



Pricing strategies and service differentiation

Michel Mandjes *

*CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands, and
University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands
E-mail: michel@cw.nl*

Received 13 May 2002; accepted 18 December 2003

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).

Keywords: congestion, differentiated services, game theory, microeconomics, negative externalities, packet networks, pricing

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. Shenker [12] further motivates this prioritization and related design issues for the Internet.

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

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 [12] 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 [12]), 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* [13].

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. Then we have a Nash equilibrium if

$$\begin{aligned} u_d(\mathbb{E}D_L) - p_L &\geq \max\{u_d(\mathbb{E}D_H) - p_H, 0\}; \\ u_v(\mathbb{E}D_H) - p_H &\geq \max\{u_v(\mathbb{E}D_L) - p_L, 0\}. \end{aligned} \tag{1}$$

Literature. The problems of price selection and incentive-compatibility in priority queues were dealt with in [9]. 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

$$\begin{aligned} v_d \cdot \mathbb{E}D_L + p_L &\leq \min\{v_d \cdot \mathbb{E}D_H + p_H, 0\}; \\ v_v \cdot \mathbb{E}D_H + p_H &\leq \min\{v_v \cdot \mathbb{E}D_L + p_L, 0\}. \end{aligned}$$

In [9] 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, [9] shows that *the optima for dedicated classes and open classes coincide*.

We believe that some aspects of the model of [9] 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 [9] is restrictive. As argued above, for low values of the delay the delay-sensitive voice users have a higher utility than the delay-tolerant, whereas for high delay the opposite holds. This cannot be modeled in the framework of [9], as it is not clear whether v_d should be larger than v_v or vice versa. In other words, the utility curves (and hence the penalty functions) should not have a monotonous relation: they should *intersect*.

Another interesting approach to service differentiation can be found in [10,11]: he proposes to offer multiple qualities by using multiple logically separated networks with different prices. The idea is that the expensive network attracts the delay-sensitive users, whereas the delay-tolerant users opt for the cheap network, see also [2]. Principles behind congestion pricing are given in, e.g., [3,6]. There are many references with more practical reflections on pricing in multiservice networks, see, for instance, [1,14] and several articles in [7].

Contribution and organization. This paper looks at the situation in which the utility curves *do* intersect: for $\mathbb{E}D \in (0, 1)$ it holds that $u_v(\mathbb{E}D) > u_d(\mathbb{E}D)$, whereas for $\mathbb{E}D > 1$ the opposite holds: $u_d(\mathbb{E}D) > u_v(\mathbb{E}D)$. We first consider voice and data users sharing a FIFO queue. In *section 2* we see that, depending on the value of the link speed μ , the network population will consist of just one class: data users for small μ , and voice users for large μ . We focus on prices that maximize the provider's profit, which is slightly different from the 'net value' maximization problem solved in [8,9] (cf. social welfare maximization).

An important conclusion of our paper is that under our utility curves the solutions of the open-classes model and the dedicated-classes model do *not* coincide (which *did* hold in the setting of [9]). Section 3 analyzes the profit maximization problem for the model with dedicated classes, whereas section 4 focuses on the situation with open classes. As could be expected, section 4 is more involved, since the customers have more options. Again, depending on the link rate μ , different regimes are optimal. The theory is illustrated by a numerical example. Section 5 concludes and summarizes.

2. No service differentiation – tragedy of the commons

Data and voice users – utility. Consider a system with an infinite population of (potential) customers. The *utility* they get depends on the level of *congestion*. Obviously, generally speaking, the larger the number of users in the network, the lower the utility. Throughout this paper we will use the *mean packet delay*, $\mathbb{E}D$, as the measure of congestion.

The price per packet transmission is p . Customers want to use the service as long as utility minus price – or *compensated* utility – is positive. When customers join the level of congestion increases. In other words, customers join as long as the compensated utility is positive, cf. [8]. A complication is that we have two types of users. In the first place there are users who strongly prefer low congestion or, equivalently, low packet delay. We will refer to these users as to *voice users*. On the other hand, there are users who do not mind so much about the delay: compared to voice users, they assign less utility to low delay, but more utility to high delay. We call these customers *data users*. To model these specific preferences, we define the (compensated) utility curves of both types of users by

$$\begin{aligned} U_d(\mathbb{E}D) &:= u_d(\mathbb{E}D) - p, & \text{with } u_d(y) &:= y^{-\alpha_d}; \\ U_v(\mathbb{E}D) &:= u_v(\mathbb{E}D) - p, & \text{with } u_v(y) &:= y^{-\alpha_v}, \end{aligned}$$

with $0 < \alpha_d < \alpha_v$. Notice that both expression are equal for $\mathbb{E}D = 1$.

A system without service differentiation. Both data and voice users generate information packets that they feed into the system. Each data (voice) user generates packets at rate λ_d (λ_v , respectively). In this section we let both types of customers use a single server queue that does not make any distinction between the packets of both sorts, a FIFO queue. We assume that the service times of the individual packets are i.i.d. exponentially distributed random variables, with mean μ^{-1} .

In an M/M/1 queue, with N (independent) customers that generate packets according to a Poisson process with rate λ , and service times that are i.i.d. exponential with mean μ^{-1} , the mean delay is

$$\mathbb{E}D = \frac{1}{\mu - \lambda N},$$

provided that $\lambda N < \mu$ [5]. We now compute how many users of each type will subscribe to the network, as a function of the packet transmission price p .

Equilibrium for fixed price. Consider first two hypothetical cases.

- Suppose there are only data users. They enter as long as their (compensated) utility is non-negative. For simplicity, we do not restrict ourselves to an integer number of customers. It is not hard to show that this number equals

$$N_d(p) = \frac{\mu - \alpha_d \sqrt{p}}{\lambda_d}. \quad (2)$$

This holds if $p < \mu^{\alpha_d}$; otherwise $N_d(p) = 0$.

- Similarly, with only voice users,

$$N_v(p) = \frac{\mu - \alpha_v \sqrt{p}}{\lambda_v}.$$

This holds if $p < \mu^{\alpha_v}$; otherwise $N_v(p) = 0$.

Now consider the situation that both groups are competing for service. Suppose $N_d(p)$ customers are present, with $N_d(p)$ given by (2). We may ask ourselves if there is any incentive for voice users to join? Notice that the utility an infinitesimally small voice user would experience is

$$U_v := \left(\mu - \lambda_d \left(\frac{\mu - \alpha_d \sqrt{p}}{\lambda_d} \right) \right)^{\alpha_v} - p = p^{\alpha_v/\alpha_d} - p.$$

Using that $\alpha_v > \alpha_d$, it is easily seen that if $p > 1$ this number is positive, so voice users would join. If $p < 1$ there is no incentive for voice users to enter when $N_d(p)$ data users are present. Conversely, if $N_v(p)$ voice customers are present, data users join if and only if $p < 1$. In fact, we have found a *Nash equilibrium* [13].

Tragedy of the commons. From the above, we conclude that if prices are low, data users dominate over voice users; the opposite happens when prices are high.

This describes, albeit it in a stylized sense, the current situation in the Internet. Prices are low, or, more precisely, there is a usually a flat fee, i.e., the amount of money charged does not depend on usage. Customers who require low packet delay (voice) are excluded. In fact, so many delay-indifferent users join, that the congestion is unacceptably high for the delay-averse users. This phenomenon is commonly referred to as the *tragedy of the commons* [4].

The price selection problem. The network operator will choose the price such that his profit is maximized. The customers pay for every packet they transmit. We define *profit* Π as the expected number of packets sent (by the users who subscribe to the network) per unit time, multiplied by the price per packet p . From the above,

$$\Pi(p, \mu) \equiv \begin{cases} \lambda_d N_d(p) p = f_d(p) := (\mu - \alpha_d \sqrt{p}) p & \text{if } p \in (0, 1]; \\ \lambda_v N_v(p) p = f_v(p) := (\mu - \alpha_v \sqrt{p}) p & \text{if } p \in (1, \infty). \end{cases}$$

Hence the provider wishes to maximize $\Pi(p, \mu)$ over $p \geq 0$.

Proposition 2.1. The profit function is given by

$$\Pi^*(\mu) := \max_{p>0} \Pi(p, \mu) = \max \left\{ \left(\frac{\mu \alpha_d}{\alpha_d + 1} \right)^{\alpha_d} \left(\frac{\mu}{\alpha_d + 1} \right), \left(\frac{\mu \alpha_v}{\alpha_v + 1} \right)^{\alpha_v} \left(\frac{\mu}{\alpha_v + 1} \right) \right\}.$$

Proof. We prove this proposition in two steps.

Step 1. We first derive an elementary expression for the profit as a function of service rate μ .

