Less-than-Best-Effort Services: Pricing and Scheduling

Yezekael Hayel
IRISA/INRIA Rennes
Campus Universitaire de Beaulieu
35042 Rennes cedex, France
Email: Yezekael.Hayel@irisa.fr

David Ros
GET/ENST Bretagne
Rue de la Châtaigneraie, CS 17607
35576 Cesson Sévigné cedex, France
Email: David.Ros@enst-bretagne.fr

Bruno Tuffin
IRISA/INRIA Rennes
Campus Universitaire de Beaulieu
35042 Rennes cedex, France
Email: Bruno.Tuffin@irisa.fr

***Abstract*— In recent years, the notion of a service offering a degraded performance with respect to the best-effort service traditionally found in IP networks has gained acceptance among network researchers. Such a *less-than-best-effort* (LBE) service may be considered as another way of providing a differentiated quality of service, following A. Odlyzko's "damaged goods for the Internet" approach. In this paper we are interested in evaluating, from a *pricing* perspective, the implications of the two scheduling models commonly proposed for building a LBE service—namely, Priority Queueing and Generalized Processor Sharing (GPS). In particular, we focus on the network operator's issue of maximizing her revenue. We wish to study, for each scheduler, how to set prices and, especially, the impact that a given queueing model may have on revenues when users are mostly sensitive to delay. Drawing on previous work by Mandjes (2003), we present analytical expressions of the revenue earned by the network operator, when a GPS scheduler is used. A comparison of optimal revenues shows that: (a) Priority Queueing is more efficient, in economic terms, than both a GPS scheduler and a simple FIFO queue, that is, a network with no service differentiation; (b) revenues are lower with a GPS scheduler than with a FIFO queue. These results may have implications both on the practical implementation of LBE services and on the Paris Metro Pricing proposal by Odlyzko (1999).**

Keywords: Less-than-best-effort, service differentiation, pricing, stochastic processes/queueing theory, economics.

I. Introduction

In the context of IP networks, the term "service differentiation" usually carries the implicit meaning of offering enhanced services. A great deal of research effort has been devoted to defining, implementing and testing network architectures and mechanisms allowing to improve the quality of service (QoS) provided to some flows. However, the notion of differentiated services does not preclude the possibility of having a (potentially) *degraded* service with respect to the ubiquitous best-effort (BE) service.

Less-than-best-effort (LBE) has been proposed as a service for carrying non-critical traffic. The goal of LBE is to exploit unused network capacity, while protecting best-effort flows (and, of course, flows transported by enhanced services, if any) from such non-critical traffic. The usual definition of the LBE service is the following: in the event of congestion, all LBE traffic must be discarded before any BE packet is dropped.

LBE may be regarded not only as a means of protecting more "important" flows from congestion (due either to less-critical or to potentially "damaging" flows), but also as a way to shift network usage towards off-peak times. Since LBE is designed to use idle network capacity, flows using LBE experience a better quality of service whenever the critical-traffic load is low. Moreover, it is intuitively clear that LBE traffic should be charged at a (much) lower rate than, say, BE traffic, so LBE offers an incentive to users to transport non-critical data at a lower cost. Therefore, a LBE service may be of interest not only to network operators, but also to end-users.

Some examples of application scenarios [1] where a LBE service may prove useful are: content mirroring and news distribution; new distributed applications that can take advantage of spare network capacity; non-time-critical, bulk-data transfer based on TCP; isolating production traffic from test traffic; isolating mission-critical traffic from other kinds of production traffic that may be disruptive (e.g., traffic generated from a student dormitory in a university campus).

The LBE concept has already been tested in academic research networks like Internet2 [2] and GÉANT [3]. Such studies have focused mainly on the impact of LBE on more-critical traffic, and on practical issues like router configuration.

A. Providing LBE in a DiffServ network

The notion of a LBE service can be easily integrated into the DiffServ architecture defined by the IETF; the Lower Effort per-domain behaviour [4] is an example of a LBE type of service that can be offered within a DiffServ domain. One interesting point of LBE is that it can be incrementally deployed[1], allowing to put a LBE per-hop behaviour where it matters—that is, in congested links. Another advantage of a LBE service is that there is no need to police or to shape traffic: since the service offers no guarantee of delivery, excess LBE traffic can simply be discarded.

The implementation of LBE using the standard DiffServ building blocks requires, among other things: (1) marking packets with a DiffServ codepoint (DSCP) selected to mark packets as "LBE", and (2) a router mechanism allowing to treat

[1]Provided, of course, that intermediate network nodes do not erase or modify the DiffServ codepoint.

IP packets differently according to the DSCP. Regarding the marking of packets, one can imagine that it is done voluntarily by end-systems (for instance, to select the lowest-cost service), or that marking is enforced by the network operator (for example, to treat all traffic coming from a given subnetwork as non-critical). Concerning the second issue, an implementor might choose between putting both LBE and BE traffic in a single queue handled by an active queue management algorithm like RIO [5], and using a separate queue for LBE traffic. Since RIO and similar algorithms may introduce excessive jitter in best-effort traffic, it may be preferable to adopt the separate-queue strategy [2], providing the desired differentiation by means of a standard scheduling algorithm. Two kinds of schedulers have been proposed to handle LBE traffic [1], [2], [4]: strict, non-preemptive priority queueing (PQ) and weighted fair queueing (WFQ) or one of its variants, like for instance weighted round-robin [6]. From a theoretical standpoint, a WFQ-like scheduler can be regarded as a packet-level version of a Generalized Processor Sharing (GPS) server [7]. As we will discuss in Section IV, each scheduler may have a different impact on the performance of the LBE service, especially when TCP traffic is considered.

B. Marking, scheduling and pricing

In this paper, we are interested in studying the implications of the two scheduling models commonly used for building a LBE service—namely, PQ and GPS—from a *pricing* perspective (for an thorough overview of pricing issues in telecommunication networks, see e.g. [8]). In particular, we consider a for-profit network provider, and we focus on the provider's issue of maximizing her revenue. We wish to study, for each scheduler, how to optimally set prices and, especially, the impact that a given queueing system may have on revenues. Our work is based on a recent paper by Mandjes [9], where optimal prices and revenues for PQ have been determined. Like Mandjes, we will use the performance of a single FIFO queue (corresponding to a network *without* service differentiation) as a benchmark.

In our model, as in [9], we will assume that delay is the main QoS parameter users care about. Hence, a user's utility is described by a strictly decreasing function of delay; the impact of packet losses is left for further study. We will consider that there exist two classes of users (or applications), which differ in their sensitivity to delay. Best-effort—and, more so, LBE—can be thought of as a service intended for transporting data flows without strong time constraints; nonetheless, for the sake of notational consistency with [9], we will call *"voice" users* those that prefer a lower packet delay, and *"data" users* those having less stringent delay requirements. Both types of flows are assumed to be *elastic* [10], in the sense that they may trade delay for price.

We assume that the network under consideration offers only two services: best-effort and less-than-best-effort[2], and that users are charged on a per-packet basis. The network is modelled as a single bottleneck node, where either a PQ server or a GPS server is used to handle two queues, one for packets marked as "BE" and another for packets marked as "LBE". BE traffic is charged at a higher rate than LBE traffic. Like Mandjes, we suppose that there can be two marking scenarios: (1) a situation of *dedicated traffic classes*, in which the network chooses to which queue packets must be sent, and (2) a situation of *open traffic classes*, in which users are able to select what class to join, according to charging rates and the expected QoS.

