Perturbation Analysis of a Variable M/M/1 Queue: A Probabilistic Approach

Nelson Antunes, Christine Fricker, Fabrice Guillemin, Philippe Robert

PERTURBATION ANALYSIS OF A VARIABLE M/M/1 QUEUE: A PROBABILISTIC APPROACH

NELSON ANTUNES, CHRISTINE FRICKER, FABRICE GUILLEMIN, AND PHILIPPE ROBERT

Abstract. Motivated by the problem of the coexistence on transmission links of telecommunication networks of elastic and unresponsive traffic, we study in this paper the impact on the busy period of an $M/M/1$ queue of a small perturbation in the server rate. The perturbation depends upon an independent stationary process $(X(t))$ and is quantified by means of a parameter $\varepsilon \ll 1$. We specifically compute the two first terms of the power series expansion in ε of the mean value of the busy period duration. This allows us to study the validity of the Reduced Service Rate (RSR) approximation, which consists in comparing the perturbed $M/M/1$ queue with the $M/M/1$ queue where the service rate is constant and equal to the mean value of the perturbation. For the first term of the expansion, the two systems are equivalent. For the second term, the situation is more complex and it is shown that the correlations of the environment process $(X(t))$ play a key role.

Contents

1. Introduction

We consider in this paper an $M/M/1$ queue with a time varying server rate. We specifically assume that the server rate depends upon a random environment represented by means of a process $(X(t))$, taking values in some (discrete or continuous) state space and assumed to be stationary. The study of this queueing system is motivated by the following engineering problem: Consider a transmission link of a telecommunication network carrying elastic traffic, able to adapt to the congestion level of the network, and a small proportion of traffic, which is unresponsive to congestion. The problem addressed in this paper is to derive quantitative results for estimating the influence of unresponsive traffic on elastic traffic.

In real implementations, elastic traffic is controlled by the so-called transmission control protocol (TCP), which has been designed in order to achieve a fair bandwidth allocation among sufficiently long flows at bottleneck links. If we assume that the link under consideration is the bottleneck, say, the access link to

Date: January 6, 2006.

Key words and phrases. Perturbation Analysis. Expansion of Cycle Formulas. $M/M/1$ queues.

the network, then it is reasonable to assume that bandwidth is distributed among the different competing elastic flows according to the processor sharing discipline (see for instance Massoulié and Roberts [10] and Delcoigne *et al.* [6]). Unresponsive traffic is then composed of small data transfers, which are too short to adapt to the congestion level of the network. Throughout the paper, it will be assumed that long flows arrive according to a Poisson process.

With the above modeling assumptions, unresponsive traffic appears for elastic flows as a small perturbation of the available bandwidth. In addition, when there is no unresponsive traffic, owing to the insensitivity property satisfied by the $M/G/1$ processor sharing queue, the number of long flows is identical to the number of customers in an $M/M/1$ queue. Hence, in order to obtain a global system able to describe the behavior of long flows in the presence of unresponsive traffic, we study an $M/M/1$ queue with a time varying server rate, which depends upon unresponsive traffic (for instance the number of small flows and their bit rate). The problem is then to estimate the impact of unresponsive traffic on the performance of the system. A classical issue is in particular to investigate the validity of the so-called reduced service rate (RSR) approximation, which states that everything happens as if the server rate for long flows were reduced by the mean load of unresponsive traffic. RSR approximation results (also called reduced load equivalence) have been shown to hold in a large number of queueing systems where some distributions are heavy tailed see Agrawal *et al.* [1], Jelenković and Momčilović [9] for example.

It is worth noting that queueing systems with time varying server rate have been studied in the literature in many different situations. In Núñez-Queija and Boxma [13], the authors consider a queueing system where priority is given to some flows driven by Markov Modulated Poisson Processes (MMPP) with finite state spaces and the low priority flows share the remaining server capacity according to the processor sharing discipline. By assuming that arrivals are Poisson and service times are exponentially distributed, the authors solve the system by means of matrix analysis methods. Similar models have been investigated in Núñez-Queija [11, 12] by still using the quasi-birth and death process associated with the system and a matrix analysis. In this setting, the characteristics of the queue at equilibrium are expressed in terms of the spectral quantities of some matrices leading to potential numerical applications. More recently, priority queueing systems with fast dynamics, which can be described by means of quasi-birth and death processes, have been studied via a perturbation analysis of a Markov chain by Altman *et al.* [2]. Boxma and Kurkova [4] studies the tail distributions of an $M/M/1$ queue with two service rates.

Getting qualitative results for queueing systems with variable service rates to study, for example, the impact of the variability of the service rate on the performances of the system is rather difficult. At the intuitive level, it is quite well known that the variability deteriorates them but, rigorously speaking, only few results are available. The main objective of this paper is to get some insight on these phenomena by considering a slightly perturbed system. As it will be seen, deriving such an expansion is already quite technical.

In this paper, it is assumed that the server rate of the $M/M/1$ queue is equal at time t to $\mu + \varepsilon p(X(t))$ for some function p, where $(X(t))$ is the process describing the environment affecting the service rate. In Fricker *et al.* [7], it has been assumed that the process $(X(t))$ is a diffusion process and that $p(x) = -x$. In this paper,

the perturbation function p is quite general and the environment process $(X(t))$ is only assumed to be stationary and Markovian. Moreover, we are specifically interested in the power series expansion of mean busy period duration in ε, which quantifies the magnitude of the perturbation. As far as the first order is concerned, the RSR approximation is valid: The time-varying server queue is identical to an equivalent $M/M/1$ queue with a fixed service rate equal to the average service rate $\mu + \varepsilon\mathbb{E}[p(X(0))]$. Combining the observation with the results obtained in Antunes *et al.* [3], one can easily conclude, via a simple regenerative argument, that the RSR holds for the mean number of customers in the queue. The analysis of the second order is much more intricate; the correlations of the process $(X(t))$ play a key role and, consequently, the RSR approximation is no more valid.

The organization of this paper is as follows: The model is described in Section 2. The first order term in the power series expansion of the mean busy period duration is computed in Section 3. The second order term is derived in Section 4. Applications of the results are discussed in Section 5. Some basic elements of the $M/M/1$ queue are recalled in Appendix.

2. Model

2.1. **Notation and Assumptions.** Throughout the paper $L(t)$ denotes the number of customers at time t in an $M/M/1$ queue with arrival rate λ and service rate μ. The variable B denotes the duration of a busy period starting with one customer: Given $L(0) = 1$,

$$B = \inf\{s \geq 0 : L(s) = 0\}.$$

It is assumed that the stability condition $\lambda < \mu$ holds. The invariant distribution of $(L(t))$ is geometrically distributed with parameter $\rho = \lambda/\mu$. For $x \geq 1$, the variable B_x denotes the duration of a busy period starting with x customers. By definition, $B_1 \stackrel{\text{dist.}}{=} B$. By convention, in the following when the variables B, B_1 and B_1' are used in the same expression, they are assumed to be independent with the same distribution as B. This queue will be referred to as the standard queue denoted, for short, by S-Queue.