- It is not hard to verify that, on $p \in \mathbb{R}_+$, the function $f_d(p)$ attains its maximum at

$$p_d := \left(\frac{\mu \alpha_d}{\alpha_d + 1} \right)^{\alpha_d}.$$

Notice that p_d is indeed smaller than μ^{α_d} , as desired. Hence, with $\mu_d := 1 + \alpha_d^{-1}$, the maximum value of $\Pi(p, \mu)$ on $p \in [0, 1]$ is

$$\begin{aligned} & \max_{p \in [0, 1]} \Pi(p, \mu) \\ &= \begin{cases} f_d(\mu) := \left(\frac{\mu \alpha_d}{\alpha_d + 1} \right)^{\alpha_d} \left(\frac{\mu}{\alpha_d + 1} \right) & \text{if } \mu < \mu_d, \text{ attained at } p = p_d; \\ \mu - 1 & \text{otherwise, attained at } p = 1. \end{cases} \end{aligned}$$

- Similarly, with $\mu_v := 1 + \alpha_v^{-1}$, the maximum value of $\Pi(p, \mu)$ on $p \in [1, \infty)$ is

$$\begin{aligned} & \max_{p \in [1, \infty)} \Pi(p, \mu) \\ &= \begin{cases} f_v(\mu) := \left(\frac{\mu \alpha_v}{\alpha_v + 1} \right)^{\alpha_v} \left(\frac{\mu}{\alpha_v + 1} \right) & \text{if } \mu < \mu_v, \text{ attained at } p = p_v; \\ \max\{0, \mu - 1\} & \text{otherwise, attained at } p = 1. \end{cases} \end{aligned}$$

Recall that $\mu_v < \mu_d$; we get that $\Pi^*(\mu) = \max_{p > 0} \Pi(p, \mu)$ equals

$$g(\mu) := \begin{cases} \max\{\mu - 1, f_d(\mu)\} & \text{if } 0 \leq \mu \leq \mu_v; \\ \max\{f_d(\mu), f_v(\mu)\} & \text{if } \mu_v \leq \mu \leq \mu_d; \\ \max\{0, \mu - 1, f_v(\mu)\} & \text{if } \mu \geq \mu_d. \end{cases}$$

Step 2. We now prove the following two properties.

- It is trivial to show that $f_d(\mu_d) = \mu_d - 1$. Also,

$$f'_d(\mu) = \left(\frac{\mu \alpha_d}{\alpha_d + 1} \right)^{\alpha_d} < 1$$

on $[0, \mu_d)$. So both curves cannot intersect. This proves that $f_d(\mu) \geq \mu - 1$ for $\mu \in [0, \mu_d]$.

- Analogously, $f_v(\mu) \geq \mu - 1$ for $\mu \in [\mu_v, \infty)$.

We arrive at $\Pi^*(\mu) = g(\mu) = \max\{f_d(\mu), f_v(\mu)\}$. This completes our proof. \square

The following corollary states that for small (large) link rates data users (voice users, respectively) dominate, and identifies the threshold service rate μ^* .

Corollary 2.2. With $\mu^* \in [\mu_v, \mu_d]$ defined by

$$\mu^* := \left(\left(\frac{\alpha_d}{\alpha_d + 1} \right)^{\alpha_d} \left(\frac{\alpha_v + 1}{\alpha_v} \right)^{\alpha_v} \frac{\alpha_v + 1}{\alpha_d + 1} \right)^{1/(\alpha_v - \alpha_d)},$$

for all

- $\mu < \mu^*$ it holds that $f_d(\mu) > f_v(\mu)$. This implies that $\lambda_d N_d = \mu / (\alpha_d + 1) > 0$ and $N_v = 0$, and the price per packet transmission p equals $p_d < 1$;
- $\mu > \mu^*$ it holds that $f_d(\mu) < f_v(\mu)$. This implies that $N_d = 0$ and $\lambda_v N_v = \mu / (\alpha_v + 1) > 0$, and the price per packet transmission p equals $p_v > 1$.

3. Service differentiation by priority queueing: dedicated classes

In the previous section we concluded that – in case of heterogeneous traffic classes – the network will serve only one of them. In this section we concentrate on ways to satisfy the demands of both classes. Adhering to the principles explained in [12], we do this by using a *priority queueing system*. We will argue that this solution is beneficial for the network (as its profit increases compared to the FIFO solution), the dominating class (as the service will be offered against a lower price), and the excluded class (as it will receive service).

A priority queueing model; dedicated and open classes. Let us assume that we are in the regime that $\mu < \mu^*$, so in a FIFO system the voice users would not get any service. Now consider a two-queue system, as depicted in figure 1. We assume that the voice packets are directed to the high priority queue, and the data packets to the low priority queue. We now suppose that voice get strict service priority over data: data traffic is served only when there is no voice traffic in the system. As identified in the Introduction, this is the *dedicated-classes* model; this is in contrast with the *open-classes* model, in which the customers themselves choose the most attractive queue (based on the expected delays in both queues and the respective prices). We return to the issue of dedicated and open classes in section 4.

Standard queueing theory [5] gives that the mean packet delay for both classes is given by

$$\mathbb{E}D_v = \frac{1}{\mu - \lambda_v N_v} \quad \text{and} \quad \mathbb{E}D_d = \frac{\mu}{(\mu - \lambda_v N_v)(\mu - \lambda_v N_v - \lambda_d N_d)}.$$

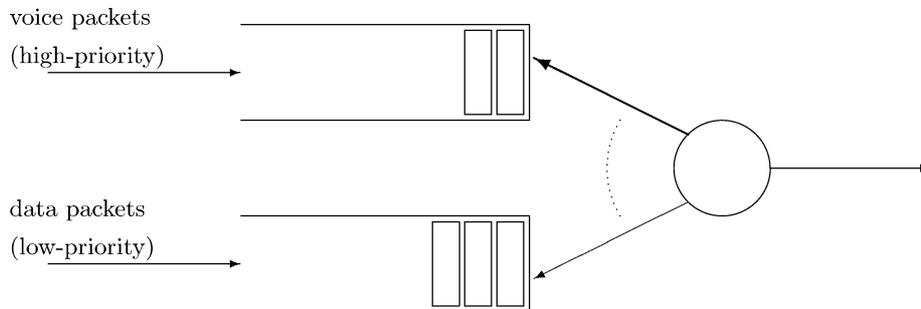


Figure 1. Priority system; voice has strict priority over data.

Here we assume that the service of a data packet can be interrupted when voice packets arrive; the service is resumed as soon as the high-priority queue gets empty.

Equilibrium for fixed price. Suppose a packet in the high priority queue is charged an amount p_H , and a packet in the low priority queue p_L . Clearly, as seen in section 2, the number of voice users joining is given by

$$N_v(p_L, p_H) = \frac{\mu - \alpha_v \sqrt[p_H]{p_H}}{\lambda_v}$$

if $p_H < \mu^{\alpha_v}$ and 0 otherwise. Similarly, data users join as long as their compensated utility exceeds 0. This results in

$$N_d(p_L, p_H) = \begin{cases} \lambda_d^{-1} \left(\alpha_v \sqrt[p_H]{p_H} - \frac{\mu \alpha_d \sqrt[p_L]{p_L}}{\alpha_v \sqrt[p_H]{p_H}} \right) & \text{if } p_L < \frac{p_H^{2\alpha_d/\alpha_v}}{\mu^{\alpha_d}} \text{ and } p_H \leq \mu^{\alpha_v}; \\ \lambda_d^{-1} (\mu - \alpha_d \sqrt[p_L]{p_L}) & \text{if } p_L < \mu^{\alpha_d} \text{ and } p_H > \mu^{\alpha_v}; \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Notice that $N_d(p_L, p_H)$ decreases in p_L and increases in p_H , as expected.

The price selection problem. Again the provider wants to achieve maximum profit. Notice that the priority system cannot lead to lower profits than the FIFO system. The reason for this is that the FIFO queue is a special case of the priority queue – this is seen by taking $p_H \equiv \mu^{\alpha_v}$ or $p_L = p_H^{2\alpha_d/\alpha_v} / \mu^{\alpha_d}$.

To obtain the optimal prices, we have to perform the following optimization:

$$\Pi_D^*(\mu) = \max_{p_L > 0, p_H > 0} \Pi_D(p_L, p_H, \mu),$$

with

$$\Pi_D(p_L, p_H, \mu) := \lambda_d N_d(p_L, p_H) p_L + \lambda_v N_v(p_L, p_H) p_H,$$

where the subscript ‘D’ denotes the regime of dedicated classes. Let us for the moment assume that both services are in a regime in which customers get service. We get

$$\max_{p_L > 0, p_H > 0} (\mu - \alpha_v \sqrt[p_H]{p_H}) p_H + \left(\alpha_v \sqrt[p_H]{p_H} - \mu \frac{\alpha_d \sqrt[p_L]{p_L}}{\alpha_v \sqrt[p_H]{p_H}} \right). \quad (4)$$

We compute this maximum in two steps. First, we find the optimizing value of p_L for given p_H . Subsequently, we maximize over p_H .

