Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Pricing strategies under heterogeneous service requirements

M.R.H. Mandjes

REPORT PNA-E0408 JUNE 2004

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

Pricing strategies under heterogeneous service requirements

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

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: Packet networks -- differentiated services -- pricing -- congestion -- microeconomics -- negative externalities -- game theory

Pricing strategies under heterogeneous service requirements

Michel Mandjes *

Abstract

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

Key words: Packet networks – differentiated services – pricing – congestion – microeconomics – negative externalities – game theory

*The author is with CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands, and Faculty of Mathematical Sciences, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands. 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. Email: michel@cwi.nl.

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

Pricing. Without an appropriate pricing scheme, any prioritization is useless; if there were no price difference between the priority classes, all users would opt for the high-priority class. In other words: the prices of the priority classes should give users an incentive to join the 'right' priority class. In terms of the delay-tolerant user (or, shortly, the *data* user) and the delay-sensitive user (or, shortly, the *voice* user): voice users are encouraged to use the high-priority class, whereas data users are given an incentive to join the low-priority class. This is done by imposing a higher charge on the high-priority class. A next question is: how should the network provider choose the prices for both classes in order to maximize its profit?
Here two models can be distinguished. In the first model the provider assigns a priority class to each user type – for instance, the provider can decide that the voice customers are directed to the high-priority queue, and the data users to the low-priority queue. This model of 'dedicated classes' (or 'implicit supply of service', in Shenker's [13] terminology) is relatively simple to analyze, as the network users have only two alternatives: joining the network or not.
The harder, but perhaps more realistic, model is the model with 'open classes' (or 'explicit supply of service', as it is called in [13]), in which the users can choose between the priority classes. It is not clear beforehand whether the prices that optimize the profit in the dedicated-classes model, are also profit optimizing for the open-classes model. The reason is that the prices found in the dedicated-classes model might lead to a situation in which data (voice) users might appreciate the high-(low-)priority class more. In other words: it is not *a priori* clear whether the optimal prices from the dedicated-classes model lead to an *incentive-compatible* situation in the open-classes model.

Incentive-compatibility. In economic terms, in the model with open classes, the users of

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

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

$$u_d(\mathbb{E}D_{\rm L}) - p_{\rm L} \geq \max\{u_d(\mathbb{E}D_{\rm H}) - p_{\rm H}, 0\}; \quad u_v(\mathbb{E}D_{\rm H}) - p_{\rm H} \geq \max\{u_v(\mathbb{E}D_{\rm L}) - p_{\rm L}, 0\}. \quad (1)$$

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

$$v_d \cdot \mathbb{E}D_{\rm L} + p_{\rm L} \leq \min\{v_d \cdot \mathbb{E}D_{\rm H} + p_{\rm H}, 0\}; \quad v_v \cdot \mathbb{E}D_{\rm H} + p_{\rm H} \leq \min\{v_v \cdot \mathbb{E}D_{\rm L} + p_{\rm L}, 0\}.$$

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

We believe that some aspects of the model of [10] do not apply to the situation of competing data and voice users described above. In the first place, clearly the choice of the penalty functions in [10] is restrictive. As argued above, for low values 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 [10], 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 Odlyzko [11, 12]: 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. Using game-theoretic techniques, [2] argues that this mechanism, known as *Paris Metro Pricing*, does not work if there are multiple competing providers: in order to maximize profits the providers rather focus on one user type. Principles behind congestion pricing are given in, e.g., [3, 6]; the former reference explicitly covers heterogeneous users. There are many references with more practical reflections on pricing in multiservice networks, see for instance [1, 15], and several articles in [8].

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: $u_d(\mathbb{E}D) > u_v(\mathbb{E}D)$. First we look at the situation in which there are large populations of 'potential' voice and data users sharing a FIFO queue. We see that, depending on the value of the link speed μ, the network population will consist of just one class. For small μ (i.e., the link is relatively slow) data will dominate, whereas for fast links voice will push aside data. This situation is considered in *Section 2*. We focus on prices that maximize the provider's profit, which is slightly different from the 'net value' maximization problem solved in [9, 10] (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 [10]). *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: the customers have more options, and therefore the incentive-compatibility requirement is more involved. We find that, depending on the value of the link rate μ, different regimes are optimal: for small μ only data users will be present, for moderate μ the high-priority class is used by voice and the low-priority class by data, whereas for large μ voice users dominate.

Strikingly, even in the cases where only one type of traffic is present (i.e., small and large μ), it is optimal (i.e., profit maximizing) to use both the high-priority and low-priority queue. In other words, even for homogeneous users it is beneficial to introduce service differentiation (and price differentiation). This somewhat counterintuitive result is further explained in *Section 5*. This section also contains a discussion on the specific shape of the utility function, as well as a numerical example.

The paper is concluded by a model in which throughput (rather than packet delay) is the main performance measure. We consider a stream of jobs that is served according to the processor sharing discipline [5, Ch. IV]. For a job of given size x, we can (given the load of the queue and the service speed) compute the required transmission time, and hence the throughput

during the transmission. The utility is an increasing function of the throughput; we assume that the utility curve $U_x(\cdot)$ is parametrized by the job size x. *Section 6* analyzes the situation in which a volume charge is imposed on the jobs (i.e., a fixed price per byte). It shows that under specific assumptions on the ordering of the utility curves, it is beneficial to discriminate the jobs on the basis of their *size*: if $U_x(\cdot)$ decreases in x, the small jobs (usually referred to as *web mice*) are preferred over the larger jobs (*elephants*).

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, unless stated otherwise.

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. [9]. 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

$$U_d(\mathbb{E}D) := u_d(\mathbb{E}D) - p, \quad \text{with} \quad u_d(y) := y^{-\alpha_d};$$

$$U_v(\mathbb{E}D) := u_v(\mathbb{E}D) - p, \quad \text{with} \quad u_v(y) := y^{-\alpha_v},$$

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

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

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

- Similarly, with only voice users,

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

  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

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

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* [14].

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* as the expected number of packets sent (by the users who subscribe to the network) per unit time, multiplied by the price per packet. From the above, this profit function $\Pi(p,\mu)$, for a given price $p > 0$ and service rate μ, reads

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

Notice that in fact this profit function $\Pi(\cdot,\cdot)$ should have been decreased by the provider's *costs*. Important components of these costs are

- The *service costs*, for instance the costs related to the billing and invoicing process. These are increasing in the usage (most notably the numbers of customers N). We neglect these costs, as taking them into account does not really provide additional insight, whereas it makes the resulting expressions less explicit.
- The *equipment costs*, i.e., the costs of (the purchase of) the router. These are increasing in the link rate μ. We assume that the time scale on which the provider can adapt his capacity μ is relatively long, so μ is *not* a decision variable.

Hence the provider wishes to maximize $\Pi(p,\mu)$ over $p \geq 0$. Notice that this function is continuous in $p = 1$.

Proposition 2.1 *The profit function is given by*

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

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

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

- Similarly, on $\mathbb{R}_+$, $f_v(p)$ is maximized by

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

which is smaller than μ^{α_v}. Hence, with $\mu_v := 1 + {\alpha_v}^{-1}$, the maximum value of $\Pi(p,\mu)$ on $p \in [1,\infty)$ is

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

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

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

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 $f_d(\mu) \ge \mu - 1$ for $\mu \in [0, \mu_d]$.

- Also $f_v(\mu_v) = \mu_v - 1$. With

$$f_v'(\mu) = \left(\frac{\mu\alpha_v}{\alpha_v+1}\right)^{\alpha_v} > 1$$

on (μ_v, ∞), this yields $f_v(\mu) \ge \mu - 1$ for $\mu \in [\mu_v, \infty)$.