For $\xi \geq 0$, $\mathcal{N}_\xi$ denotes a Poisson process with intensity ξ and for $0 \leq a < b$, $\mathcal{N}_\xi([a, b])$ denotes the number of points of this point process in the interval $[a, b]$. In particular, $\mathcal{N}_\lambda$ will represent the arrival process and $\mathcal{N}_\mu$ the process of the services of the S-Queue. The Poisson processes $\mathcal{N}_\lambda$ and $\mathcal{N}_\mu$ will be assumed to be independent one of each other and independent of the modulating Markov process $(X(t))$. The process $(L(t))$ can be represented as the solution of the stochastic differential equation

$$\begin{aligned} dL(t) \stackrel{\text{def.}}{=} L(t) - L(t-) &= \mathcal{N}_\lambda([t, t+dt]) - \mathbb{1}_{\{L(t-)>0\}}\mathcal{N}\mu([t, t+dt]) \\ &= d\mathcal{N}_\lambda(t) - \mathbb{1}_{\{L(t-)>0\}} d\mathcal{N}_\mu(t), \end{aligned} \tag{1}$$

where $L(t-)$ is the left limit of $L(s)$ at $s \nearrow t$. For the representation of queueing Markov processes as solutions of stochastic differential equations, see Robert [14].

The perturbed queue. In the following, we consider an $M/M/1$ queue with a service rate varying in time as a function of some process $(X(t))$ taking values in some space, denoted by $\mathcal{S}$. We assume that the process $(X(t))$ is an ergodic Markov process on $\mathcal{S}$. Typically, the state space of the environment $\mathcal{S}$ is a finite or countable set when $(X(t))$ is a Markov Modulated Poisson Process (MMPP) or

$\mathcal{S} = \mathbb{R}$ in the case of a diffusion, for instance an Ornstein-Uhlenbeck process (see Fricker *et al.* [7]). The invariant measure of the process $(X(t))$ is denoted by ν. The Markovian notation $\mathbb{E}_x(\cdot)$ will refer only to the initial state x of the Markov process $(X(t))$, therefore $\mathbb{E}_\nu(\cdot)$ will denote the expected value when the process $(X(t))$ is at equilibrium.

The variable $\widetilde{L}^\varepsilon(t)$ denotes the number of customers at time t in the $M/M/1$ queue with time-varying service rate. The process $(\widetilde{L}^\varepsilon(t), X(t))$ is a Markov process. The transitions of the process $(\widetilde{L}^\varepsilon(t))$ are given by: If $\widetilde{L}^\varepsilon(t) = l$ and $X(t) = x$ at time t,

$$l \to \begin{cases} l+1 & \text{at rate } \lambda \\ l-1 & '' \qquad (\mu + \varepsilon p(x))\mathbb{1}_{\{l>0\}} \end{cases}$$

for some function $p(x)$ on the state space of the environment $\mathcal{S}$ and some small parameter $\varepsilon \geq 0$. When $p(x) > 0$, this implies that there is an additional capacity of service when compared to the S-Queue. On the contrary, when $p(x) < 0$, the server is with a slower rate than in the S-Queue. The quantity $p^+(a)$ (respectively $p^-(a)$) is defined as $\max(p(a), 0)$ (respectively $\max(0, -p(a))$). At time $t \geq 0$, the additional capacity is therefore $\varepsilon p^+(X(t))$ and $-\varepsilon p^-(X(t))$ is the lost capacity. The perturbation considered in this paper is regular, see Altman *et al.* [2].

The variable $\widetilde{B}^\varepsilon$ is the duration of a busy period starting with one customer, that is, given $\widetilde{L}^\varepsilon(0) = 1$,

$$\widetilde{B}^\varepsilon = \inf\{s \geq 0 : \widetilde{L}^\varepsilon(s) = 0\}.$$

For $x \geq 1$, the variable $\widetilde{B}^\varepsilon_x$ denotes the duration of a busy period starting with x customers ($\widetilde{B}^\varepsilon_1 \overset{\text{dist.}}{=} \widetilde{B}^\varepsilon$). In the rest of this paper, we make the two following assumptions:

(H_1) the function $|p(x)|$ is bounded by a constant $M > 0$

(H_2) $\varepsilon \sup\{|p(x)| : x \in \mathcal{S}\} < \mu$.

The following proposition establishes that the length of the busy cycle is indeed integrable. The rest of the paper is devoted to the expansion of its expected value with respect to ε.

Proposition 1. *Under the condition $\lambda < \mu$, there exist some constants K and $\varepsilon_0 > 0$ such that for any $\varepsilon < \varepsilon_0$ and $n \geq 1$,*

$$\sup_{x \in \mathcal{S}} \mathbb{E}\left(\widetilde{B}^\varepsilon_n \mid X(0) = x\right) \leq Kn.$$

Proof. If one chooses ε_0 so that

$$\mu_0 \overset{\text{def.}}{=} \mu - \varepsilon_0 \inf\{p^-(x) : x \in \mathcal{S}\} > \lambda,$$

then clearly the number of customers of the P-Queue is certainly smaller than the number of customers of an $M/M/1$ queue with arrival rate λ and service rate μ_0. Consequently, the corresponding busy periods compare in the same way, hence it is enough to take $K = 1/(\mu_0 - \lambda)$. □

The queue with time-varying service rate as defined above will be referred to as the perturbed queue, denoted, for short, by P-Queue. The case $\varepsilon = 0$ obviously corresponds to the S-Queue.

2.2. **Adding and Canceling Departures.** The basic idea of the perturbation analysis carried out in this paper is to construct a coupling of the busy periods of the processes $(L(t))$ and $(\widetilde{L}^{\varepsilon}(t))$. This is done as follows, provided that for both queues the arrival process is $\mathcal{N}_\lambda$.

Additional departures. We denote by $\mathcal{N}^+$ the non-homogeneous Poisson process whose intensity is given by $t \to \varepsilon p^+(X(t))$. Conditionally on $(X(t))$, the number of points of $\mathcal{N}^+$ in the interval $[a, b]$, $0 \le a \le b$ is Poisson with parameter

$$\varepsilon \int_a^b p^+(X(s))\, ds.$$

The points of $\mathcal{N}^+$ are denoted by $0 < t_1^+ \le t_2^+ \le \ldots \le t_n^+ \le \ldots$ and are called additional departures. In particular the distribution of the location t_1^+ of the first point of $\mathcal{N}^+$ after 0 is given by, for $x \ge 0$,

$$\mathbb{P}(t_1^+ \ge x) = \mathbb{P}(\mathcal{N}^+([0,x]) = 0) = \mathbb{E}\left(\exp\left(-\varepsilon \int_0^x p^+(X(s))\, ds\right)\right). \tag{2}$$

See Grandell [8] for an account on non-homogeneous Poisson processes, referred to as doubly stochastic Poisson processes.

Canceling Departures. We denote by $\mathcal{N}^-$ the point process obtained by *thinning* the point process $\mathcal{N}_\mu$ (see Robert [14]). It is defined as follows: A point at $s > 0$ of the Poisson process $\mathcal{N}_\mu$ is a point of $\mathcal{N}^-$ with probability $\varepsilon p^-(X(s))/\mu$. In this way, $\mathcal{N}^-$ is a stationary point process with intensity $\varepsilon p^-(X(s))$. A point of $\mathcal{N}^-$ is called a canceled departure. The points of the point process $\mathcal{N}^-$ are denoted by $0 < t_1^- \le t_2^- \le \ldots \le t_n^- \le \ldots$. For $x > 0$, by definition,

$$\mathbb{P}(t_1^- \ge x) = \mathbb{E}\left(\prod_{i=1}^{\mathcal{N}_\mu([0,x])} \left(1 - \frac{\varepsilon p^-(X(s_i))}{\mu}\right)\right), \tag{3}$$

where (s_i) are the points of the point process $\mathcal{N}_\mu$.

