A Mathematical Model of the Paris Metro Pricing Scheme for Charging Packet Networks

David Ros [a,*] Bruno Tuffin [b]

[a] *GET/ENST Bretagne, rue de la Châtaigneraie, CS 17607, 35576 Cesson Sévigné Cedex, France*

[b] *IRISA/INRIA Rennes, Campus de Beaulieu, 35042 Rennes Cedex, France*

Abstract

Pricing has become one of the main challenges of the networking community and is receiving a great deal of interest in the literature. In this paper, we analyze the so-called Paris Metro Pricing scheme which separates the network into different and independent subnetworks, each behaving equivalently, except that they charge their customers at different rates. In our model, each subnetwork is represented by a single bottleneck queue, and the "customers" (data packets) choose their subnetwork taking into account not only the prices, but also the expected delay, which is supposed to have an economic impact. We obtain some necessary and sufficient conditions for the stability of the system; we analyze the problem of maximizing the network revenue and compare it with the case of a single network, and present several extensions of the model. Numerical results illustrating some key aspects of the system are provided throughout the paper.

Key words: Internet, packet networks, pricing, queueing models.

1 Introduction

Pricing in IP networks is becoming a main concern of Internet Service Providers (ISPs). Classical pricing schemes, based on a subscription fee and/or a (possibly unlimited) number of hours of network access, depending on the country and the type of usage, are becoming obsolete due to three major reasons:

* Corresponding author. Phone: +(33) 2 99 12 70 46; fax: +(33) 2 99 12 70 30
Email addresses: David.Ros@enst-bretagne.fr (David Ros), Bruno.Tuffin@irisa.fr (Bruno Tuffin).

(1) Such simple schemes were an incentive to stimulate utilization. Even if they probably helped to develop the network and allowed ISPs to gain market share, nowadays they may prove sub-optimal. Indeed, demand in access networks may easily exceed the installed capacity, leading to congestion.
(2) Some users might wish to pay more than others to avoid congestion and get an improved service. Why not serve them first and (possibly) increase the ISP revenue?
(3) Unlike the telephone network, the current Internet—and, more so, the next-generation Internet—has to deal with different kinds of applications with diverse quality of service (QoS) requirements:
   - Interactive video needs very small delays and fairly low packet losses. On the other hand, high-quality video-on-demand may tolerate a higher delay and jitter (within some bounds) but lower packet losses than interactive video.
   - Voice over IP requires small delays but can afford some packet losses.
   - Electronic mail can afford delay (within some bounds).
   - File transfer needs a good average throughput, whereas web browsing demands also a relatively low latency.
   - Remote login requires small round-trip times.

   Including these requirements in a pricing scheme may, in some way, improve users' satisfaction.

A wide range of charging methods for packet networks has been devised and discussed in the literature; see for instance the survey papers [1–6]. The main classes include resource reservation, priorities between packets, auctions for bandwidth or pricing based on transfer rates.

In this paper, we study another method, called *Paris Metro Pricing* (PMP) and introduced by A. Odlyzko in [7,8]. Under the PMP scheme, the network is split into independent subnetworks; more precisely, a fixed fraction of the capacity of each link is totally allocated to each subnetwork. The tariff will be different for each subnetwork, so that (hopefully) congestion will be alleviated in the most expensive ones. This method does not offer any QoS guarantees, so that it is somehow weak with respect to the abovementioned techniques; however, it is very attractive because of its implementation and management simplicity. In this paper, we develop a mathematical model which may be helpful in setting prices in a PMP-based network. Particularly, our model will use the delay introduced by the network as a cost.

It has been argued by Gibbens et al. [9] that network segmentation might not work under competition (i.e., no Nash equilibrium exists); nonetheless, their result applies to a model quite different from ours. Moreover, note that even if the optimization problem in [9] looks similar to that presented in this paper, Gibbens et al. deal with an additional constraint representing the competition

between two subsets of subnetworks.

Note also that in [10], PMP has been adapted in such a way that, instead of logically separating the subnetworks, a round-robin service discipline is used to share the bandwidth among them. This "improved" version of PMP, even though it deserves attention, is not considered here.

The layout of the paper is as follows. The PMP scheme is introduced in Section 2. In Section 3 the mathematical model is described. Necessary and sufficient conditions for stability are then provided, and an optimization problem (maximizing the ISP revenue) is introduced. Section 4 provides some numerical illustrations of the results of Section 3. In Section 5, we compare the revenues of a PMP network with those of an equivalent, non-PMP network. Section 6 is devoted to some extensions of the model: the case of multiple applications with different delay sensitivities, the case where the performance measures of interest include the loss probabilities, and time-of-day pricing. Finally, conclusions are given in Section 7.

2 PMP: a brief description

The original PMP proposal in [8] consists in partitioning a network into several logically separate networks (or classes), each having a fixed fraction of the capacity of the entire network. Every subnetwork would route and handle packets according to the current Internet protocols. There is no formal guarantee of QoS, but by charging different rates for different classes (served in the same way), it is supposed that the most expensive classes will be less congested by way of self-regulation, hence delivering a better QoS. The name given to this model, *Paris Metro Pricing*, stems from the rules of the Paris Metro about 20 years ago, where trains were composed of cars of two classes, offering exactly the same quality of seats. As tickets prices were different, the cars for the most expensive class were less congested, leading to a better perceived QoS.

As pointed out by Odlyzko, the advantage of PMP is that, even if we do not have any strict QoS guarantee using this scheme, but rather a *statistical* guarantee, it would permit dispensing with complex, non-scalable mechanisms (like, say, resource reservation protocols) and keep the simpler and cheaper current model of the Internet, while improving the QoS experience. Indeed, no signalling protocols would have to be used, and the tariff would be fixed, which is preferred by most users [11].

From the preceding discussion, it is intuitively clear that prices play in PMP an important role in controlling congestion and, notably, in distributing the

load among subnetworks: prices are the key to quality-of-service differentiation. The problem of finding the "right" set of prices for a given bandwidth partition may be informally stated as follows: for a PMP network to work in an efficient manner [1], the charge for using the most expensive subnetwork should be high enough, so that this subnetwork is lightly loaded, but not "too high"—otherwise, the subnetwork would remain empty, because the high quality offered would not compensate for the (very) high monetary cost incurred by users.

In the next section, we present a mathematical model of PMP that allows us to analyze its stability properties with respect to prices, as well as to solve the revenue maximization problem.

3 Mathematical model

For the sake of clarity, we begin our study of PMP by presenting a simplified model in which all packets are generated by the same kind of application and all users have the same valuation of QoS (represented in our model by the mean packet delay). Later, in Section 6, we will describe a more realistic model corresponding to a multi-application scenario where each application may have a different QoS valuation.

3.1 Model presentation

Assume that there are I classes (i.e., subnetworks) in a PMP network, and that the class-i per-packet price at the entrance of the network is p_i ($1 \leq i \leq I$). We assume that there is a *potential total* arrival rate $\tilde{\lambda}$ of packets at the network, corresponding to the arrival rate when prices are set to 0.

A utility measure is associated to each packet. It is assumed to be a random variable U that follows the same distribution for every packet, and that it is independent of other packets' utility.

Also, a *total* cost function

$$p_i + \gamma d_i$$

is associated to a class i, where d_i is the mean delay for a packet in the network and γ is a constant converting delay in money.

[1] Efficiency may be defined, for instance, in terms of optimizing the network operator's revenue and providing an adequate usage of each subnetwork's resources.

A packet enters (i.e., chooses) network i if $i = \text{argmin}_j(p_j + \gamma d_j)$ and $U \geq p_i + \gamma d_i$, that is, it chooses the less expensive subnetwork *in terms of total cost* (which is a linear combination of delay and price). If $U < \min_j(p_j + \gamma d_j)$, the packet does not enter at all, meaning that the network is too expensive for it. Traffic *elasticity* is then a central assumption of our model since a larger delay can be accepted by a user, provided that the price is decreased proportionally.

The actual total arrival rate is then

$$\lambda = \tilde{\lambda} P(U \geq \min_j p_j + \gamma d_j).$$

We define by $\bar{F}$ the complementary cumulative distribution function of U, i.e., $\bar{F}(x) = P(U \geq x)$. Finally, denote by λ_i $(1 \leq i \leq I)$ the actual arrival rate at subnetwork i, so that $\lambda = \sum_{i=1}^{I} \lambda_i$.

3.2 Stability: existence and uniqueness

Using this model, in equilibrium, the distribution of packets among classes has to be stable, meaning that the total cost $p_j + \gamma d_j$ is the same for all classes j. Indeed, if for a given class j the value $p_j + \gamma d_j$ were smaller than the total cost of the other classes, then new packets entering the network would choose class j until its total cost reaches that of other classes. This corresponds to a Wardrop equilibrium [12]. Let us call p_{tot} the value $p_j + \gamma d_j$ (identical for all j). As we want all subnetworks to be used, we have the following set of $I + 1$ equations with $I + 1$ unknown variables λ_i $(1 \leq i \leq I)$ and p_{tot}:

$$\begin{cases} \sum_{i=1}^{I} \lambda_i = \tilde{\lambda} P(U \geq p_{tot}) \\ p_i + \gamma d_i = p_{tot}, \text{ for } 1 \leq i \leq I \end{cases} \tag{1}$$

where d_i is a function of λ_i.

The remaining of this section is separated into two parts. In the first part, the d_i represent the mean *waiting* times (i.e., service excluded). In the second part, results are obtained as a corollary when the d_i represent the *response* times (i.e., service included). What users and applications care about is total delay, so the latter case seems more realistic; nonetheless, we will begin by treating the waiting time case for the sake of clarity.

In order to determine p_{tot} and λ_i $\forall i$, we have the following theorem.