The main contributions of this paper are as follows. First, we extend Mandjes' model of a PQ server to the case of a GPS server. To the best of our knowledge, there are no closed-form theoretical results concerning delay in GPS queues (for a discussion, see for instance [11]). However, one can argue that there is interest in offering some form of differentiated services (like LBE) to elastic flows only if congestion may arise. Hence, for modelling purposes it makes sense to consider a network under the assumption of a *heavy-traffic regime*, as in [12]. In such a case, a two-queue GPS server behaves approximately like a "partitioned" server, that is, two independent FIFO queues, each with a service rate equal to the corresponding minimum guaranteed rate in the GPS system. This amounts to saying that, under the heavy-traffic hypothesis, a GPS server with a different charging rate for each queue can be regarded as a *Paris Metro Pricing* (PMP) network [13], in which the capacity of the link is logically split in two. Second, we compare the optimal revenues earned with PQ, FIFO and GPS scheduling. Our main conclusion is that a network offering two different services (i.e., BE and LBE) may yield higher revenues than a network with no service differentiation, and also that the type of scheduler used may play an important role in maximizing revenues. In particular, we show that: (a) Priority Queueing is more efficient, in economic terms, than both a GPS scheduler and a simple FIFO queue; (b) revenues are lower with a GPS scheduler than with a FIFO queue.

C. Outline of the paper

This paper is organized as follows. Section II presents a mathematical model of a DiffServ node supporting both BE and LBE service classes, using two different scheduling policies. Section III compares the performance of each scheduling policy in terms of the network provider's revenue. Section IV provides a discussion on the possible economic and practical implications of our main results. Finally, Section V concludes the paper.

II. Mathematical Model

Let us begin by a brief presentation of the model introduced in [9] by Mandjes[3]. Next, we will formalize the heavy-traffic assumption that we will use to treat a GPS scheduler as a set

[2]The main motivation of our work was to study the case of less-than-best-effort services; however, the model presented in this paper may of course be applied also to a network offering, say, EF and AF services.

[3]In order to make our paper as self-contained as possible, we provide in the Appendix a summary of the main results of [9], concerning the optimal prices for both a FIFO and a PQ server.

of independent FIFO queues, before stating our results on the optimal prices for the GPS queue that will be necessary for the revenue comparison.

A. Basic model

Consider an infinite population of potential users. Two types of flows are considered, differing in their sensitivity to delay; we will call these traffic classes type-v ("voice") and type-d ("data") traffic.

The *utility* users get depends not only on the mean packet delay $I\!E D$, but also on the price per packet p, in the following way:

$$U_d(I\!E D) = \frac{1}{(I\!E D)^{\alpha_d}} - p, \tag{1}$$

$$U_v(I\!E D) = \frac{1}{(I\!E D)^{\alpha_v}} - p, \tag{2}$$

with $0 < \alpha_d < \alpha_v$, so that type-v flows have a higher preference for small delays. α_d and α_v may be regarded as the "delay-sensitivity parameter" for each traffic class. It is assumed that users enter the network whenever their utility is positive.

Throughout the paper, we are going to consider a M/M/1 queue with service rate μ and with N independent users, each user generating packets according to a Poisson process with rate λ. The delay is then

$$I\!E D = \frac{1}{\mu - N\lambda}$$

as long as $N\lambda < \mu$ [14]. Let λ_d and λ_v be the packet arrival rate for type-d and type-v users, respectively.

If there were only type-d users, the maximum number of users entering the network would be

$$N_d(p) = \frac{\mu - \sqrt[\alpha_d]{p}}{\lambda_d}, \tag{3}$$

if $p \leq \mu^{\alpha_d}$, and $N_d(p) = 0$ otherwise; note that $N_d(p)$ represents the highest number of users corresponding to a non-negative utility. Similarly, if there were only type-v users, we would have a maximum of $N_v(p)$ users, where

$$N_v(p) = \frac{\mu - \sqrt[\alpha_v]{p}}{\lambda_v}, \tag{4}$$

if $p \leq \mu^{\alpha_v}$, and $N_v(p) = 0$ otherwise.

Regarding the case in which there is competition between both types of traffic, let us quote the following key result from [9].

Proposition 1 (Mandjes [9]): Consider a FIFO M/M/1 queue and a per-packet price p. If $p < 1$ only type-d users will have an incentive to join the network, whereas if $p > 1$ only type-v users have that incentive.

Indeed, if $N_d(p)$ users are already present, with $N_d(p)$ given by (3), an infinitesimal type-v user will enter the network if and only if

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

Since $\alpha_v > \alpha_d$, type-v users would join if and only if $p > 1$. Conversely, if there are $N_v(p)$ type-v users in the queue, a type-d user will have an incentive to join if and only if $p < 1$.

B. GPS scheduling in a heavy-traffic scenario

Offering a differentiated treatment to elastic flows makes sense mainly in the case where long-term congestion may occur. The heavy-traffic hypothesis can then be regarded as a central assumption for service differentiation. To illustrate this point, let us recall how a multi-class queue served by a GPS scheduler with I classes works. Class i is served with a given proportion γ_i of the resource, such that $\sum_{i=1}^{I} \gamma_i = 1$; however, when a queue is empty, its idle server capacity is shared among the other classes (i.e., the server is work-conserving). In a heavy-traffic scenario, it is assumed that, for all the classes, the probability to have an empty buffer is close to zero; in this situation, the GPS queue can be seen as I logically separate M/M/1 queues (with queue i having a service rate $\gamma_i\mu$, where μ is the service rate of the "global" M/M/1 queueing system). If we consider that traffic in each queue is charged at a different rate, this logical split of the server actually results in the Paris Metro Pricing (PMP) model [13], whose performance has been investigated elsewhere [15], [16].

Assuming independence between the M/M/1 queues, it is easy to show that the steady-state probability of having at least one empty queue is $P_0 = 1 - \prod_{i=1}^{I} \rho_i$, with $\rho_i = \lambda/(\gamma_i\mu)$. An *$\epsilon$-heavy-traffic* regime is such that $P_0 < \epsilon$, for an arbitrarily small value of ϵ. We will provide later an interpretation of this probability, when prices are set to their optimal values for the GPS queue.

From now on, let us focus on the $I = 2$ case, corresponding to two traffic classes. Let $0 \leq \gamma \leq 1$ denote the proportion of bandwidth allocated to queue $i = 2$; service rates are then $(1-\gamma)\mu$ and $\gamma\mu$ for queues 1 and 2, respectively.

We will begin by looking at the optimal prices and revenues when a given type of traffic is directed to a given queue (i.e., the dedicated classes scenario described in Section I-B). Afterwards, we will look at the case of open classes.

1) Dedicated classes: Assume that queue 1 is dedicated to type-v traffic (the most stringent one) and that queue 2 is devoted to type-d traffic. Therefore, $\gamma \times 100\%$ and $(1-\gamma) \times 100\%$ of the bandwidth is assigned to type-d and type-v traffic, respectively. The number of type-d users is

$$N_2(p) = N_d(p) = \frac{\gamma\mu - \sqrt[\alpha_d]{p}}{\lambda_d}, \tag{5}$$

when $p < (\gamma\mu)^{\alpha_d}$, and 0 otherwise. Similarly, the number of type-v users is

$$N_1(p) = N_v(p) = \frac{(1-\gamma)\mu - \sqrt[\alpha_v]{p}}{\lambda_v}, \tag{6}$$

when $p < ((1-\gamma)\mu)^{\alpha_v}$, and 0 otherwise.

a) Optimal prices: Suppose that the per-packet price is different for each queue. Let p_1 denote the price in the queue dedicated to type-v traffic, and p_2 the price in the queue reserved to type-d traffic. The revenue $\Pi_{\text{GPS}}^{(\gamma)}$ is defined as the product of the mean input packet rate and the per-packet price:

$$\begin{aligned}\Pi_{\text{GPS}}^{(\gamma)}(p_1,p_2) &= \lambda_1 N_1(p_1)p_1 + \lambda_2 N_2(p_2)p_2,\\ &= \lambda_v N_v(p_1)p_1 + \lambda_d N_d(p_2)p_2,\\ &= (1-\gamma)\mu p_1 - p_1^{1+1/\alpha_v}\\ &\quad + \gamma\mu p_2 - p_2^{1+1/\alpha_d}\end{aligned} \tag{7}$$

for $(p_1,p_2) \in D_{\text{D}} = [0,((1-\gamma)\mu)^{\alpha_v}] \times [0,(\gamma\mu)^{\alpha_d}]$. The optimal prices for a fixed value of γ are given by the following theorem.

