations for calculating  $PL_a$  based on the buffer contents  $m_{rt}$  and  $m_{nrt}$  of the real time and non-real-time queues, respectively, given in the proposal are,

$$\begin{aligned} PL_a &= a + b \cdot \left( \frac{m_{rt}}{K_{rt}} + \frac{m_{nrt}}{K_{nrt}} \right) \\ PL_a &= a + b \cdot \max \left( \frac{m_{rt}}{K_{rt}}, \frac{m_{nrt}}{K_{nrt}} \right) \\ PL_a &= a + b \cdot \sqrt{\left( \frac{m_{rt}}{K_{rt}} \right)^2 + \left( \frac{m_{nrt}}{K_{nrt}} \right)^2}, \end{aligned} \quad (9)$$

where  $K_{nrt}$  and  $K_{rt}$  are the sizes of the non real-time and real-time buffers, respectively and  $a$  and  $b$  are constants to be determined by the implementors.

If the packet priority level is equal to or higher than  $PL_a$ , the packet is accepted to the scheduling unit and placed in the appropriate queue. Otherwise, the packet is discarded. In the case of a TCP flow, the packet discarding signals the TCP source that it should drop (half) its sending rate, which in terms of the priority level would mean an increase by one in the flow's priority level, thus decreasing the probability of packets being discarded.

Note that the accepted level of priority,  $PL_a$ , is not a fixed value, but varies in time according to variations in the buffer contents and thus in response to congestion build up. This variation also affects the TCP mechanism through the loss probability feedback. Stability questions of the threshold mechanism have been considered in [19].