Step 1. First, find the optimal p_L for a given value of p_H . Differentiation to p_L and equating to 0 yields

$$p_L(p_H) = \left(\frac{\alpha_d}{\mu(\alpha_d + 1)} \right)^{\alpha_d} p_H^{2\alpha_d/\alpha_v},$$

which is indeed smaller than $p_H^{2\alpha_d/\alpha_v}/\mu^{\alpha_d}$. Directly from (3) and (4), we get that $\Pi_d^*(\mu)$ equals

$$\max_{0 < p_H \leq \mu^{\alpha_v}} g(p_H) \text{ with } g(p) := (\mu - \sqrt[\alpha_v]{p})p + \left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} \frac{1}{\alpha_d + 1} p^{(2\alpha_d+1)/\alpha_v}. \quad (5)$$

Step 2. Now we find the profit-maximizing value of p_H . It is straightforward that $g(0) = 0$ and $g'(0) > 0$. Moreover,

$$\begin{aligned} g'(p) &= \mu - \frac{\alpha_v + 1}{\alpha_v} \sqrt[\alpha_v]{p} + \left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} \frac{2\alpha_d + 1}{\alpha_d + 1} \cdot \frac{1}{\alpha_v} p^{(2\alpha_d+1)/\alpha_v-1}; \\ g''(p) &= -\frac{\alpha_v + 1}{\alpha_v} \cdot \frac{1}{\alpha_v} p^{1/\alpha_v-1} \\ &\quad + \left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} \frac{2\alpha_d + 1}{\alpha_d + 1} \cdot \frac{2\alpha_d + \alpha_v + 1}{\alpha_v} \cdot \frac{1}{\alpha_v} p^{(2\alpha_d+1)/\alpha_v-2}. \end{aligned}$$

It is not hard to verify that the function $g''(\cdot)$ changes sign at

$$\bar{p} := \left(\left(\frac{\mu(\alpha_d + 1)}{\alpha_d} \right)^{\alpha_d} \frac{\alpha_d + 1}{2\alpha_d + 1} \cdot \frac{\alpha_v + 1}{2\alpha_d - \alpha_v + 1} \right)^{\alpha_v/(2\alpha_d - \alpha_v)}$$

if $2\alpha_d + 1 > \alpha_v$; if $2\alpha_d + 1 \leq \alpha_v$ there is not such a point. More detailed inspection yields the following corollary.

Corollary 3.1. The function $g(\cdot)$, as defined in (5), increases in the origin. Also,

- if $2\alpha_d < \alpha_v < 2\alpha_d + 1$ the function $g(\cdot)$ shifts from convexity to concavity at \bar{p} ;
- if $2\alpha_d > \alpha_v$ the function $g(\cdot)$ shifts from concavity to convexity at \bar{p} ;
- if $2\alpha_d + 1 \leq \alpha_v$ the function $g(\cdot)$ is concave on $[0, \infty)$.

We are now in a position to characterize the optimizing p_H ; we do this in lemmas 3.2 and 3.3. We first define

$$\mu_-^* := \left(\left(\frac{\alpha_d}{\alpha_d + 1} \right)^{\alpha_d} \frac{2\alpha_d + 1}{\alpha_d + 1} \right)^{1/(\alpha_v - \alpha_d)}.$$

It is easy to verify that $\mu_-^* < \mu^*$ and that $\mu_-^* < 1$.

Lemma 3.2. For $\mu \in (\mu_-^*, \mu^*)$ function $g(\cdot)$ is first increasing and then decreasing on the interval $p \in [0, \mu^{\alpha_v}]$.

Proof. Applying corollary 3.1, it suffices to show that $g'(\mu^{\alpha_v}) < 0$ for $\mu \in (\mu_-^*, \mu^*)$. This is a matter of straightforward calculus. \square

Lemma 3.3. For $\mu \in (0, \mu_-^*)$ the function $g(\cdot)$ is nondecreasing on the interval $p \in [0, \mu^{\alpha_v}]$.

Proof. We prove this lemma by considering the cases that $2\alpha_d$ is smaller and larger than α_v , separately.

First, observe that for $2\alpha_d < \alpha_v$, corollary 3.1 entails that $g'(\mu^{\alpha_v}) > 0$ for $\mu \in (0, \mu_-^*)$ implies the stated. This is easy to verify.

Now consider $2\alpha_d > \alpha_v$. Write for ease $p \equiv \beta^{\alpha_v} \mu^{\alpha_v}$. We have to show that $g'(\beta^{\alpha_v} \mu^{\alpha_v}) \geq 0$ for all $\beta \in [0, 1]$. Elementary calculations give that an equivalent condition is

$$(\beta - (1 - \beta)\alpha_v)\mu^{\alpha_v - \alpha_d} \leq \left(\frac{\alpha_d}{\alpha_d + 1}\right)^{\alpha_d} \frac{2\alpha_d + 1}{\alpha_d + 1} \beta^{2\alpha_d - \alpha_v + 1}, \quad (6)$$

for all $\beta \in [0, 1]$ and $\mu \in (0, \mu_-^*)$. The stated is clearly true for $\beta < \beta_v := \alpha_v / (\alpha_v + 1)$; in this case the left hand side of (6) is negative, whereas the right hand side is positive.

Now concentrate on $\beta \in [\beta_v, 1]$. Because the left-hand side of condition (6) is increasing in μ , we have to verify it only for $\mu = \mu_-^*$. For this value of μ , the condition reduces to

$$\eta(\beta) := (\beta - (1 - \beta)\alpha_v) - \beta^{2\alpha_d - \alpha_v + 1} \leq 0.$$

As $\eta(\beta_v) = -\beta_v^{2\alpha_d - \alpha_v + 1} < 0$ and $\eta(1) = 0$, it is sufficient to prove that $\eta'(\beta) \geq 0$ for $\beta \in [\beta_v, 1]$. Since $2\alpha_d > \alpha_v$,

$$\begin{aligned} \eta'(\beta) &= 1 + \alpha_v - (2\alpha_d - \alpha_v + 1)\beta^{2\alpha_d - \alpha_v} \\ &\geq 1 + \alpha_v - (2\alpha_d - \alpha_v + 1) = 2(\alpha_v - \alpha_d) > 0. \end{aligned}$$

This proves the lemma. \square

The following proposition follows immediately from the lemmas 3.2 and 3.3.

Proposition 3.4. Assume $\mu \in (0, \mu^*)$ and suppose that the provider can prioritize voice. We distinguish between two cases.

- $\mu \in (0, \mu_-^*)$: A FIFO queue is optimal for the provider. Only data users enter. On the interval $[0, \mu^{\alpha_v}]$, the function $g(\cdot)$ attains its maximum at the upper limit μ^{α_v} . The profit-maximizing prices are

$$p_H := \mu^{\alpha_v} \quad \text{and} \quad p_L := \left(\frac{\mu\alpha_d}{\alpha_d + 1}\right)^{\alpha_d}.$$

- $\mu \in (\mu_-^*, \mu^*)$: The provider gives voice priority over data. Both types of users enter. On the interval $[0, \mu^{\alpha_v}]$, the function $g(\cdot)$ attains its maximum in the interior; there is a unique $\bar{p}_H \in [0, \mu^{\alpha_v}]$ with $g'(\bar{p}_H) = 0$. The profit-maximizing prices are

$$p_H := \bar{p}_H \quad \text{and} \quad p_L := \left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} \bar{p}_H^{-2\alpha_d/\alpha_v}.$$

The proposition implies that for $\mu \in (0, \mu_-^*)$ the provider maximizes profit by having just a FIFO queue. Prices will be relatively low, so that only data users enter the

system. In fact, the system is so slow that prioritizing voice does not help increasing the provider's profit. For $\mu \in (\mu_*, \mu^*)$ profit is increased by giving voice priority over data.

A similar analysis can be done for the situation in which voice is dominant, i.e., $\mu > \mu^*$. Again we find that it is not always beneficial to prioritize traffic: for very fast link rates a 'voice-only solution' generates higher profit; there is a threshold link speed μ_+ .

4. Service differentiation by priority queueing: open classes

In the previous section, an essential assumption was that the network (i.e., the provider) selects the queue for both types of users; more specifically: the voice customers are forced to use the high-priority queue, whereas the data users are directed to the low-priority queue. In other words: we focused on the situation of *dedicated classes*.