Theorem 1: For a fixed γ, the prices that maximize the revenue $\Pi_{\text{GPS}}^{(\gamma)}$ for a GPS queue under an ϵ-heavy-traffic regime are given by:

$$p_1^* = \left(\frac{(1-\gamma)\mu}{1+\frac{1}{\alpha_v}}\right)^{\alpha_v} \text{ and } p_2^* = \left(\frac{\gamma\mu}{1+\frac{1}{\alpha_d}}\right)^{\alpha_d}. \tag{8}$$

Proof: The revenue function $\Pi_{\text{GPS}}^{(\gamma)}(p_1,p_2)$ as given by (7) is a continuous, derivable function on the domain D_{D}. To find its maximum, we first equate the following partial derivatives to 0:

$$\frac{\partial \Pi_{\text{GPS}}^{(\gamma)}}{\partial p_1}(p_1,p_2) = (1-\gamma)\mu - (1+\frac{1}{\alpha_v})p_1^{1/\alpha_v}$$

and

$$\frac{\partial \Pi_{\text{GPS}}^{(\gamma)}}{\partial p_2}(p_1,p_2) = \gamma\mu - (1+\frac{1}{\alpha_d})p_2^{1/\alpha_d}.$$

Hence, the critical point (p_1^*,p_2^*) is:

$$(p_1^*,p_2^*) = \left(\left(\frac{(1-\gamma)\mu}{1+\frac{1}{\alpha_v}}\right)^{\alpha_v}, \left(\frac{\gamma\mu}{1+\frac{1}{\alpha_d}}\right)^{\alpha_d}\right).$$

Remark that (p_1^*,p_2^*) is inside D_{D}. From the second-order derivatives, it is easy to verify that (p_1^*,p_2^*) corresponds to a maximum.

Finally, it is easy to verify from (7) that, on the frontier of D_{D}, the revenue $\Pi_{\text{GPS}}^{(\gamma)}$ is lower (because either the price or the number of users is 0 for at least one of the queues). Therefore, the point (p_1^*,p_2^*) is a global optimum. ∎

We may now compute from (7) the optimal revenue for a given value of γ:

$$\Pi_{\text{GPS}}^*(\gamma) = (1-\gamma)^{1+\alpha_v}A(\alpha_v) + \gamma^{1+\alpha_d}A(\alpha_d) \tag{9}$$

with: $A(x) = \mu^{1+x}(\frac{x}{1+x})^x\frac{1}{1+x}$.

b) Optimal bandwidth sharing: Once the optimal prices have been found, we wish to optimally share the bandwidth among traffic classes, so as to maximize $\Pi_{\text{GPS}}^*(\gamma)$. The following theorem gives the corresponding value of γ.

Theorem 2: For a GPS queue under an ϵ-heavy-traffic regime, the maximum revenue is given by:

$$\Pi_{\text{GPS}}^* = \max(A(\alpha_v), A(\alpha_d)),$$

which implies that either $\gamma = 0$ or $\gamma = 1$. This is equivalent to considering that only one queue is served. Moreover, we have the following particular cases:

- If $\alpha_d > \frac{1}{\mu-1}$, then $\Pi_{\text{GPS}}^* = A(\alpha_v)$.
- If $\alpha_v < \frac{1}{\mu-1}$, then $\Pi_{\text{GPS}}^* = A(\alpha_d)$.

Proof: From (9), we get that the second derivative of $\Pi_{\text{GPS}}^*(\gamma)$ with respect to γ is of the form:

$$\begin{aligned}\Pi_{\text{GPS}}^{*''}(\gamma) &= (1+\alpha_v)\alpha_v(1-\gamma)^{\alpha_v-1}A(\alpha_v)\\ &\quad + (1+\alpha_d)\alpha_d\gamma^{\alpha_d-1}A(\alpha_d) \geq 0.\end{aligned}$$

Therefore, $\Pi_{\text{GPS}}^*(\gamma)$ is convex in γ. Moreover, we have:

$$\Pi_{\text{GPS}}^{*'}(0) = -(1+\alpha_v)A(\alpha_v) < 0$$

and

$$\Pi_{\text{GPS}}^{*'}(1) = (1+\alpha_d)A(\alpha_d) > 0.$$

Hence, $\Pi_{\text{GPS}}^*(\gamma)$ is a convex function which reaches its maximum either at $\gamma = 0$ or $\gamma = 1$. Because of the shape of the optimal revenue function, we conclude that a single type of traffic should be served if we wish to optimize the revenue:

$$\begin{aligned}\max_\gamma \Pi_{\text{GPS}}^*(\gamma) &= \max(\Pi_{\text{GPS}}^*(0), \Pi_{\text{GPS}}^*(1)),\\ &= \max(A(\alpha_v), A(\alpha_d)),\end{aligned}$$

Let us study the behavior of the function $A(.)$ defined over $[0,+\infty[$. Its derivative is:

$$A'(x) = \left(\frac{\mu}{1+x}\right)^{1+x} x^x \left(\ln\left(\frac{x}{1+x}\right) + \ln\mu\right). \tag{10}$$

This expression is negative if and only if:

$$\ln\left(\frac{x}{1+x}\right) + \ln\mu < 0 \quad \Leftrightarrow \quad (\mu-1)x < 1. \tag{11}$$

However, if $0 < \mu \leq 1$ then $(\mu-1)x \leq 0$, $\forall x > 0$, so the function $A(.)$ is decreasing over $[0,+\infty[$. Since $\alpha_v > \alpha_d$, we have that $A(\alpha_v) < A(\alpha_d)$, so the maximum revenue is attained at $\gamma = 1$; in other words, only "data" traffic is handled whenever $0 < \mu \leq 1$.

Finally, when $\mu > 1$ we may deduce from (10) and (11) that:

$$\begin{aligned}&A'(x) = 0 \text{ only at } x = 1/(\mu-1),\\ &A'(x) < 0, \forall x < 1/(\mu-1),\\ &A'(x) > 0, \forall x > 1/(\mu-1).\end{aligned}$$

Hence, $A(x)$ reaches its minimum at $x = 1/(\mu - 1)$. Depending on the values of μ, α_v and α_d, we have three possible cases (recall that $\alpha_d < \alpha_v$):

- If $\alpha_d > \frac{1}{\mu-1}$, then the revenue is maximized by taking $\gamma = 0$ (i.e., only "voice" traffic is accepted).
- If $\alpha_v < \frac{1}{\mu-1}$, then the revenue is maximized by taking $\gamma = 1$ (i.e., only "data" traffic is accepted).
- Otherwise, if $\alpha_d < \frac{1}{\mu-1} < \alpha_v$, then one has to numerically compare $A(\alpha_d)$ and $A(\alpha_v)$ to find the maximum revenue.