Theorem 1 *Assume that $d_i = f_i(\lambda_i) \in \mathbb{R}^+$ is a strictly increasing and continuous function of the arrival rate, representing the mean waiting time of a*

class-i user. Without loss of generality, we suppose that $p_1 > p_2 > \cdots > p_I$. *Assume also that the distribution of* U *is absolutely continuous and strictly increasing. Then the solution of Eq. (1) exists and is unique if and only if*

$$p_1 \leq \bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}}\sum_{i=1}^{I} f_i^{-1}\left(\frac{p_1 - p_i}{\gamma}\right)\right). \tag{2}$$

Remark 1 *Note that if condition (2) is not verified, then it means that the highest price* p_1 *is too high to obtain an equilibrium between classes. Consequently, this class would be ignored (i.e., the subnetwork would remain empty) and the computation would have to be carried out again with the* $I-1$ *remaining classes.*

Thus, condition (2) has to be taken as an assumption to get the equilibrium for I *classes.*

Proof of the theorem: Denote by $\bar{F}$ the complementary cumulative distribution function of random variable U. The system (1) can be re-written as

$$\begin{cases} p_{\text{tot}} = \bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}}\sum_{i=1}^{I}\lambda_i\right) \\ \lambda_i = f_i^{-1}\left(\frac{p_{\text{tot}}-p_i}{\gamma}\right), \text{ for } 1 \leq i \leq I. \end{cases}$$

In particular we can get

$$p_{\text{tot}} = \bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}}\sum_{i=1}^{I} f_i^{-1}\left(\frac{p_{\text{tot}} - p_i}{\gamma}\right)\right). \tag{3}$$

If we are able to prove that there exists a unique p_{tot} satisfying this equation, then the existence and uniqueness of the λ_i will be straightforward to show.

It is important to note that, by definition, $p_{\text{tot}} \geq p_i \ \forall i$ (i.e., $p_{\text{tot}} \geq p_1$) because $d_i \geq 0$.

Since f_i^{-1} is strictly increasing $\forall i$ and $\bar{F}^{-1}$ is strictly decreasing (and both are continuous), $\bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}}\sum_{i=1}^{I} f_i^{-1}\left(\frac{p_{\text{tot}}-p_i}{\gamma}\right)\right)$ is a continuous and strictly decreasing function of p_{tot}. Thus the solution of (3), if it exists, is unique.

Existence depends on border values. Let $E(S_i)$ denote the mean service time for class i. The delay goes to infinity when the mean arrival rate approaches the mean service rate $1/E(S_i)$, hence: $\lim_{d_i \to \infty} f_i^{-1}(d_i) = 1/E(S_i)$. Therefore, since the left-hand side of (3) tends to infinity and the right-hand side tends to $\bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}}\sum_{i=1}^{I}\frac{1}{E(S_i)}\right)$ when p_{tot} tends to infinity, the existence depends on whether, at the minimal value of p_{tot} (i.e., p_1), the right hand side of (3) is

greater than or equal to p_1. Both situations are represented in Fig. 1. The condition is expressed by (2). In this case, Eq. (3) has a unique solution. □

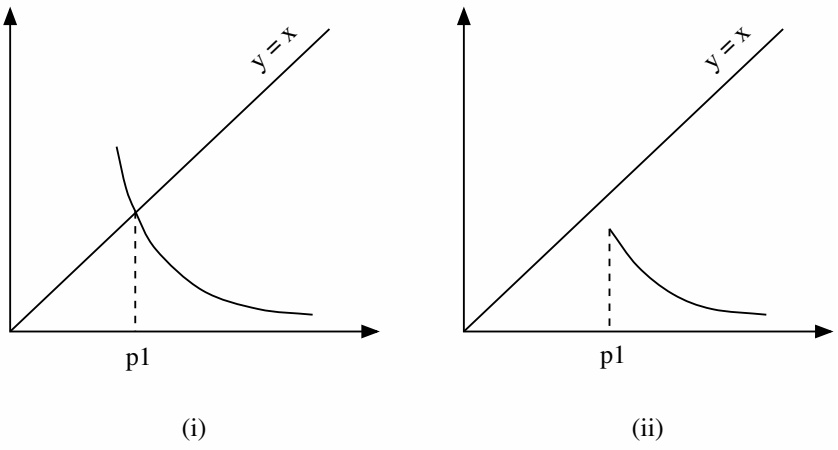


Fig. 1. Two situations for the border value. The solution (intersection of $y = \bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}}\sum_{i=1}^{I} f_i^{-1}\left(\frac{x-p_i}{\gamma}\right)\right)$ with the $y = x$ line) exists in case (i) only.

Corollary 1 *Assume now that $d_i = f_i^{(R)}(\lambda_i)$ represents the system's response time. Then $d_i \in [E(S_i), \infty)$, with $E(S_i)$ denoting the mean service time.*

In this case, and without loss of generality, the classes are supposed to be ordered by their minimum total cost $p_1 + \gamma E(S_1) \geq p_2 + \gamma E(S_2) \geq \cdots \geq p_I + \gamma E(S_I)$.

Then the solution of (1) exists and is unique if and only if

$$p_1 + \gamma E(S_1) \leq \bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}}\sum_{i=1}^{I}(f_i^{(R)})^{-1}\left(\frac{p_1 - p_i}{\gamma} + E(S_1)\right)\right), \tag{4}$$

or equivalently, using the waiting times,

$$p_1 + \gamma E(S_1) \leq \bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}}\sum_{i=1}^{I} f_i^{-1}\left(\frac{p_1 - p_i}{\gamma} + E(S_1) - E(S_i)\right)\right),$$

Proof: Since the mean waiting time cannot be negative, we have to use the fact that $p_{\text{tot}} \geq \max_i(p_i + \gamma E(S_i)) = p_1 + \gamma E(S_1)$. By replacing p_1 in Eq. (2) using $f_i^{(R)}$ instead of f_i, we get Eq. (4).

Besides, using the fact that $f_i^{(R)}$ is given by $f_i^{(R)} = f_i + E(S_i)$, we can easily obtain the equation with the waiting times. □

Remark 2 (Special case: upper-bounded utility) *In the case where the users' utility is bounded by a maximum utility U_{max} (that is, $\bar{F}(x) = 0\ \forall x \geq U_{max}$), we can derive a simpler, necessary condition for stability as follows.*

The inequality can be satisfied (meaning that some customers will enter the network) if and only if $\bar{F}(p_1 + \gamma E(S_1)) > 0$, but this can only be true if $p_1 + \gamma E(S_1) < U_{max}$. Hence, in order for the system to be stable, prices must satisfy the following condition:

$$U_{max} > p_i + \gamma E(S_i),\ \forall i \tag{5}$$

Notice that $U_{max} \leq \gamma E(S_i)$ means the cost due to fixed delay only is so high that class i would never be chosen by users, regardless of the price p_i.

Remark 3 *It is easy to show that, for* any *I-tuple of prices $(p_1, \ldots, p_I)$ such that $p_i + \gamma E(S_i) = p_j + \gamma E(S_j)\ \forall i \neq j$, a unique solution of (1) exists as long as the utility is* not *upper-bounded by a finite U_{max}. We will call $\mathcal{E}_0$ the set of such prices.*

Remark 4 *Finally, note that (4) can be written as:*

$$p_1 + \gamma E(S_1) \leq \bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}}\sum_{i=2}^{I}(f_i^{(R)})^{-1}\left(\frac{p_1 - p_i}{\gamma} + E(S_1)\right)\right),$$

because $(f_1^{(R)})^{-1}(E(S_1)) = 0$.

3.3 Optimization problem

The idea, from a network provider's perspective, is to find the p_i maximizing the revenue of the network, i.e.,

$$R = \max_{p_i, \forall i} \sum_{i=1}^{I} \lambda_i p_i$$

subject to $p_i \geq 0\ \forall i$.

Based on the existence and uniqueness condition of the arrival rates and total cost (defined in terms of p_i), and using the previous notations, the problem can be reformulated as:

$$\max_{p_i, \forall i} \sum_{i=1}^{I} f_i^{-1}\left(\frac{p_{\text{tot}} - p_i}{\gamma}\right) p_i$$

subject to

$$
\begin{aligned}
p_{\text{tot}} &\geq p_i \ \forall i \\
p_{\text{tot}} &= \bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}}\sum_{i=1}^{I} f_i^{-1}\left(\frac{p_{\text{tot}}-p_i}{\gamma}\right)\right) \\
p_i &\geq 0 \ \forall i
\end{aligned}
$$

or its equivalent form, if we rather talk about the response time instead of the waiting time.

As this problem looks analytically intractable in the general case, in the next section we are going to present numerical results illustrating the stability domain, as well as the impact of prices and bandwidth partition on delay, total cost and revenue.

Remark that, in some sense, our approach can be related to that of Honig and Stieglitz [13] (albeit applied to a different model), where revenue, prices and performance were studied.

4 Numerical results

In this section we will present some numerical results obtained with the above model [2]. In what follows we will use the response time of the system as a measure of delay, that is, $d_i = f^{(R)}(\lambda_i)$. We will also relax the restriction $p_1 + \gamma E(S_1) \geq p_2 + \gamma E(S_2)$ imposed in the previous section.

For the sake of simplicity, let us consider the case in which the ISP partitions its network in $I = 2$ subnetworks. Indeed, even if the case $I > 2$ is, in principle, not harder to solve numerically than the $I = 2$ case, we will only consider the latter in order to being able of graphically presenting the results.

Each subnetwork i is viewed as a single bottleneck, modeled as a M/G/1 FIFO queue with service rate $\mu_i = 1/E(S_i)$, so:

$$
f_i^{(R)}(\lambda_i) = \frac{1}{\mu_i} + \frac{\left(1+\text{CV}^2\right)\,\rho_i}{2\,\mu_i\,(1-\rho_i)}
$$

where $\rho_i = \lambda_i/\mu_i$ and CV^2 is the squared coefficient of variation of the service law.