With the above notation, it is not difficult to show that the Markov process $(\widetilde{L}^{\varepsilon}(t))$ has the same distribution as the solution of the stochastic differential equation

$$d\widetilde{L}^{\varepsilon}(t) = d\mathcal{N}_\lambda(t) - \mathbb{1}_{\{\widetilde{L}^{\varepsilon}(t-)>0\}} d\left(\mathcal{N}_\mu + \mathcal{N}^+ - \mathcal{N}^-\right)(t), \tag{4}$$

which is the analogue of Equation (1) for the P-Queue.

3. Busy period analysis: First order term

Let us assume that a busy period with one customer starts at time 0 in the S-Queue and P-Queue. In this section, we determine the first term of the power series expansion in ε of the expected value of $\widetilde{B}^{\varepsilon}$, the duration of the busy period in the P-Queue. This derivation allows us in addition to lay down part of the material needed in the next section to compute the more intricate second term of the power series expansion in ε.

For the first order term, we only have to consider the cases when there is either a single additional departure or else a single canceled departure. The probability that both events occur in the same busy period is clearly of the order of magnitude of ε^2 since the intensities of the associated Poisson processes are proportional to ε.

For $x \geq 1$, the stability assumptions ensure that the expected values of the busy periods starting with x customers, namely $\mathbb{E}(B_x)$ and $\mathbb{E}(\widetilde{B}^\varepsilon_x)$, are both finite. When the first additional and canceled departures are such that $t_1^+ > \widetilde{B}^\varepsilon$ and $t_1^- > \widetilde{B}^\varepsilon$ then $B = \widetilde{B}^\varepsilon$. We now consider the different possibilities.

A single additional departure. If there is only one additional departure and no canceled departure in $(0, \widetilde{B}^\varepsilon)$ then at time $\widetilde{B}^\varepsilon$, the P-queue is empty and the S-queue is with one customer (see Figure 1).

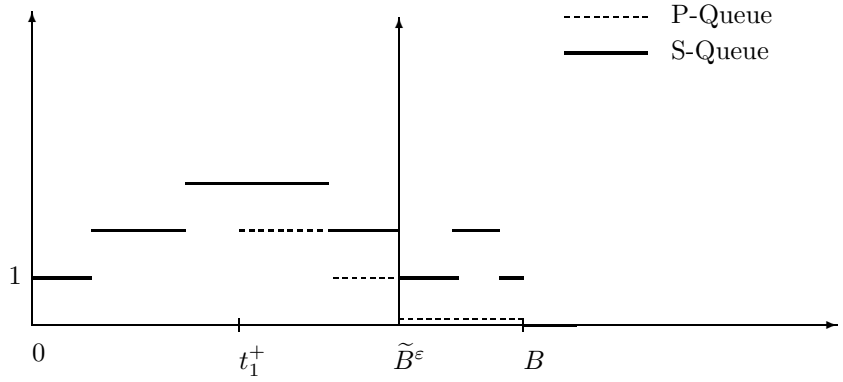

FIGURE 1. A busy Period with an Additional Departure

We specifically prove the following lemma.

Lemma 2. *In the case of a single departure, we have*

$$\mathbb{E}\left((B - \widetilde{B}^\varepsilon)\mathbb{1}_{\{t_1^+ < B\}}\right) = \varepsilon \frac{\mathbb{E}_\nu[p(X(0))^+]}{(\mu - \lambda)^2} + o(\varepsilon), \tag{5}$$

where ν is the equilibrium distribution of the environment $(X(t))$.

Proof. When there is only one additional departure, the variable $\widetilde{B}^\varepsilon$ is between t_1^+ and t_2^+. We can write

$$\mathbb{E}\left((B - \widetilde{B}^\varepsilon)\mathbb{1}_{\{t_1^+ < B\}}\right) = \mathbb{E}\left((B - \widetilde{B}^\varepsilon)\mathbb{1}_{\{t_1^+ < \widetilde{B}^\varepsilon < t_2^+, t_1^- > \widetilde{B}^\varepsilon\}}\right) + \Delta, \tag{6}$$

where the offset term Δ can be bounded as follows

$$\Delta \leq \mathbb{E}\left(|B - \widetilde{B}^\varepsilon| \left(\mathbb{1}_{\{t_2^+ < \widetilde{B}^\varepsilon, t_1^- > \widetilde{B}^\varepsilon\}} + \mathbb{1}_{\{t_1^- \leq \widetilde{B}^\varepsilon, t_1^+ \leq \widetilde{B}^\varepsilon\}}\right)\right). \tag{7}$$

Let us estimate the first term of the right-hand side of (6). Equation (2) and the boundedness of p give that

$$\begin{aligned}
\mathbb{P}(t_1^+ \leq B) &= 1 - \mathbb{E}\left(\exp\left(-\varepsilon \int_0^B p^+(X(s))\, ds\right)\right) \\
&= \varepsilon \mathbb{E}\left(\int_0^B p^+(X(s))\, ds\right) + o(\varepsilon) \\
&= \varepsilon \mathbb{E}(B)\, \mathbb{E}_\nu\left[p^+(X(0))\right] + o(\varepsilon) = \frac{\varepsilon}{\mu - \lambda} \mathbb{E}_\nu\left[p^+(X(0))\right] + o(\varepsilon),
\end{aligned}$$

by independence between B and $(X(t))$ and the stationarity of $(X(t))$. By the strong Markov property at the stopping time $\widetilde{B}^\varepsilon$, conditionally on the event $\{t_1^+ <$

$\widetilde{B}^\varepsilon < t_2^+, \widetilde{B}^\varepsilon < t_1^-\}$, the S-Queue starts at $\widetilde{B}^\varepsilon$ an independent busy period with one customer, therefore

$$\mathbb{E}\left((B-\widetilde{B}^\varepsilon)\mathbb{1}_{\{t_1^+<\widetilde{B}^\varepsilon<t_2^+,t_1^->\widetilde{B}^\varepsilon\}}\right) = \mathbb{P}(t_1^+ < \widetilde{B}^\varepsilon < t_2^+, t_1^- > \widetilde{B}^\varepsilon)$$
$$\times\, \mathbb{E}\left((B-\widetilde{B}^\varepsilon) \mid t_1^+ < \widetilde{B}^\varepsilon < t_2^+, t_1^- > \widetilde{B}^\varepsilon\right) = \mathbb{P}(t_1^+ < \widetilde{B}^\varepsilon < t_2^+, t_1^- > \widetilde{B}^\varepsilon)\mathbb{E}(B_1).$$

