PACKET SCHEDULING ALGORITHM WITH WEIGHT OPTIMIZATION

Ari Viinikainen, Jyrki Joutsensalo, Mikko Pääkkönen and Timo Hämäläinen
Department of Mathematical Information Technology, University of Jyväskylä
P.O. Box 35 (Agora)
40014 University of Jyväskylä, FINLAND

Keywords: Packet scheduling, flat pricing, revenue maximization, quality of service (QoS).

Abstract: In this paper we present a scheduling algorithm for traffic allocation. In our model, we use a flat pricing scenario in which the weights of the queues are updated using revenue as a target function. Due to the algorithm's closed form nature, it is capable of operating in non-stationary environments. In addition, the algorithm is nonparametric and deterministic in the sense that any assumptions about the call density functions or duration distributions are not made.

1 INTRODUCTION

Numerous pricing proposals have been developed for the efficient use of packet–switched networks. Here, we review the research work that has similar features as in our model. Several researchers have proposed a pricing approach in which a small number of service classes (between 2 and 4) should be offered on the network in order to achieve service differentiation (Gubta et al., 1997; Odlyzko, 1999; Gibbens et al., 2000).

The pricing of a single network which provides multiple services at different performance levels is studied in (Cocchi et al., 1993). They present a good example which shows that in comparison to flat rate pricing for all services, a price schedule based on performance objectives can enable every customer to derive a higher surplus from the service, and at the same time, generate a bigger revenue for the service provider. Reference (Gibbens and Kelly, 1999) describes yet another scheme for packet-based pricing as an incentive for more efficient flow control.

However, it seems that packet-based pricing schemes are not always appropriate, because real-time traffic requires QoS (Quality of Service) measures that are hard to analyze (Kelly, 1996; Bertsimas et al., 1998b; Bertsimas et al., 1998a; Paschalidis, 1999).

The pricing of multiple services in single network, with guaranteed QoS requirements, is studied in (Keon and Anandalingam, 2003), where the optimal pricing problem is formulated as a nonlinear integer expected revenue optimization problem.

A static pricing schedule, based on dynamic programming is proposed in (Paschalidis and Tsitsiklis, 2000) for maximization of the revenue and social welfare and is extended to multiple service loss networks in (Paschalidis and Liu, 2002).

Our research differs from the above studies by linking pricing and queuing issues together; in addition, our model does not need any excess information about user behavior, utility functions etc. (as most pricing and game-theoretic approaches do). This paper extends our previous pricing and QoS research (Joutsensalo et al., 2003a; Joutsensalo et al., 2003b; Joutsensalo et al., tion), where linear pricing algorithm has been investigated. However, the flat pricing algorithm seems more realistic and is now the scope of our studies.

We consider pricing and scheduling issues to manage multiple communications services in single telecommunication network. Different services are grouped into service classes, differing in resource requirements and tolerating different quality limitations such as loss of data and transmission delay. The proposed flat algorithm takes also into account the variable average data packet sizes in different traffic classes. The QoS and revenue aware scheduling algorithm is studied in a single node case. It is derived from a Lagrangian optimization problem, and an optimal closed form solution is presented.

The rest of the paper is organized as follows. In

Viinikainen A., Joutsensalo J., Pääkkönen M. and Hämäläinen T. (2004).
PACKET SCHEDULING ALGORITHM WITH WEIGHT OPTIMIZATION.
In *Proceedings of the First International Conference on E-Business and Telecommunication Networks*, pages 127-133
DOI: 10.5220/0001390901270133

Section 2, the proposed flat pricing scenario is presented and generally defined. A closed form scheduling algorithm is also derived in this section. Section 3 contains experimental part justifying theorems. Conclusions are presented and discussions about the future work is made in Section 4.

2 THE FLAT PRICING SCENARIO AND ALGORITHM

The scheduling mechanism is one of the major component for providing differentiated service levels. In our scheduling model, the weights of the queues are dynamically updated based on the QoS and pricing criterions of the service classes. In other word, the weights for different classes can be assigned in a way that the performance of high priority classes is guaranteed and no starvation of low priority classes occur.