The opposite situation relates to *open classes*. There the provider offers a network with a certain queueing discipline and prices, and the customers have to decide themselves what class to join. In the situation of a priority queue, the customers can select the queue (or decide not to join any queue at all) based on the prices of high priority and low priority p_H and p_L , and the expected quality of service (i.e., delay): they select the queue with the highest compensated utility.

It is easy to check that if $p_H < 1$ the high priority queue will be used exclusively by data users, and if $p_H > 1$ by voice users; the same holds for the low priority queue. The procedure of section 3 does not guarantee that $p_L < 1$ and $p_H > 1$. For that reason, if the customers were to choose the most attractive queue themselves, the solution of proposition 3.4 would not persist. Put in a game-theoretic language [13]: unilateral changes may lead to increase of the compensated utility, and for that reason the solution is possibly not a Nash equilibrium. If for instance both p_L and p_H are smaller than 1, in the model with open classes, the data users would drive away the voice users from both queues.

Equilibrium for fixed price. Like in sections 2 and 3, we first analyze the network population for given prices. For all combinations of prices (p_L, p_H) we analyze the type of equilibrium.

- R_- . Trivially if both prices are smaller (larger) than 1, the network will be populated exclusively by data (voice) users. Hence in this regime both types of users will not coexist in the system. If both prices are smaller than 1, the profit function reads

$$\Pi_O(p_L, p_H, \mu) := (\mu - \alpha \sqrt[p_H]{p_H}) p_H + \left(\alpha \sqrt[p_H]{p_H} - \mu \sqrt[p_H]{\frac{p_L}{p_H}} \right) p_H,$$

to be maximized over the set R_- with prices $p_L < 1$ and $p_H < 1$ such that $p_L \leq p_H^2 / \mu^{\alpha_d}$ and $p_H \leq \mu^{\alpha_d}$. The subscript O refers to the situation of open classes.

- R_+ . If both prices are larger than 1, α_d is replaced by α_v :

$$\Pi_O(p_L, p_H, \mu) := (\mu - \alpha_v \sqrt{\alpha_v p_H}) p_H + \left(\alpha_v \sqrt{\alpha_v p_H} - \mu \sqrt{\frac{p_L}{p_H}} \right) p_H,$$

to be maximized over R_+ with prices $p_L > 1$ and $p_H > 1$ such that $p_L \leq p_H^2 \mu^{\alpha_v}$ and $p_H \leq \mu^{\alpha_v}$.

- Now consider the situation in which $p_L > 1$ and $p_H < 1$. Hence, the low-priority queue will be used by voice customers, and the high-priority queue by data customers. Similarly to the analysis of section 3, both types of users are present if

$$p_H < \mu^{\alpha_d} \quad \text{and} \quad p_L < \frac{p_H^{2\alpha_v/\alpha_d}}{\mu^{\alpha_v}}. \quad (7)$$

If μ is smaller than 1, suppose that the first condition in (7) is met. Then it is easy to conclude that the second requirement is violated:

$$\frac{p_H^{2\alpha_v/\alpha_d}}{\mu^{\alpha_v}} < \frac{\mu^{2\alpha_v}}{\mu^{\alpha_v}} = \mu^{\alpha_v} < 1.$$

If μ is larger than 1, the first condition in (7) is automatically satisfied, whereas the second is violated:

$$\frac{p_H^{2\alpha_v/\alpha_d}}{\mu^{\alpha_v}} < \mu^{-\alpha_v} < 1.$$

- R_0 . The remaining regime is $p_L < 1$ and $p_H > 1$. It is not hard to verify that in this case the an equilibrium is possible in which voice users (in the high-priority queue) and data users (in the low-priority queue) coexist only $\mu > 1$. We have to maximize

$$\Pi_O(p_L, p_H, \mu) := (\mu - \alpha_v \sqrt{\alpha_v p_H}) p_H + \left(\alpha_v \sqrt{\alpha_v p_H} - \mu \frac{\alpha_d \sqrt{p_L}}{\alpha_v \sqrt{p_H}} \right) p_L,$$

over a region R_0 that is given by

$$p_L \in \left(0, \min \left\{ 1, \frac{p_H^{2\alpha_d/\alpha_v}}{\mu^{\alpha_d}} \right\} \right), \quad p_H \in (1, \mu^{\alpha_v}). \quad (8)$$

The resulting admissible region is depicted in figure 2, with the R_- , R_+ , and R_0 defined above. Notice that R_0 and R_+ are empty for $\mu < 1$; then a situation with only data users (R_-) is the only persistent solution.

The price selection problem. From the above, it is clear that we have to evaluate

$$\Pi_O^*(\mu) := \max \{ \Pi_{O,-}^*(\mu), \Pi_{O,+}^*(\mu), \Pi_{O,0}^*(\mu) \},$$

with $\Pi_{O,i}^*(\mu) := \max_{(p_L, p_H) \in R_i} \Pi_O(p_L, p_H, \mu)$, for $i \in \{-, +, 0\}$. We now compute these three maxima subsequently, in sections 4.1–4.3. First, an auxiliary result is proven in lemma 4.1.

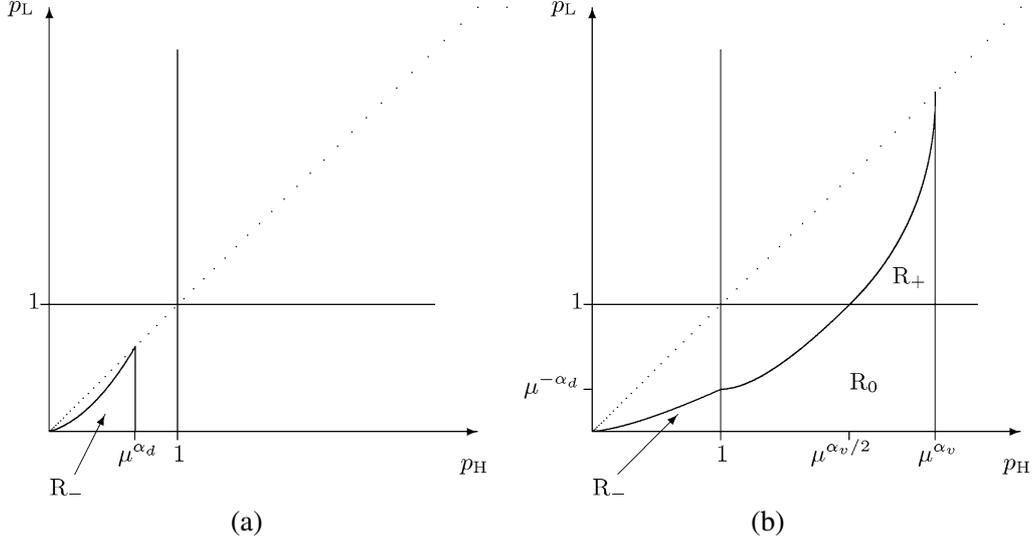


Figure 2. Admissible region: (a) corresponds to $\mu < 1$, (b) to $\mu > 1$.

Lemma 4.1. $\Pi_O^*(\mu)$ is nondecreasing and convex in μ .

Proof. It suffices to prove that the $\Pi_{O,i}^*(\mu)$ are nondecreasing and convex in μ , $i \in \{-, +, 0\}$.

- First, notice that $\Pi_O(p_L, p_H, \mu)$ is a linear function of μ .
- The $\Pi_O(p_L, p_H, \mu)$ are nondecreasing in μ . This is seen as follows for R_0 (a similar reasoning applies to R_- and R_+). In R_0 the coefficient of μ is given by

$$p_H - \frac{\alpha_d \sqrt[p_L]{p_L}}{\alpha_v \sqrt[p_H]{p_H}} p_L \geq p_H - p_H^{\alpha_d/\alpha_v} > 0,$$

as follows from $p_L \leq p_H^{2\alpha_d/\alpha_v} / \mu^{\alpha_d} = p_H^{\alpha_d/\alpha_v} (p_H^{\alpha_d/\alpha_v} / \mu^{\alpha_v}) \leq p_H^{\alpha_d/\alpha_v}$, in conjunction with $p_H \geq 1$.