■

Remark 1: Theorem 2 states that revenue is optimized by serving only one queue of the GPS system. This amounts to say that, in order to maximize the revenue, *only one traffic class* should be accepted into the network.

c) Heavy-traffic hypothesis and optimal prices: Let $0 < \gamma < 1$, that is, both traffic classes are handled by the network. The load of queues 1 and 2 is given by:

$$\rho_1 = \rho_v = \frac{\lambda_v N_v}{(1-\gamma)\mu} \text{ and } \rho_2 = \rho_d = \frac{\lambda_d N_d}{\gamma\mu}.$$

Assume that we have the maximum number of users of each class, as given by (5) and (6). Hence, ρ_1 and ρ_2 can be expressed as:

$$\rho_1 = 1 - \frac{\sqrt[\alpha_v]{p_1}}{(1-\gamma)\mu} \text{ and } \rho_2 = 1 - \frac{\sqrt[\alpha_d]{p_2}}{\gamma\mu}.$$

If we set prices to their revenue-optimizing values p_1^* and p_2^*, as given by (8), we get

$$\rho_2 = \frac{1}{1+\alpha_d} \text{ and } \rho_1 = \frac{1}{1+\alpha_v}.$$

Hence, the probability of having at least an empty queue in the system is:

$$P_0 = 1 - \frac{1}{(1+\alpha_v)(1+\alpha_d)}.$$

Note that, interestingly enough, when prices are optimal the load of each queue (and so the probability P_0 related to the heavy-traffic regime) depends only on the delay sensitivities α_v and α_d.

2) Open classes: Let us suppose now that users are free to select to which queue packets are sent, irrespective of the traffic class; this means that there *might* be both type-v and type-d packets in the same queue. We will keep the same notation as in the dedicated-class scenario. In particular, γ denotes the proportion of total bandwidth μ allocated to queue 2.

Let p_1 be the per-packet price for queue 1, that is, the queue receiving a service rate $(1-\gamma)\mu$. Likewise, let p_2 be the per-packet price for queue 2, that is, the queue receiving a service rate equal to $\gamma\mu$. The revenue is given by:

$$\Pi^{(\gamma)}_{\text{GPS}}(p_1, p_2) = \lambda_1 N_1(p_1) p_1 + \lambda_2 N_2(p_2) p_2. \qquad (12)$$

Proposition 2: If $p_1 < 1$ then only type-d packets will enter queue 1. If $p_2 < 1$ then only type-d packets will enter queue 2. A similar situation will arise with respect to type-v packets if either $p_1 > 1$ or $p_2 > 1$.

Proof: The system is composed of two queues in parallel, so we may use (3) and (4) and follow the same reasoning as in Section II-A. Consider first queue 1, and assume that there are $N_d(p_1)$ type-d users in this queue, where:

$$N_d(p_1) = \frac{(1-\gamma)\mu - \sqrt[\alpha_d]{p_1}}{\lambda_d}.$$

For an infinitesimal type-v user to enter this queue, her utility $U_v(p_1)$ has to be positive:

$$\begin{aligned} U_v(p_1) &= \left((1-\gamma)\mu - \lambda_d N_d(p_1)\right)^{\alpha_v} - p_1 \\ &= p_1^{\alpha_v/\alpha_d} - p_1 > 0, \end{aligned}$$

which happens if and only if $p_1 > 1$. Conversely, if there are already $N_v(p_1)$ type-v users in this queue, an infinitesimal type-d user will only enter the queue if $p_1 < 1$.

The same analysis can be done for queue 2. Assuming $N_v(p_2)$ type-v users are already in this queue, an infinitesimal type-d user will enter this queue if

$$\begin{aligned} U_d(p_2) &= \left(\gamma\mu - \lambda_v N_v(p_2)\right)^{\alpha_d} - p_2 \\ &= p_2^{\alpha_d/\alpha_v} - p_2 > 0, \end{aligned}$$

which happens if and only if $p_2 < 1$.

Therefore, there can be only one class of traffic in each queue. ■

a) Optimal prices: As a consequence of the price-induced separation of traffic classes, the revenue function $\Pi^{(\gamma)}_{\text{GPS}}$ takes a different form depending on the composition of incoming traffic. For a fixed γ, the optimization domain can be decomposed in three sub-domains: V_{O} (only type-v traffic), D_{O} (only type-d traffic) and M_{O} (both types of traffic are present), as follows; see also Fig. 1.

$$\begin{aligned} V_{\text{O}} &= [1, ((1-\gamma)\mu)^{\alpha_v}] \times [1, (\gamma\mu)^{\alpha_v}] \\ D_{\text{O}} &= [0, \min(1, ((1-\gamma)\mu)^{\alpha_d})] \times [0, \min(1, (\gamma\mu)^{\alpha_d})] \\ M_{\text{O}} &= M_{\text{O}}^{(1)} \cup M_{\text{O}}^{(2)} \end{aligned}$$

where:

$$\begin{aligned} M_{\text{O}}^{(1)} &= [1, ((1-\gamma)\mu)^{\alpha_v}] \times [0, \min(1, (\gamma\mu)^{\alpha_d})] \\ M_{\text{O}}^{(2)} &= [0, \min(1, ((1-\gamma)\mu)^{\alpha_v})] \times [1, (\gamma\mu)^{\alpha_d}] \end{aligned}$$

When using the notation $[a, b]$, we follow the convention: $[a, b] = \emptyset$ whenever $a > b$.

For each sub-domain V_{O}, D_{O} and M_{O}, we will now find $\Pi_{\text{GPS}}(p_1, p_2)$.

b) V_O (only type-v traffic): In this case, only "voice" traffic enters the network, and this homogeneous traffic is split among the two queues. Suppose that $((1-\gamma)\mu)^{\alpha_v} \geq 1$ and $(\gamma\mu)^{\alpha_v} \geq 1$, so that the set V_{O} is non-empty. Note that this happens if and only if:

$$0 \leq \frac{1}{\mu} \leq \gamma \leq 1 - \frac{1}{\mu} \leq 1. \qquad (13)$$

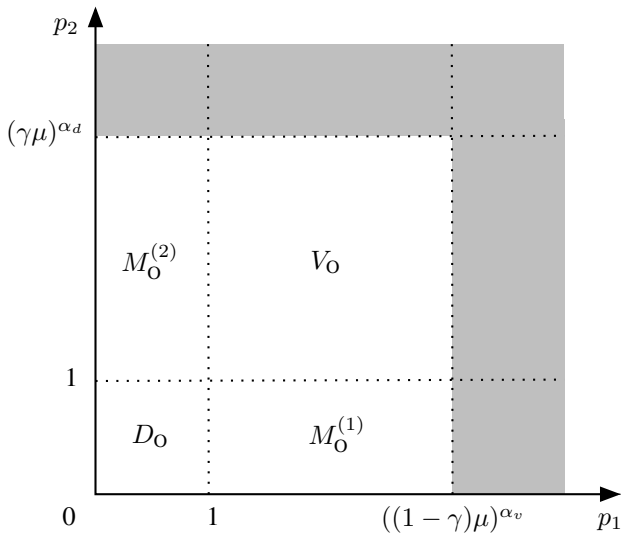

Fig. 1. GPS queue in an ϵ-heavy-traffic scenario: optimization domains for the revenue function.