Now, since $\{t_1^+ < \widetilde{B}^\varepsilon\} = \{t_1^+ < B\}$ on the event $\{t_1^+ < \widetilde{B}^\varepsilon < t_2^+, \widetilde{B}^\varepsilon < t_1^-\}$, then

$$\begin{aligned}\mathbb{P}(t_1^+ < \widetilde{B}^\varepsilon < t_2^+, t_1^- > \widetilde{B}^\varepsilon) &= \mathbb{P}(t_1^+ < B) - \mathbb{P}(t_1^+ < \widetilde{B}^\varepsilon,\, t_2^+ < \widetilde{B}^\varepsilon) - \\ &\quad \mathbb{P}(t_1^+ < \widetilde{B}^\varepsilon,\, t_1^- < \widetilde{B}^\varepsilon) + \mathbb{P}(t_1^+ < \widetilde{B}^\varepsilon,\, t_2^+ < \widetilde{B}^\varepsilon,\, t_1^+ < \widetilde{B}^\varepsilon) \\ &= \mathbb{P}(t_1^+ < B) + o(\varepsilon),\end{aligned}$$

since two or more extra jumps in the same busy period is $o(\varepsilon)$. Similarly, by using again the strong Markov property, one gets the following estimation

$$\begin{aligned}\mathbb{E}\left(|B-\widetilde{B}^\varepsilon|\mathbb{1}_{\{t_2^+<\widetilde{B}^\varepsilon,t_1^->\widetilde{B}^\varepsilon\}}\right) &\le \sum_{n\ge 2}\mathbb{E}(B_n)\mathbb{P}(t_n^+ \le \widetilde{B}^\varepsilon \le t_{n+1}^+, t_1^- \ge \widetilde{B}^\varepsilon) \\ &\le \frac{1}{(\mu-\lambda)}\sum_{n\ge 2} n\mathbb{P}(\mathcal{N}^+([0,B]) = n).\end{aligned}$$

Indeed, given the S-Queue, $\mathcal{N}^+([0,B])$ has a Poisson distribution with parameter $\int_0^B \varepsilon p^+(X(s))\,ds$, which implies that

$$\sum_{n\ge 2} n\mathbb{P}(\mathcal{N}^+([0,B]) = n) =$$
$$\mathbb{E}\left(\int_0^B \varepsilon p^+(X(s))\,ds\right) - \mathbb{E}\left(\int_0^B \varepsilon p^+(X(u))\,du\, e^{-\varepsilon\int_0^B p^+(X(s))\,ds}\right) = o(\varepsilon)$$

and the first term in the right hand side of Inequality (7) is thus negligible at the first order in ε.

To estimate the second term in the right hand side of Inequality (7), we need to consider the different possibilities for the location of the points t_1^+ and t_1^-. In the case that t_1^+ and t_1^- occur during $[0,B]$ and $\widetilde{B}^\varepsilon \ge B$, at time B the P-Queue has at most $p \ge 0$ customers if there have been $p+1$ canceled departures. If $\mathcal{D}([0,B])$ is the number of customers during the busy period of the S-Queue, then certainly

$$\begin{aligned}&\mathbb{E}\left((\widetilde{B}^\varepsilon - B)\ \mathbb{1}_{\{\widetilde{B}^\varepsilon\ge B, t_1^-\le B, t_1^+\le B\}}\right) \\ &\qquad \le \mathbb{E}\left(\mathbb{E}_{X(B)}\left(B_{\mathcal{D}([0,B])}\right)\right)\mathbb{P}(t_1^- < B, t_1^+ \le B \le t_2^+) \\ &\qquad \le K\mathbb{E}\left(\mathcal{D}([0,B])\right)\mathbb{P}(t_1^- < B, t_1^+ \le B \le t_2^+) = o(\varepsilon),\end{aligned}$$

by Proposition 1. On the other hand,

$$\mathbb{E}\left(\left|\widetilde{B}^\varepsilon - B\right|\ \mathbb{1}_{\{\widetilde{B}^\varepsilon<B, t_1^-\le B, t_1^+\le B\}}\right) \le \mathbb{E}\left(B\mathbb{1}_{\{t_1^-\le B, t_1^+\le B\}}\right) = o(\varepsilon).$$

Finally,

$$\begin{aligned}&\mathbb{E}\left(\left|\widetilde{B}^\varepsilon - B\right|\mathbb{1}_{\{t_1^-\le\widetilde{B}^\varepsilon, t_1^+\le\widetilde{B}^\varepsilon\}}\right) \\ &\qquad \le \mathbb{E}\left(\left|\widetilde{B}^\varepsilon - B\right|\mathbb{1}_{\{t_1^-\le B, t_1^+\le B\}}\right) + \mathbb{E}\left(B\mathbb{1}_{\{t_1^-\le B, B\le t_1^+\le\widetilde{B}^\varepsilon\}}\right),\end{aligned}$$

where it can be shown in a similar way as before that the last term is $o(\varepsilon)$. One concludes that the term Δ is $o(\varepsilon)$ as ε goes to 0. By using Equation (6), we obtain the desired result. □

The estimation of the right hand side of Equation (6) may appear quite cumbersome. It is however worth noting that the environment $(X(t))$ of the P-Queue introduces delicate dependences, which have to be handled with care. This is why we have chosen to explicitly write the precise setting in which the strong Markov property is used to get the first order term. In the following, similar arguments will not be explicitly formulated.

A single canceled departure. Suppose now that there is only one canceled departure, i.e. a departure of the S-Queue is canceled for the P-Queue, and no additional jumps during the busy period of the S-Queue. In this case, at the end of the busy period of the S-Queue, at time B, the P-queue has one customer and thus starts a busy period. Provided that there are no more canceled and additional departures during $(B, \widetilde{B}^\varepsilon)$ in the P-Queue then the difference between both busy periods has the same distribution as the length B_1 of a standard busy period. See Figure 2.

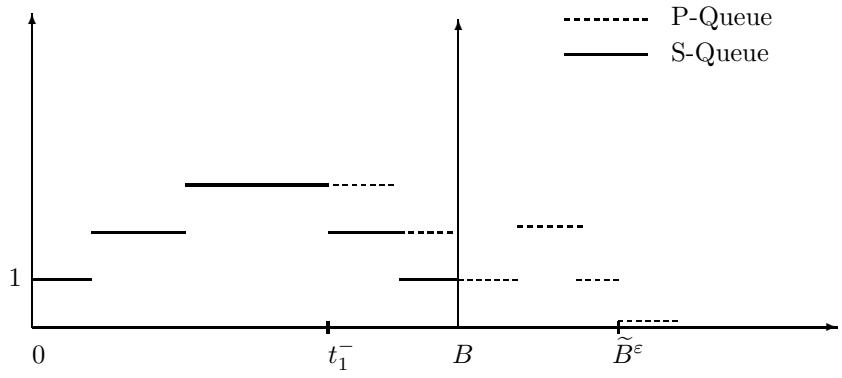

FIGURE 2. A Busy Period with a Canceled Departure

Lemma 3. *In the case of a single canceled departure, we have*

$$\mathbb{E}\left((\widetilde{B}^\varepsilon - B)\mathbb{1}_{\{t_1^- \le B\}}\right) = \varepsilon \frac{\mathbb{E}_\nu[p^-(X(0))]}{(\mu-\lambda)^2} + o(\varepsilon). \tag{8}$$

Proof. By using the same arguments as before, one obtains the relation

$$\begin{aligned}\mathbb{E}\left((\widetilde{B}^\varepsilon - B)\mathbb{1}_{\{t_1^- \le B\}}\right) &= \mathbb{E}\left(B_1 \mathbb{1}_{\{t_1^- \le B,\, B+B_1 < \min(t_1^+, t_2^-)\}}\right) + o(\varepsilon) \\ &= \mathbb{E}(B_1)\mathbb{P}(t_1^- \le B) + o(\varepsilon).\end{aligned}$$