[2] The standard numerical and optimization packages included in, say, Matlab or Mathematica can be used to solve this problem.

Unless stated otherwise, the following parameters will be used throughout this section:

- Total capacity of the network: $c = 2$.
- Potential total arrival rate: $\tilde{\lambda} = 3$.
- Utility distribution function $\bar{F}$: exponential[3] with mean $\bar{U} = 1$.
- Cost per unit of delay: $\gamma = 1$.
- Service time distribution: exponential (i.e., each queue is a M/M/1 queue and so $\mathrm{CV}^2 = 1$). Note that the M/D/1 queue gives similar results.

Note that, in the case of TCP flows, bandwidth sharing among flows has been successfully modeled as a M/G/1 *processor sharing* (PS) queue—see for instance [14]—, so a M/M/1 PS queueing model would seem more fit. However, since the response time of a M/M/1 FIFO queue is identical to that of a M/G/1 PS queue (as long as the mean service time is the same for both queues), results obtained with both models should be qualitatively similar.

We assume, without loss of generality, that the total bandwidth c of the network is shared among subnetworks according to:

$$\mu_1 = \alpha c,\ \mu_2 = (1-\alpha)c \quad \text{with } 0.5 \leq \alpha < 1. \tag{6}$$

4.1 Equilibrium region

It is interesting to look at the shape of the *equilibrium region*, defined as the set $\mathcal{E}$ of all pairs (p_1, p_2) such that the system of equations (1) has a unique solution, that is:

$$\mathcal{E} = \mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2 \tag{7}$$

with $\mathcal{E}_0$ defined by the set of prices such that $p_1 + \gamma/\mu_1 = p_2 + \gamma/\mu_2$ that also verify the stability condition and, for $i \in \{1, 2\}$:

$$\mathcal{E}_i = \left\{ (p_1, p_2) : p_{-i} + \frac{\gamma}{\mu_{-i}} < p_i + \frac{\gamma}{\mu_i} \leq \bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}} (f_{-i}^{(R)})^{-1} \left(\frac{p_i - p_{-i}}{\gamma} + \frac{1}{\mu_i} \right) \right) \right\}$$

where $-i$ denotes the element of $\{1, 2\}$ which is not i. $\mathcal{E}_i$, for $i \in \{1, 2\}$, can be

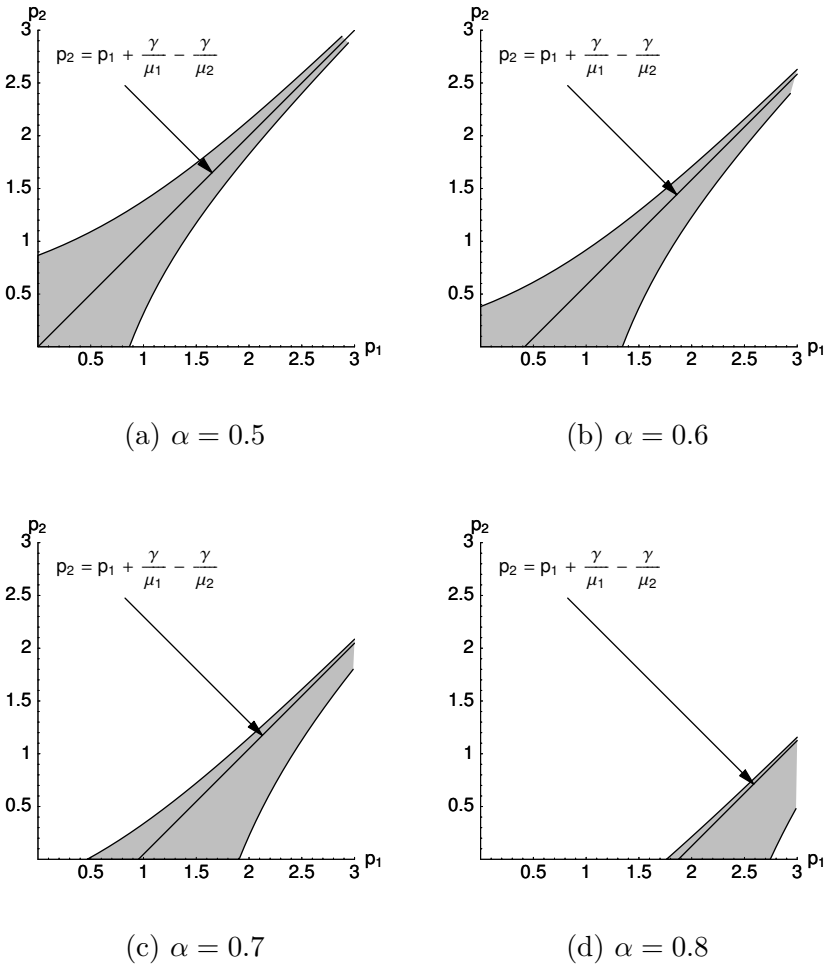
[3] Qualitatively similar results (not shown for space reasons) were obtained by taking $\bar{F}$ uniformly distributed.

regarded as the price area such that the charge plus the cost of service, i.e., the total cost minus the waiting cost, is more expensive for class i (and such that the stability condition is verified).

In what follows we will present the equilibrium region corresponding to an exponentially-distributed utility. For comparison purposes, we will also show the equilibrium region when the utility is uniformly distributed in $[0, U_{max}]$, where $U_{max} = 2\,\bar{U}$.

4.1.1 Exponentially-distributed utility

Figure 2 shows the equilibrium region for four distinct bandwidth allocations, in the case where the utility is exponentially distributed. The equilibrium region is shown in gray; the straight line pointed to by an arrow corresponds to the "frontier" $\mathcal{E}_0$ where the total costs (minus the waiting cost) are the same for both classes.



(a) $\alpha = 0.5$ (b) $\alpha = 0.6$

(c) $\alpha = 0.7$ (d) $\alpha = 0.8$

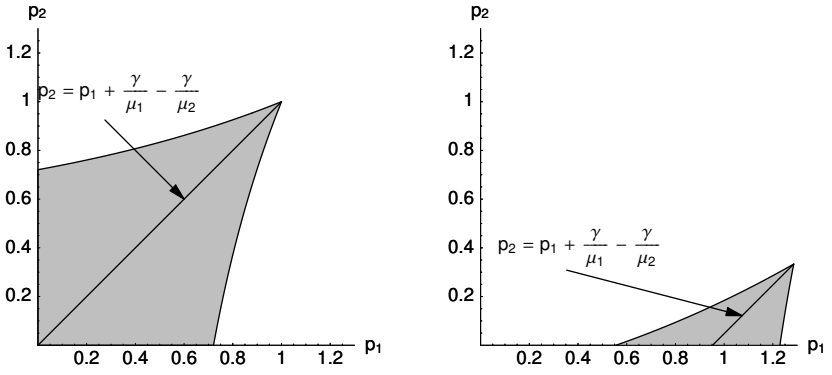
Fig. 2. Equilibrium region, $\bar{F}$ exponential.

Note how the equilibrium region above the frontier gets narrower as the parameter α increases. This may be intuitively interpreted as follows: for a given p_1, as the service rate μ_2 of subnetwork 2 gets lower (and so μ_1 gets higher), subnetwork 1 would tend to attract more traffic than subnetwork 2; hence, total delay d_1 would eventually tend to increase so, for a fixed p_1, the maximum value p_2 can take must get lower in order to attain an equilibrium.

Remark also that, as prices get higher and higher, the equilibrium region degenerates into the straight line given by the frontier condition, no matter the value of α. Since there is no upper bound to the utility (meaning that some users are willing to tolerate a very high cost), and the fixed cost $p_i + \gamma/\mu_i$ increases for both subnetworks, queueing delay tends to zero (together with the arrival rate), so equilibrium can only be reached if the fixed cost is about the same in each subnetwork; otherwise, traffic would switch entirely to the less-expensive one.

4.1.2 Uniformly-distributed utility

Figure 3 shows the equilibrium region for bandwidth allocations $\alpha = 0.5$ and $\alpha = 0.7$, as in Figs. 2(a) and 2(c), but with a uniformly-distributed utility (with $U_{max} = 2$).



(a) $\alpha = 0.5$ (b) $\alpha = 0.7$

Fig. 3. Equilibrium region, $\bar{F}$ uniform.

First, note that values of p_i cannot be arbitrarily high because of the equilibrium condition (5). For instance, for $\alpha = 0.6$ we have that $p_1 < U_{max} - \gamma/\mu_1 \approx 1.167$ and $p_2 < U_{max} - \gamma/\mu_2 = 0.75$ in order to have stability.

Moreover, remark that if $\alpha \geq 0.75$ equilibrium *cannot* be attained, irrespective of the prices. This is so because, in this case, $U_{max} \leq \gamma/\mu_2$; in other words, the fixed delay of subnetwork 2 is higher than the highest delay users can tolerate.

4.2 Total delay

Delay in each subnetwork is shown in Figure 4. Observe that, for subnetwork i, delay is always minimal along the frontier of the equilibrium region $\mathcal{E}_i$.

The tradeoff between prices and delay is fairly evident in these figures. When one of the prices (say, p_1) is kept fixed, delay in the corresponding subnetwork (i.e., d_1) decreases as the other subnetwork's price (p_2, in this case) decreases, while delay d_2 increases. This is due to the fact that a lower p_2 will tend to attract more customers to subnetwork 2, increasing queueing delay in it—and so decreasing queueing delay in subnetwork 1.

4.3 Equilibrium cost

Figure 5 shows that the total price p_{tot} is strictly increasing with p_i. This, indeed, can easily be proved analytically.