We arrive at

$$\Pi^\star(\mu) = g(\mu) = \max\left\{f_d(\mu), f_v(\mu)\right\}.$$

This completes our proof. □

The following corollary states that for small (large) link rates data users (voice users, respectively) dominate. We can compute the critical service rate $\mu^\star$ at which the system changes from the data-regime to the voice-regime.

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

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

for all

- *$\mu < \mu^\star$ 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 $p_d < 1$;*
- *$\mu > \mu^\star$ 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. It depends on the specific values of the link rate μ and the 'utility-parameters' α_d and α_v which type of customers will dominate. In this section we concentrate on ways to satisfy the demands of both classes. Adhering to the principles explained in [13], 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 $\mu < \mu^\star$, so in a FIFO system the voice users would not get any service. We now suppose that they get strict service priority over the data sources. We assume that the voice users are directed to the high priority queue, and the data users to the low priority queue. 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

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

Here we assume that the service of a low-priority packet can be interrupted when high-priority 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

$$N_v(p_{\mathrm{L}}, p_{\mathrm{H}}) = \frac{\mu - \sqrt[\alpha_v]{p_{\mathrm{H}}}}{\lambda_v}$$

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

$$N_d(p_{\mathrm{L}}, p_{\mathrm{H}}) = \begin{cases} {\lambda_d}^{-1} \left(\sqrt[\alpha_v]{p_{\mathrm{H}}} - \mu \sqrt[\alpha_d]{p_{\mathrm{L}}} / \sqrt[\alpha_v]{p_{\mathrm{H}}} \right) & \text{if } p_{\mathrm{L}} < p_{\mathrm{H}}^{2\alpha_d/\alpha_v}/\mu^{\alpha_d} \text{ and } p_{\mathrm{H}} \le \mu^{\alpha_v}; \\ {\lambda_d}^{-1} \left(\mu - \sqrt[\alpha_d]{p_{\mathrm{L}}} \right) & \text{if } p_{\mathrm{L}} < \mu^{\alpha_d} \text{ and } p_{\mathrm{H}} > \mu^{\alpha_v}; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Notice that $N_d(p_{\mathrm{L}}, p_{\mathrm{H}})$ 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 $p_{\mathrm{H}} \equiv \mu^{\alpha_v}$ or $p_{\mathrm{L}} = p_{\mathrm{H}}^{2\alpha_d/\alpha_v}/\mu^{\alpha_d}$.
To obtain the optimal prices, we have to perform the following optimization:

$$\Pi_{\mathrm{D}}^{\star}(\mu) = \max_{p_{\mathrm{L}}>0, p_{\mathrm{H}}>0} \Pi_{\mathrm{D}}(p_{\mathrm{L}}, p_{\mathrm{H}}, \mu),$$

with

$$\Pi_{\mathrm{D}}(p_{\mathrm{L}}, p_{\mathrm{H}}, \mu) := \lambda_d \cdot N_d(p_{\mathrm{L}}, p_{\mathrm{H}}) \cdot p_{\mathrm{L}} + \lambda_v \cdot N_v(p_{\mathrm{L}}, p_{\mathrm{H}}) \cdot p_{\mathrm{H}},$$

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

$$\max_{p_{\mathrm{L}}>0, p_{\mathrm{H}}>0} (\mu - \sqrt[\alpha_v]{p_{\mathrm{H}}}) p_{\mathrm{H}} + \left(\sqrt[\alpha_v]{p_{\mathrm{H}}} - \mu \frac{\sqrt[\alpha_d]{p_{\mathrm{L}}}}{\sqrt[\alpha_v]{p_{\mathrm{H}}}} \right) p_{\mathrm{L}}. \tag{4}$$

We compute this maximum in two steps. First we find the optimizing value of $p_{\rm L}$ for given $p_{\rm H}$. Subsequently, we maximize over $p_{\rm H}$.

STEP 1. First find the optimal $p_{\rm L}$ for a given value of $p_{\rm H}$. Differentiation to $p_{\rm L}$ and equating to 0 yields

$$p_{\rm L}(p_{\rm H}) = \left(\frac{\alpha_d}{\mu(\alpha_d+1)}\right)^{\alpha_d} p_{\rm H}^{2\alpha_d/\alpha_v},$$

which is indeed smaller than $p_{\rm H}^{2\alpha_d/\alpha_v}/\mu^{\alpha_d}$. Directly from (3) and (4), we get that $\Pi_{\rm d}^{\star}(\mu)$ equals

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

STEP 2. Now we find the profit-maximizing value of $p_{\rm H}$. It is straightforward that $g(0) = 0$ and $g'(0) > 0$. Moreover,

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

$$g''(p) = -\frac{\alpha_v+1}{\alpha_v} \cdot \frac{1}{\alpha_v} \cdot p^{1/\alpha_v-1} +$$

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

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

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

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 Lemma 3.2 and Lemma 3.3. We first define

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

It is easy to verify that $\mu^\star_- < \mu^\star$ and that $\mu^\star_- < 1$.

Lemma 3.2 *For $\mu \in (\mu^\star_-, \mu^\star)$ the 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 $\mu \in (\mu^\star_-, \mu^\star)$. This is a matter of straightforward calculus. □

Lemma 3.3 *For $\mu \in (0, \mu^\star_-)$ the function $g(\cdot)$ is non-decreasing 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 $\mu \in (0, \mu^\star_-)$ 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

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

  for all $\beta \in [0, 1]$ and $\mu \in (0, \mu^\star_-)$. 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 $\mu = \mu^\star_-$. 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 $2\alpha_d > \alpha_v$,

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

This proves the lemma. □

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

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

- $\mu \in (0, \mu^\star_-)$*: 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

  $$p_{\rm H} := \mu^{\alpha_v} \quad \textit{and} \quad p_{\rm L} := \left(\frac{\mu\alpha_d}{\alpha_d + 1}\right)^{\alpha_d}.$$

- $\mu \in (\mu^\star_-, \mu^\star)$*: 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 $\bar{p}_{\rm H} \in [0, \mu^{\alpha_v}]$ with $g'(\bar{p}_{\rm H}) = 0$. The profit-maximizing prices are

  $$p_{\rm H} := \bar{p}_{\rm H} \quad \textit{and} \quad p_{\rm L} := \left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} \bar{p}_{\rm H}^{\,2\alpha_d/\alpha_v}.$$

The proposition implies that for $\mu \in (0, \mu^\star_-)$ 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 $\mu \in (\mu^\star_-, \mu^\star)$ 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., $\mu > \mu^\star$. Again we find that it is not always benificial to prioritize traffic: for very fast link rates a 'voice-only solution' generates higher profit; there is a threshold link speed $\mu^\star_+$.

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 $p_{\rm H}$ and $p_{\rm L}$, 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 $p_{\rm H} < 1$ the high priority queue will be used exclusively by data users, and if $p_{\rm H} > 1$ by voice users; the same holds for the low priority queue. The procedure of Section 3 does not guarantee that $p_{\rm L} < 1$ and $p_{\rm H} > 1$. 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 [14]: 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 $p_{\rm L}$ and $p_{\rm H}$ 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 Section 2 and 3, we first analyze the network population for given prices. For all combinations of prices $(p_{\rm L}, p_{\rm H})$ we analyze the type of equilibrium.