To estimate $\mathbb{P}(t_1^- \le B)$, denote by (D_i) the sequence of departures times in the S-Queue and N the number of customers served during the busy period of length

B, then Equation (3) gives the identity

$$\begin{aligned}\mathbb{P}(t_1^- \leq B) &= \mathbb{E}\left(\sum_{i=1}^{N} \frac{\varepsilon p^-(X(D_i))}{\mu} \prod_{j=1}^{i-1}\left(1 - \frac{\varepsilon p^-(X(D_j))}{\mu}\right)\right) \\ &= \frac{\varepsilon}{\mu}\mathbb{E}\left(\sum_{i=1}^{N} p^-(X(D_i))\right) + o(\varepsilon) \\ &= \frac{\varepsilon}{\mu}\mathbb{E}(N)\mathbb{E}\left(p^-(X(D_1))\right) + o(\varepsilon)\end{aligned}$$

by stationarity of $(X(t))$ and Wald's Formula. Since $\mathbb{E}(N) = \mu/(\mu - \lambda)$ (see Appendix), Equation (8) follows. □

In the expansion of the busy period of the P-Queue, the term in ε is given by the two events consisting in only one canceled or only one additional departure during the busy period of the S-queue. The next proposition follows from Equations (5) and (8).

Proposition 4 (First Order Expansion)**.**

$$\mathbb{E}(\widetilde{B}^{\varepsilon}) = \frac{1}{\mu - \lambda} - \varepsilon\frac{\mathbb{E}_{\nu}[p(X(0))]}{(\mu - \lambda)^2} + o(\varepsilon). \tag{9}$$

Equation (9) is consistent with the so-called Reduced Service Rate approximation. As a matter of fact, everything happens as if we had a classical $M/M/1$ queue with service rate $\mu + \varepsilon\mathbb{E}_{\nu}[p(X(0))]$ and arrival rate λ. In that queue, the mean length of the busy period is given by

$$\frac{1}{\mu + \varepsilon\mathbb{E}_{\nu}[p(X(0))] - \lambda} = \frac{1}{\mu - \lambda} - \varepsilon\frac{\mathbb{E}_{\nu}[p(X(0))]}{(\mu - \lambda)^2} + o(\varepsilon),$$

which coincides with Equation (9). In the following section, we investigate the second order term and show that the RSR approximation is no more valid.

4. Busy Period: Second order term

In this section, the coefficient of ε^2 of the mean busy period $\mathbb{E}(\widetilde{B}^{\varepsilon})$ is calculated. In the same way as for the first order, this coefficient is related to the event that two extra jumps occur during a busy period of the perturbed $M/M/1$ queue. Since extra jumps can be either additional departures or canceled departures, there are three cases to investigate. As it will be seen, this coefficient stresses the importance of the evolution of the varying capacity, in particular through its correlation function. This was not the case for the first order term, since only the average value of the capacity shows up there.

In the following, in order to get the ε^2 coefficient, one has to consider the different possibilities for the location of the points t_1^+, t_2^+ and t_1^-, t_2^-. By using similar arguments as in Section 3, it is not difficult to show that any event involving t_3^+ or t_3^- yields a term of the order ε^3 in the expansion of $\mathbb{E}(\widetilde{B}^{\varepsilon} - B)$.

Define

$$\mathcal{A}_+ = \{t_1^+ \leq B, t_1^- \geq t_1^+ + B_{L(t_1^+)-1}\}.$$

On this event, at least one departure is added and the busy period of the P-Queue finishes before a departure is canceled (note that $B_{L(t_1^+)-1}$ is the length of a busy

period of S-Queue starting at time t_1 with $L(t_1^+)-1$ customers). On the event

$$\mathcal{A}_{\pm} = \{t_1^- \leq B,\, B \leq t_1^+ \leq B + B_1\},$$

a canceled departure occurs and another departure is added before the completion of the busy period of the P-queue, where B_1 denotes the duration of the additional busy period due to the canceled departure. Finally, on the event

$$\mathcal{A}_{-} = \{t_1^- \leq B,\, B + B_1 \leq t_1^+\},$$

at least a canceled departure occurs and no additional departures are added before the completion of the busy period B_1.

By checking all the different cases, it is not difficult to see that if $\mathcal{A} = \mathcal{A}_+ \cup \mathcal{A}_{\pm} \cup \mathcal{A}_-$, the expression $\mathbb{E}((\widetilde{B}^\varepsilon - B)\mathbb{1}_{\mathcal{A}^c})$ is $o(\varepsilon^2)$ (and even equal to 0 in some cases, for instance when there are a canceled departure and an additional departure in such a way that $\widetilde{B}^\varepsilon = B$). The following sections are devoted to the estimation of $\mathbb{E}((\widetilde{B}^\varepsilon - B)\mathbb{1}_A)$ for $A \in \{\mathcal{A}_+, \mathcal{A}_{\pm}, \mathcal{A}_-\}$.

In a first step, we analyze the case when there are only additional departures before B, that is, we consider the term $\mathbb{E}((\widetilde{B}^\varepsilon - B)\mathbb{1}_{\mathcal{A}_+})$. When no canceled departure occurs, at most two additional departures in the time interval $[0, B]$, occurring at times t_1^+ and t_2^+ respectively, may play a role in the computation of the coefficient of ε^2 of $\mathbb{E}(B - \widetilde{B}^\varepsilon)$. In this case, the difference between $B - \widetilde{B}^\varepsilon$ is equal to the busy period of an S-Queue which starts with either one or two customers, depending on the fact that, on the event $\{t_1^+ \leq B\}$, the busy period of the P-Queue is already completed at time t_2^+ or not. See Figure 3.

As before, B_2 denotes a random variable with the same distribution as the sum of two independent variables distributed as B_1 and independent of B, t_1^+ and t_2^+. One gets

$$\begin{aligned}
\mathbb{E}\left((B - \widetilde{B}^\varepsilon)\mathbb{1}_{\mathcal{A}_+}\right) &= \mathbb{E}\left((B - \widetilde{B}^\varepsilon)\mathbb{1}_{\{t_1^+ \leq B, t_1^- \geq t_1^+ + B_{L(t_1^+)-1}\}}\right) \\
&= \mathbb{E}(B_2)\,\mathbb{P}\left(t_1^+ < B, t_2^+ < t_1^+ + B_{L(t_1^+)-1}, t_1^- \geq t_1^+ + B_{L(t_1^+)-1}\right) \\
&+ \mathbb{E}(B_1)\,\mathbb{P}\left(t_1^+ < B, t_2^+ \geq t_1^+ + B_{L(t_1^+)-1}, t_1^- \geq t_1^+ + B_{L(t_1^+)-1}\right) + o(\varepsilon^2).
\end{aligned}$$

This decomposition entails that

$$\begin{aligned}
\mathbb{E}\left((B - \widetilde{B}^\varepsilon)\mathbb{1}_{\mathcal{A}_+}\right) &= (\mathbb{E}(B_2) - \mathbb{E}(B_1))\,\mathbb{P}\left(t_1^+ < B, t_2^+ < t_1^+ + B_{L(t_1^+)-1}\right) \\
&+ \mathbb{E}(B_1)\left(\mathbb{P}\left(t_1^+ < B\right) - \mathbb{P}\left(t_1^+ < B, t_1^- \leq t_1^+ + B_{L(t_1^+)-1}\right)\right) + o(\varepsilon^2).
\end{aligned} \tag{10}$$

From Equation (10), one has to expand three expressions with respect to ε. This is done by proving the three following lemmas.