4.4 Revenue and optimality

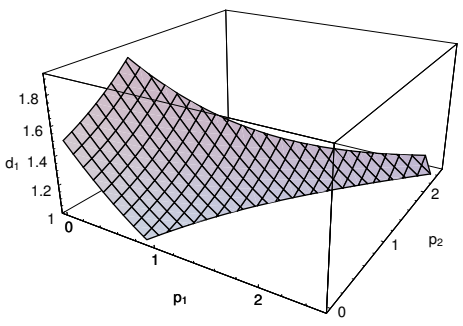
Figure 6 shows the revenue of the network, R, as a function of prices, for three different bandwidth allocations. It is interesting to remark that there appears to be a single maximum (i.e., there is a single pair of prices which is *optimal*, in the sense that revenue is maximized). As α increases, the position of the maximum value of R moves closer to the rear "edge" of the surface, which corresponds to the upper edge of the equilibrium region $\mathcal{E}_2$ in Figure 2.

4.5 Sensitivity of the optimal revenue

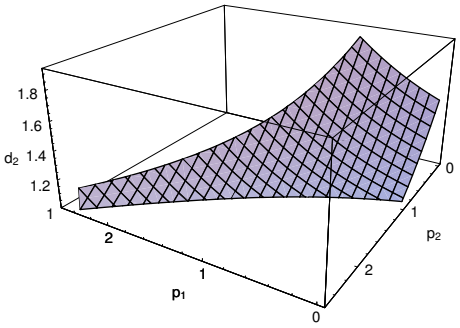
4.5.1 Sensitivity to the bandwidth partition

Let us now examine how the *optimal* revenue, as well as the corresponding prices, delay and total cost, depend on the bandwidth allocation parameter α; see Fig. 7.

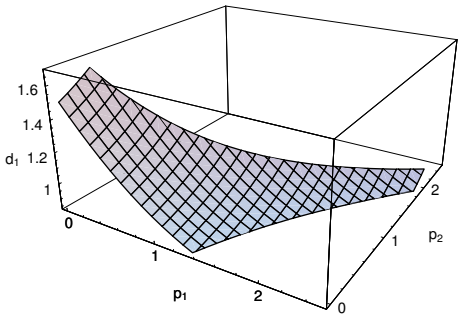
First, remark that the optimal revenue $R_{\text{opt}} = \max_{p_1,p2\geq 0} R$ is highly sensitive to the value of the bandwidth allocation parameter α. Figure 7(a) suggests that the ISP may find interesting, from an economic point of view, to operate at the maximum-revenue allocation α_{opt}. However, since R_{opt} (and, hence, the revenue for any non-optimal set of prices) rapidly decreases for values of α



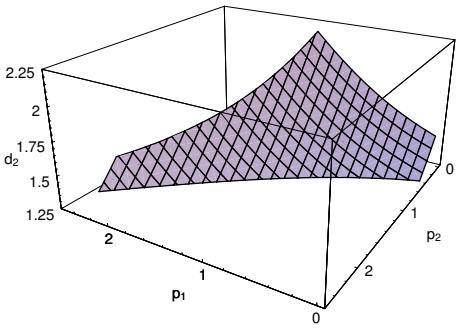
(a) $\alpha = 0.5$, subnetwork 1.



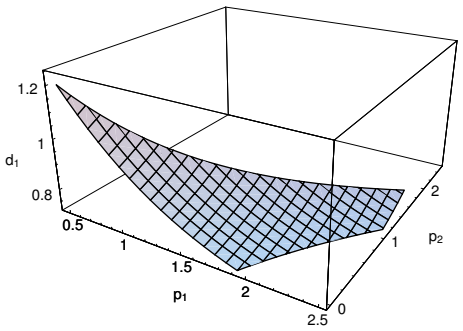
(b) $\alpha = 0.5$, subnetwork 2.



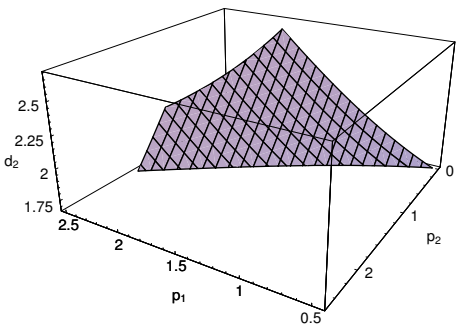
(c) $\alpha = 0.6$, subnetwork 1.



(d) $\alpha = 0.6$, subnetwork 2.



(e) $\alpha = 0.7$, subnetwork 1.

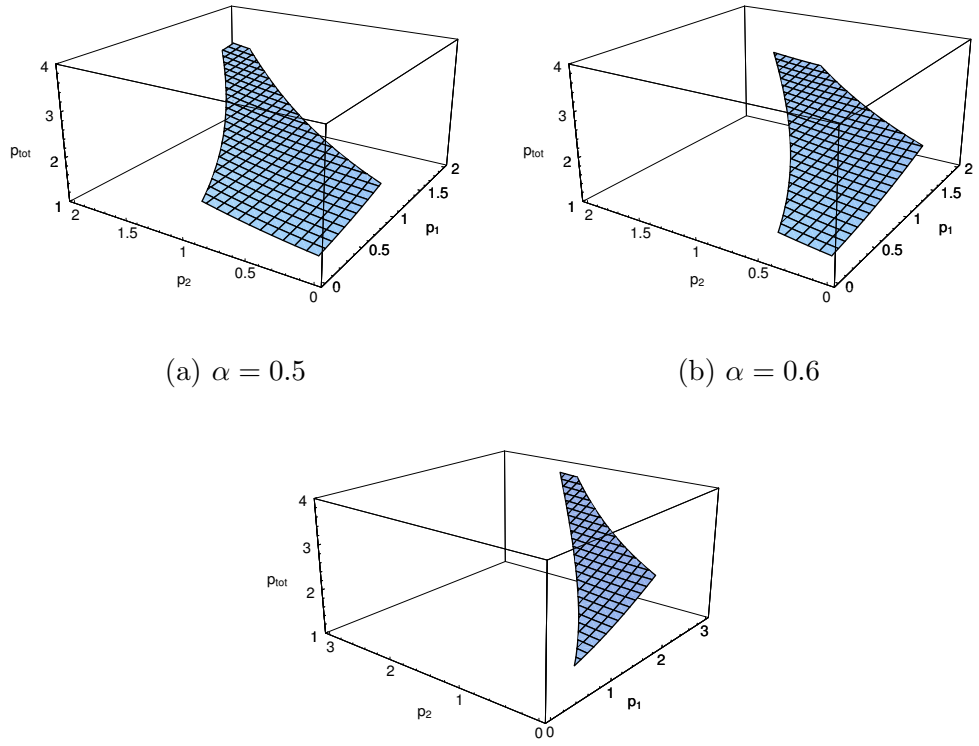


(f) $\alpha = 0.7$, subnetwork 2.

Fig. 4. Delay in each subnetwork, $\bar{F}$ exponential.

above α_{opt}, a too-unequal bandwidth allocation policy may result in lower income. On the other hand, a "safe" bandwidth allocation (say, $\alpha = 0.5$) may result in a maximum revenue as low as $\approx 70\%$ of the highest R_{opt}.

Note that, interestingly enough, the optimal revenue is maximized when the price p_2 of the lowest-capacity subnetwork drops to zero. For higher values of α, since p_2 is at its lowest-possible value, the ISP cannot compensate for the



(a) $\alpha = 0.5$ (b) $\alpha = 0.6$

(c) $\alpha = 0.7$

Fig. 5. Equilibrium cost, $\bar{F}$ exponential.

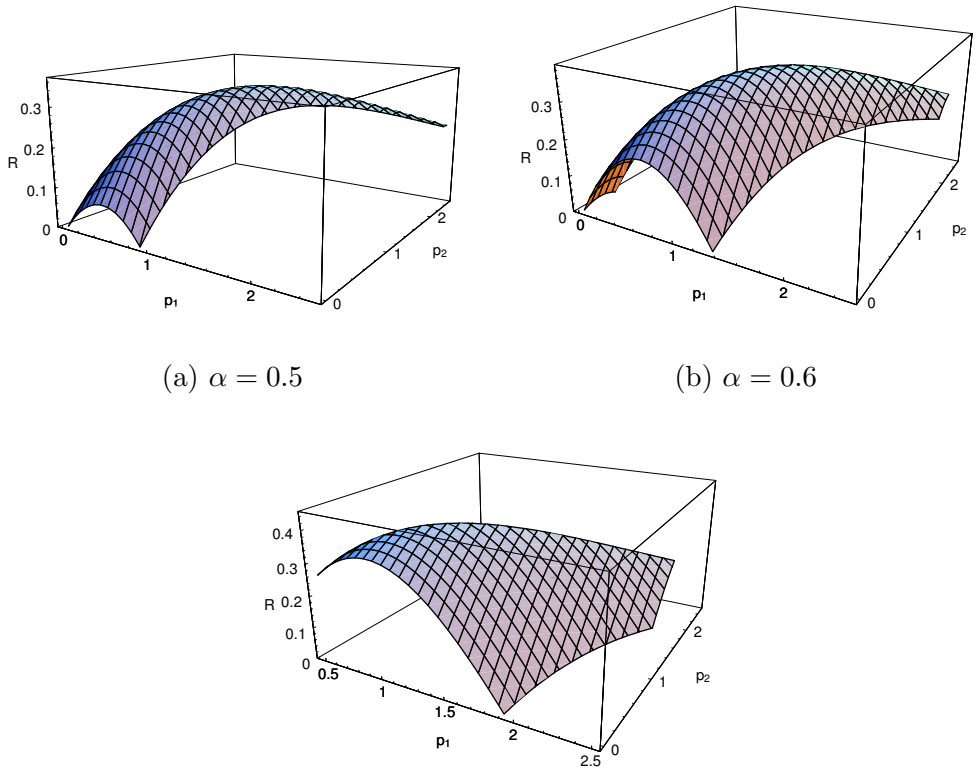
poorer performance of subnetwork 2 by lowering its price, so more and more traffic tends to flow through subnetwork 1 and the price p_1 must be raised in order to have a stable network. For $\alpha < \alpha_{\text{opt}}$, a falling p_2 tends to attract more traffic to subnetwork 2 in spite of the degradation in delay d_2 when α increases, resulting in an increasing revenue.