- **$\mathbf{R}_-$.** 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

$$\Pi_{\rm O}(p_{\rm L}, p_{\rm H}, \mu) := (\mu - \sqrt[\alpha_d]{p_{\rm H}})p_{\rm H} + \left(\sqrt[\alpha_d]{p_{\rm H}} - \mu \sqrt[\alpha_d]{\frac{p_{\rm L}}{p_{\rm H}}} \right) p_{\rm L},$$

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

- **$\mathbf{R}_+$.** If both prices are larger than 1, α_d is replaced by α_v:

$$\Pi_{\rm O}(p_{\rm L}, p_{\rm H}, \mu) := (\mu - \sqrt[\alpha_v]{p_{\rm H}})p_{\rm H} + \left(\sqrt[\alpha_v]{p_{\rm H}} - \mu \sqrt[\alpha_v]{\frac{p_{\rm L}}{p_{\rm H}}} \right) p_{\rm L},$$

  to be maximized over R_+ with prices $p_{\rm L} > 1$ and $p_{\rm H} > 1$ such that $p_{\rm L} \leq p_{\rm H}^2/\mu^{\alpha_v}$ and $p_{\rm H} \leq \mu^{\alpha_v}$.

- Now consider the situation in which $p_{\rm L} > 1$ and $p_{\rm H} < 1$. 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

$$p_{\rm H} < \mu^{\alpha_d} \quad \text{and} \quad p_{\rm L} < \frac{p_{\rm H}^{2\alpha_v/\alpha_d}}{\mu^{\alpha_v}}. \tag{7}$$

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:

$$\frac{p_{\rm H}^{2\alpha_v/\alpha_d}}{\mu^{\alpha_v}} < \frac{\mu^{2\alpha_v}}{\mu^{\alpha_v}} = \mu^{\alpha_v} < 1.$$

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

$$\frac{p_{\rm H}^{2\alpha_v/\alpha_d}}{\mu^{\alpha_v}} < \mu^{-\alpha_v} < 1.$$

- **$\mathbf{R}_0$.** The remaining regime is $p_{\rm L} < 1$ and $p_{\rm H} > 1$. 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 if $\mu > 1$. We have to maximize

$$\Pi_{\rm O}(p_{\rm L}, p_{\rm H}, \mu) := (\mu - \sqrt[\alpha_v]{p_{\rm H}})p_{\rm H} + \left(\sqrt[\alpha_v]{p_{\rm H}} - \mu \frac{\sqrt[\alpha_d]{p_{\rm L}}}{\sqrt[\alpha_v]{p_{\rm H}}} \right) p_{\rm L},$$

over a region R_0 that is given by

$$p_{\rm L} \in \left(0, \min\left\{ 1, \frac{p_{\rm H}^{2\alpha_d/\alpha_v}}{\mu^{\alpha_d}} \right\} \right), \quad p_{\rm H} \in (1, \mu^{\alpha_v}]. \tag{8}$$

The resulting admissible region is depicted in Figure 1, 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

$$\Pi^\star_{\rm O}(\mu) := \max\left\{ \Pi^\star_{{\rm O},-}(\mu), \Pi^\star_{{\rm O},+}(\mu), \Pi^\star_{{\rm O},0}(\mu) \right\},$$

with $\Pi^\star_{{\rm O},i}(\mu) := \max_{(p_{\rm L}, p_{\rm H}) \in \mathrm{R}_i} \Pi_{\rm O}(p_{\rm L}, p_{\rm H}, \mu)$, for $i \in \{-, +, 0\}$. We now compute these three maxima subsequently, in Subsections A, B, and C. First an auxiliary result is proven in Lemma 4.1.

Lemma 4.1 $\Pi^\star_{\rm O}(\mu)$ *is* non decreasing *and* convex *in* μ.

Proof. It suffices to prove that the $\Pi^\star_{{\rm O},i}(\mu)$ are non decreasing and convex in μ, $i \in \{-, +, 0\}$.

- First notice that $\Pi_{\rm O}(p_{\rm L}, p_{\rm H}, \mu)$ is a linear function of μ.

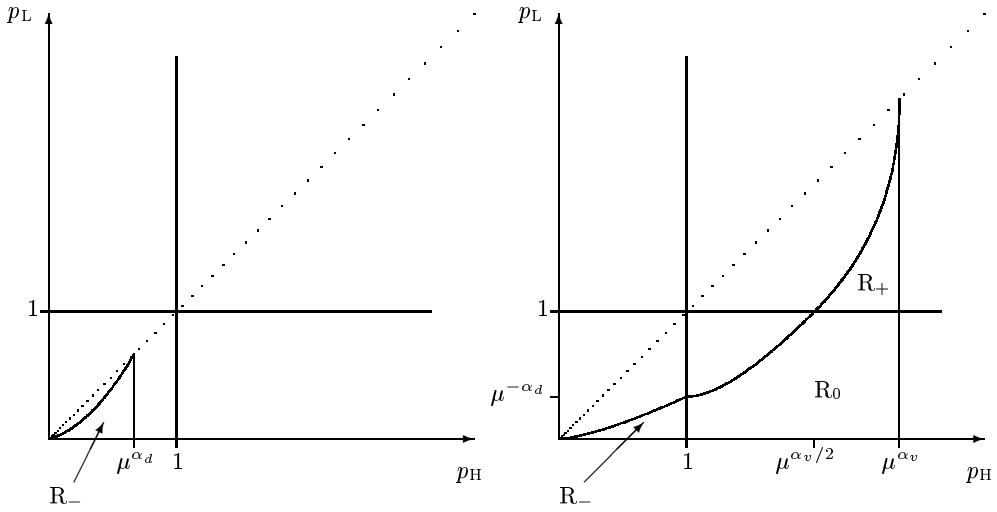

Figure 1: Admissible region. Left panel corresponds to $\mu < 1$, right panel to $\mu > 1$.

- The $\Pi_{\rm O}(p_{\rm L}, p_{\rm H}, \mu)$ are non decreasing in μ. This is seen as follows for $\rm R_0$ (a similar reasoning applies to $\rm R_-$ and $\rm R_+$). In $\rm R_0$ the coefficient of μ is given by

$$p_{\rm H} - \frac{\sqrt[\alpha_d]{p_{\rm L}}}{\sqrt[\alpha_v]{p_{\rm H}}} p_{\rm L} \geq p_{\rm H} - p_{\rm H}^{\alpha_d/\alpha_v} > 0,$$

  as follows from $p_{\rm L} \leq p_{\rm H}^{2\alpha_d/\alpha_v}/\mu^{\alpha_d} = p_{\rm H}^{\alpha_d/\alpha_v} \cdot (p_{\rm H}^{\alpha_d/\alpha_v}/\mu^{\alpha_v}) \leq p_{\rm H}^{\alpha_d/\alpha_v}$, in conjunction with $p_{\rm H} \geq 1$.