Lemma 5. *The quantity* $\mathbb{P}\left(t_1^+ < B, t_2^+ < t_1^+ + B_{L(t_1^+)-1}\right)$ *can be expanded as*

$$\begin{aligned}
&\mathbb{P}\left(t_1^+ < B, t_2^+ < t_1^+ + B_{L(t_1^+)-1}\right) \\
&\qquad = \rho\varepsilon^2 E\left(\int_0^B (B - v)\mathbb{E}_\nu\left(p^+(X(0))p^+(X(v))\right)\right) dv + o(\varepsilon^2).
\end{aligned} \tag{11}$$

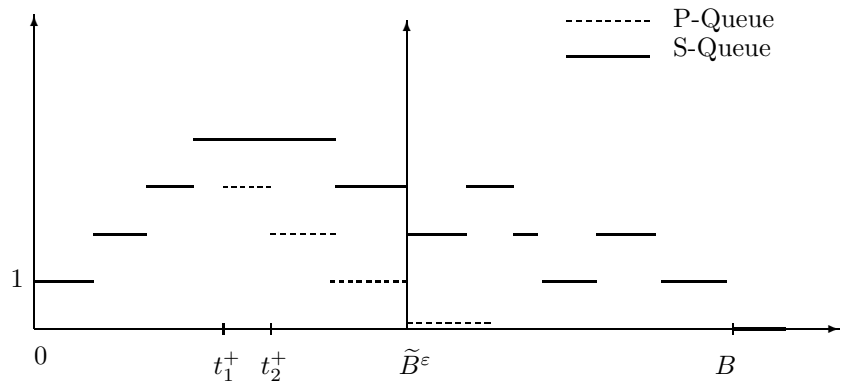

FIGURE 3. Two Additional Departures

Proof. Let us recall the regenerative description of a busy period starting at time 0 with one customer: At time E_1 (exponentially distributed with parameter $\lambda+\mu$), with probability $\mu/(\lambda+\mu)$ the busy period is finished. Otherwise, with probability $\lambda/(\lambda+\mu)$, a new customer arrives and a sub-busy period of duration B_1^1 (with the same distribution as B_1) begins until the number of customers reaches 1 again. In this way, the variable B can be represented as follows

$$B = E_0 + \sum_{i=1}^{H} \left(E_i + B_1^i\right), \tag{12}$$

where H is geometrically distributed with parameter $\lambda/(\lambda+\mu)$, (E_i) are i.i.d exponentially distributed with parameter $\lambda+\mu$ and (B_1^i) are i.i.d. All these random variables are independent. For $0 \le i \le H$,

- s_i denotes the end of the ith sub-busy cycle: $s_0 = 0$ and, for $i \ge 1$, $s_i = s_{i-1} + E_i + B_1^i$, $B = s_H + E_0$;
- N_i denotes the number of arrivals during the ith sub-busy cycle;
- $s_{i-1}+D_1^i, \ldots, s_{i-1}+D_{N_i}^i$ are the instants of departures of customers during the ith sub-busy cycle.

For the joint distribution of the vector $(N_i, D_1^i, \ldots, D_{N_i}^i)$, see the Appendix. Figure 4 gives an illustration of the above definitions.

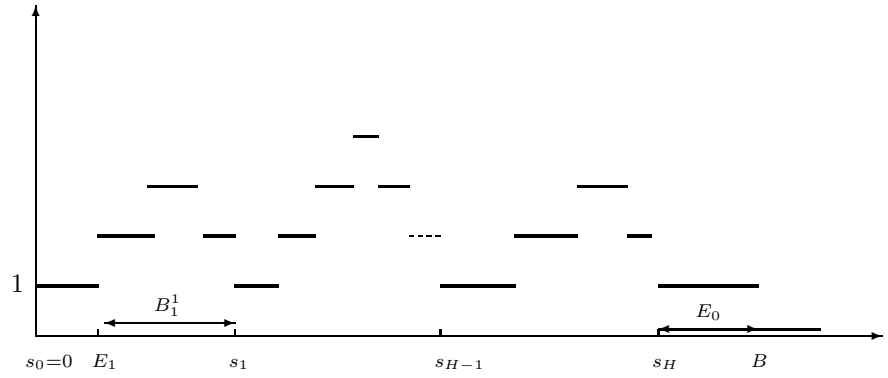

FIGURE 4. Decomposition of a Busy Period

It is easy to see that for the event $\{t_1^+ \leq B\,, t_2^+ < t_1^+ + B_{L(t_1^+)-1}\}$ to occur, t_1^+ and t_2^+ have to be in the same sub-busy period, $[s_{i-1}+E_i, s_i]$, for some $i \in \{1,\ldots,H\}$. For a fixed i, the probability that the first two additional jumps are in the ith sub-busy period, is

$$\mathbb{E}\left(\int_{s_{i-1}+E_i}^{s_i} \varepsilon p^+(X(u)) e^{-\varepsilon\int_0^u p^+(X(s))\,ds}\left(1-e^{-\varepsilon\int_u^{s_i} p^+(X(s))\,ds}\right)\,du\right)$$
$$= \varepsilon^2\mathbb{E}\left(\int_{s_{i-1}+E_i}^{s_i} p^+(X(u))\int_u^{s_i} p^+(X(s))\,ds\,du\right) + o(\varepsilon^2).$$

Since, $B_1^i = s_i - s_{i-1} - E_{i-1}$ has the same distribution as B and by the stationarity of $((X(t))$, the coefficient of ε^2 can be expressed as follows,

$$\mathbb{E}\left(\int_{0\leq u\leq v\leq B} p^+(X(u))p^+(X(v))\,du\,dv\right)$$
$$= \mathbb{E}\left(\int_{0\leq u\leq v\leq B} \mathbb{E}_\nu\left[p^+(X(0))p^+(X(v-u))\right]\,du\,dv\right).$$

Finally, since H is geometrically distributed with parameter $\lambda/(\lambda+\mu)$, Equation (11) follows. □

We turn now to the expansion of the quantity $\mathbb{P}(t_1^+ \leq B)$, which is of course a refinement of what has been done in Section 3.

Lemma 6. *The quantity $\mathbb{P}(t_1^+ \leq B)$ can be expanded as*

$$\mathbb{P}(t_1^+ \leq B) = \varepsilon\frac{\mathbb{E}_\nu\left[p^+(X(0))\right]}{\mu-\lambda} - \varepsilon^2\mathbb{E}\left(\int_0^B (B-v)\,\mathbb{E}_\nu\left(p^+(X(0))p^+(X(v))\right)\,dv\right) + o(\varepsilon^2). \tag{13}$$

Proof. We clearly have

$$\mathbb{P}(t_1^+ \leq B) = \mathbb{E}\left(1-e^{-\varepsilon\int_0^B p^+(X(s))\,ds}\right)$$
$$= \varepsilon\frac{\mathbb{E}_\nu\left[p^+(X(0))\right]}{\mu-\lambda} - \frac{\varepsilon^2}{2}\mathbb{E}\left(\left(\int_0^B p^+(X(s))\,ds\right)^2\right) + o(\varepsilon^2).$$

The second moment of the integral can be expressed as follows, by symmetry,

$$\mathbb{E}\left(\left(\int_0^B p^+(X(s))\,ds\right)^2\right) = 2\mathbb{E}\left(\int_{0\leq u\leq v\leq B} p^+(X(u))p^+(X(v))\,du\,dv\right)$$
$$= 2\mathbb{E}\left(\int_{0\leq u\leq v\leq B} \mathbb{E}_\nu\left(p^+(X(0))p^+(X(v-u))\right)\,du\,dv\right),$$

by stationarity of the process $(X(t))$ and Equation (13) follows. □

Finally, we examine the expansion of $\mathbb{P}(t_1^+ < B, t_1^- \leq t_1^+ + B_{L(t_1^+)-1})$. This term is more delicate to expand, because of the canceled departure.