4.5.2 Sensitivity to the users' valuation of delay

The γ parameter, which expresses the cost per unit of delay, can be regarded as how much users value having a good quality of service (in terms of delay): the higher the value of γ, the higher the impact of delay on the total cost $p_i + \gamma d_i$, and the lower the probability that a given packet enters the network.

Figure 8 illustrates the sensitivity of the optimal revenue to the value of γ, for α in the $[0.5, 0.95]$ range. For $\gamma = 0.25$, R_{opt} is fairly stable for a wide range of bandwidth allocations: $R_{\text{opt}}(0.5)/R_{\text{opt}}(\alpha_{\text{opt}}) \approx 0.87$, with $\alpha_{\text{opt}} \approx 0.91$.

On the other hand, when $\gamma = 2$ (meaning that users' valuation of delay is eight times higher than in the previous case) we have that $R_{\text{opt}}(0.5)/R_{\text{opt}}(\alpha_{\text{opt}}) \approx$



(a) $\alpha = 0.5$ (b) $\alpha = 0.6$

(c) $\alpha = 0.7$

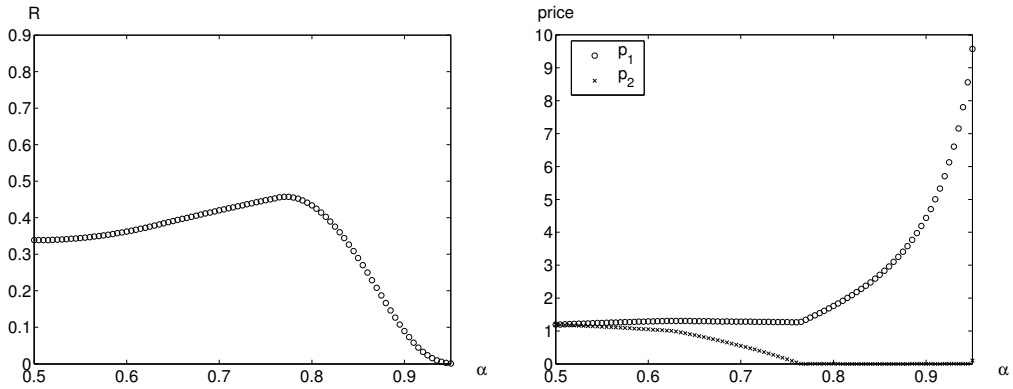
Fig. 6. Revenue, $\bar{F}$ exponential.

0.66, i.e., the "loss" incurred by operating the network at a safe bandwidth allocation is higher than in the $\gamma = 0.25$ case. In other words, the model predicts that revenues are higher and fairly insensitive to bandwidth allocation when users are more tolerant of delay (which is an intuitively appealing result).

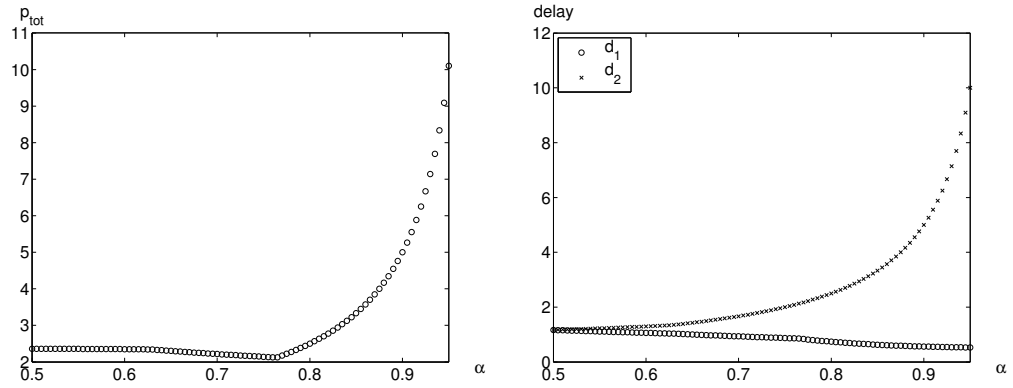
Notice also how the optimal allocation α_{opt} is dependent on γ, going from $\alpha_{\text{opt}} \approx 0.91$ to $\alpha_{\text{opt}} \approx 0.67$. In practical terms, this means that an ISP would have to operate its network on the "safe" side (i.e., with α close to 0.5) if the value of γ cannot be accurately estimated.

5 Comparison with the case of a single network

An interesting question is how well a PMP network performs with respect to a one-tiered, non-PMP network. We will begin by presenting a numerical example, followed by a mathematical demonstration of the results that were observed.



(a) Optimal revenue R_{opt} (b) Optimal prices p_i



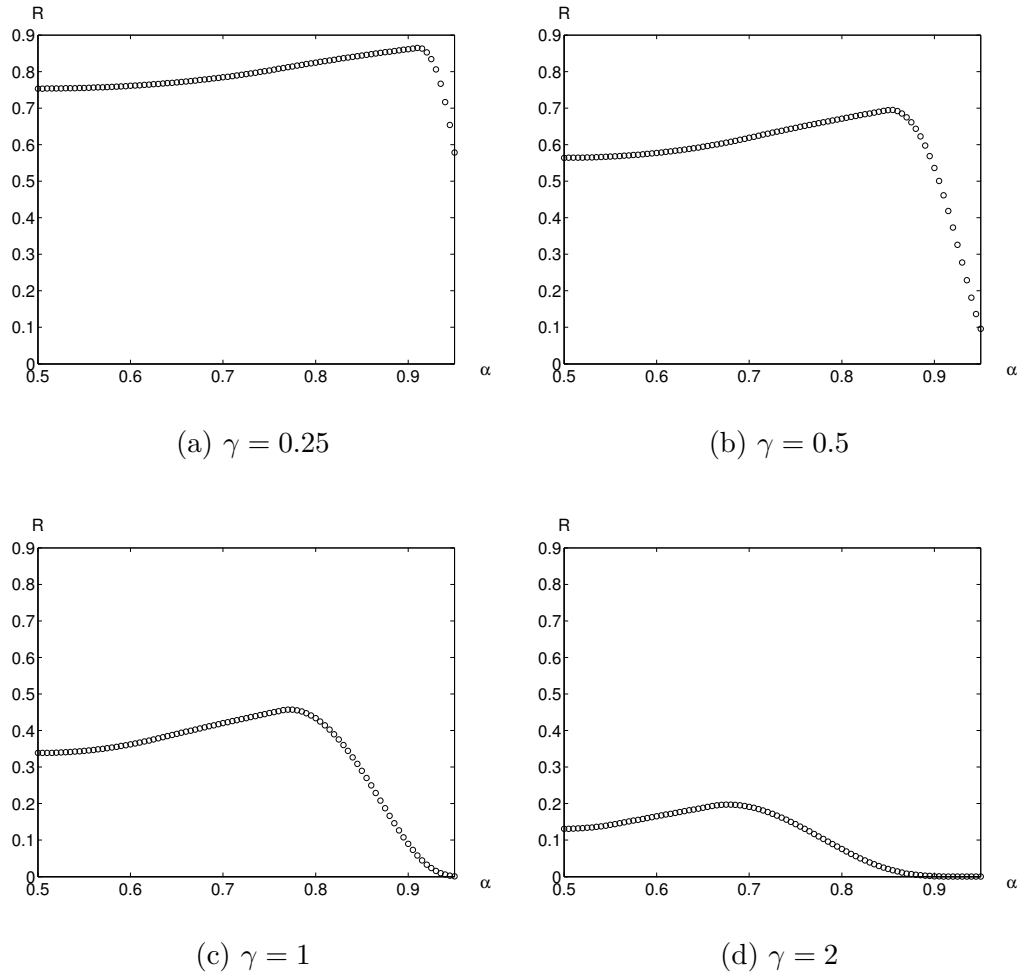
(c) Total cost p_{tot} (d) Delay d_i

Fig. 7. Optimal revenue and prices (and corresponding total cost and delay) as a function of α, with an exponentially-distributed utility.

5.1 Numerical comparison

We are interested in comparing the performance of the two-tiered ($I = 2$) case to that of the one-tiered ($I = 1$) case, under the same set of conditions, both in terms of revenue and network performance measures like delay and input rate. We will revisit here the numerical example studied in Section 4, under the assumption of an exponentially-distributed utility.

We will suppose that both the one-tiered network and the two-tiered PMP network have *the same link capacity* c; in the latter case, this capacity is shared among subnetworks according to Eq. (6). This would correspond to the scenario in which an ISP wants to evaluate the consequences of introducing PMP in its network by deploying the appropriate mechanisms in routers (e.g., packet classification and scheduling mechanisms), but *without increasing the*



(a) $\gamma = 0.25$ (b) $\gamma = 0.5$

(c) $\gamma = 1$ (d) $\gamma = 2$

Fig. 8. Effect of the γ parameter on the optimal revenue, with an exponentially-distributed utility.

installed capacity.