The pricing scenario consists of m different traffic classes for different applications and priorities. In our scenario we have four classes, $m = 4$, which are referred to as the platinum, gold, silver and bronze classes (Table 1). The platinum and gold classes are reserved for the realtime and the lower priority classes (silver and bronze) for non–realtime purposes. As an example of the four classes, Voice over IP (VoIP) and video conference traffic streams belong to the platinum and gold classes, respectively. Video streams (e.g. MPEG4) are considered as a silver class customers and File Transfer Protocol (FTP) flows belong to the lowest priority (bronze) service class. The platinum class customers are willing to pay for some guaranteed bandwidth, delay, jitter and packet loss probability. The gold class customers are ready to pay for guaranteed bandwidth, delay and delay variance (jitter). The silver class customers are not so tight on the QoS requirements. In the silver class bandwidth and jitter should be guaranteed. In the bronze class the guaranteed bandwidth and packet loss are the most important QoS parameters, while the delay can vary a lot.

Let us consider a packet scheduler which receives packets to be delivered from m different queues (i.e. classes). Now, let d_0 be the processing time of the classifier for transmitting data from one queue to the output of a packet scheduler. The data packets have variable sizes and the average packet size for class i is $E(b_i), i = 1, \ldots, m$.

In our scheduling model, the real processing time (delay) for class i in the packet scheduler is

$$d_i = \frac{N_i E(b_i) d_0}{w_i} = \frac{N_i E(b_i)}{w_i}, \qquad (1)$$

where $w_i(t) = w_i, i = 1, \ldots, m$ are weights allotted for each class, $N_i(t) = N_i$ is the number of customers and $E(b_i)(t) = E(b_i)$ is the average data packet size in the ith queue. Here, the time index t has been dropped for convenience until otherwise stated and d_0 can be scaled to $d_0 = 1$ without loss of generality. The natural constraints for the weights are

$$w_i > 0 \qquad (2)$$

and

$$\sum_{i=1}^{m} w_i = 1. \qquad (3)$$

Without loss of generality, only non-empty queues are considered, and therefore $w_i \neq 0, i = 1, \ldots, m$. If some weight is $w_i = 1$, then $m = 1$, the packet size can be scaled to $E(b_i) = 1$ and the only class to be served has the minimum processing time $d_0 = 1$, if $N_i = 1$. For each service class, a *pricing function*

$$r_i(d_i) = r_i(\frac{N_i E(b_i)}{w_i} + c_i) \qquad (4)$$

(euros/minute) is non-increasing with respect to the delay d_i. Here $c_i(t) = c_i$ includes insertion delay, transmission delay etc., and it is assumed to be constant.

In the *flat* pricing scenario, the pricing function is defined via *maximum delay* for each class and queue as a QoS parameter.

The *Gain factor* r_i of class i is measured by money paid by one customer to the service provider per unit time, e.g. euros/minute. Hence, the pricing function in (4) reduces to the piecewise flat function

$$r_i(d_i) = r_i, \qquad (5)$$

under the constraint

$$\frac{N_i E(b_i)}{w_i} \leq d_{i,max}, \quad i = 1, \ldots, m, \qquad (6)$$

where $d_{i,max}$ are preselected maximum delays to be guaranteed. When N_i customers are in the class (or in the queue) i, the revenue achieved from that class is

$$F_i = N_i r_i \qquad (7)$$

euros/minute. Therefore, the total price paid by the N_i customers in m classes is

$$F = \sum_{i=1}^{m} F_i = \sum_{i=1}^{m} N_i r_i \qquad (8)$$

under the constraint that the pre-selected maximum delays $d_{i,max}$ are not exceeded. By using Lagrangian approach, the revenue can be presented in the form

$$F = \sum_{i=1}^{m} N_i r_i + \lambda(1 - \sum_{i=1}^{m} w_i), \qquad (9)$$

where

$$w_i = \frac{N_i E(b_i)}{d_i}. \qquad (10)$$

Table 1: QoS parameters for different traffic classes.