Lemma 7. *The quantity* $\mathbb{P}(t_1^+ < B, t_1^- \leq t_1^+ + B_{L(t_1^+)-1})$ *can be expanded as*

$$\begin{aligned}(14)\quad &\mathbb{P}(t_1^+ < B, t_1^- \leq t_1^+ + B_{L(t_1^+)-1}) \\ &\qquad = \frac{\varepsilon^2}{\mu}\mathbb{E}\left(\sum_{i=1}^{H}\sum_{j=1}^{N_i}\int_0^{A_i} p^+(X(u))p^-(X(D_j^i))\,du\right) + o(\varepsilon^2),\end{aligned}$$

where H *is geometric distributed with parameter* $\lambda/(\mu+\lambda)$, $(N_i, D_1^i, \dots D_N^i)$ *denotes the number of departures and the departures times in a busy period of length* B^i, *and*

$$A_i = B_1^i + E_0 + \sum_{k=i+1}^{H}(E_k + B_1^k),$$

where (E_i) *are i.i.d is exponentially distributed with parameter* $\mu+\lambda$ *and* (B_1^i) *are i.i.d with the same distribution as* B.

Proof. Using the regenerative description of a standard busy period introduced in the proof of Lemma 5, the variable t_1^- has to occur in some sub-busy period $[s_{i-1}+E_i, s_i]$ of B for some $1 \leq i \leq H$. A little thought show that if $t_1^- \in [s_{i-1}+E_i, s_i]$ then t_1^+ has to be in $[s_{i-1}+E_i, B]$ for the event $\{t_1^+ < B, t_1^- \leq t_1^+ + B_{L(t_1^+)-1}\}$ to occur. The probability that t_1^- and t_1^+ are located in $[s_{i-1}+E_i, s_i]$ and $[s_{i-1}+E_i, B]$, respectively, is

$$\begin{aligned}\mathbb{E}\Bigg(&\int_{s_{i-1}+E_i}^{B} \varepsilon p^+(X(u))e^{-\varepsilon\int_0^u p^+(X(s))\,ds}\,du \sum_{j=1}^{N_i}\varepsilon\frac{p^-(X(s_{i-1}+D_j^i))}{\mu} \\ &\prod_{k=1}^{j-1}\left(1-\varepsilon\frac{p^-(X(s_{i-1}+D_k^i))}{\mu}\right)\prod_{l=1}^{i-1}\prod_{r=1}^{N_l}\left(1-\varepsilon\frac{p^-(X(s_{l-1}+D_r^l))}{\mu}\right)\Bigg),\end{aligned}$$

where the coefficient of ε^2 is

$$\frac{1}{\mu}\mathbb{E}\left(\sum_{j=1}^{N_i}\int_{s_{i-1}+E_i}^{B} p^+(X(u))p^-(X(s_{i-1}+D_j^i))\,du\right).$$

Considering the different sub-cycles during B and by the stationarity of $(X(t))$, Equation (14) follows. □

We are now able to compute the coefficient of ε^2 in the power series expansion of $\mathbb{E}((\widetilde{B}^\varepsilon - B)\mathbb{1}_{\mathcal{A}_+})$ in ε.

Proposition 8. *The coefficient of* ε^2 *in the expansion of* $\mathbb{E}((B-\widetilde{B}^\varepsilon)\mathbb{1}_{\mathcal{A}_+})$ *with respect to* $\varepsilon > 0$ *is given by*

$$\begin{aligned}(15)\quad a_+ = &-\frac{1}{\mu}\mathbb{E}\left(\int_0^B (B-v)\,\mathbb{E}_\nu\left(p^+(X(0))p^+(X(v))\right)\,dv\right) \\ &-\frac{1}{\mu^2(1-\rho)}\mathbb{E}\left(\sum_{i=1}^{H}\sum_{j=1}^{N_i}\int_0^{A_i} p^+(X(u))p^-(X(D_j))\,du\right).\end{aligned}$$

To complete the analysis, we now turn to the expansion of $\mathbb{E}((\widetilde{B}^\varepsilon - B)\mathbb{1}_{\mathcal{A}_\pm})$ and $\mathbb{E}((\widetilde{B}^\varepsilon - B)\mathbb{1}_{\mathcal{A}_-})$. In the calculations, it appears more convenient to consider the sum of both terms and we then have the following result.

Proposition 9. *The coefficient of ε^2 in the expansion of $\mathbb{E}((\widetilde{B}^\varepsilon - B)\mathbb{1}_{\mathcal{A}_\pm \cup \mathcal{A}_-})$ with respect to $\varepsilon > 0$ is given by*

$$a_- = \frac{1}{\mu^2(1-\rho)} \left(-\mathbb{E}\left(\sum_{i=1}^{N} \int_0^{B+B_1} p^-(X(D_i)) p^+(X(s))\, ds \right) + \frac{1}{\mu} \mathbb{E}\left(\sum_{i=1}^{N} \sum_{k=1}^{N'} p^-(X(0)) p^-(X(B - D_i + D'_k)) \right) \right), \tag{16}$$

where $(N, D_1, \ldots, D_N)$ and $(N', D'_1, \ldots, D'_{N'})$ denote the number of departures and the departure times in the busy periods of length B and B_1, respectively.

Proof. When a single canceled departure occurs (at time t_1^-) before B, an additional busy period of length B_1 has to be added to take into account the canceled departure.

By the strong Markov property, with the same method as in Section 3, one obtains the relation

$$\mathbb{E}\left((B + B_1 - \widetilde{B}^\varepsilon) \mathbb{1}_{\{t_1^- \le B,\, B \le t_1^+ \le B + B_1\}} \right) = \mathbb{E}\,(B'_1)\, P\left(t_1^- \le B,\, B \le t_1^+ \le B + B_1\right) + o(\varepsilon^2),$$

where the random variable B'_1 has the same distribution as the random variable B_1, hence,

$$\begin{aligned} \mathbb{E}\left((\widetilde{B}^\varepsilon - B) \mathbb{1}_{\mathcal{A}_\pm} \right) &= \mathbb{E}\left((\widetilde{B}^\varepsilon - B) \mathbb{1}_{\{t_1^- \le B,\, B \le t_1^+ \le B + B_1\}} \right) \\ &= \mathbb{E}\left(B_1 \mathbb{1}_{\{t_1^- \le B,\, B \le t_1^+ \le B + B_1\}} \right) - \mathbb{E}\,(B'_1)\, \mathbb{P}\left(t_1^- \le B,\, B \le t_1^+ \le B + B_1\right) + o(\varepsilon^2). \end{aligned} \tag{17}$$