The number of users $N_1(p_1)$ and $N_2(p_2)$ in queues 1 and 2, respectively, is:

$$N_1(p_1) = \frac{(1-\gamma)\mu - \sqrt[\alpha_v]{p_1}}{\lambda_1} \tag{14}$$

and

$$N_2(p_2) = \frac{\gamma\mu - \sqrt[\alpha_v]{p_2}}{\lambda_2}. \tag{15}$$

The revenue function to be optimized, with μ and γ fixed, is obtained from (12), (14) and (15):

$$\begin{aligned}\Pi_{\rm GPS}^{(\gamma)}(p_1,p_2) = (1-\gamma)\mu p_1 - p_1^{1+1/\alpha_v} \\ + \gamma\mu p_2 - p_2^{1+1/\alpha_v}.\end{aligned} \tag{16}$$

Notice that $\Pi_{\rm GPS}^{(\gamma)}(p_1,p_2)$ is continuous in p_2 and in p_1. Let us find the maximum of $\Pi_{\rm GPS}^{(\gamma)}$ for $p_1 \in [1,((1-\gamma)\mu)^{\alpha_v}]$ and $p_2 \in [1,(\gamma\mu)^{\alpha_v}]$.

Proposition 3: The optimal prices over $V_{\rm O}$ are given by:

$$p_1^* = \max\left(1, \left(\frac{(1-\gamma)\mu}{1+1/\alpha_v}\right)^{\alpha_v}\right),$$

and

$$p_2^* = \max\left(1, \left(\frac{\gamma\mu}{1+1/\alpha_v}\right)^{\alpha_v}\right).$$

Note that $(p_1^*,p_2^*) \in V_{\rm O}$.

Proof: Note that the revenue function (16), with γ fixed, can be decomposed as the sum of two functions, one depending only on p_1 and the other depending only on p_2:

$$\Pi_{\rm GPS}^{(\gamma)}(p_1,p_2) = f(p_1) + g(p_2),$$

with:

$$\begin{aligned}f(p_1) &= (1-\gamma)\mu p_1 - p_1^{1+1/\alpha_v}, \\ g(p_2) &= \gamma\mu p_2 - p_2^{1+1/\alpha_v}.\end{aligned}$$

From:

$$\begin{aligned}f'(p_1) &= (1-\gamma)\mu - (1+1/\alpha_v)p_1^{1/\alpha_v}, \\ g'(p_2) &= \gamma\mu - (1+1/\alpha_v)p_2^{1/\alpha_v},\end{aligned}$$

we get that, over $[1,((1-\gamma)\mu)^{\alpha_v}]$, f attains its maximum at $\max\left(1,\left(\frac{(1-\gamma)\mu}{1+1/\alpha_v}\right)^{\alpha_v}\right)$ and, over $[1,(\gamma\mu)^{\alpha_v}]$, g attains its maximum at $\max\left(1,\left(\frac{\gamma\mu}{1+1/\alpha_v}\right)^{\alpha_v}\right)$, which leads to the result. ■

c) D_O (only type-d traffic): This case is similar to the preceding one, except that only "data" packets are present.

Proposition 4: The optimal prices over $D_{\rm O}$ are given by:

$$p_1^* = \min\left(1, \left(\frac{(1-\gamma)\mu}{1+1/\alpha_d}\right)^{\alpha_d}\right),$$

and

$$p_2^* = \min\left(1, \left(\frac{\gamma\mu}{1+1/\alpha_d}\right)^{\alpha_d}\right).$$

Note that $(p_1^*,p_2^*) \in D_{\rm O}$.

Proof: The revenue function to be optimized is:

$$\Pi_{\rm GPS}^{(\gamma)}(p_1,p_2) = (1-\gamma)\mu p_1 - p_1^{1+1/\alpha_d} + \gamma\mu p_2 - p_2^{1+1/\alpha_d}.$$

Using the same kind of reasoning as for the previous case, we can obtain the result of Proposition 4. ■

d) M_O (both types of traffic are present): The condition for $M_{\rm O}^{(1)}$ to be non-empty is:

$$1 \le p_1 \le ((1-\gamma)\mu)^{\alpha_v} \text{ and } 0 \le p_2 \le \min(1,(\gamma\mu)^{\alpha_d}).$$

which is verified if

$$0 \le \gamma \le 1 - \frac{1}{\mu}.$$

Similarly, the condition for $M_{\rm O}^{(2)}$ to be non-empty is:

$$0 \le p_1 \le \min(1,((1-\gamma)\mu)^{\alpha_v}) \text{ and } 1 \le p_2 \le (\gamma\mu)^{\alpha_d}$$

which is verified if

$$\gamma \ge \frac{1}{\mu}.$$

Over $M_{\rm O}^{(1)}$, the revenue function is:

$$\Pi_{\rm GPS}^{(\gamma)}(p_1,p_2) = (1-\gamma)\mu p_1 - p_1^{1+1/\alpha_v} + \gamma\mu p_2 - p_2^{1+1/\alpha_d}.$$

In a similar fashion, we obtain the revenue function over $M_{\rm O}^{(2)}$, which is:

$$\Pi_{\rm GPS}^{(\gamma)}(p_1,p_2) = (1-\gamma)\mu p_1 - p_1^{1+1/\alpha_d} + \gamma\mu p_2 - p_2^{1+1/\alpha_v}.$$

Proposition 5: The optimal prices over $M_{\rm O}^{(1)}$ are given by:

$$p_1^{(1)*} = \max\left(1, \left(\frac{(1-\gamma)\mu}{1+1/\alpha_v}\right)^{\alpha_v}\right),$$

and

$$p_2^{(1)*} = \min\left(1, \left(\frac{\gamma\mu}{1+1/\alpha_d}\right)^{\alpha_d}\right).$$

Similarly, the optimal prices over $M_{\mathrm{O}}^{(2)}$ are given by:

$$p_1^{(2)*} = \min\left(1, \left(\frac{(1-\gamma)\mu}{1+1/\alpha_d}\right)^{\alpha_d}\right),$$

and

$$p_2^{(2)*} = \max\left(1, \left(\frac{\gamma\mu}{1+1/\alpha_v}\right)^{\alpha_v}\right).$$

Note that $(p_1^{(1)*}, p_2^{(1)*}) \in M_{\mathrm{O}}^{(1)}$ and $(p_1^{(2)*}, p_2^{(2)*}) \in M_{\mathrm{O}}^{(2)}$.

Proof: The proof follows along the same lines as the proof of Proposition 3. ∎

Finally, in order to find the optimal prices, one simply has to compare the maximum revenue over the three sub-domains V_{O}, D_{O} and M_{O}.

e) Optimal bandwidth sharing: In order to find the optimal revenue for the open-classes case, it remains to find over which sub-domain this maximum is attained and for which value of γ. However, for the sake of clarity we will postpone this analysis to Section III-B (we will find that, anyway, the best solution is to take $\gamma = 0$ or $\gamma = 1$, which is equivalent to the single-queue, FIFO case).

III. Comparison of optimal revenues

In this section, we will compare the maximum revenue that may be obtained with the two chosen server types, PQ and GPS. We will use the maximum revenue yielded by a single FIFO queue, given by Mandjes in [9], as a benchmark. First, we will study the dedicated classes scenario, then the case of open classes.

A. Dedicated classes

In this case, type-v packets are sent to queue 1, whereas type-d packets are sent to queue 2. Let Π_{D}^* denote the optimal revenue for a PQ system with dedicated classes (Appendix B.1). We have the following main result.

Theorem 3: In the dedicated-classes context, maximum revenues always verify:

$$\Pi_{\mathrm{GPS}}^* = \Pi_{\mathrm{FIFO}}^* \leq \Pi_{\mathrm{D}}^*.$$

In words, the maximum revenue in a Priority Queueing system is always higher or equal than that of a "partitioned" server (i.e., a GPS server under heavy load). Moreover, a FIFO queue yields the same optimal revenue as a GPS server under heavy traffic.

The proof of this theorem is divided in two parts. First, we will show the relationship between the optimal revenues for a GPS server and for a FIFO queue, then we will prove that the latter is less than or equal to the optimal revenue of a PQ server. Each result is given as a lemma.