Traffic class	**Type**	**bandwidth**	**e-to-e delay**	**packet loss**	**jitter**
Platinum	VoIP	x	x	x	x
Gold	Video conference (H.263)	x	x		x
Silver	Video stream (MPEG4)	x			x
Bronze	FTP	x		x	

From (9) and (10), the revenue can be presented as

$$F = \sum_{i=1}^{m} N_i r_i + \lambda(1 - \sum_{i=1}^{m} \frac{N_i E(b_i)}{d_i}), \qquad (11)$$

or

$$F = \sum_{i=1}^{m} \frac{r_i d_i w_i}{E(b_i)} + \lambda(1 - \sum_{i=1}^{m} w_i). \qquad (12)$$

Optimal weights are obtained from the first derivative

$$\frac{\partial F}{\partial w_i} = \frac{r_i d_i}{E(b_i)} - \lambda = 0. \qquad (13)$$

$$\lambda = \frac{r_i d_i}{E(b_i)} = \frac{r_i N_i E(b_i)}{w_i E(b_i)} = \frac{r_i N_i}{w_i} \qquad (14)$$

$$w_i = \frac{r_i N_i}{\lambda} \qquad (15)$$

Because $\sum_i w_i = 1$, then

$$w_i = \frac{r_i N_i}{\lambda \sum_l w_l} = \frac{r_i N_i}{\lambda \sum_l \frac{r_l N_l}{\lambda}} = \frac{r_i N_i}{\sum_l r_l N_l}. \qquad (16)$$

From (10) and (16) one obtains

$$d_i = \frac{E(b_i)}{r_i} \sum_{l=1}^{m} r_l N_l \qquad (17)$$

Equation (16) expresses the closed form solution to the weights. Interpretation of (16) is obvious:

- Larger the gain factor r_i is, larger is the corresponding weight w_i.
- Larger the number of users N_i is, larger is the corresponding weight w_i.

By using optimal weights, revenue F can be expressed as follows:

$$\begin{aligned} F &= \sum_{i=1}^{m} N_i r_i = \sum_{i=1}^{m} \frac{w_i d_i r_i}{E(b_i)} = \sum_{i=1}^{m} \frac{r_i N_i d_i r_i}{E(b_i) \sum_l r_l N_l} \\ &= \sum_{i=1}^{m} \frac{r_i^2 d_i N_i}{E(b_i) \sum_l r_l N_l} = \frac{1}{F} \sum_{i=1}^{m} \frac{r_i^2 d_i N_i}{E(b_i)}, \qquad (18) \end{aligned}$$

and F is

$$F = \sqrt{\sum_{i=1}^{m} \frac{r_i^2 d_i N_i}{E(b_i)}} \qquad (19)$$

with constraints

$$\sum_{i=1}^{m} \frac{N_i E(b_i)}{d_i} \leq 1, \qquad (20)$$

and

$$d_i = \frac{E(b_i)}{r_i} \sum_{l=1}^{m} r_l N_l \leq d_{i,max}, \quad i = 1, \ldots, m. \qquad (21)$$

From (19) and (21) it is seen that gain factors r_i, maximum allowed delays $d_{i,max}$, as well as number of users N_i increase the revenue, which is a plausible result. Also, as seen from (19), smaller packet sizes increase the revenue.

Call Admission Control mechanism can be made by simple hypothesis testing without assumptions about call or dropping rates. Let the state at the time t be $N_i(t)$, $t = 1, \ldots, m$. Let the new hypothetical state at the time $t+1$ be $\tilde{N}_i(t+1)$, $t = 1, \ldots, m$, when one or several calls appear in some class/classes. In hypothesis testing, revenue formula (19) is applied as follows:

$$F(t) = \sqrt{\sum_{i=1}^{m} \frac{r_i^2 d_i N_i(t)}{E(b_i)}} \qquad (22)$$

$$\tilde{F}(t+1) = \sqrt{\sum_{i=1}^{m} \frac{r_i^2 d_i \tilde{N}_i(t+1)}{E(b_i)}} \qquad (23)$$

$$d_i(t) = \frac{E(b_i)}{r_i} \sum_{l=1}^{m} r_l N_l(t) \qquad (24)$$

$$\tilde{d}_i(t+1) = \frac{E(b_i)}{r_i} \sum_{l=1}^{m} r_l \tilde{N}_l(t+1) \qquad (25)$$

If $F(t) > \tilde{F}(t+1)$ or maximum delay is exceeded ($\tilde{d}_i(t+1) > d_{i,max}$), then call is rejected, otherwise it is accepted.

