Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT®

Computer Networks 42 (2003) 231–249

COMPUTER NETWORKS

www.elsevier.com/locate/comnet

Pricing strategies under heterogeneous service requirements ☆

Michel Mandjes [a,b]

[a] Faculty of Mathematical Sciences, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands
[b] CWI, P.O. Box 94079, 1090 GB Amsterdam, Netherlands

Received 13 May 2002; received in revised form 23 January 2003; accepted 30 January 2003

Responsible Editor: N. Shroff

Abstract

This paper analyzes a communication network, used by customers with heterogeneous service requirements. We investigate priority queueing as a way to establish service differentiation. It is assumed that there is an infinite population of customers, who join the network as long as their utility (which is a function of the queueing delay) is larger than the price of the service. We focus on the specific situation with two types of users: one type is delay-sensitive ('voice'), whereas the other is delay-tolerant ('data'); these preferences are reflected in their utility curves. Two models are considered: in the first the *network* determines the priority class of the users, whereas the second model leaves this choice to the *users*. For both models we determine the prices that maximize the provider's profit. Importantly, these situations do *not* coincide. Our analysis uses elements from queueing theory, but also from microeconomics and game theory (e.g., the concept of a Nash equilibrium). We conclude the paper by considering a model in which throughput (rather than delay) is the main performance measure. Again the pricing strategy exploits the heterogeneity in service requirements and willingness-to-pay.

Keywords: Packet networks; Differentiated services; Pricing; Congestion; Microeconomics; Negative externalities; Game theory

1. Introduction

Current usage of data-networks, such as the *Internet*, is still dominated by 'traditional' data services: web browsing, file transfer, remote terminal, electronic mail, etc. These applications do not impose severe requirements on the network, in that they tolerate relatively large packet delays. New Internet applications, e.g., real-time applications such as interactive voice and video, can be characterized as delay-sensitive, and are consequently considerably more demanding. This heterogeneity of the service requirements makes it necessary that the delay-tolerant and delay-sensitive users are handled differently—otherwise all traffic must be handled according to the requirements of the *most* demanding class, i.e., the real-time class, which will inevitably lead to a network running at a relatively poor utilization level. A possible solution is to give *priority* to the delay-sensitive traffic in the queues of the network.

☆ Part of this work was done while the author was at Bell Laboratories/Lucent Technologies, 600 Mountain Ave., P.O. Box 636, Murray Hill, NJ 07974-0636, USA.

E-mail addresses: michel@cwi.nl, michel.mandjes@cwi.nl (M. Mandjes).

1389-1286/03/$ - see front matter
doi:10.1016/S1389-1286(03)00191-9

Shenker [13] further motivates this prioritization and related design issues for the Internet.

Pricing. Without an appropriate pricing scheme, any prioritization is useless; if there were no price difference between the priority classes, all users would opt for the high-priority class. In other words: the prices of the priority classes should give users an incentive to join the 'right' priority class. In terms of the delay-tolerant user (or, shortly, the *data* user) and the delay-sensitive user (or, shortly, the *voice* user): voice users are encouraged to use the high-priority class, whereas data users are given an incentive to join the low-priority class. This is done by imposing a higher charge on the high-priority class. A next question is: how should the network provider choose the prices for both classes in order to maximize its profit?

Here two models can be distinguished. In the first model the provider assigns a priority class to each user type—for instance, the provider can decide that the voice customers are directed to the high-priority queue, and the data users to the low-priority queue. This model of 'dedicated classes' (or 'implicit supply of service', in Shenker's [13] terminology) is relatively simple to analyze, as the network users have only two alternatives: joining the network or not.

The harder, but perhaps more realistic, model is the model with 'open classes' (or 'explicit supply of service', as it is called in [13]), in which the users can choose between the priority classes. It is not clear beforehand whether the prices that optimize the profit in the dedicated-classes model, are also profit optimizing for the open-classes model. The reason is that the prices found in the dedicated-classes model might lead to a situation in which data (voice) users might appreciate the high-(low-) priority class more. In other words: it is not a priori clear whether the optimal prices from the dedicated-classes model lead to an *incentive-compatible* situation in the open-classes model.

Incentive-compatibility. In economic terms, in the model with open classes, the users of the network are *agents*, who individually choose between the three alternatives offered, that is, joining the high-priority class, joining the low-priority class, or not using the network at all. The situation in which no user has any incentive to unilaterally change his policy is called a *Nash equilibrium* [14].

It is not obvious that by making high-priority transfer more expensive than low-priority transfer the voice customers will use the high-priority class and the data customers will use the low priority class; this strongly depends on the price difference between the queues, and the delay performance of both queues. This statement can be made more precise as follows. Let for both types of traffic the *mean* delay determine the utility experienced by the users. Now the utility curves for data and voice are denoted by $u_d(\cdot)$ and $u_v(\cdot)$, respectively, and are decreasing in their argument, i.e., the mean delay. Clearly, this mean delay is affected by the number of customers of both types who join both service classes. Suppose that data (voice) customers are assigned to the low-(high-)priority class, leading to mean delays $\mathbb{E}D_L$ and $\mathbb{E}D_H$, respectively. Assume that customers are 'infinitely divisible', i.e., we do not restrict ourselves to integer numbers of customers. Then we have a Nash equilibrium if

$$u_d(\mathbb{E}D_L) - p_L \geqslant \max\{u_d(\mathbb{E}D_H) - p_H, 0\};$$
$$u_v(\mathbb{E}D_H) - p_H \geqslant \max\{u_v(\mathbb{E}D_L) - p_L, 0\}. \quad (1)$$

Literature. The problems of price selection and incentive-compatibility in priority queues were dealt with in Mendelson and Whang [10]. They consider the special case in which the penalty functions—which can be interpreted as minus the utility functions—are *linear* in the mean delays. Conditions (1) become

$$v_d \cdot \mathbb{E}D_L + p_L \leqslant \min\{v_d \cdot \mathbb{E}D_H + p_H, 0\};$$
$$v_v \cdot \mathbb{E}D_H + p_H \leqslant \min\{v_v \cdot \mathbb{E}D_L + p_L, 0\}.$$

In [10] prices are derived which are optimal and incentive compatible: the prices maximize the system's 'net value', where the choice what class to join is left to the individual users (and the solution is a Nash equilibrium). Importantly, [10] shows that *the optima for dedicated classes and open classes coincide.*

We believe that some aspects of the model of [10] do not apply to the situation of competing data and voice users described above. In the first place, clearly the choice of the penalty functions in [10] is restrictive. As argued above, for low values