As the $\Pi_{O,i}^*(\mu)$ are maxima (over $(p_L, p_H) \in R_i$) of nondecreasing, linear (and hence convex) functions, they are nondecreasing and convex as well. \square

4.1. Maximization over R_-

The maximum over R_- reduces to maximizing $g_-(p_H)$ over $0 < p_H \leq \min\{\mu^{\alpha_d}, 1\}$, where

$$g_-(p) := (\mu - p^{1/\alpha_d})p + \left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} \frac{1}{\alpha_d + 1} p^{2+1/\alpha_d}. \tag{9}$$

The price for the low priority service is given by

$$p_L(p_H) = \beta_d(\mu) p_H^2, \quad \text{with } \beta_d(\mu) := \left(\frac{\alpha_d}{\mu(\alpha_d + 1)} \right)^{\alpha_d}.$$

With $\zeta_-(\mu) := g'_-(1)$, it is shown in appendix A.1 that

- A₁. For $\mu \in (0, 1]$ the optimum over \mathbb{R}_- is reached at a price \bar{p}_H in the interior of $(0, \mu^{\alpha_d}]$; this \bar{p}_H is the unique solution of $g'_-(p) = 0$ in $(0, \mu^{\alpha_d}]$. Also, $\bar{p}_L = \beta_d(\mu) \bar{p}_H^2$.
- A₂. Let v_- be the unique solution to $\zeta_-(\mu) = 0$ in $(1, \infty)$. For $\mu \in (1, \infty)$ the optimum over \mathbb{R}_- is reached at a price \bar{p}_H
 - (i) in the interior of $(0, 1]$ if $\mu \in (1, v_-)$; this \bar{p}_H is the unique solution of $g'_-(p) = 0$ in $(0, 1]$. Also, $\bar{p}_L = \beta_d(\mu) \bar{p}_H^2$;
 - (ii) equal to 1 if $\mu \in (v_-, \infty)$. Also, $\bar{p}_L = \beta_d(\mu)$.

4.2. Maximization over \mathbb{R}_+

The region \mathbb{R}_+ is empty if $\mu \leq 1$; we therefore concentrate on $\mu > 1$. We have that

$$p_L(p_H) = \max\{1, \beta_v(\mu) p_H^2\}, \quad \text{with } \beta_v(\mu) := \left(\frac{\alpha_v}{\mu(\alpha_v + 1)} \right)^{\alpha_v}.$$

With $\mu_v := 1 + \alpha_v^{-1}$, two cases need to be considered.

- If $\mu \in (1, \mu_v)$, it is not hard to see that the optimal p_L equals 1. The maximizing p_H should be found from

$$\max_{\mu^{\alpha_v/2} < p_H < \mu^{\alpha_v}} \bar{g}_+(p_H) \quad \text{with } \bar{g}_+(p) := (\mu - \alpha_v \sqrt{p})p + \alpha_v \sqrt{p} - \mu \frac{1}{\alpha_v \sqrt{p}}.$$

- If $\mu \in (\mu_v, \infty)$, it turns out that

$$p_L(p_H) = \left(\frac{\alpha_v}{\mu(\alpha_v + 1)} \right)^{\alpha_v} p_H^2 \quad \text{if } p_H \in [q_+(\mu), \mu^{\alpha_v}],$$

$$\text{with } q_+(\mu) := \left(\frac{\mu(\alpha_v + 1)}{\alpha_v} \right)^{\alpha_v/2},$$

and $p_L(p_H) = 1$ if $p_H \in [\mu^{\alpha_v/2}, q_+(\mu)]$. Define $g_+(\cdot)$ as in (9), but with α_d replaced by α_v . We have to solve

$$\max \left\{ \max_{\mu^{\alpha_v/2} < p_H < q_+(\mu)} \bar{g}_+(p_H), \max_{q_+(\mu) < p_H < \mu^{\alpha_v}} g_+(p_H) \right\}.$$

With $\zeta_+(\mu) := \bar{g}'_+(\mu^{\alpha_v})$ and $\xi_+(\mu) = g'_+(q_+(\mu))$, it is shown in appendix A.2 that

- B₁. Let $v_{+,1}$ be the unique solution to $\zeta_+(\mu) = 0$ in $(1, \mu_v)$. For $\mu \in (1, \mu_v)$ the optimum over \mathbb{R}_+ is reached at a price \bar{p}_H

- (i) equal to μ^{α_v} if $\mu \in (1, v_{+,1})$. Also, $\bar{p}_L = 1$;
- (ii) in the interior of $[\mu^{\alpha_v/2}, \mu^{\alpha_v}]$ if $\mu \in (v_{+,1}, \mu_v)$; this \bar{p}_H is the unique solution of $\bar{g}'_+(p) = 0$ in $[\mu^{\alpha_v/2}, \mu^{\alpha_v}]$. Also, $\bar{p}_L = 1$.
- B₂. Let $v_{+,2}$ be the unique solution to $\xi_+(\mu) = 0$ in (μ_v, ∞) . For $\mu \in (\mu_v, \infty)$ the optimum over \mathbf{R}_+ is reached at a price \bar{p}_H
 - (i) in the interior of $[\mu^{\alpha_v/2}, q_+(\mu)]$ if $\mu \in (\mu_v, v_{+,2})$; this \bar{p}_H is the unique solution of $\bar{g}'_+(p) = 0$ in $[\mu^{\alpha_v/2}, q_+(\mu)]$. Also, $\bar{p}_L = 1$;
 - (ii) in the interior of $[q_+(\mu), \mu^{\alpha_v}]$ if $\mu \in (v_{+,2}, \infty)$; this \bar{p}_H is the unique solution of $g'_+(p) = 0$ in $[q_+(\mu), \mu^{\alpha_v}]$. Also $\bar{p}_L = \beta_v(\mu)\bar{p}_H^2$.

4.3. Maximization over \mathbf{R}_0

Again we first perform the optimization over p_L for given p_H . It is straightforward to obtain that the optimum is attained at

$$p_L(p_H) = \min\{1, \beta_d(\mu)p_H^{2\alpha_d/\alpha_v}\}.$$

Defining $\mu_d := 1 + \alpha_d^{-1}$, two cases need to be distinguished.

- If $\mu \leq \mu_d$, it is not hard to verify that it holds that

$$\left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} p_H^{2\alpha_d/\alpha_v} \leq 1 \quad \text{for all } p_H \in (1, \mu^{\alpha_v}],$$

so that, with $g_0(\cdot)$ defined as $g(\cdot)$ in (5), the optimization reduces to

$$\max_{1 < p_H \leq \mu^{\alpha_v}} g_0(p_H).$$

- If $\mu \in (\mu_d, \infty)$, we have that

$$\beta_d(\mu)p_H^{2\alpha_d/\alpha_v} \leq 1 \quad \text{if and only if} \quad p_H \leq q_0(\mu) := \left(\frac{\mu(\alpha_d + 1)}{\alpha_d}\right)^{\alpha_v/2}.$$

Notice that $q_0(\mu)$ is smaller than μ^{α_v} (as $\mu_d = (\alpha_d + 1)/\alpha_d < \mu$). So we get the optimization

$$\max\left\{\max_{1 < p_H \leq q_0(\mu)} g_0(p_H), \max_{q_0(\mu) < p_H \leq \mu^{\alpha_v}} \bar{g}_0(p_H)\right\},$$

where $\bar{g}_0(\cdot) = \bar{g}_+(\cdot)$.

With $\zeta_0(\mu) := g'_0(1)$ and $\xi_0(\mu) = g'_0(q_0(\mu))$, it is shown in appendix A.3 that

- C₁. Let $v_{0,1}$ be the unique solution to $\zeta_0(\mu) = 0$ in $(1, \mu_d)$. For $\mu \in (1, \mu_d)$ the optimum over \mathbf{R}_0 is reached at a price \bar{p}_H
 - (i) equal to 1 if $\mu \in (1, v_{0,1})$. Also, $\bar{p}_L = \beta_d(\mu)$;

- (ii) in the interior of $[1, \mu^{\alpha_v}]$ if $\mu \in (\nu_{0,1}, \mu_d)$; this \bar{p}_H is the unique solution of $g'_0(p) = 0$ in $[1, \mu^{\alpha_v}]$. Also, $\bar{p}_L = \beta_d(\mu) \bar{p}_H^{2\alpha_d/\alpha_v}$.
- C_2 . Let $\nu_{0,2}$ be the unique solution to $\xi_0(\mu) = 0$ in (μ_d, ∞) . For $\mu \in (\mu_d, \infty)$ optimum over R_0 is reached at a price \bar{p}_H
 - (i) in the interior of $[1, q_0(\mu)]$ if $\mu \in (\mu_d, \nu_{0,2})$; this \bar{p}_H is the unique solution of $g'_0(p) = 0$ in $[1, q_0(\mu)]$. Also, $\bar{p}_L = \beta_d(\mu) \bar{p}_H^{2\alpha_d/\alpha_v}$;
 - (ii) in the interior of $[q_0(\mu), \mu^{\alpha_v}]$ if $\mu \in (\nu_{0,2}, \infty)$; this \bar{p}_H is the unique solution of $\bar{g}'_0(p) = 0$ in $[q_0(\mu), \mu^{\alpha_v}]$. Also, $\bar{p}_L = 1$.