Lemma 1: In a system with dedicated classes, we have that:

$$\Pi_{\mathrm{GPS}}^* = \Pi_{\mathrm{FIFO}}^*.$$

Proof: We showed in Theorem 2 that the optimal revenue Π_{GPS}^* for the GPS server is:

$$\Pi_{\mathrm{GPS}}^* = \max(A(\alpha_v), A(\alpha_d)),$$

with $A(x) = \mu^{1+x}(\frac{x}{1+x})^x \frac{1}{1+x}$, which amounts to taking either $\gamma = 0$ or $\gamma = 1$. That is, when the revenue is optimal the system becomes a single M/M/1 queue served in FIFO fashion with service rate μ. As shown in Appendix A, the above expression is exactly that of the optimal revenue for a FIFO queue. Hence, $\Pi_{\mathrm{GPS}}^* = \Pi_{\mathrm{FIFO}}^*$. ∎

In other words, revenue is optimal when there is a single traffic type in the GPS system with dedicated classes but, as Mandjes [9] proved, optimizing the revenue in a single FIFO queue also requires having a single traffic class (because, for a given price, flows from different classes do not mix).

Lemma 2: In a system with dedicated classes, we have that:

$$\Pi_{\mathrm{D}}^* \geq \Pi_{\mathrm{FIFO}}^*.$$

Proof: Let us quote the following argument from [9]. A FIFO queue can be regarded as a special case of a PQ system—indeed, it suffices to take $p_1 = \mu^{\alpha_v}$ or $p_2 = p_1^{2\alpha_d/\alpha_v}/\mu^{\alpha_d}$, in which case only one queue is "active". Hence, the PQ system cannot yield lower revenues than the FIFO queue. ∎

B. Open classes

In this scenario, users are able to select to which queue packets are sent. Let us recall the form of the revenue function that we wish to maximize:

$$\Pi(p_1, p_2) = \lambda_1 N_1(p_1, p_2) p_1 + \lambda_2 N_2(p_1, p_2) p_2,$$

with λ_i and N_i denoting the arrival rate and the number of users, respectively, for queue i. Let Π_{O}^* denote the optimal revenue for a PQ system with open classes (Appendix B.2). We then have the following main result.

Theorem 4: In the open-classes context, maximum revenues always verify:

$$\Pi_{\mathrm{GPS}}^* = \Pi_{\mathrm{FIFO}}^* \leq \Pi_{\mathrm{O}}^*.$$

That is, the maximum revenue in a Priority Queueing system is always higher or equal than that of a "partitioned" server (i.e., a GPS server under heavy load). Moreover, a FIFO queue also yields higher or equal revenues than a GPS server.

To prove this theorem, we will proceed as in the dedicated-classes case; two intermediate lemmas will allow us to show the main theorem.

Lemma 3: In a system with open classes, we have that:

$$\Pi_{\mathrm{GPS}}^* = \Pi_{\mathrm{FIFO}}^*.$$

In words, the highest revenue that we can get with a GPS server corresponds to that of a single FIFO queue.

Proof: Let us compare, over each sub-domain V_{O}, D_{O} and M_{O}, the revenue of a GPS server under heavy load with that of a FIFO queue.

Let us focus first on V_{O}. Recall from Proposition 3 that the optimal prices are:

$$p_1^* = \max\left(1, \left(\frac{(1-\gamma)\mu}{1+1/\alpha_v}\right)^{\alpha_v}\right),$$
$$p_2^* = \max\left(1, \left(\frac{\gamma\mu}{1+1/\alpha_v}\right)^{\alpha_v}\right).$$

Hence, there are four possible cases to study.

1) Suppose that $p_1^* = (\frac{(1-\gamma)\mu}{1+1/\alpha_v})^{\alpha_v}$ and $p_2^* = (\frac{\gamma\mu}{1+1/\alpha_v})^{\alpha_v}$. From (16), we readily obtain the optimal revenue for a fixed γ:

$$\Pi^*_{\text{GPS}}(\gamma) = (1-\gamma)^{1+1/\alpha_v} A(\alpha_v) + \gamma^{1+1/\alpha_v} A(\alpha_v).$$

This function is convex in γ, and its maximum is at one of the edges of the interval given by (13):

$$\Pi^*_{\text{GPS}} = \max_{\gamma \in [(1+1/\alpha_v)/\mu, 1-(1+1/\alpha_v)/\mu]} \Pi^*_{\text{GPS}}(\gamma),$$

since we are assuming that V_{O} is non-empty, i.e. $\left(\frac{(1-\gamma)\mu}{1+1/\alpha_v}\right)^{\alpha_v} > 1$ and $\left(\frac{\gamma\mu}{1+1/\alpha_v}\right)^{\alpha_v} > 1$. So the maximum revenue is

$$\begin{aligned}\Pi^*_{\text{GPS}} &= \max\left(\Pi^*_{\text{GPS}}(\tfrac{1+1/\alpha_v}{\mu}), \Pi^*_{\text{GPS}}(1-\tfrac{1+1/\alpha_v}{\mu})\right) \\ &\leq \max\left(\Pi^*_{\text{GPS}}(0), \Pi^*_{\text{GPS}}(1)\right) \\ &\leq \Pi^*_{\text{FIFO}}.\end{aligned}$$

2) Suppose now that $p_1^* = 1$ and $p_2^* = 1$. We obtain that the revenue for a fixed γ is

$$\Pi^*_{\text{GPS}}(\gamma) = \mu - 2,$$

which is independent of γ. However, for a FIFO queue with $p = 1$ we have a revenue $\mu - 1$. Therefore, we also have that $\Pi^*_{\text{GPS}} \leq \Pi^*_{\text{FIFO}}$.

3) Suppose now that $p_1^* = (\frac{(1-\gamma)\mu}{1+1/\alpha_v})^{\alpha_v}$ but $p_2^* = 1$. The optimal revenue for a fixed γ is now:

$$\Pi^*_{\text{GPS}}(\gamma) = (1-\gamma)^{1+1/\alpha_v} A(\alpha_v) + \gamma\mu - 1,$$

which is also convex in γ. As before, we have that

$$\begin{aligned}\Pi^*_{\text{GPS}} &\leq \max\left(\Pi^*_{\text{GPS}}(0), \Pi^*_{\text{GPS}}(1)\right) \\ &\leq \max\left(A(\alpha_v), \mu - 1\right) \\ &\leq \Pi^*_{\text{FIFO}}.\end{aligned}$$

4) Finally, the case $p_1^* = 1$ and $p_2^* = (\frac{\gamma\mu}{1+1/\alpha_v})^{\alpha_v}$ is similar to the previous one.

We follow a similar approach to deal with the sub-domains D_{O} and M_{O}. For instance, regarding D_{O} (i.e., only "data" traffic is present), it is easy to check that the previous results apply by simply changing α_v by α_d, with the optimal prices given by Proposition 4. By similar convexity arguments, we also get that $\Pi^*_{\text{GPS}} \leq \Pi^*_{\text{FIFO}}$.

Since taking $\gamma = 0$ or $\gamma = 1$ in the GPS case is equivalent to having a single FIFO queue, we deduce that $\Pi^*_{\text{GPS}} = \Pi^*_{\text{FIFO}}$. ■

Let us now state the second lemma, concerning the revenues of the PQ and the FIFO system.

Lemma 4: In a system with open classes, we have that:

$$\Pi^*_{\text{FIFO}} \leq \Pi^*_{\text{O}}.$$

Proof: As described in Appendix B.2, we decompose the optimization domain of Π_{O} in three sub-domains V, D and M (see Fig. 2).