3 EXPERIMENTS

The experiments demonstrate by simulation the effects of the maximum delays $d_{i,max}$ on the revenue. In the experiments, calls and durations are Poisson and exponentially distributed, respectively. In addition, the number of classes is $m = 4$. The different classes have different average data packet sizes $E(b_1) = 1$, $E(b_2) = 3$, $E(b_3) = 5$ and $E(b_4) = 10$ kilobytes, with the platinum class (e.g delivering VoIP traffic) having the smallest size and the size increases towards the bronze class (e.g delivering FTP traffic). Call rates per unit time for the platinum, gold, silver, and bronze classes are $\alpha_1 = 0.20$, $\alpha_2 = 0.30$, $\alpha_3 = 0.40$ and $\alpha_4 = 0.50$, respectively. The duration parameters (i.e. "decay rates") are $\beta_1 = 0.001$, $\beta_2 = 0.003$, $\beta_3 = 0.005$ and $\beta_4 = 0.01$, where probability density functions for duration are

$$f_i(t) = \beta_i e^{-\beta_i t}, \quad ,i = 1, 2, 3, 4 \quad t \geq 0. \qquad (26)$$

The number of unit times in the experiments was $T = 2000$.

Experiment 1. In the first experiment, the four service classes have the gain factors $r_1 = 40$, $r_2 = 30$, $r_3 = 20$ and $r_4 = 10$. The maximum delays for the classes are $d_{1,max} = 15$, $d_{2,max} = 100$, $d_{3,max} = 250$ and $d_{4,max} = 1000$. Fig. 1 shows the evolution of the delays $d_1(t)$, $d_2(t)$, $d_3(t)$ and $d_4(t)$ as well as the maximum boundaries for them as a function of time in the first experiment. It is seen that the delays are always below the maximum values, as guaranteed by the constraint (21). Also, the delay boundary of the platinum class is limiting the number of users in the other classes as their delays are well beyond the limits (i.e. the gold, silver and bronze classes could support more users). Fig. 2 shows the number of users and Fig. 3 the revenue in this experiment.

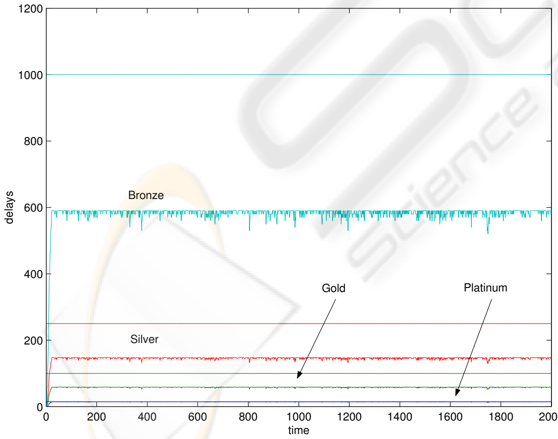

Figure 1: Evolution of the delays $d_1(t)$ (platinum), $d_2(t)$ (gold), $d_3(t)$ (silver) and $d_4(t)$ (bronze) as a function of time in the first experiment. The straight lines show the maximum delays.

Experiment 2. In the second experiment, the parameters are as in experiment 1, except for the delay of the platinum class, which is increased to $d_{1,max} = 25$. By comparing Fig. 2 and Fig. 5 it is seen that the number of users can be increased, while keeping the delays below the defined maximum values (Fig. 1 and Fig. 4). In Fig. 3 and Fig. 6 the increasing of the revenue is due to the increase of the maximum delay of the platinum class. This increases the number of highly charged users more in the highest-priority classes (Fig. 2 and Fig. 5) and decreases the overall packet sizes in the queues, thus the revenue increases.

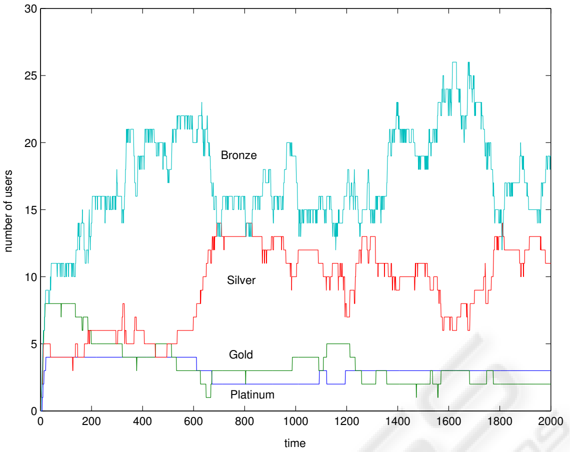

Figure 2: Evolution of the number of users $N_1(t)$ (platinum), $N_2(t)$ (gold), $N_3(t)$ (silver) and $N_4(t)$ (bronze) as a function of time in the first experiment.