As the $\Pi^\star_{{\rm O},i}(\mu)$ are maxima (over $(p_{\rm L}, p_{\rm H}) \in {\rm R}_i$) of non decreasing, linear (and hence convex) functions, they are non decreasing and convex as well. □

A. Maximization over $\rm R_-$

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

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

The price for the low priority service is given by

$$p_{\rm L}(p_{\rm H}) = \beta_d(\mu) p_{\rm H}^2, \quad \text{with} \quad \beta_d(\mu) := \left(\frac{\alpha_d}{\mu(\alpha_d+1)}\right)^{\alpha_d}.$$

With $\zeta_-(\mu) := g'_-(1)$, it is shown in the Appendix (Section 7.A) that

▷ A_1. *For $\mu \in (0,1]$ the optimum over* R_- *is reached at a price $\bar{p}_\mathrm{H}$ in the interior of $(0, \mu^{\alpha_d}]$; this $\bar{p}_\mathrm{H}$ is the unique solution of $g'_-(p) = 0$ in $(0, \mu^{\alpha_d}]$. Also, $\bar{p}_\mathrm{L} = \beta_d(\mu)\bar{p}_\mathrm{H}^2$.*

▷ A_2. *Let ν_- be the unique solution to $\zeta_-(\mu) = 0$ in $(1,\infty)$. For $\mu \in (1,\infty)$ the optimum over* R_- *is reached at a price $\bar{p}_\mathrm{H}$*

(i) *in the interior of $(0,1]$ if $\mu \in (1,\nu_-)$; this $\bar{p}_\mathrm{H}$ is the unique solution of $g'_-(p) = 0$ in $(0,1]$. Also, $\bar{p}_\mathrm{L} = \beta_d(\mu)\bar{p}_\mathrm{H}^2$;*
(ii) *equal to 1 if $\mu \in (\nu_-,\infty)$. Also, $\bar{p}_\mathrm{L} = \beta_d(\mu)$.*

B. Maximization over R_+

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

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

With $\mu_v := 1 + \alpha_v^{-1}$, we 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

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

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

$$p_\mathrm{L}(p_\mathrm{H}) = \left(\frac{\alpha_v}{\mu(\alpha_v+1)}\right)^{\alpha_v} p_\mathrm{H}^2 \quad \text{if} \quad p_\mathrm{H} \in [q_+(\mu), \mu^{\alpha_v}], \quad \text{with } q_+(\mu) := \left(\frac{\mu(\alpha_v+1)}{\alpha_v}\right)^{\alpha_v/2},$$

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

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

With $\zeta_+(\mu) := \bar{g}'_+(\mu^{\alpha_v})$ and $\xi_+(\mu) = g'_+(q_+(\mu))$, it is shown in the Appendix (Section 7.B) that

▷ B_1. *Let $\nu_{+,1}$ be the unique solution to $\zeta_+(\mu) = 0$ in $(1,\mu_v)$. For $\mu \in (1,\mu_v)$ the optimum over* R_+ *is reached at a price $\bar{p}_\mathrm{H}$*

(i) *equal to* μ^{α_v} *if* $\mu \in (1, \nu_{+,1})$. *Also,* $\bar{p}_{\rm L} = 1$*;*
(ii) *in the interior of* $[\mu^{\alpha_v/2}, \mu^{\alpha_v}]$ *if* $\mu \in (\nu_{+,1}, \mu_v)$*; this* $\bar{p}_{\rm H}$ *is the unique solution of* $\bar{g}'_+(p) = 0$ *in* $[\mu^{\alpha_v/2}, \mu^{\alpha_v}]$. *Also,* $\bar{p}_{\rm L} = 1$.

▷ B_2. *Let* $\nu_{+,2}$ *be the unique solution to* $\xi_+(\mu) = 0$ *in* (μ_v, ∞). *For* $\mu \in (\mu_v, \infty)$ *the optimum over* R_+ *is reached at a price* $\bar{p}_{\rm H}$

(i) *in the interior of* $[\mu^{\alpha_v/2}, q_+(\mu)]$ *if* $\mu \in (\mu_v, \nu_{+,2})$*; this* $\bar{p}_{\rm H}$ *is the unique solution of* $\bar{g}'_+(p) = 0$ *in* $[\mu^{\alpha_v/2}, q_+(\mu)]$. *Also,* $\bar{p}_{\rm L} = 1$*;*
(ii) *in the interior of* $[q_+(\mu), \mu^{\alpha_v}]$ *if* $\mu \in (\nu_{+,2}, \infty)$*; this* $\bar{p}_{\rm H}$ *is the unique solution of* $g'_+(p) = 0$ *in* $[q_+(\mu), \mu^{\alpha_v}]$. *Also* $\bar{p}_{\rm L} = \beta_v(\mu)\bar{p}_{\rm H}^2$.

C. Maximization over R_0

Again we first perform the optimization over $p_{\rm L}$ for given $p_{\rm H}$. It is straightforward to obtain that the optimum is attained at

$$p_{\rm L}(p_{\rm H}) = \min\left\{1, \beta_d(\mu) p_{\rm H}^{2\alpha_d/\alpha_v}\right\}.$$

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

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

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

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

$$\max_{1 < p_{\rm H} \le \mu^{\alpha_v}} g_0(p_{\rm H}).$$

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

$$\beta_d(\mu) p_{\rm H}^{2\alpha_d/\alpha_v} \le 1, \quad \text{if and only if } p_{\rm H} \le q_0(\mu) := \left(\frac{\mu(\alpha_d+1)}{\alpha_d}\right)^{\alpha_v/2}.$$

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

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

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 the Appendix (Section 7.C) that

▷ C_1. *Let* $\nu_{0,1}$ *be the unique solution to* $\zeta_0(\mu) = 0$ *in* $(1, \mu_d)$. *For* $\mu \in (1, \mu_d)$ *the optimum over* R_0 *is reached at a price* $\bar{p}_{\rm H}$

(i) *equal to 1 if* $\mu \in (1, \nu_{0,1})$. *Also,* $\bar{p}_{\rm L} = \beta_d(\mu)$;
(ii) *in the interior of* $[1, \mu^{\alpha_v}]$ *if* $\mu \in (\nu_{0,1}, \mu_d)$; *this* $\bar{p}_{\rm H}$ *is the unique solution of* $g_0'(p) = 0$ *in* $[1, \mu^{\alpha_v}]$. *Also,* $\bar{p}_{\rm L} = \beta_d(\mu)\bar{p}_{\rm H}^{2\alpha_d/\alpha_v}$.

▷ C_2. *Let* $\nu_{0,2}$ *be the unique solution to* $\xi_0(\mu) = 0$ *in* (μ_d, ∞). *For* $\mu \in (\mu_d, \infty)$ *the optimum over* R_0 *is reached at a price* $\bar{p}_{\rm H}$

(i) *in the interior of* $[1, q_0(\mu)]$ *if* $\mu \in (\mu_d, \nu_{0,2})$; *this* $\bar{p}_{\rm H}$ *is the unique solution of* $g_0'(p) = 0$ *in* $[1, q_0(\mu)]$. *Also,* $\bar{p}_{\rm L} = \beta_d(\mu)\bar{p}_{\rm H}^{2\alpha_d/\alpha_v}$;
(ii) *in the interior of* $[q_0(\mu), \mu^{\alpha_v}]$ *if* $\mu \in (\nu_{0,2}, \infty)$; *this* $\bar{p}_{\rm H}$ *is the unique solution of* $\bar{g}_0'(p) = 0$ *in* $[q_0(\mu), \mu^{\alpha_v}]$. *Also,* $\bar{p}_{\rm L} = 1$.

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 $\nu_-^\star$ and $\nu_+^\star$ such that (for the profit-maximizing prices) voice will dominate in the network for all $\mu < \nu_-^\star$, data will dominate for $\mu > \nu_+^\star$, and there is a 'mixed scenario' (with priority for voice) for $\mu \in (\nu_-^\star, \nu_+^\star)$. The second possibility data dominates for μ smaller than some $\nu^\star$, and voice dominates otherwise.