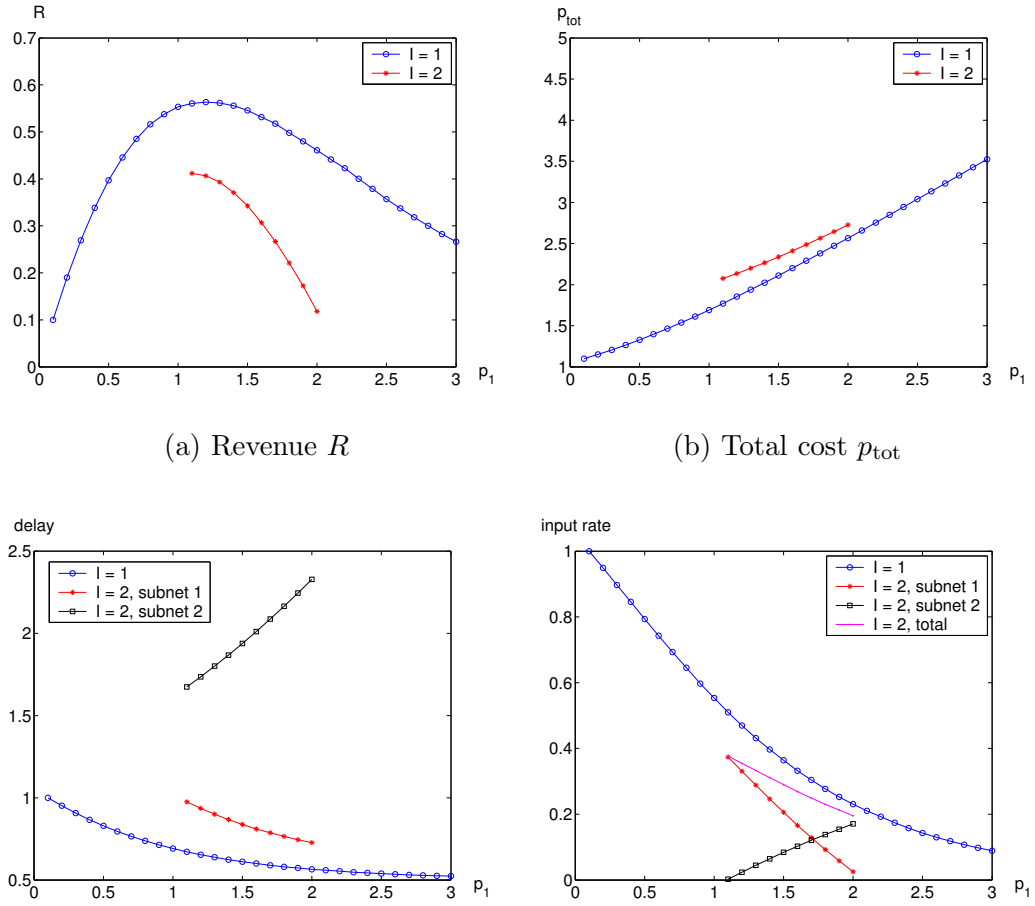
Let us recall the parameters used in this example:

- Total capacity of the network: $c = 2$.
- Potential total arrival rate: $\tilde{\lambda} = 3$.
- Utility: exponentially distributed, with mean $\bar{U} = 1$.
- Cost per unit of delay: $\gamma = 1$.
- Service time distribution: exponential.

For the $I = 2$ case, Figs. 2, 4 5 and 6 illustrate the equilibrium region, delay in subnetworks 1 and 2, total cost and revenue, respectively, as a function of the prices p_1 and p_2.

We will only present here the results for a bandwidth partition $\alpha = 0.7$, because similar results were obtained when using other values of α.

Figure 9 plots the revenue R, the total cost p_{tot}, per-queue delays and input rates, for both the $I = 1$ case and the $I = 2$ case. In the latter, for the sake of clarity we show the values of these as a function of p_1, for a *single* value of p_2 such that the curve of R contains the maximum revenue R_{opt}. Hence, the curve for $I = 2$ corresponds to a "slice" of the surface depicted in Fig. 6, taken at $p_2 \approx 0.4$.



(a) Revenue R

(b) Total cost p_{tot}

(c) Delay d

(d) Input rate λ

Fig. 9. Performance as a function of the price p_1, for $I = 1$ (single network) and $I = 2$ (PMP).

Remark 5 *Note that, in Fig. 9, the range of values for which the revenue and other performance measures are shown is larger in the $I = 1$ case than in the $I = 2$ case. This is due to the fact that an equilibrium condition (in the sense of Eqs. (2) and (4)) must be satisfied in the PMP network, which restricts the allowed values for the p_i, while the single network does not face such equilibrium problems.*

Observe that the maximum revenue R_{opt} is lower (Fig. 9(a)) and the total cost p_{tot} (Fig. 9(b)) is higher when $I = 2$. Total delays, shown in Fig. 9(c),

are higher in both PMP subnetworks. Both per-queue (λ_i) and total (λ) input rates are lower when $I = 2$ (Fig. 9(d)), which accounts for the lower revenues.

Note also that, even if we choose the partition parameter α maximizing the revenue for $I = 2$ in Fig. 8(c), the revenue is still lower than the optimal one obtained for $I = 1$ (Fig. 9(a)).

5.2 Formal comparison

We observed in our preliminary numerical experiments that the revenue generated by a single network ($I = 1$) is greater than the revenue generated by two subnetworks ($I = 2$) when the total network capacity—i.e., the service rate—is unchanged, regardless of the partition parameter α. Let us formally prove this result for the M/M/1 queue in the following theorem.

Theorem 2 *Consider a M/M/1 queue (representing the network) with service rate μ and the response time as the delay cost. The revenue generated by this single queue is greater than the revenue generated by two separate queues with service rates $\alpha\mu$ and $(1-\alpha)\mu$, respectively.*

Proof: Consider first the case $I = 1$ with a price p per packet. Since the response time $d = 1/(\mu - \lambda)$ is also given by $(p_{\text{tot}} - p)/\gamma$, we can write $\lambda = \mu - \frac{\gamma}{p_{\text{tot}} - p}$, which yields the stability equation

$$\bar{F}^{-1}(p_{\text{tot}}) = \frac{1}{\tilde{\lambda}} \left(\mu - \frac{\gamma}{p_{\text{tot}} - p} \right).$$

In the case $I = 2$ with prices p_1 and p_2, we have

$$\begin{aligned} \bar{F}^{-1}(p_{\text{tot}}) &= \frac{1}{\tilde{\lambda}} \left(\alpha\mu - \frac{\gamma}{p_{\text{tot}} - p_1} + (1-\alpha)\mu - \frac{\gamma}{p_{\text{tot}} - p_2} \right) \\ &= \frac{1}{\tilde{\lambda}} \left(\mu - \frac{\gamma}{p_{\text{tot}} - p_1} - \frac{\gamma}{p_{\text{tot}} - p_2} \right). \end{aligned}$$

The proof of the theorem is in two steps: we first show that, for a given p_{tot}, the revenue generated is greater in the case $I = 1$; next, we show that, for all p_{tot} obtained in the case $I = 2$, there exists a p in the case $I = 1$ giving the same p_{tot}.

Hence, for a given p_{tot}, from the above equations in p_{tot} we have that, necessarily,

$$\frac{1}{p_{\text{tot}} - p} = \frac{1}{p_{\text{tot}} - p_1} + \frac{1}{p_{\text{tot}} - p_2},$$

that is,

$$p_2 = p_{\text{tot}} - \frac{(p_{\text{tot}} - p_1)(p_{\text{tot}} - p)}{p - p_1}.$$

As $p_2 \geq 0$ and $p, p_1, p_2 \leq p_{\text{tot}}$, we necessarily have that $p \geq p_1$. By symmetry, we also get that $p \geq p_2$. Since, for a given p_{tot}, the total arrival rate is the same in both the $I = 1$ and the $I = 2$ cases, we obtain that the revenue $(\lambda_1 + \lambda_2)p$ in the case $I = 1$ is greater than $\lambda_1 p_1 + \lambda_2 p_2$ in the case $I = 2$.

Consider now a p_{tot} obtained in the case $I = 2$. Then a price

$$p = p_{\text{tot}} - \frac{1}{\frac{1}{p_{\text{tot}} - p_1} + \frac{1}{p_{\text{tot}} - p_2}}$$

gives the same value of p_{tot} in the case $I = 1$. The only thing that remains to be proven is that such a p is non negative. Since $1/(p_{\text{tot}} - p_1) + 1/(p_{\text{tot}} - p_2) > 1/p_{\text{tot}}$, this is always the case. □

Remark 6 *This result can be related to the well-known problem of "splitting" a server in two. Such separation introduces a supplementary mean waiting time as one of the two servers may by idle while the other has some customers waiting for service, something that does not occur in the single-server case (see, for instance, [15]).*

Remark 7 *The same result applies if we consider the waiting time instead of the response time. Moreover, in the former, the revenue generated in the case $I = 2$ tends to the revenue in the case $I = 1$ if α tends to 1. The reason why this does not happen when considering the response time is that the first class (which gets almost all the service rate when α gets close to 1) has to match the total cost of the second class, which tends to infinity with its mean service time $1/((1 - \alpha)\mu)$.*

6 Extensions

6.1 A multi-application extension

Let us now consider different kinds of applications/customers, each having different requirements, i.e., utility variables. Let K be the number of such classes. Note that we will use the superscript k for denoting *application* classes,

which are different from the *PMP* classes—i.e., subnetworks—noted by the subscript i.

Instead of a total potential arrival rate $\tilde{\lambda}$ and a utility random variable U, we now have a collection of such inputs $(\tilde{\lambda}^{(k)}, U^{(k)})$ for $1 \leq k \leq K$.