Assume that $\mu > 1$ (otherwise V and M would be empty).

- Over V, the revenue function is

$$\begin{aligned}\Pi^{(V)}_{\text{O}}(p_1,p_2) = (\mu - \sqrt[\alpha_v]{p_1})p_1 \\ + \left(\sqrt[\alpha_v]{p_1} - \mu\sqrt[\alpha_v]{\frac{p_2}{p_1}}\right)p_2,\end{aligned}$$

with $p_1 \leq \mu^{\alpha_d}$ and $p_2 \leq p_1^2/\mu^{\alpha_d}$.

- Over D, the revenue function is

$$\begin{aligned}\Pi^{(D)}_{\text{O}}(p_1,p_2) = (\mu - \sqrt[\alpha_d]{p_1})p_1 \\ + \left(\sqrt[\alpha_d]{p_1} - \mu\sqrt[\alpha_d]{\frac{p_2}{p_1}}\right)p_2,\end{aligned}$$

with $p_1 \leq \mu^{\alpha_v}$ and $p_2 \leq p_1^2/\mu^{\alpha_v}$.

- Over M, the revenue function is

$$\begin{aligned}\Pi^{(M)}_{\text{O}}(p_1,p_2) = (\mu - \sqrt[\alpha_v]{p_1})p_1 \\ + \left(\sqrt[\alpha_v]{p_1} - \mu\frac{\sqrt[\alpha_d]{p_2}}{\sqrt[\alpha_v]{p_1}}\right)p_2.\end{aligned}$$

with $p_1 \in [1, \mu^{\alpha_v}]$ and $p_2 \in [0, \min(1, \frac{p_1^{2\alpha_d/\alpha_v}}{\mu^{\alpha_d}})]$.

We have:

$$\begin{aligned}&\Pi^*_{\text{FIFO}} = \\ &= \max\left(\max_{1\leq p\leq \mu^{\alpha_v}}(\mu - \sqrt[\alpha_v]{p})p, \max_{p\leq \min(1,\mu^{\alpha_d})}(\mu - \sqrt[\alpha_d]{p})p\right) \\ &= \max\left(\max_{1\leq p\leq \mu^{\alpha_v}}\Pi^{(M)}_{\text{O}}(p,0), \max_{p\leq \min(1,\mu^{\alpha_d})}\Pi^{(M)}_{\text{O}}(\mu^{\alpha_v},p)\right)\end{aligned}$$

The optimization domain of the FIFO queue is a subset of the optimization domain of the PQ server. Therefore, we have necessarily that $\Pi^*_{\text{FIFO}} \leq \Pi^*_{\text{O}}$.

On the other hand, if $\mu \leq 1$, then V and M are empty, so we simply have to find the maximum of $\Pi^{(D)}_{\text{O}}$. We have that:

$$\begin{aligned}\Pi^*_{\text{FIFO}} &= \max_{p\in(0,\mu^{\alpha_d})}((\mu - \sqrt[\alpha_d]{p})p) \\ &= \max_{p\in(0,\mu^{\alpha_d})}\Pi^{(D)}_{\text{O}}(0,p).\end{aligned}$$

which also leads to $\Pi^*_{\text{FIFO}} \leq \Pi^*_{\text{O}}$. ■

IV. Discussion

The results of Mandjes [9] imply that it is interesting, from an economic point of view, to offer some kind of service differentiation; in fact, as shown in [9], having more than one queue may increase revenues even if there is a *single* traffic class. Hence, adding LBE to her service offer may help a for-profit network provider to increase her income.

Nonetheless, as we have seen in Sections II and III, the choice of the scheduling mechanism for building a LBE service may have an impact on the network provider's revenues. Indeed, our results seem to suggest that revenues may be *lower*, not higher, when a differential treatment is offered to flows—it depends on what kind of queueing system is used. It is noteworthy that we have arrived at similar conclusions as in [15], in spite of the fact that the model analyzed in this paper is fairly different from that of [15].

On the other hand, the scheduling mechanism may also strongly affect the *performance* of traffic flows and applications. We have assumed that flows are "infinitely elastic", in the sense that they may accept an unbounded queueing delay, as long as the per-packet price is decreased accordingly. However, in general real flows cannot tolerate arbitrarily large delays—such delays may result either in an unacceptable quality for the end-user or in an interrupted or broken flow. Nonetheless, in the absence of techniques such as admission control, very large delays may occur with a PQ server during heavy-load periods, leading to the starvation of the low-priority (i.e., LBE) flows[4].

To summarize, economic considerations suggest using a PQ server, whereas technical considerations suggest using a GPS server with a (quite) small value of γ. Note that our results, as well as those of Mandjes, are based on a Poisson model of incoming traffic, i.e., a model that does not take into account the adaptive nature of protocols such as TCP [17]. In several application scenarios, LBE is viewed as a service for transporting TCP (or TCP-friendly) flows [2]. In the case of a TCP connection, very long delays may result in the connection being broken. Further study is needed to clarify the implications of the type of server and the pricing mechanism on TCP(-friendly) flows.

V. Conclusions and Future Work

In this paper we have evaluated the issue of deploying a less-than-best-effort service for revenue-maximization purposes. Based on a model by Mandjes, we have studied the economic implications of the two scheduling models, Priority Queueing and Generalized Processor Sharing, which are commonly proposed to build such a service. We have extended Mandjes' model to cover the case of a GPS server in a heavy-traffic regime. Our main results suggest that, in order to increase revenues with respect to a single-service network, it may be necessary to carefully choose the scheduler used in routers to deploy a LBE service.

Concerning our future work, the following issues are worth exploring:

- Incorporating a model of TCP traffic into the queueing and pricing model. This may allow us to study, for instance, the effect of starvation due to the use of a PQ server.
- Concerning the utility function, it may be interesting to take packet losses into account.
- We would like to extend our results to a whole network, and investigate the associated routing problems.
- Finally, we wish to study the case where LBE flows are more bandwidth-sensitive than delay-sensitive, by means of a modified utility function.

Acknowledgment

This work was partially supported by INRIA's "Prixnet" Cooperative Research Action.

[4]Indeed, it is for this very reason that the definition of the Scavenger service in the Internet2 project explicitly discourages the use of PQ [2].

Appendix
Results for FIFO and PQ servers

We recapitulate here the main results of Mandjes [9] on the optimal prices for both a FIFO and a PQ server.

A. FIFO scheduling

We consider a single queue with FIFO (*First In First Out*) scheduling policy, and we look at the price p optimizing the revenue

$$\Pi_{\text{FIFO}}(p) = \lambda_d N_d(p)p + \lambda_v N_v(p)p.$$

From Proposition 1, if $p > 1$ we only have type-v packets, and only type-d packets otherwise. It is shown in [9] that the optimal revenue is given by

$$\Pi^*_{\text{FIFO}} = \max_{p\in I\!R_+} \Pi_{\text{FIFO}}(p) = \max(A(\alpha_d), A(\alpha_v)),$$

where $A(x) = \frac{\mu}{x+1}(\frac{\mu x}{x+1})^x$. Also, for a service rate less that μ^*, with

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

we only have type-d traffic, whereas if $\mu > \mu^*$ we only have type-v traffic.

B. PQ scheduling

The main results of [9] deal with finding optimal prices in the cases where there are two classes of traffic with a non-preemptive strict priority for queue 1. We are going to summarize first the case of dedicated classes, then the case of open classes.