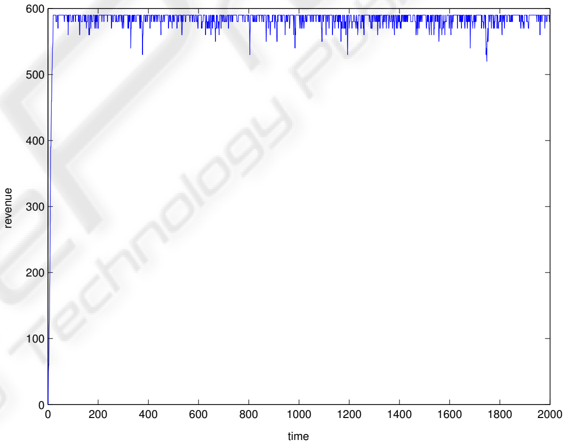

Figure 3: Evolution of the revenue as a function of time in the first experiment.

Experiment 3. In the third experiment we demonstrate the effect of gain factors r_i, by changing the gain factor of the platinum class from $r_1 = 40$ to $r_1 = 100$. The other parameters are the same as in Experiment 1 (i.e. the platinum class delay boundary is limiting the resources of the other classes). Now, by comparing Fig. 1 and Fig. 7, it is seen that the delays of the gold, silver and bronze classes are increased

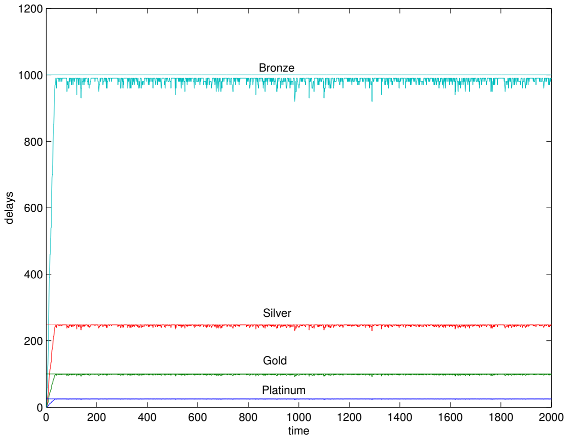

Figure 4: Evolution of the delays $d_1(t)$ (platinum), $d_2(t)$ (gold), $d_3(t)$ (silver) and $d_4(t)$ (bronze) as a function of time in the second experiment. The straight lines show the maximum delays.

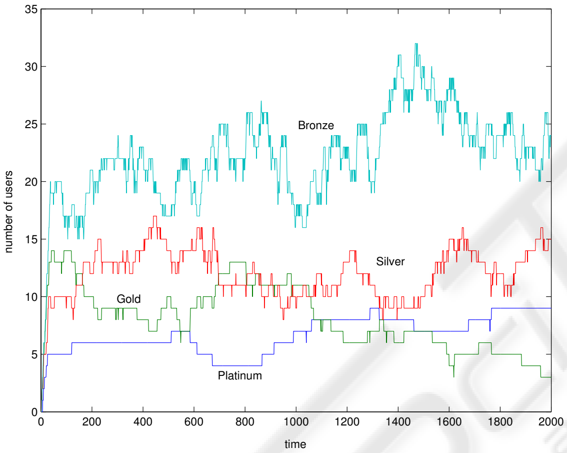

Figure 5: Evolution of the number of users $N_1(t)$ (platinum), $N_2(t)$ (gold), $N_3(t)$ (silver) and $N_4(t)$ (bronze) as a function of time in the second experiment.

(but remaining under the constraints). This is due to the fact that, now the number of users in the gold, silver and bronze classes is increased while the number of users in the platinum class is decreased (Fig. 2 and Fig. 8). In this case, by changing one of the gain factors, which lead to a different distribution of users in the classes, the revenue is increased (Fig. 3 and Fig. 9).

These kinds of results give valuable information on the tuning the model to work at the most optimal way under different kind of traffic scenarios e.g. at the situations when the connection arrival rate varies at different times scales - say night, morning, day, and evening. This can also lead to the situation where the customers of the highest traffic classes move to use resources of the lower traffic classes at night time. This is because there is enough capacity available, and thus needed QoS level can be achieved with the price of the lower class.