Characterization of the solution. We are now in a position to prove that there are two possible situations. In the first there are service rates ν_-^* and ν_+^* such that (for the profit-maximizing prices) voice will dominate in the network for all $\mu < \nu_-^*$, data will dominate for $\mu > \nu_+^*$, and there is a ‘mixed scenario’ (with priority for voice) for $\mu \in (\nu_-^*, \nu_+^*)$. The second possibility data dominates for μ smaller than some ν^* , and voice dominates otherwise.

Theorem 4.2. For $\mu < \nu_{\min} := \min\{\nu_{0,1}, \nu_{+,2}\}$, ‘data-only’ maximizes the profit: $\Pi_O^*(\mu) = \Pi_{O,-}^*(\mu)$; for $\mu > \nu_{\max} := \max\{\nu_-, \nu_{0,2}\}$ ‘voice-only’ maximizes the profit: $\Pi_O^*(\mu) = \Pi_{O,+}^*(\mu)$.

Proof. First, notice that $\nu_{0,1} < \nu_{0,2}$, implying that $\nu_{\min} < \nu_{\max}$. The stated follows immediately from the inequalities (i) $\Pi_{O,0}^*(\mu) \geq \Pi_{O,-}^*(\mu)$ for $\mu > \nu_-$, (ii) $\Pi_{O,0}^*(\mu) \geq \Pi_{O,+}^*(\mu)$ for $\mu < \nu_{+,2}$, (iii) $\Pi_{O,0}^*(\mu) \leq \Pi_{O,-}^*(\mu)$ for $\mu < \nu_{0,1}$, and (iv) $\Pi_{O,0}^*(\mu) \leq \Pi_{O,+}^*(\mu)$ for $\mu > \nu_{0,2}$. These inequalities are almost trivial to prove from the maximizations over R_- , R_+ and R_0 that were described above.

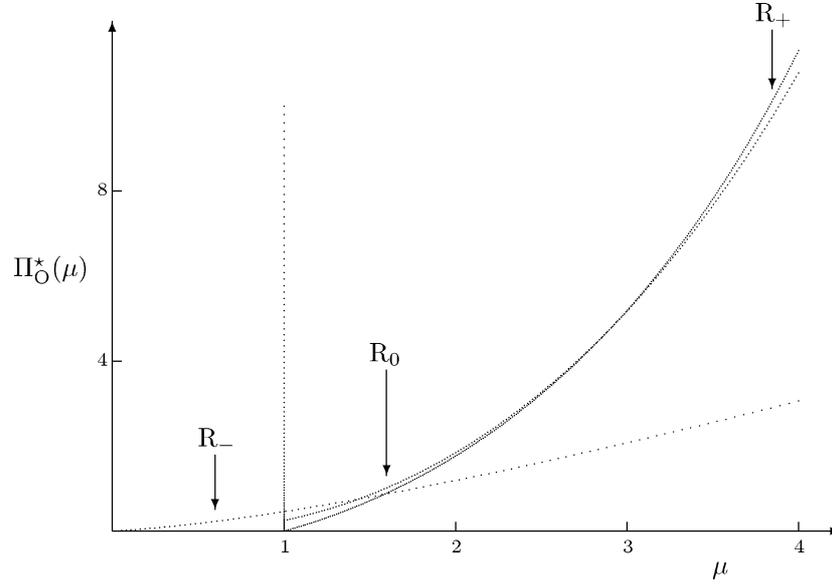
Consider, for instance, the first inequality. For μ larger than ν_- , the optimum $\Pi_{O,-}^*(\mu)$ over R_- is attained at $\bar{p}_H = 1$ and a $\bar{p}_L < 1$, with profit $g_-(1)$. As this price vector lies on the boundary of R_- and R_0 , it equals $g_0(1)$, which is, by definition, majorized by $\Pi_{O,0}^*(\mu)$. The other inequalities are proven similarly. \square

Theorem 4.2 in conjunction with lemma 4.1 (i.e., the convexity of the functions $\Pi_{O,i}^*(\cdot)$ with $i \in \{-, +, 0\}$) implies the following corollary.

Corollary 4.3. The global profit maximization can be characterized as follows. Two regimes are possible:

- There exist service rates ν_-^* and ν_+^* such that

$$\Pi_O^*(\mu) = \begin{cases} \Pi_{O,-}^*(\mu), & \mu \in (0, \nu_-^*); \\ \Pi_{O,0}^*(\mu), & \mu \in (\nu_-^*, \nu_+^*); \\ \Pi_{O,+}^*(\mu), & \mu \in (\nu_+^*, \infty). \end{cases}$$

Figure 3. Profit as a function of link speed, for $\mu \in (0, 4]$.

- There exists a service rate v^* such that

$$\Pi_O^*(\mu) = \begin{cases} \Pi_{O,-}^*(\mu), & \mu \in (0, v^*); \\ \Pi_{O,+}^*(\mu), & \mu \in (v^*, \infty). \end{cases}$$

Example. We now give a numerical example that demonstrates the theory of the previous sections. It gives numerical results for the model with open classes. We choose $\alpha_v = 2\alpha_d = 2$. The values of the ‘critical’ service rates, as introduced in section 4, are given by $v_- = 1.500$; $v_{+,1} = 1.325$; $v_{+,2} = 2.422$; $v_{0,1} = 1.183$; and $v_{0,2} = 3.948$. Applying the inequalities used in the proof of theorem 4.2, it is not so hard to prove that, due to $v_{0,1} < v_- < v_{+,2} < v_{0,2}$, five regimes can be distinguished:

$$\Pi_O^*(\mu) = \begin{cases} \Pi_{O,-}^*(\mu), & \mu \in (0, v_{0,1}); \\ \max\{\Pi_{O,-}^*(\mu), \Pi_{O,0}^*(\mu)\}, & \mu \in (v_{0,1}, v_-); \\ \Pi_{O,0}^*(\mu), & \mu \in (v_-, v_{+,2}); \\ \max\{\Pi_{O,0}^*(\mu), \Pi_{O,+}^*(\mu)\}, & \mu \in (v_{+,2}, v_{0,2}); \\ \Pi_{O,+}^*(\mu), & \mu \in (v_{0,2}, \infty). \end{cases}$$

In figure 3 the three lines depict the maxima over R_- , R_+ and R_0 , respectively. For $\mu < 1.31$ ‘data-only’ is optimal (i.e., maximum profit is achieved in R_-), for $\mu > 2.96$ ‘voice-only’ is optimal (i.e., maximum profit is achieved in R_+), and in between a ‘mixed scenario’ – with priority for voice – is optimal (i.e., maximum profit is achieved in R_0). Figure 4 displays the optimizing prices for various values of μ .

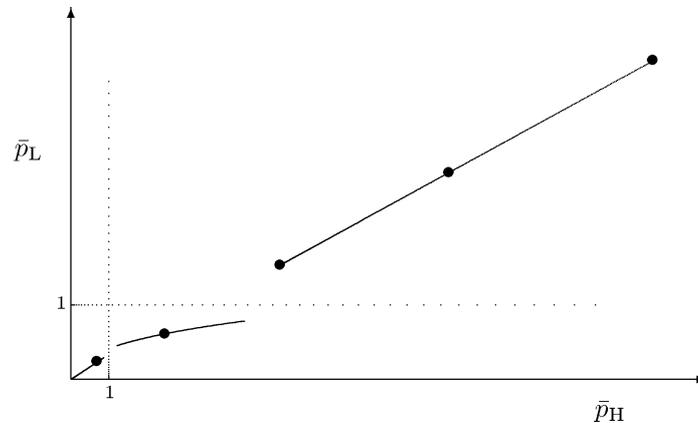


Figure 4. Relation between \bar{p}_L and \bar{p}_H . The bullets correspond to $\mu = 1, \dots, 5$.

We can now use the above example to compare the priority setting (with open classes) to the FIFO system of section 2. There we saw that the lack of differentiation led to a situation in which one type of traffic pushes aside the other, the ‘tragedy of the commons’. We also saw that it depended on the actual value of the link rate which class of users dominated; as could be expected, for high (low) link speeds the delay-sensitive (delay-tolerant) group is served. The example shows that the introduction of priority service partly solves the ‘tragedy of the commons’. We still have voice-only and data-only regimes, but now there is also an intermediate regime in which both classes are served.

5. Summary and concluding remarks

In this paper we have considered a network, to be used by two groups of users, that differ in their preferences regarding quality of service (say, voice and data users). First, we considered a situation in which the two classes share the network resources without any prioritization (‘FIFO’). We showed that such a policy leads to the undesirable situation that one group of users pushes aside the other. This motivated the search for alternative policies.