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

Proof. First notice that $\nu_{0,1} < \nu_{0,2}$, implying that $\nu_{\min} < \nu_{\max}$. The stated follows immediately from the inequalities (i) $\Pi_{{\rm O},0}^\star(\mu) \geq \Pi_{{\rm O},-}^\star(\mu)$ for $\mu > \nu_-$, (ii) $\Pi_{{\rm O},0}^\star(\mu) \geq \Pi_{{\rm O},+}^\star(\mu)$ for $\mu < \nu_{+,2}$, (iii) $\Pi_{{\rm O},0}^\star(\mu) \leq \Pi_{{\rm O},-}^\star(\mu)$ for $\mu < \nu_{0,1}$, and (iv) $\Pi_{{\rm O},0}^\star(\mu) \leq \Pi_{{\rm O},+}^\star(\mu)$ 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 $\Pi_{{\rm O},-}^\star(\mu)$ over R_- is attained at $\bar{p}_{\rm H} = 1$ and a $\bar{p}_{\rm L} < 1$, 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 $\Pi_{{\rm O},0}^\star(\mu)$. The other inequalities are proven similarly. □

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

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

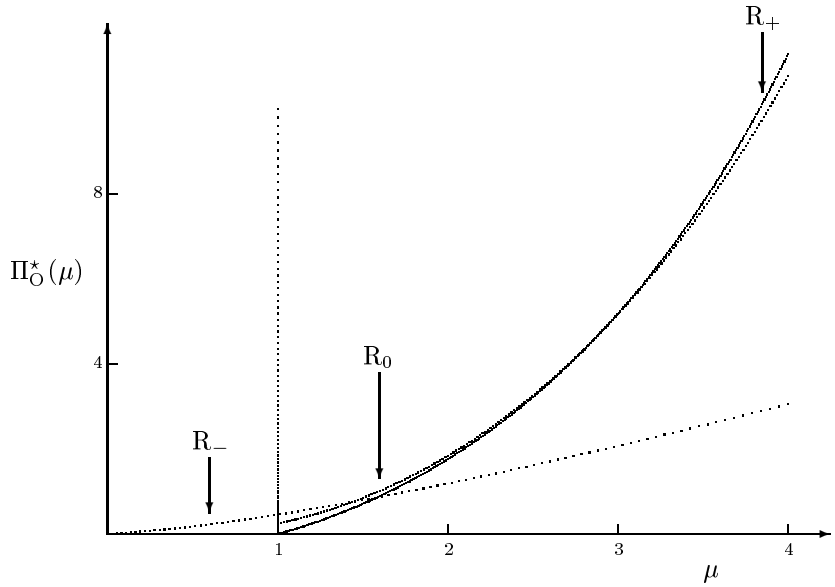

Figure 2: Profit as a function of link speed, for $\mu \in (0, 4]$.

- *There exist service rates* $\nu_-^\star$ *and* $\nu_+^\star$ *such that*

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

- *There exists a service rate* $\nu^\star$ *such that*

$$\Pi_{\mathrm{O}}^\star(\mu) = \begin{cases} \Pi_{\mathrm{O},-}^\star(\mu), & \mu \in (0, \nu^\star); \\ \Pi_{\mathrm{O},+}^\star(\mu), & \mu \in (\nu^\star, \infty). \end{cases}$$

5 Discussion and example

In this section we start by giving a numerical example that demonstrates the theory of the previous sections. Then we motivate the somewhat paradoxical fact that in this model it is beneficial to use both queues, *even if the user population is homogeneous*. Finally we provide some reflections on the utility functions and the queueing model.

Example. This example 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 $\nu_- = 1.500$; $\nu_{+,1} = 1.325$; $\nu_{+,2} = 2.422$; $\nu_{0,1} = 1.183$; and $\nu_{0,2} = 3.948$. Applying the inequalities used in the proof of Theorem 4.2, it is not so hard to prove that, due to

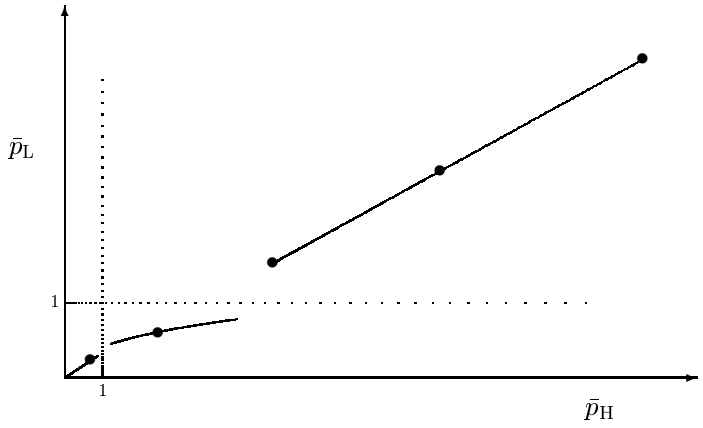

Figure 3: Relation between $\bar{p}_{\rm L}$ and $\bar{p}_{\rm H}$. The bullets correspond to $\mu = 1, \cdots, 5$.

$\nu_{0,1} < \nu_- < \nu_{+,2} < \nu_{0,2}$, five regimes can be distinguished:

$$\Pi^\star_{\rm O}(\mu) = \begin{cases} \Pi^\star_{{\rm O},-}(\mu), & \mu \in (0, \nu_{0,1}); \\ \max\{\Pi^\star_{{\rm O},-}(\mu), \Pi^\star_{{\rm O},0}(\mu)\}, & \mu \in (\nu_{0,1}, \nu_-); \\ \Pi^\star_{{\rm O},0}(\mu), & \mu \in (\nu_-, \nu_{+,2}); \\ \max\{\Pi^\star_{{\rm O},0}(\mu), \Pi^\star_{{\rm O},+}(\mu)\}, & \mu \in (\nu_{+,2}, \nu_{0,2}); \\ \Pi^\star_{{\rm O},+}(\mu), & \mu \in (\nu_{0,2}, \infty). \end{cases}$$

In Figure 2 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 3 displays the optimizing prices for various values of μ.

A paradox. Consider a single type of traffic, with utility function $U(\cdot)$; for ease, assume that the packet arrival rate λ equals 1. $U(\cdot)$ is a positive, decreasing function of $\mathbb{E}D$. Assume that $U(\cdot)$ has inverse $V(\cdot)$ with derivative $V'(\cdot)$. Then

$$N_{\rm H}(p_{\rm L}, p_{\rm H}) = \mu - \frac{1}{V(p_{\rm H})},$$

under the proviso that $V(p_{\rm H}) \geq \mu^{-1}$, or $p_{\rm H} \leq U(\mu^{-1})$. Similarly,

$$N_{\rm L}(p_{\rm L}, p_{\rm H}) = \frac{1}{V(p_{\rm H})} - \mu \cdot \frac{V(p_{\rm H})}{V(p_{\rm L})},$$

requiring that $V(p_{\rm L}) \geq \mu V^2(p_{\rm H})$. So, for a given value of $p_{\rm H}$, the largest admissible $p_{\rm L}$ equals $p^\star_{\rm L}(p_{\rm H}) := U(\mu V^2(p_{\rm H}))$. We get the admissible region

$$\mathrm{R} := \left\{ (p_{\rm L}, p_{\rm H}) \,|\, p_{\rm L} \leq U\big(\mu V^2(p_{\rm H})\big) \,, p_{\rm H} \leq U\big(\mu^{-1}\big) \right\}.$$

Notice that for all $(p_{\rm L}, p_{\rm H}) \in {\rm R}$ it holds that $p_{\rm L} \leq p_{\rm H}$, due to the fact that $V(\cdot)$ decreases and

$$V(p_{\rm L}) \geq \mu V^2(p_{\rm H}) \geq V(p_{\rm H}).$$

This is of course what was expected: the price of the low-priority service is lower than the high-priority service.