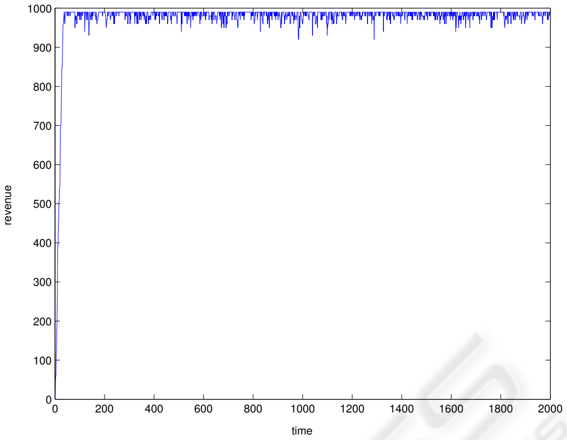

Figure 6: Evolution of the revenue as a function of time in the second experiment.

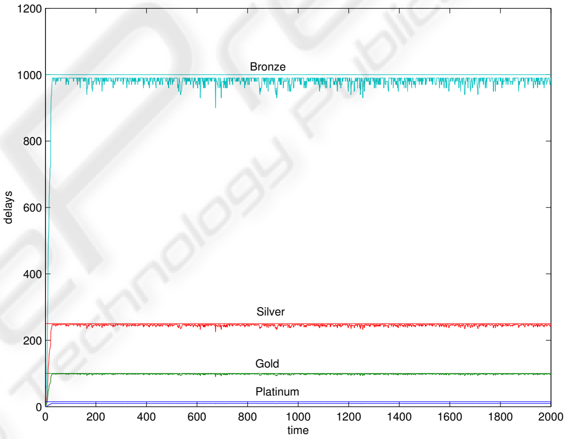

Figure 7: Evolution of the delays $d_1(t)$ (platinum), $d_2(t)$ (gold), $d_3(t)$ (silver) and $d_4(t)$ (bronze) as a function of time in the third experiment. The straight lines show the maximum delays.

Next, we present summary of our approach and the experiments.

- The proposed weight updating algorithm is computationally inexpensive in our scope of study.
- The algorithm uses variable average data packet sizes, taking into consideration the relative average packet sizes in different classes.
- Experiments clearly justify the performance of the algorithm, i.e. revenue curves are positive and the delays remain below the predefined limits.
- Some of the statistical and deterministic algorithms

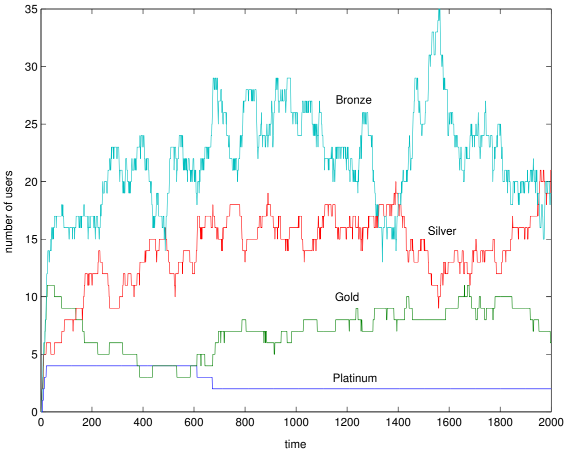

Figure 8: Evolution of the number of users $N_1(t)$ (platinum), $N_2(t)$ (gold), $N_3(t)$ (silver) and $N_4(t)$ (bronze) as a function of time in the third experiment.

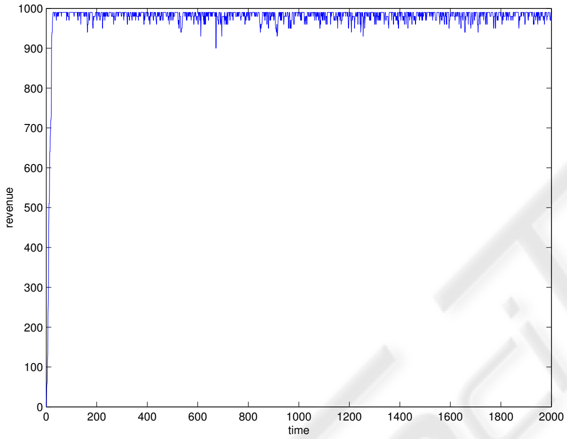

Figure 9: Evolution of the revenue as a function of time in the third experiment.

presented in the literature assume quite strict *a priori* information about parameters or statistical behavior such as call densities, duration or distributions. However, such methods usually are - in addition to computationally complex - not robust against erroneous assumptions or estimates. On the contrary, our algorithm is deterministic and non-parametric, ie. it uses only the information about the number of customers, and thus we believe that in practical environments it is competitive candidate due to the robustness.