Denote by $\lambda^{(k)}$ the total arrival rate of packets for application k, $\lambda_i^{(k)}$ the arrival of such packets in subnetwork i and, as before, λ_i the total arrival rate in subnetwork i. We have $\lambda_i = \sum_{k=1}^{K} \lambda_i^{(k)}$ and $\lambda^{(k)} = \sum_{i=1}^{I} \lambda_i^{(k)}$.

The set of Wardrop equilibrium equations (1) (i.e., stability conditions) is now

$$\begin{cases} \lambda^{(k)} = \sum_{i=1}^{I} \lambda_i^{(k)} = \tilde{\lambda}^{(k)} P(U^{(k)} \geq p_{\text{tot}}) & \text{for } 1 \leq k \leq K, \\ p_i + \gamma d_i = p_{\text{tot}} & \text{for } 1 \leq i \leq I \end{cases} \tag{8}$$

with $d_i = f_i(\lambda_i)$.

Remark 8 *Uniqueness, in a strict sense, will not be valid anymore since two different application classes can exchange their (or some of their) traffic without modifying Eqs. (8).*

Formally, consider a solution of (8) such that $\lambda_{i_1}^{(k_1)}, \lambda_{i_2}^{(k_1)}, \lambda_{i_1}^{(k_2)}, \lambda_{i_2}^{(k_2)} > 0$, with $k_1, k_2 \in \{1, \cdots, K\}$ and $i_1, i_2 \in \{1, \cdots, I\}$. Then replacing $\lambda_{i_1}^{(k_1)}$ by $\lambda_{i_1}^{\prime(k_1)}$, $\lambda_{i_2}^{(k_1)}$ by $\lambda_{i_2}^{\prime(k_1)} = \lambda_{i_1}^{(k_1)} + \lambda_{i_2}^{(k_1)} - \lambda_{i_1}^{\prime(k_1)}$, $\lambda_{i_1}^{(k_2)}$ by $\lambda_{i_1}^{\prime(k_2)} = \lambda_{i_1}^{(k_1)} + \lambda_{i_1}^{(k_2)} - \lambda_{i_1}^{\prime(k_1)}$ and $\lambda_{i_2}^{(k_2)}$ by $\lambda_{i_2}^{\prime(k_2)} = \lambda_{i_1}^{(k_2)} + \lambda_{i_2}^{(k_2)} - \lambda_{i_1}^{\prime(k_2)}$, still provides a solution of (8) because this system depends only on $\lambda^{(k)}$ and λ_i $\forall k, i$ and those total rates are kept unchanged.

Nonetheless, we have the following result on the existence and uniqueness of equilibrium, in terms of p_{tot}, $\lambda^{(k)}$ and λ_i $\forall k, i$:

Proposition 1 *Assume that $p_1 > \cdots > p_I$. Existence and uniqueness of equilibrium is verified, i.e., the $\lambda^{(k)}$ and λ_i $\forall k, i$ exist and are unique provided that*

$$\sum_{k=1}^{K} \tilde{\lambda}^{(k)} \bar{F}^{(k)}(p_1) \geq \sum_{i=1}^{I} f_i^{-1}\left(\frac{p_1 - p_i}{\gamma}\right). \tag{9}$$

Proof: We have $\forall 1 \leq k \leq K$,

$$p_{\text{tot}} = (\bar{F}^{(k)})^{-1}(\lambda^{(k)}/\tilde{\lambda}^{(k)}) \tag{10}$$

and $\forall 1 \leq i \leq I$,

$$p_{\text{tot}} = p_i + \gamma d_i. \tag{11}$$

Note that, according to (10), $\lambda^{(k)}$ decreases continuously when p_{tot} increases, so that $\lambda = \sum_{k=1}^{K} \lambda^{(k)}$ decreases continuously when p_{tot} increases.

Also, according to (11), λ_i increases continuously when p_{tot} increases, so that $\lambda = \sum_{i=1}^{I} \lambda_i$ increases continuously with p_{tot}. Thus, there exists a unique p_{tot} such that both λ (obtained from Eq. (10) and (11)) are equal, as long as the value of $\sum_{k=1}^{K} \lambda^{(k)}$ (from Eq. (10)) is greater that the value of $\sum_{i=1}^{I} \lambda_i$ (from Eq. (11)) when $p_{\text{tot}} = p_1 = \max\{p_i\}$. This gives condition (9).

λ_i and $\lambda^{(k)}$ $\forall k, i$ immediately follow. $\square$

Note that the strict non-uniqueness result is not a problem for computing the revenue since the objective function depends only on λ_i and p_i, and not directly on $\lambda_i^{(k)}$.

Remark 9 *The multi-application case can be seen as a single-application case if, instead of considering the set* $\{(\tilde{\lambda}^{(k)}, F^{(k)}) | 1 \leq k \leq K\}$ *with* $F^{(k)}$ *the cumulative distribution of random variable* $U^{(k)}$*, we consider the "mixture"*

$$\left(\sum_{k=1}^{K} \tilde{\lambda}^{(k)}, \frac{\sum_{k=1}^{K} \tilde{\lambda}^{(k)} F^{(k)}}{\sum_{k=1}^{K} \tilde{\lambda}^{(k)}} \right).$$

6.2 A multi-application extension with per-application delay valuations

6.2.1 Model description and analysis

Now let us suppose that, as in the previous subsection, there are K types of applications/customers, but with a delay valuation $\gamma^{(k)}$ for the class-k application. In this case, the set of stability equations (1) becomes

$$\begin{cases} \lambda^{(k)} = \sum_{i=1}^{I} \lambda_i^{(k)} = \tilde{\lambda}^{(k)} P(U^{(k)} \geq p_{\text{tot}}^{(k)}) & \text{for } 1 \leq k \leq K, \\ p_i + \gamma^{(k)} d_i = p_{\text{tot}}^{(k)} & \text{for } 1 \leq i \leq I \text{ and } 1 \leq k \leq K \end{cases} \tag{12}$$

with $d_i = f_i(\lambda_i)$.

This means that the set (12) has to be verified for all k, but with a different p_{tot} for each k. Note however that the equations cannot be solved independently $\forall k$ since they are related through the queueing delays d_i $(1 \leq i \leq I)$.

Remark 10 *In this case, the system behaves like a game where each player k (i.e., each application) tries to stabilize its traffic, that is, it tries to equalize*

$p_i + \gamma^{(k)} d_i$ *for all* i*, meaning that* $p_{\text{tot}}^{(k)}$ *exists. In this sense, a solution of (12), if it exists, is a Nash equilibrium between all the classes.*

Proving the existence and uniqueness of the solution is more complicate in this case, since we do not have a single p_{tot}, so that the proof in Section 6.1 does not stand anymore. Therefore, finding out conditions for existence and uniqueness is still an open question.

Nevertheless, based on the results in previous sections we can still find a *necessary*, but not sufficient, condition for equilibrium.

Proposition 2 *Assume that* $p_1 \geq p_2 \geq \cdots \geq p_I$*. A necessary condition for solving (12) is that*

$$\sum_{k=1}^{K} \tilde{\lambda}^{(k)} \bar{F}^{(k)}(p_1) \geq \sum_{i=1}^{I} f_i^{-1} \left(\frac{p_1 - p_i}{\gamma^{(k)}} \right). \tag{13}$$

Proof: We have, $\forall k$,

$$p_{\text{tot}}^{(k)} = (\bar{F}^{(k)})^{-1} \left(\frac{\lambda^{(k)}}{\tilde{\lambda}^{(k)}} \right). \tag{14}$$

If (1) is verified, then we also have that, $\forall i$,

$$p_{\text{tot}}^{(k)} = p_i + \gamma^{(k)} d_i. \tag{15}$$

From (14), we can see that $\lambda^{(k)}$ decreases when $p_{\text{tot}}^{(k)}$ increases, so $\lambda = \sum_{k=1}^{K} \lambda^{(k)}$ decreases when $p_{\text{tot}}^{(k)}$ increases. From (15), we see that λ_i increases when $p_{\text{tot}}^{(k)}$ increases, so $\lambda = \sum_{i=1}^{I} \lambda_i$ increases as well. In order to ensure that both quantities be equal, the inequality given in the proposition must hold when $p_{\text{tot}}^{(k)} = \max_i p_i = p_1$, meaning that $\sum_{k=1}^{K} \lambda^{(k)} = \sum_{i=1}^{I} \lambda_i$ at the minimum $p_{\text{tot}}^{(k)}$ value. □

The reason why this proof does not provide a sufficient condition for the existence and uniqueness of equilibrium is that, in the proof, we look at the existence of $p_{\text{tot}}^{(k)}$ for a given k, and not at the existence of the $p_{\text{tot}}^{(k)}$ all together.