We are now faced with the following profit maximization problem:

$$\max_{p_{\rm L}, p_{\rm H} \in {\rm R}} p_{\rm H} N_{\rm H}(p_{\rm L}, p_{\rm H}) + p_{\rm L} N_{\rm L}(p_{\rm L}, p_{\rm H}).$$

To find the optimum we first differentiate (for given $p_{\rm L}$) to $p_{\rm H}$:

$$M(p_{\rm L}, p_{\rm H}) := N_{\rm L}(p_{\rm L}, p_{\rm H}) + p_{\rm L} \cdot \frac{\partial N_L}{\partial p_{\rm L}} = N_{\rm L}(p_{\rm L}, p_{\rm H}) + \mu p_{\rm L} \cdot \frac{V(p_{\rm H})}{V^2(p_{\rm L})} \cdot V'(p_{\rm L}).$$

As $N(p_{\rm L}^\star(p_{\rm H}), p_{\rm H}) = 0$, inserting $p_{\rm L} = p_{\rm L}^\star(p_{\rm H})$ gives

$$M(p_{\rm L}^\star(p_{\rm H}), p_{\rm H}) = \mu p_{\rm L} \cdot \frac{V(p_{\rm H})}{V^2(p_{\rm L}^\star(p_{\rm H}))} \cdot V'(p_{\rm L}^\star(p_{\rm H}));$$

notice that this quantity is *negative*, as $V(\cdot)$ is positive and decreasing, just like its inverse $U(\cdot)$. Apparently $M(\cdot,\cdot)$ is negative in the neighborhood of $p_{\rm L} = p_{\rm L}^\star(p_{\rm H})$. In other words: a value of $p_{\rm L}$ smaller than $p_{\rm L}^\star(p_{\rm H})$ gives a higher profit than $p_{\rm L}^\star(p_{\rm H})$ itself. This entails that *under the profit-maximizing prices both queues will be used!* In other words: even if the customers have a homogeneous aversion to delay, it is beneficial to create performance differentiation. Of course, the compensated utility at both queues equals 0.

On the choice of the utility curves. As was stated in the Introduction, one of the main purposes of the paper, is to show that the result by Mendelson and Whang [10] relies critically on the specific assumptions imposed in that paper. We do this by choosing hyperbolic (rather than linear) utility curves, where it is noted that these new utility curves are more realistic in the context of delay-sensitive and delay-tolerant users (as they *intersect* at some positive delay value). We find that indeed the solutions under open classes and dedicated classes do not coincide. We also identify a number of other interesting phenomena, such as the tragedy of the commons, the way the nature of the solution depends on the link speed, etc. The assumption of hyperbolic utility curves can be considered restrictive, but it nicely captures the crucial features of customer behavior, while maintaining tractability.
As a next step, it would be interesting to see under what *minimal* conditions these phenomena are valid. The analysis of the above paradox (featuring the situation with only one type of users) suggests that the qualitative results of this paper might not depend critically on the

utility curves chosen. An interesting question could be: is it possible to characterize the solution to the profit maximization problem (both for the model with dedicated classes and the model with open classes), if we somewhat relax the requirements on the utility curves. Rather than assuming a hyperbolic shape *a priori*, we could for instance consider the class of utility curves that are such that the voice users appreciate the service more for $\mathbb{E}D < 1$, and the data users for $\mathbb{E}D > 1$. It can be expected that the resulting models are more involved, and constitute an interesting direction for further research.

It is noted that straightforward calculations yield that the analysis remains tractable for a broad class of functional forms. The major drawback of these alternative function forms, however, is that the resulting expressions become less transparent. This motivates our choice for hyperbolic curves. Examples of other functional forms are given in the Appendix (Section 7.D).

On the choice of the queueing model. It can be expected that our results can be extended to M/G/1 priority systems (rather than M/M/1), given the results on the mean delay in this type of networks [5]. Another interesting direction is the extension of the number of priority classes (i.e., queues). Above, we already saw that having two queues, rather than just one, increases the profit (given that there are no costs associated with having an additional queue), even for homogeneous users. The next question is, of course, is it beneficial to have even more queues? We expect that the profit increases as the number of priority classes increases, but remains bounded. It would be of interest to verify this conjecture.

6 Competition between elastic flows

This last section focuses on the situation in which the users' utility is primarily determined by *throughput*, rather than packet delay. One could think of files arriving at a network node, competing for service. Loosely speaking, in the Internet the Transfer Control Protocol (TCP) divides the available capacity equally among the active users. Hence, during congestion the transmission rate will be low, whereas jobs can claim a more significant part of the link bandwidth during more quiet periods. The corresponding type of traffic is commonly called *elastic*, as the source transmission rate adapts to the level of congestion. Due to these properties, the node can be modeled as an M/G/1 *processor sharing* queue [5, Ch. IV], if the jobs arrive according to a Poisson process, and their sizes are i.i.d. samples (independent of the arrival process), as argued by Massoulié and Roberts [7].

Where in the model of Sections 2-5 the heterogeneity between the delay-tolerant and delay-

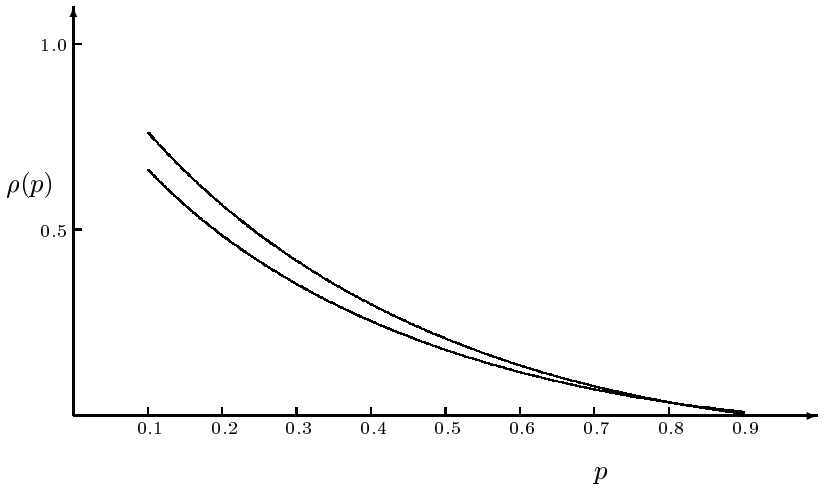

Figure 4: Load ρ as a function of the price p; the upper (lower) curve corresponds to exponential (Pareto) files

sensitive users played a crucial role, we here assume that the utility curves depend on the *size of the job*. The underlying idea here is the following. When retrieving an extremely large file the user's expectation of the throughput will be different than when retrieving small files. One could imagine that the retrieval of large files (for instance non-real-time audio and video) usually does not bring about the expectation of a 'real-time response' — this might be in contrast with the retrieval of smaller files (for instance web surfing) for which some kind of direct response is required. If the user preferences are indeed ordered in this way, it would suggest that the $U_x(\cdot)$ curves are decreasing in x. Loosely speaking, it would mean that users are willing to pay more per byte for small jobs than for large jobs.
However, notice that the (decreasing) ordering of the utility curves suggested above is just one possibility; we emphasize that the framework below does not require any specific ordering.

The parametrization by job size is inspired by a phenomenon recently observed in traffic measurements: most of the Internet connections are short in terms of the amount of traffic they carry (commonly referred to as 'mice'), while a small fraction of the connections are carrying a large portion of the traffic ('elephants'). If there is no usage charge, the presence of the (few) long files might deteriorate the performance experienced by the (many) short files. This gives rise to the idea of somehow 'protecting' the small files by imposing (for instance) a volume charge (i.e., a fixed price per volume unit, say, byte).