We have presented two alternative schemes, that both rely on a priority queue. In the first of these (‘dedicated classes’), voice users are led into the high-priority queue, whereas data packets use the low-priority queue. This model, however, has the disadvantage that it might result in situations in which the solution found is not incentive-compatible: if the data users could choose themselves they might have chosen to join the voice queue (i.e., the high-priority class), and vice versa.

Therefore we focused on a priority system with ‘open classes’, in which all users can choose themselves which queue to join. It can be argued that this situation is more realistic than ‘dedicated classes’. If, for instance, in an IP (Internet-protocol) network a voice-over-IP user notices that he is better off using the (cheaper) data class, rather than

sticking to the voice class, he will do so. We have seen that the introduction of priority service partly solves the ‘tragedy of the commons’. For specific range of values of the link rate, voice-only and data-only solution appear, but, interestingly, there is now also an intermediate regime in which both classes are served.

The current tendency in communication networking is to integrate a broad range of services in a single network – think of the Internet supporting voice and video. Our results indicate that prioritization and pricing play a crucial role in this integration. Without adequate prioritization and pricing the network might (technologically) be capable of supporting many different applications, but in practice a few applications – usually those with the least stringent quality-of-service requirements – will dominate. This situation can be avoided by adopting a priority/pricing scheme that reflects the users’ preferences regarding quality of service.

Appendix A

Lemma A.1. For all $x > 0$,

$$f(x) := \left(\frac{x}{x+1}\right)^x \frac{2x+1}{x+1} < 1.$$

Proof. Some tedious calculus yields for $x > 0$,

$$f'(x) := f(x) \left(\frac{1}{x+1} + \log\left(\frac{x}{x+1}\right) \right) < f(x) \left(\frac{1}{x+1} + \left(\frac{x}{x+1} - 1\right) \right) = 0;$$

here the standard inequality $\log x < x - 1$ is applied, in conjunction with $f(x) > 0$ for $x > 0$. The stated now follows from $f(0) = 1$ and $f'(x) < 0$ for all $x > 0$. \square

Lemma A.2. For all $\mu > 0$ and $\alpha > 0$, the function $\bar{f}(\cdot)$ defined by

$$\bar{f}(x) := (\mu - \sqrt[\alpha]{x})x + \sqrt[\alpha]{x} - \frac{\mu}{\sqrt[\alpha]{x}}$$

is concave on $[1, \infty)$. In addition, $\bar{f}'(\mu^{\alpha/2}) > 0$.

Proof. Differentiating twice yields

$$\bar{f}''(x) = -\frac{x^{1/\alpha-2}}{\alpha} \left(\frac{x-1}{\alpha} + x+1 \right) - \frac{\mu}{\alpha} \left(\frac{1}{\alpha} + 1 \right) x^{-1/\alpha-2}.$$

Notice that this is negative for $x > 1$, which proves the first part of the lemma.

Furthermore, we have to prove that

$$\frac{\alpha}{\sqrt{\mu}} \bar{f}'(\mu^{\alpha/2}) = \alpha\sqrt{\mu} + \frac{2}{\mu^{\alpha/2}} - \alpha - 1 \quad (\text{A.1})$$

is positive. This inequality clearly holds for $\mu = 1$. The second claim in the lemma follows from the fact that (A.1) increases in μ , as is checked easily. \square

Appendix A.1. Case A. Maximization over R_-

The maximum over R_- reduces to maximizing $g_-(p_H)$ over $0 < p_H \leq \min\{\mu^{\alpha_d}, 1\}$, where

$$g_-(p) := \left(\mu - p^{1/\alpha_d}\right)p + \left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} \frac{1}{\alpha_d + 1} p^{2+1/\alpha_d}.$$

The price for the low priority service is given by

$$p_L(p_H) = \beta_d(\mu) p_H^2, \quad \text{with } \beta_d(\mu) := \left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d}.$$

Now distinguish between μ smaller and larger than 1.

- Case A₁: $\mu \in (0, 1]$. In this case the optimum has to be computed over the interval $(0, \mu^{\alpha_d}]$. It is straightforward to derive that $g'_-(0) > 0$ and $g'_-(\mu^{\alpha_d}) = (f(\alpha_d) - 1)\mu/\alpha_d < 0$, by lemma A.1. It also follows that $g_-(\cdot)$ is concave on $(0, \mu^{\alpha_d}]$, as

$$g''_-(p) = \frac{\alpha_d + 1}{\alpha_d^2} \cdot p^{1/\alpha_d - 1} \left(\left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} \frac{2\alpha_d + 1}{\alpha_d + 1} p - 1 \right) < 0$$

for all $p \in (0, \mu^{\alpha_d}]$, again invoking lemma A.1. This proves A₁.

- Case A₂: $\mu \in (1, \infty)$. Now the optimum has to be computed over the interval $(0, 1]$. Recall that $g_-(\cdot)$ is concave on $(0, \mu^{\alpha_d}]$, and $g'_-(0) > 0$ and $g'_-(\mu^{\alpha_d}) < 0$. Hence the optimum is reached at $p_H = 1$ if $g'_-(1) \geq 0$, and in the interior of $(0, \mu^{\alpha_d}]$ if $g'_-(1) < 0$. Denote

$$\zeta_-(\mu) := g'_-(1) = \mu - 1 + \frac{1}{\alpha_d} \left(\left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} \frac{2\alpha_d + 1}{\alpha_d + 1} - 1 \right).$$

Now note that $\zeta_-(1) = (f(\alpha_d) - 1)/\alpha_d < 0$ (apply lemma A.1!), and $\zeta_-(\mu) \rightarrow \infty$ as $\mu \rightarrow \infty$. Applying lemma A.1 again, together with the fact that $\mu > 1$, we see that $\zeta_-(\cdot)$ increases:

$$\zeta'_-(\mu) = 1 - \mu^{-\alpha_d - 1} \left(\frac{\alpha_d}{\alpha_d + 1}\right)^{\alpha_d} \frac{2\alpha_d + 1}{\alpha_d + 1} > 0.$$

This immediately implies A₂.

Appendix A.2. Case B. Maximization over R_+

The region R_+ is empty if $\mu \leq 1$; we therefore concentrate on $\mu > 1$. We have that

$$p_L(p_H) = \max\{1, \beta_v(\mu) p_H^2\}, \quad \text{with } \beta_v(\mu) := \left(\frac{\alpha_v}{\mu(\alpha_v + 1)}\right)^{\alpha_v}.$$

With $\mu_v := 1 + \alpha_v^{-1}$, we consider two cases.

- Case B₁: $\mu \in (1, \mu_v)$. It is not hard to see that the optimal p_L equals 1. The maximizing p_H should be found from

$$\max_{\mu^{\alpha_v/2} < p_H < \mu^{\alpha_v}} \bar{g}_+(p_H) \quad \text{with } \bar{g}_+(p) := (\mu - \sqrt[\alpha_v]{p})p + \sqrt[\alpha_v]{p} - \mu \frac{1}{\sqrt[\alpha_v]{p}}.$$

By lemma A.2, $\bar{g}_+(\cdot)$ is concave on $(\mu^{\alpha_v/2}, \mu^{\alpha_v})$ and $\bar{g}'_+(\mu^{\alpha_v/2}) > 0$. Hence, the maximum is attained in μ^{α_v} if $\bar{g}'_+(\mu^{\alpha_v}) > 0$; otherwise it is attained in the interior of the interval. Define

$$\zeta_+(\mu) := \bar{g}'_+(\mu^{\alpha_v}) = \frac{1}{\alpha_v}(\mu^{-\alpha_v}(\mu + 1) - \mu).$$

First, observe that $\zeta_+(1) = \alpha_v^{-1} > 0$, and that $\zeta_+(\mu_v) = (f(\alpha_v) - 1)\mu_v/\alpha_v < 0$ (lemma A.1). Elementary calculus shows that $\zeta_+(\cdot)$ decreases on $[1, \mu_v)$ is equivalent to

$$(1 - \alpha_v)\mu - \alpha_v \leq \mu^{\alpha_v+1}.$$

This last inequality follows from the fact that there is equality at $\mu = 1$, in conjunction with the fact that the derivative of the right-hand side (i.e., $(\alpha_v + 1)\mu^{\alpha_v}$) majorizes the derivative of the left-hand side (i.e., $1 - \alpha_v$) for all $\mu > 1$. We arrive at B₁.