4 CONCLUSION

In this paper we designed a QoS- aware scheduling and pricing model, that takes into account the variable packet sizes in different service classes, the user's satisfaction (price vs. received QoS) and the optimal use of the limited network resources. The presented solution gives the service provider and consumers a new way in which to use and get services from the networks. Another issue concerns the model's simplicity for the network operators.

As can be seen from the results, presented model shares limited network resources in such a way that QoS requirements of service classes are fulfilled. In addition, flat pricing scenario operates well (the total revenue will increase when the optimal values for the different QoS parameters are found).

In the future work, a multinode case will be investigated. It is important to develop such a distributed approximation, which does not suffer from the large dimensionality and computational complexity of the optimal global approach. We have also started to work with Linux routers, and the goal is to implement the presented algorithm into a real router environment.

REFERENCES

Bertsimas, D., Paschalidis, I. C., and Tsitsiklis, J. N. (1998a). Asymptotic buffer overflow probabilities in multiclass multiplexers: an optimal control approach. *IEEE Trans. Automat. Contr.*, 43:315–335.

Bertsimas, D., Paschalidis, I. C., and Tsitsiklis, J. N. (1998b). On the large deviations behavior of acyclic networks of G/G/1 queues. *Ann. Appl. Prob.*, 8(4):1027–1069.

Cocchi, R., Shenker, S., Estrin, D., and Zhang, L. (1993). Pricing in computer networks: Motivation, formulation and example. *IEEE/ACM Trans. Networking*, 1(6):614–627.

Gibbens, R., Mason, R., and Steinberg, R. (2000). Internet service classes under competition. *IEEE J. Select. Areas Commun.*, 18(12):2490–2498.

Gibbens, R. J. and Kelly, F. P. (1999). Resource pricing and the evolution of congestion control. *Automatica*, 35(12):1969–1985.

Gubta, A., Stahl, D., and Whinston, A. (1997). A stochastic equilibrium model of internet pricing. *J. Economics Dynamics Contr.*, 21(4–5):697–722.

Joutsensalo, J., Gomzikov, O., Hämäläinen, T., and Luostarinen, K. (2003a). Enhancing revenue maximization with adaptive WRR. In *Proc. Eighth IEEE International Symposium on Computers and Communication (ISCC'03)*, pages 175–180.

Joutsensalo, J., Hämäläinen, T., Pääkkönen, M., and Sayenko, A. (2003b). Adaptive weighted fair scheduling method for channel allocation. In *Proc. IEEE International Conference on Communications (ICC'03)*, volume 1, pages 228–232.

Joutsensalo, J., Hämäläinen, T., Pääkkönen, M., and Sayenko, A. (submitted for publication). Revenue aware scheduling algorithm in the single node case. *Journal of Communications and Networks*.

Kelly, F. P. (1996). *Stochastic Networks: Theory and Applications*, volume 4 of *Royal Statistical Society Lecture Notes Series*, chapter Notes on effective bandwidths, pages 141–168. Oxford University Press, London.

Keon, N. J. and Anandalingam, G. (2003). Optimal pricing for multiple services in telecommunications networks offering quality-of-service guarantees. *IEEE/ACM Trans. Networking*, 11(1):66–80.

Odlyzko, A. M. (1999). Paris metro pricing for the internet. In *ACM Conference on Electronic Commerce (EC'99)*, pages 140–147.

Paschalidis, I. C. (1999). Class-specific quality of service guarantees in multimedia communication networks. *Automatica*, 35(12):1951–1968.

Paschalidis, I. C. and Liu, Y. (2002). Pricing in multiservice loss networks: Static pricing, asymptotic optimality, and demand substitution effects. *IEEE/ACM Trans. Networking*, 10(3):425–438.

Paschalidis, I. C. and Tsitsiklis, J. N. (2000). Congestion-dependent pricing of network services. *IEEE/ACM Trans. Networking*, 8(2):171–183.