Suppose that, without any pricing, jobs would arrive at the network node according to a Poisson process with rate λ. The job sizes are i.i.d. samples from a general distribution F (with finite first moment $\mathbb{E}F$). The queue has a constant service speed C. We do not explicitly

impose the stability condition $\lambda \cdot \mathbb{E}F < C$.
Suppose there is a volume charge of p. In the M/G/1 processor sharing queue, the throughput is uniform for all users, namely $C - \rho$, where ρ is the offered load, see for instance [5]. This entails that a customer of file size x joins if $U_x(C - \rho) > p$. Assume for ease that F has a continuous distribution; then these customers cause a load of

$$\rho = \lambda \int_0^\infty 1\{U_x(C-\rho) > p\} x \, \mathrm{d}\mathbb{P}(F \leq x).$$

The value of $\rho = \rho(p)$ can be solved from this (fixed-point) equation — notice that the right hand side is decreasing in ρ. In fact the provider wants to optimize $\Pi(p) \equiv p \cdot \rho(p)$ over all prices $p > 0$. If $\lambda \cdot \mathbb{E}F \ll C$ this might lead to a situation in which all traffic is accepted.

It is clear that the 'policy' (i.e., which jobs join the queue, and which jobs do not) strongly depends on the behavior of $U_x(\cdot)$ as a function of x. If indeed the utility curves decrease in x, as suggested above, the pricing policy entails that there is a certain threshold, up to which files join the queue. In other words, it excludes the 'elephants'. The reason is that these large files 'yield lower utility per byte'. Obviously, other orderings of the utility curves lead to different policies.

Example. Assume that $U_x(t) \equiv \beta_x t$, with $\beta_x := (x+1)^{-1}$, i.e., β_x is decreasing in x. Then files join according to a threshold policy, as argued above. The fixed point equation for ρ (with p given) reduces to

$$\rho = \lambda \int_0^{f(\rho,p)} x \, \mathrm{d}\mathbb{P}(F \leq x), \quad \text{with} \quad f(\rho,p) := \max\left\{\frac{c-\rho}{p} - 1, 0\right\}.$$

We take $\lambda = 2$, $C = 1$, and $\mathbb{E}F = 1$, i.e., a situation of severe overload. We compare two different file size distributions: (i) Pareto file sizes with density $b_{\mathrm{p}}(x) = (x+1)^{-\alpha}$, and (ii) exponential file sizes with density $b_{\mathrm{e}}(x) = e^{-x}$. Notice that $\alpha = (3+\sqrt{5})/2$ to get $\mathbb{E}F = 1$.
As illustrated in Figures 4 and 5, for exponential job sizes the optimum 0.132 is attained at $p = 0.32$ (with $\rho = 0.41$), whereas for Pareto files the optimum is 0.105, attained at $p = 0.33$ (with $\rho = 0.32$). In the former case all files up to size 0.84 have the incentive to join the queue, whereas in the latter case all files up to 1.06. For uniformly distributed files (with unit mean), the solution can be computed explicitly: $p = \sqrt{2/3}$ (attained at $\rho = 1/3$), yielding profit 0.27.

References

[1] D. Clark (1997). Internet cost allocation and pricing. In: *Internet Economics*, MIT Press, pp. 215-252.

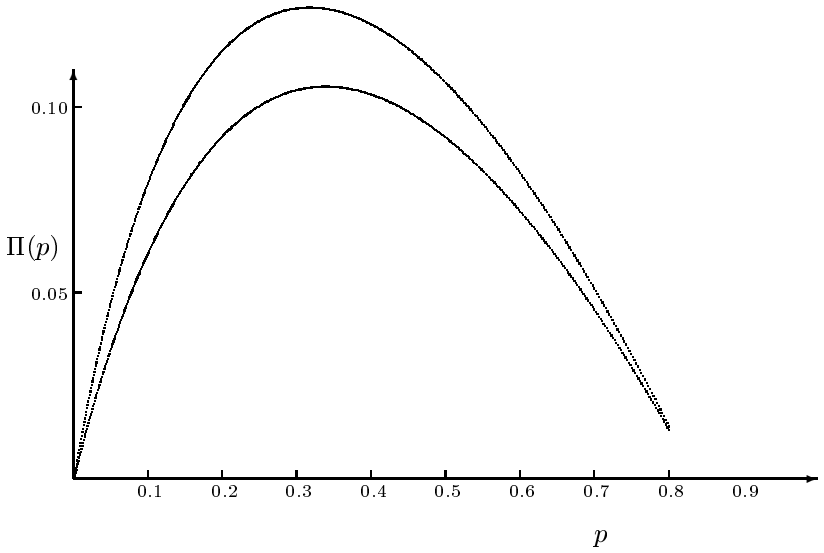

Figure 5: Profit $\Pi(p) \equiv p \cdot \rho(p)$ as a function of the price p; the upper (lower) curve corresponds to exponential (Pareto) files

[2] R. Gibbens, R. Mason, and R. Steinberg (2000). Internet Service Classes under Competition. *IEEE Journal on Selected Areas in Communications*, Vol. 18, pp. 2490-2498.

[3] A. Gupta, D. Stahl, and A. Whinston (1997). A stochastic equilibrium model of Internet pricing. *Journal of Economic Dynamics and Control*, Vol. 21, pp. 697-722.

[4] G. Hardin (1968). The tragedy of the commons, *Science*, Vol. 162, pp. 1243-1248.

[5] L. Kleinrock (1976). *Queueing Systems, Vol. 2: Computer Applications.* Wiley, New York.

[6] J. MacKie-Mason and H. Varian (1995). Pricing congestible network resources. *IEEE Journal on Selected Areas in Communications,* Vol. 13, pp. 1141-1149.

[7] L. Massoulié and J. Roberts (2000). Bandwidth sharing and admission control for elastic traffic, *Telecommunication Systems*, Vol. 15, pp. 185-201.

[8] L. McKnight and J. Bailey (Ed.) (1997). Internet Economics, MIT Press, Cambridge MA, USA.

[9] H. Mendelson (1985). Pricing computer services: queueing effects. *Communications of the ACM,* Vol. 28, pp. 312-321.

[10] H. Mendelson and S. Whang (1990). Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research,* Vol. 38, pp. 870-883.

[11] A. Odlyzko (1999). Paris Metro Pricing: The minimalist differentiated services solution. Proceedings 1999 Seventh International Workshop on Quality of Service (IWQoS '99), IEEE, pp. 159-161.

[12] A. Odlyzko (1999). Paris Metro Pricing for the Internet, Proceedings ACM Conference on Electronic Commerce (EC'99), ACM, pp. 140-147.

[13] S. Shenker (1995). Fundamental design issues for the future Internet. *IEEE Journal on Selected Areas in Communications,* Vol. 13, pp. 1176-1188.

[14] J. Tirole (1989). *The Theory of Industrial Organization*, MIT Press, Cambridge MA, USA.

[15] D. Walker, F. Kelly, and J. Solomon (1997). Tariffing in the new IP/ATM environment. *Telecommunications Policy*, Vol. 21, pp. 283-295.