- Case B₂: $\mu \in (\mu_v, \infty)$. It turns out that

$$p_L(p_H) = \left(\frac{\alpha_v}{\mu(\alpha_v + 1)}\right)^{\alpha_v} p_H^2 \quad \text{if } p_H \in [q_+(\mu), \mu^{\alpha_v}], \quad \text{with}$$

$$q_+(\mu) := \left(\frac{\mu(\alpha_v + 1)}{\alpha_v}\right)^{\alpha_v/2},$$

and $p_L(p_H) = 1$ if $p_H \in [\mu^{\alpha_v/2}, q_+(\mu)]$. Define $g_+(\cdot)$ as in (9), but with α_d replaced by α_v ; the concavity of $g_+(\cdot)$ and $g'_+(\mu^{\alpha_v}) < 0$ follow like in case A₁. We have to solve

$$\max \left\{ \max_{\mu^{\alpha_v/2} < p_H < q_+(\mu)} \bar{g}_+(p_H), \max_{q_+(\mu) < p_H < \mu^{\alpha_v}} g_+(p_H) \right\}.$$

From (i) the concavity of both functions, (ii) $\bar{g}_+(\mu^{\alpha_v/2}) > 0$, (iii) $g'_+(\mu^{\alpha_v}) < 0$, and (iv) $g'_+(q_+(\mu)) = g'_+(q_+(\mu))$, we find that the optimal p_H lies in $(q_+(\mu), \mu^{\alpha_v})$ if $g'_+(p_H) > 0$, whereas $p_H \in (\mu^{\alpha_v/2}, q_+(\mu))$ otherwise. Define

$$\xi_+(\mu) := g'_+(q_+(\mu)) = \mu - \left(\frac{\alpha_v + 1}{\alpha_v}\right)^{3/2} \sqrt{\mu} + \left(\frac{\alpha_v}{\mu(\alpha_v + 1)}\right)^{(\alpha_v-1)/2} \frac{2\alpha_v + 1}{\alpha_v + 1} \cdot \frac{1}{\alpha_v}.$$

We now prove that $\xi_+(\mu) = 0$ has a unique zero in (μ_v, ∞) . To this end, first observe that $\xi_+(\mu_v) = (f(\alpha_v) - 1)\mu_v/\alpha_v < 0$ (due to lemma A.1) and $\xi_+(\mu) \rightarrow \infty$ as $\mu \rightarrow \infty$. Also, using lemma A.1, it is straightforward to prove that $\xi''_+(\mu_v) > 0$.

If $\alpha_v > 1$ the function $\xi_+''(\cdot)$ does not change sign at all; so $\xi_+(\cdot)$ is convex. Therefore concentrate on $\alpha_v \leq 1$. In this case, $\xi_+''(\cdot)$ changes sign at

$$\bar{\mu}_v := \frac{\alpha_v}{\alpha_v + 1} \alpha_v \sqrt{\left(\frac{2\alpha_v + 1}{\alpha_v + 1}\right)^2 (1 - \alpha_v)^2}.$$

Some calculus gives that verifying that $\bar{\mu}_v < \mu_v$ reduces to checking if

$$(1 - \alpha_v) \left(\frac{\alpha_v}{\alpha_v + 1}\right)^{\alpha_v} \frac{2\alpha_v + 1}{\alpha_v + 1} < 1,$$

which holds due to lemma A.1. We conclude that $\xi_+(\cdot)$ is convex on the domain $[\mu_v, \infty)$. Notice that a function $F(\cdot)$, convex (or concave) on interval $[a, b]$, has exactly one zero in this interval if $F(a)F(b) < 0$. We have proven B₂.

Appendix A.3. Case C. Maximization over R_0

Again we first perform the optimization over p_L for given p_H . It is straightforward to obtain that the optimum is attained at

$$p_L(p_H) = \min\{1, \beta_d(\mu) p_H^{2\alpha_d/\alpha_v}\}.$$

Defining $\mu_d := 1 + \alpha_d^{-1}$, two cases need to be distinguished.

- Case C₁: $\mu \leq \mu_d$ It is not hard to verify that for these μ it holds that

$$\left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} p_H^{2\alpha_d/\alpha_v} \leq 1 \quad \text{for all } p_H \in (1, \mu^{\alpha_v}],$$

so that the optimization reduces to

$$\max_{1 < p_H \leq \mu^{\alpha_v}} g_0(p_H),$$

where $g_0(\cdot)$ is defined as $g_0(\cdot)$ in (5). Using that $\mu_-^* < 1 < \mu$, and invoking proposition 3.4, we know that $g_0(\cdot)$ first increases and then decreases on $(0, \mu^{\alpha_v}]$. In other words: a price $p_H > 1$ is optimal iff

$$\zeta_0(\mu) := g_0'(1) = \mu - 1 + \frac{1}{\alpha_v} \left(\left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} \frac{2\alpha_d + 1}{\alpha_d + 1} - 1 \right) > 0; \quad (\text{A.2})$$

otherwise the maximum is attained at $p_H = 1$.

For $\mu = 1$, condition (11) is not met; this is because of lemma A.1. Notice that

$$\zeta_0(\mu_d) = \left(\frac{1}{\alpha_d} - \frac{1}{\alpha_v}\right) + \frac{1}{\alpha_v} \left(\frac{\alpha_d}{\alpha_d + 1}\right)^{2\alpha_d} \frac{2\alpha_d + 1}{\alpha_d + 1} > 0,$$

and $\zeta_0'(\mu_d) > 0$ for all $\mu > 1$. Hence, $\zeta_0(\cdot)$ has a unique root in $(1, \mu_d)$. This proves C₂.

- Case C₂: $\mu \in (\mu_d, \infty)$. For these μ we have that

$$\beta_d(\mu) p_H^{2\alpha_d/\alpha_v} \leq 1 \quad \text{if and only if} \quad p_H \leq q_0(\mu) := \left(\frac{\mu(\alpha_d + 1)}{\alpha_d} \right)^{\alpha_v/2}.$$

Notice that $q_0(\mu)$ is smaller than μ^{α_v} (as $\mu_d = (\alpha_d + 1)/\alpha_d < \mu$). So we get the optimization

$$\max \left\{ \max_{1 < p_H < q_0(\mu)} g_0(p_H), \max_{q_0(\mu) < p_H < \mu^{\alpha_v}} \bar{g}_0(p_H) \right\},$$

where $\bar{g}_0(\cdot) = \bar{g}_+(\cdot)$. Define

$$\xi_0(\mu) := g'_0(q_0(\mu)) = \mu - \frac{\alpha_v + 1}{\alpha_v} \sqrt{\frac{\mu(\alpha_d + 1)}{\alpha_d}} + \left(\frac{\alpha_d}{\mu(\alpha_d + 1)} \right)^{(\alpha_v - 1)/2} \frac{2\alpha_d + 1}{\alpha_d + 1} \cdot \frac{1}{\alpha_v}.$$

With an analysis that is analogous to case B₂, we prove C₂.

References

- [1] D. Clark, Internet cost allocation and pricing, in: *Internet Economics* (MIT Press Cambridge, MA, 1997) pp. 215–252.
- [2] R. Gibbens, R. Mason and R. Steinberg, Internet service classes under competition, *IEEE J. Selected Areas Commun.* 18 (2000) 2490–2498.
- [3] A. Gupta, D. Stahl and A. Whinston, A stochastic equilibrium model of Internet pricing, *J. Economic Dynamics Control* 21 (1997) 697–722.
- [4] G. Hardin, The tragedy of the commons, *Science* 162 (1968) 1243–1248.
- [5] L. Kleinrock, *Queueing Systems*, Vol. 2, *Computer Applications* (Wiley, New York, 1976).
- [6] J. MacKie-Mason and H. Varian, Pricing congestible network resources, *IEEE J. Selected Areas Commun.* 13 (1995) 1141–1149.
- [7] L. McKnight and J. Bailey, eds., *Internet Economics* (MIT Press, Cambridge MA, 1997).
- [8] H. Mendelson, Pricing computer services: Queueing effects, *Commun. ACM* 28 (1985) 312–321.
- [9] H. Mendelson and S. Whang, Optimal incentive-compatible priority pricing for the M/M/1 queue, *Oper. Res.* 38 (1990) 870–883.
- [10] A. Odlyzko, Paris Metro Pricing: The minimalist differentiated services solution, in: *Proc. of the 1999 7th Internat. Workshop on Quality of Service (IWQoS '99)* (IEEE Press, New York, 1999) pp. 159–161.
- [11] A. Odlyzko, Paris Metro Pricing for the Internet, in: *Proc. of ACM Conf. on Electronic Commerce (EC'99)* (ACM, New York, 1999) pp. 140–147.
- [12] S. Shenker, Fundamental design issues for the future Internet, *IEEE J. Selected Areas Commun.* 13 (1995) 1176–1188.
- [13] J. Tirole, *The Theory of Industrial Organization* (MIT Press, Cambridge, MA, 1989).
- [14] D. Walker, F. Kelly and J. Solomon, Tariffing in the new IP/ATM environment, *Telecommunications Policy* 21 (1997) 283–295.