1) Dedicated classes: Assume that priority-class 1 is dedicated to type-v traffic (the most stringent one) and priority-class 2 is dedicated to type-d traffic. The revenue is given by

$$\Pi_{\text{D}}(p_1, p_2) = \lambda_v N_v(p_1, p_2)p_1 + \lambda_d N_d(p_1, p_2)p_2,$$

where p_1 and p_2 are the per-packet price for priority class 1 (the highest priority) and 2 (the lowest priority), respectively. We have

$$N_v(p_2, p_1) = N_v(p_1) = \frac{\mu - \sqrt[\alpha_v]{p_1}}{\lambda_v},$$

if $p_1 \le \mu^{\alpha_v}$ and 0 otherwise. Also,

$$N_d(p_2, p_1) = \lambda_d^{-1}\left(\sqrt[\alpha_v]{p_1} - \mu\sqrt[\alpha_d]{p_2}/\sqrt[\alpha_v]{p_1}\right)$$

if $p_2 < p_1^{2\alpha_d/\alpha_v}/\mu^{\alpha_d}$ and $p_1 \le \mu^{\alpha_v}$,

$$N_d(p_2, p_1) = \lambda_d^{-1}\left(\mu - \sqrt[\alpha_d]{p_2}\right)$$

if $p_2 < \mu^{\alpha_d}$ and $p_1 > \mu^{\alpha_v}$, and $N_d(p_2, p_1) = 0$ otherwise.

In [9], Mandjes obtains the optimal prices p_1^* and p_2^*, and then the optimal revenue, depending on the value of the service rate μ. Let μ^*_- and μ^*_+ be defined by

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

and

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

Then,

- If $\mu \in [0, \mu_-^*]$, we only have type-d traffic and

$$p_1^* = \mu^{\alpha_v} \text{ and } p_2^* = \left(\frac{\mu \alpha_d}{\alpha_d + 1} \right)^{\alpha_d}.$$

- If $\mu \in [\mu_-^*, \mu^*]$, both types of traffic enter and

$$p_1^* = \overline{p}_1 \text{ and } p_2^* = \left(\frac{\mu \alpha_d}{\mu(\alpha_d + 1)} \right)^{\alpha_d} \overline{p}_1^{2\alpha_d/\alpha_v},$$

where $\overline{p}_1$ is the unique solution of $g'_-(p) = 0$ with

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

- If $\mu \in [\mu^*, \mu_+^*]$, both types of traffic enter and

$$p_1^* = \overline{p}_1 \text{ and } p_2^* = \left(\frac{\mu \alpha_v}{\mu(\alpha_v + 1)} \right)^{\alpha_v} \overline{p}_1^{2\alpha_v/\alpha_d},$$

with $\overline{p}_1$ unique solution of $g'_+(p) = 0$ with

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

- If $\mu \geq \mu_+^*$,we only have type-v traffic and

$$p_1^* = \mu^{\alpha_d} \quad \text{and} \quad p_2^* = \left(\frac{\mu \alpha_v}{\alpha_v + 1} \right)^{\alpha_v}.$$

2) Open classes: We assume now that each user is free to choose her priority class. The optimisation domain can be decomposed in three sub-domains V, D and M like shown in Fig. 2.

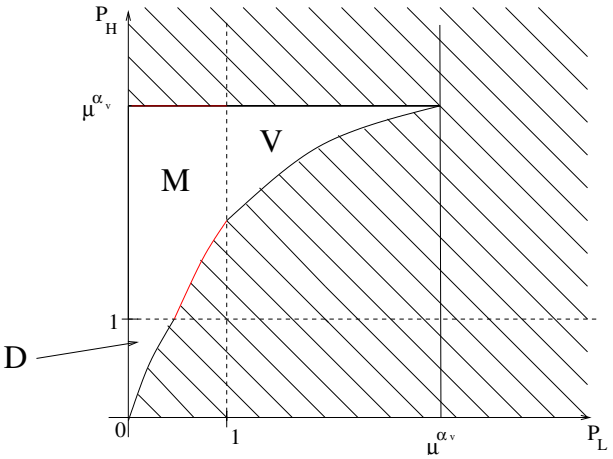

Fig. 2. Optimization domain for PQ and open classes. V, D and M represent the areas where we have only type-v traffic, only type-d traffic, and both types of traffic in the system, respectively.

Indeed, using Proposition 1 we can get that:

- On V, $1 \geq p_2 \geq p_1^2/\mu^{\alpha_v}$ and $1 \leq p_1 \leq \mu^{\alpha_v}$, meaning that we only have type-v traffic.
- On D, where $p_1 \leq \min(1, \mu^{\alpha_d})$ and $p_2 \leq \min(1, p_1^2/\mu^{\alpha_d})$, we only have type-$d$ traffic.
- On M, where $p_2 \in (0, \min(1, \frac{p_1^{2\alpha_d/\alpha_v}}{\mu^{\alpha_d}}))$ and $p_1 \in (1, \mu^{\alpha_v}]$, type-$v$ traffic uses the higher priority queue and type-d traffic uses the lower priority queue.

Note that we have assumed that $\mu > 1$ so that V and M are non-empty.

Like in the dedicated classes model, an algorithm for computing the optimal prices depending on the value of μ is provided in [9]. Due to space limits, we do not reproduce it here.

REFERENCES

[1] T. Chown, T. Ferrari, S. Leinen, R. Sabatino, N. Simar, and S. Venaas, "Less than Best Effort: Application Scenarios and Experimental Results," in *Proceedings of QoS-IP 2003*, ser. Lecture Notes in Computer Science, no. 2601, 2003, pp. 131–144.

[2] "QBone Scavenger Service (QBSS)," Internet2 QBone Initiative, http://qbone.internet2.edu/qbss/.

[3] "Analysis of Less-than-Best-Effort Services," TF-NGN LBE working group, http://www.cnaf.infn.it/~ferrari/tfngn/lbe/.

[4] R. Bless, K. Nichols, and K. Wehrle, "A Lower Effort Per-Domain Behavior for Differentiated Services," Internet Draft draft-bless-diffserv-pdb-le-01.txt, work in progress, Nov. 2002.

[5] D. Clark and W. Fang, "Explicit Allocation of Best-Effort Packet Delivery Service," *IEEE/ACM Transactions on Networking*, vol. 6, no. 4, pp. 362–373, Aug. 1998.

[6] Z. Wang, *Internet QoS — Architectures and Mechanisms for Quality of Service*. Morgan Kaufman Publishers, 2001.

[7] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Case," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 344–357, June 1993.

[8] C. Courcoubetis and R. Weber, *Pricing Communication Networks—Economics, Technology and Modelling*. Wiley, 2003.

[9] M. Mandjes, "Pricing Strategies under Heterogeneous Service Requirements," in *Proceedings of IEEE INFOCOM*, 2003.

[10] S. Shenker, "Fundamental Design Issues for the Future Internet," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1176–1188, Sept. 1995.

[11] S. Borst, O. Boxma, and R. Núñez Queija, "Heavy Tails: The Effect of the Service Discipline," in *Proceedings of Performance TOOLS 2002*, ser. Lecture Notes in Computer Science, no. 2324. Springer, Apr. 2002, pp. 1–30.

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

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

[14] L. Kleinrock, *Queuing Systems: Theory*. J. Wiley & Sons, 1975, vol. I.

[15] D. Ros and B. Tuffin, "A Mathematical Model of the Paris Metro Pricing Scheme for Charging Packet Networks," INRIA/IRISA, Tech. Rep. 1512, 2003, http://www.irisa.fr/bibli/publi/pi/2003/1512/1512.html, submitted.

[16] R. Gibbens, R. Mason, and R. Steinberg, "Internet service classes under competition," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 12, pp. 2490–2498, 2000.

[17] W. Stevens, *TCP/IP Illustrated, vol. 1: The Protocols*. Addison-Wesley, 1994.