7 Appendix

Lemma 7.1 *For all* $x > 0$,

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

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

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

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 $x > 0$. As an aside we remark that $f(x) \to 2/e$ for $x \to \infty$. □

Lemma 7.2 *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

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

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

Furthermore, we have to prove that

$$\frac{\alpha}{\sqrt{\mu}}\bar{f}'\left(\mu^{\alpha/2}\right) = \alpha\sqrt{\mu} + \frac{2}{\mu^{\alpha/2}} - \alpha - 1 > 0. \tag{10}$$

This inequality clearly holds for $\mu = 1$. The second claim in the lemma follows from the fact that (10) increases in μ, as is checked easily. □

A. Maximization over R_-

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

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

The price for the low priority service is given by

$$p_{\mathrm{L}}(p_{\mathrm{H}}) = \beta_d(\mu)p_{\mathrm{H}}^2, \quad \text{with} \quad \beta_d(\mu) := \left(\frac{\alpha_d}{\mu(\alpha_d+1)}\right)^{\alpha_d}.$$

Now distinguish between μ smaller and larger than 1.

• **Case A_1**: $\mu \in (0, 1]$. 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 7.1. It also follows that $g_-(\cdot)$ is concave on $(0, \mu^{\alpha_d}]$, as

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

for all $p \in (0, \mu^{\alpha_d}]$, again invoking Lemma 7.1. This proves A_1.

• **Case A_2**: $\mu \in (1, \infty)$. 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 $p_{\mathrm{H}} = 1$ if $g_-'(1) \geq 0$, and in the interior of $(0, \mu^{\alpha_d}]$ if $g_-'(1) < 0$. Denote

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

Now note that $\zeta_-(1) = (f(\alpha_d) - 1)/\alpha_d < 0$ (apply Lemma 7.1!), and $\zeta_-(\mu) \to \infty$ as $\mu \to \infty$. Applying Lemma 7.1 again, together with the fact that $\mu > 1$, we see that $\zeta_-(\cdot)$ increases:

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

This immediately implies A_2.

B. Maximization over R_+

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

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

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

• **Case B_1:** $\mu \in (1, \mu_v)$. It is not hard to see that the optimal p_{L} equals 1. The maximizing p_{H} should be found from

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

By Lemma 7.2, $\bar{g}_+(\cdot)$ 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

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

First observe that $\zeta_+(1) = \alpha_v^{-1} > 0$, and that $\zeta_+(\mu_v) = (f(\alpha_v) - 1)\mu_v/\alpha_v < 0$ (Lemma 7.1). 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 B_1.

• **Case B_2:** $\mu \in (\mu_v, \infty)$. It turns out that

$$p_{\mathrm{L}}(p_{\mathrm{H}}) = \left(\frac{\alpha_v}{\mu(\alpha_v+1)}\right)^{\alpha_v} p_{\mathrm{H}}^2 \quad \text{if} \quad p_{\mathrm{H}} \in [q_+(\mu), \mu^{\alpha_v}], \quad \text{with } q_+(\mu) := \left(\frac{\mu(\alpha_v+1)}{\alpha_v}\right)^{\alpha_v/2},$$

and $p_{\mathrm{L}}(p_{\mathrm{H}}) = 1$ if $p_{\mathrm{H}} \in [\mu^{\alpha_v/2}, q_+(\mu)]$. 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 A_1. We have to solve

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

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

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

We now prove that $\xi_+(\mu) = 0$ has a unique zero in (μ_v, ∞). To this end, first observe that that $\xi_+(\mu_v) = (f(\alpha_v) - 1)\mu_v/\alpha_v < 0$ (due to Lemma 7.1) and $\xi_+(\mu) \to \infty$ as $\mu \to \infty$. Also, using Lemma 7.1, 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

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

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

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

which holds due to Lemma 7.1. 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 $F(a) \cdot F(b) < 0$. We have proven B_2.

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

$$p_{\mathrm{L}}(p_{\mathrm{H}}) = \min\left\{1, \beta_d(\mu) p_{\mathrm{H}}^{2\alpha_d/\alpha_v}\right\}.$$

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

- **Case $\mathbf{C}_1$:** $\mu \leq \mu_d$. It is not hard to verify that for these μ it holds that

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

so that the optimization reduces to

$$\max_{1 < p_{\mathrm{H}} \leq \mu^{\alpha_v}} g_0(p_{\mathrm{H}}),$$

where $g_0(\cdot)$ is defined as $g(\cdot)$ in (5). Using that $\mu_-^\star < 1 < \mu$, 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 $p_{\mathrm{H}} > 1$ is optimal iff

$$\zeta_0(\mu) := g_0'(1) = \mu - 1 + \frac{1}{\alpha_v}\left(\left(\frac{\alpha_d}{\mu(\alpha_d + 1)}\right)^{\alpha_d} \cdot \frac{2\alpha_d + 1}{\alpha_d + 1} - 1\right) > 0; \tag{11}$$

otherwise the maximum is attained at $p_{\mathrm{H}} = 1$.
For $\mu = 1$, condition (11) is not met; this is because of Lemma 7.1. Notice that

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

and $\zeta_0'(\mu) > 0$ for all $\mu > 1$. Hence, $\zeta_0(\cdot)$ has a unique root in $(1, \mu_d)$. This proves C_2.

• **Case C_2**: $\mu \in (\mu_d, \infty)$. For these μ we have that

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

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

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

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

$$\xi_0(\mu) := g_0'(q_0(\mu)) = \mu - \frac{\alpha_v+1}{\alpha_v} \cdot \sqrt{\frac{\mu(\alpha_d+1)}{\alpha_d}} + \left(\frac{\alpha_d}{\mu(\alpha_d+1)}\right)^{(\alpha_v-1)/2} \cdot \frac{2\alpha_d+1}{\alpha_d+1} \cdot \frac{1}{\alpha_v}.$$

With an analysis that is analogous to Case B_2, we prove C_2.

D. Other utility curves

In this section we argue that the analysis remains tractable for a broad class of functional forms. First assume exponentially (rather than hyperbolically) decaying curves, i.e., curves of the type

$$u_d(\mathbb{E}D) = e^{-\alpha_d(\mathbb{E}D-1)},$$

and $u_v(\cdot)$ analogously. First consider the context of Section 2. With $p \leq \exp(\alpha_d)$,

$$N_d(p) = \lambda_d^{-1}\left(\mu - \frac{\alpha_d}{\alpha_d - \log p}\right).$$

When maximizing $\lambda_d \cdot N_d(p) \cdot p$, the maximizing price is

$$p_d = \exp\left(\alpha_d - \frac{\alpha_d + \sqrt{\alpha_d^2 + 4\mu\alpha_d}}{2\mu}\right);$$

p_v is found analogously. It can be checked easily that the analysis of Section 2 goes through.
Now consider the situation of Section 3, in which voice has priority. Then

$$\lambda_d N_d(p_{\mathrm{L}}, p_{\mathrm{H}}) = \frac{\alpha_v}{\alpha_v - \log p_{\mathrm{H}}} - \frac{\alpha_d}{\alpha_v} \cdot \mu \cdot \frac{\alpha_v - \log p_{\mathrm{H}}}{\alpha_d - \log p_{\mathrm{L}}}.$$

It is not hard to see that the maximization over p_{L} (for given p_{H}) can be done explicitly, but the maximization over p_{H} remains implicit. The same holds for the open classes model of Section 4. Notice that for the hyperbolic utility curves, as used in this paper, we had just the same level of 'explicitness'. A similar conclusion can be drawn if one of the users has a hyperbolic curve and the other an exponential curve (where the parameters are chosen such that the data utility-curve majorizes the voice utility-curve iff the mean delay is larger than 1).
In other words, it turns out that the analysis with other 'standard' functional forms is just as tractable as with hyperbolic utility curves. The major drawback of these alternative functional forms, however, is that the resulting expressions become less transparent.