6.2.2 Numerical example

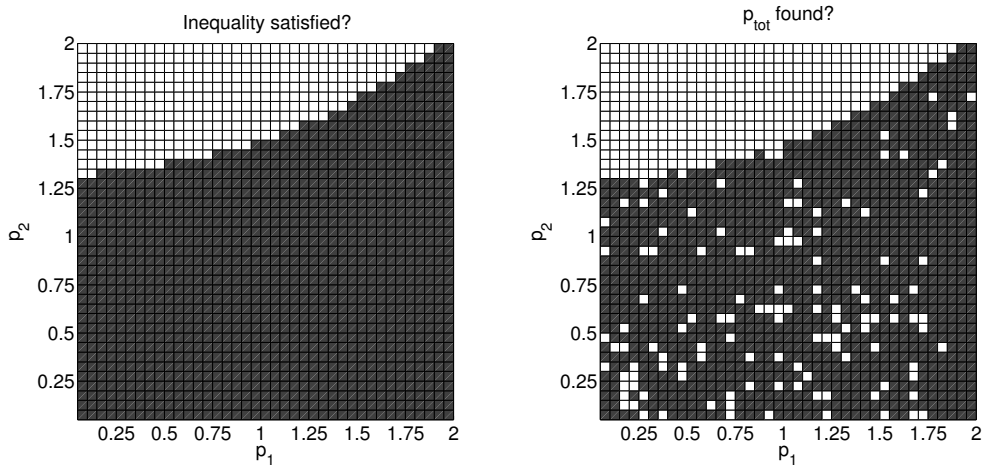
We will present a numerical example that illustrates the problems described in Section 6.2.1. Let us consider a two-tiered PMP network ($I = 2$) with $K = 2$ application classes having different delay valuations $\gamma^{(k)}$; we will suppose that

application 1 is more delay-sensitive than application 2. The parameters take the following values:

- Total capacity of the network: $c = 10$.
- Bandwidth partition: $\alpha = 2/3$.
- Potential total arrival rate, per application: $\tilde{\lambda}^{(1)} = \tilde{\lambda}^{(2)} = 10$.
- Utility: exponentially distributed, with mean $\bar{U}^{(1)} = 1$ and $\bar{U}^{(2)} = 2$.
- Cost per unit of delay: $\gamma^{(1)} = 2, \gamma^{(2)} = 1$.
- Service time distribution: exponential.

The results below were obtained by trying to solve Eq. (12) for $\lambda^{(k)}$, λ_i and $p_{\text{tot}}^{(k)}$, with $k = 1, 2$ and $i = 1, 2$, while taking into account the necessary condition (13). A solution of (12) was searched for every price pair (p_1, p_2) such that $p_i = 0.05 \cdot n$ with $n = 1, \ldots, 40$ and $i = 1, 2$.

Figure 10(a) shows, for every pair of prices, whether the necessary condition (13) is satisfied or not, whereas Fig. 10(b) indicates whether the set (12) could be solved or not—that is, whether $p_{\text{tot}}^{(k)}, k = 1, 2$ could be found. Each square on the grid corresponds to a given (p_1, p_2). A gray square denotes that the condition is satisfied and that the $p_{\text{tot}}^{(k)}$ are found, respectively, while white squares denote the opposite situation.



(a) Necessary condition for equilibrium.

(b) The set of equations (12) may be solved.

Fig. 10. Per-application delay valuations: an example.

The gray region in Fig. 10(a) is equivalent to the equilibrium region in the single-application case. Note however that, for some prices (p_1, p_2) satisfying (13), a solution to (12) could not be found: these are the white "holes" in the gray region of Fig. 10(b).

The sparsity of the set of solutions illustrates the difficulty of finding a necessary and sufficient condition over the set of prices, corresponding to the general difficulty of finding out conditions for Nash equilibrium.

6.3 A model including losses and delays as QoS requirements

Another interesting extension of our model would be to consider other QoS metrics than delay only. We consider here that quality, as perceived by a user, is a mixture of delay and loss probability.

For class i, the actual *total* cost function is

$$p_i + \gamma d_i + \zeta B_i$$

where B_i is the loss probability (for class i) and ζ is a constant converting loss in money. A packet enters network i if $i = \text{argmin}_j p_j + \gamma d_j + \zeta B_j$ and $U \geq p_i + \gamma d_i + \zeta B_i$.

In equilibrium, $p_{\text{tot}} = p_i + \gamma d_i + \zeta B_i$, $\forall i$. The actual total arrival rate is still

$$\lambda = \tilde{\lambda} P(U \geq p_{\text{tot}}).$$

The system of equations (1) becomes

$$\begin{cases} \sum_{i=1}^{I} \lambda_i = \tilde{\lambda} P(U \geq p_{\text{tot}}) \\ p_i + \gamma d_i + \zeta B_i = p_{\text{tot}}, \text{ for } 1 \leq i \leq I \end{cases} \tag{16}$$

where $d_i = f_i(\lambda_i)$ and $B_i = g_i(\lambda_i)$ depend both on λ_i. Define the function h_i as $h_i = \gamma f_i + \zeta g_i$.

For simplicity, we consider here only the waiting time; extending this model to take into account the response time is straightforward.

p_{tot} is then given by the equation:

$$p_{\text{tot}} = \bar{F}^{-1}\left(\frac{\sum_{i=1}^{I} h_i^{-1}(p_{\text{tot}} - p_i)}{\tilde{\lambda}}\right) \tag{17}$$

and, $\forall 1 \leq i \leq I$,

$$\lambda_i = h_i^{-1}(p_{\text{tot}} - p_i).$$

Proposition 3 *Assume that $p_1 > \cdots > p_I$. The system of equations (16) has a unique solution if and only if*

$$p_1 \leq \bar{F}^{-1}\left(\frac{1}{\tilde{\lambda}}\sum_{i=1}^{I} h_i^{-1}\left(p_1 - p_i\right)\right). \tag{18}$$

This corresponds to the stability condition described in Section 3.

Proof: The proof follows closely that of Theorem 1. Functions f_i and g_i are strictly increasing functions of λ_i. Then, h_i and h_i^{-1} are also strictly increasing. Following the arguments of Theorem 1, the left-hand side of Eq. 17 is strictly increasing whereas the right-hand side is decreasing which means that, if a solution exists, it is unique. By the same kind of arguments than in Theorem 1, a necessary and sufficient condition for existence is Eq. (18). □

Thus, including losses in the model does not add any theoretical complexity. Even if inverting h_i analytically is harder than inverting f_i (when considering the case of delay only), it is still numerically simple due to the monotonicity of h_i.

6.4 Time-of-day Pricing

In practice, demand is varying over time. Time-of-day pricing, as practiced for instance in electrical power pricing, is an interesting way to manage the traffic flow.

Time-of-day pricing is modeled as follows. We assume that a day is decomposed in different periods of time during which the demand (meaning the utility) follows a relatively constant distribution. In each period j of time, the random variable representing user's utility is U_j, with cumulative distribution function F_j. For every j, the prices are *independently* determined like done for problem (1).

7 Conclusions

In this paper, we have introduced a mathematical model of the Paris Metro Pricing (PMP) scheme for charging packet networks. This pricing method looks convenient for Internet Service Providers since it is fairly easy to implement and deploy in an ISP network using current, off-the-shelf technologies. Even if PMP does not provide strict QoS guarantees, users would probably

appreciate it since the total charge is linear in the volume of data, hence predictable.

Our model has allowed us to find some necessary and sufficient conditions on the prices required to obtain an equilibrium. Extensions to the model, allowing to take into account the multi-application case and multiple QoS requirements, as well as time-of-day pricing, are also included. Many numerical illustrations are provided.

The model has pointed out a possible drawback of the PMP scheme, namely, that for a given network capacity revenues may be lower in a network implementing PMP. This limitation, coming from the adopted bandwidth-sharing policy among subnetworks, could be alleviated by means of more efficient scheduling mechanisms.

As directions for future research, we are interested in trying to carry out the same kind of analysis for the round-robin scheme of [10] (to tackle the problem of revenue maximization). Also, the case of strict QoS requirements could be investigated.

Acknowledgements

This work was partially funded by INRIA's Cooperative Research Action "Prixnet" (`http://www.irisa.fr/armor/Armor-Ext/RA/prixnet/ARC.htm`).

References

[1] L. DaSilva, Pricing of QoS-Enabled Networks: A Survey, IEEE Communications Surveys & Tutorials 3 (2).

[2] P. Dolan, Internet Pricing. Is the End of the World Wide Wait in View?, Communications & Strategies 37 (2000) 15–46.

[3] M. Falkner, M. Devetsikiotis, I. Lambadaris, An Overview of Pricing Concepts for Broadband IP Networks, IEEE Communications Surveys & Tutorials 3 (2).

[4] T. Henderson, J. Crowcroft, S. Bhatti, Congestion Pricing. Paying Your Way in Communication Networks, IEEE Internet Computing September/October (2001) 85–89.

[5] B. Stiller, P. Reichl, S. Leinen, Pricing and Cost Recovery for Internet Services: Practical Review, Classification, and Application of Relevant Models, Netnomics 2 (1).

[6] B. Tuffin, Charging the Internet Without Bandwidth Reservation: An Overview and Bibliography of Mathematical Approaches, Tech. Rep. 1434, IRISA (January 2002).

[7] P. Fishburn, A. Odlyzko, Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet, in: Proceedings of ICE98, ACM, 1998, pp. 128–139.

[8] A. Odlyzko, Paris Metro Pricing for the Internet, in: ACM Conference on Electronic Commerce (EC'99), 1999, pp. 140–147.

[9] R. Gibbens, R. Mason, R. Steinberg, Internet Service Classes Under Competition, IEEE Journal on Selected Areas in Communications 18 (12) (2000) 2490–2498.

[10] P. Dube, V. Borkar, D. Manjunath, Differential Join Prices for Parallel Queues: Social Optimality, Dynamic Pricing Algorithms and Application to Internet Pricing, in: Proceedings of IEEE INFOCOM 02, 2002.

[11] A. Odlyzko, The History of Communications and its Implications for the Internet, Tech. rep., AT&T Labs (2000).

[12] E. Altman, L. Wynter, Equilibrium, Games, and Pricing in Transportation and Telecommunication Networks, Tech. Rep. 4632, INRIA (2002).

[13] M. Honig, K. Steiglitz, Usage-Based Pricing of Packet Data Generated by a Heterogeneous User Population, in: Proceedings of IEEE INFOCOM 95, 1995, pp. 867–874.

[14] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié, J. Roberts, Statistical Bandwidth Sharing: A Study of Congestion at Flow Level, in: Proceedings of ACM SIGCOMM'01, 2001, pp. 111–122.

[15] K. Trivedi, Probability and Statistics with Reliability, Queuing, and Computer Science Applications, John Wiley & Sons, 2002, second edition.