Now, two canceled departures in the same busy period gives two additional independent busy periods starting with one customer,

$$\begin{aligned} \mathbb{E}\left((\widetilde{B}^\varepsilon - B) \mathbb{1}_{\mathcal{A}_-} \right) &= \mathbb{E}\left((\widetilde{B}^\varepsilon - B) \mathbb{1}_{\{t_1^- \le B,\, B + B_1 \le t_1^+\}} \right) \\ &= \mathbb{E}\left(B_1 \mathbb{1}_{\{t_1^- \le B,\, B + B_1 \le \min(t_1^+, t_2^-)\}} \right) \\ &\quad + \mathbb{E}\left((B_1 + B'_1)\, \mathbb{1}_{\{t_1^- \le B,\, B \le t_2^- \le B + B_1,\, B + B_1 + B'_1 \le t_1^+\}} \right) \\ &\quad + \mathbb{E}\left(B_2 \mathbb{1}_{\{t_1^- \le B,\, t_2^- \le B,\, B + B_1 + B'_1 \le t_1^+\}} \right) + o(\varepsilon^2). \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}\left((\widetilde{B}^\varepsilon - B) \mathbb{1}_{\mathcal{A}-} \right) &= \mathbb{E}\left(B_1 \mathbb{1}_{\{t_1^- \le B,\, t_2^- > B + B_1\}} \right) - \mathbb{E}\left(B_1 \mathbb{1}_{\{t_1^- \le B,\, t_1^+ \le B + B_1\}} \right) \\ &\quad + \mathbb{E}\left(B_1 \mathbb{1}_{\{t_1^- \le B,\, B \le t_2^- \le B + B_1\}} \right) + \mathbb{E}\,(B'_1)\, \mathbb{E}\left(\mathbb{1}_{\{t_1^- \le B,\, B \le t_2^- \le B + B_1\}} \right) \\ &\quad + \mathbb{E}\left(B_2 \mathbb{1}_{\{t_2^- \le B\}} \right) + o(\varepsilon^2). \end{aligned}$$

Finally,

$$\begin{aligned}(18)\quad \mathbb{E}\left((\widetilde{B}^{\varepsilon}-B)\mathbb{1}_{\mathcal{A}_-}\right) &= \mathbb{E}\left(B_1\right)\mathbb{P}\left(t_1^- \le B, t_2^- > B\right) - \mathbb{E}\left(B_1\mathbb{1}_{\{t_1^- \le B, t_1^+ \le B+B_1\}}\right)\\ &+ \mathbb{E}\left(B_1'\right)\mathbb{P}\left(t_1^- \le B,\, B \le t_2^- \le B+B_1\right) + 2\mathbb{E}\left(B_1\right)\mathbb{P}\left(t_2^- \le B\right) + o(\varepsilon^2),\end{aligned}$$

From Section 2, it is not difficult to see that the expression

$$\mathbb{P}\left(t_1^- \le B, t_2^- > B\right) + 2\mathbb{P}\left(t_2^- \le B\right)$$

has no term in ε^2 in its power series expansion. Thus the first term and the last term of the right hand side of Equation (18) cancel out for the expansion.

The following expansions are obtained in a similar way,

$$\begin{aligned}&\mathbb{E}\left(B_1\mathbb{1}_{\{t_1^- \le B,\, t_1^+ \le B+B_1\}}\right)\\ &\qquad = \frac{\varepsilon^2}{\mu}\mathbb{E}\left(B_1\sum_{i=1}^{N}\int_0^{B+B_1} p^-(X(D_i))p^+(X(s))\,ds\right) + o(\varepsilon^2),\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}\left(t_1^- \le B, B < t_2^- \le B+B_1\right)\\ &\qquad = \frac{\varepsilon^2}{\mu^2}\mathbb{E}\left(\sum_{i=1}^{N}\sum_{k=1}^{N'} p^-(X(0))p^-(X(B-D_i+D_k'))\right) + o(\varepsilon^2),\end{aligned}$$

where $(N, D_1, \ldots, D_N)$ and $(N', D_1', \ldots, D_{N'}')$ denote the number of departures and the departures times in two independent busy periods of lengths B and B_1, respectively.

If we sum up the expansions obtained for canceled departures and one canceled and one additional departures (Equations (17) and (18)), with standard manipulations, one gets the second term of the expansion $\mathbb{E}((\widetilde{B}^{\varepsilon}-B)(\mathbb{1}_{\mathcal{A}_\pm}+\mathbb{1}_{\mathcal{A}_-}))$ in ε. □

To summarize the results obtained in this section, we can state the following theorem.

Theorem 10. *The coefficient of ε^2 is the power series expansion of $\mathbb{E}(\widetilde{B}^{\varepsilon}-B)$ in ε is equal to $a_- - a_+$, where the coefficients a_+ and a_- are given by Equations* (15) *and* (16)*, respectively.*

It should be noted that the distributions involved in Equations (15) and (16) can be explicited by using the classical results concerning the $M/M/1$ queue. See the Appendix where they are recalled. In the next section, we examine some applications of the above result.

5. Applications

5.1. **Non-Negative Perturbation Functions.** Equations (11) and (13) give that the expansion

$$\mathbb{E}\left(B-\widetilde{B}^{\varepsilon}\right) = \delta_1\varepsilon + \delta_2\varepsilon^2 + o(\varepsilon^2)$$

holds, with $\delta_1 = \mathbb{E}_\nu\left(p(X(0))\right)/(\mu-\lambda)^2$ and

$$\delta_2 = -\frac{1}{\mu}\mathbb{E}\left(\int_0^B (B-v)\,\mathbb{E}_\nu\left(p(X(0))p(X(v))\right)\,dv\right).$$

Denote by $C_p(u) = \mathbb{E}_\nu\left[p(X(0))p(X(u))\right] - \mathbb{E}_\nu\left[p(X(0))\right]^2$, the covariance of the extra capacity. The second term of the expansion can be expressed as

$$\delta_2 = -\frac{1}{\mu}\mathbb{E}\left(\int_0^B (B-v)C_p(v)\,dv\right) - \frac{\mathbb{E}_\nu\left[p(X(0))\right]^2}{(\mu-\lambda)^3},$$

hence,

$$\mathbb{E}\left(B-\widetilde{B}^\varepsilon\right) = \varepsilon\,\frac{\mathbb{E}_\nu\left[p(X(0))\right]}{(\mu-\lambda)^2} - \varepsilon^2\,\frac{\mathbb{E}_\nu\left[p(X(0))\right]^2}{(\mu-\lambda)^3} - \frac{\varepsilon^2}{\mu}\mathbb{E}\left(\int_0^B (B-v)C_p(v)\,dv\right) + o(\varepsilon^2).$$

The following proposition which readily follows, compares the length of the busy period of the P-Queue with an $M/M/1$ queue with service rate $\mu + \varepsilon\mathbb{E}_\nu(p(X(0)))$.

Proposition 11 (Comparison with reduced service rate)**.** *If $\widehat{B}$ is the length of a busy period of an $M/M/1$ queue with service rate $\mu + \varepsilon\mathbb{E}_\nu(p(X(0)))$ then*

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon^2}\mathbb{E}\left(\widehat{B}-\widetilde{B}^\varepsilon\right) = -\frac{1}{\mu}\mathbb{E}\left(\int_0^B (B-v)\,C_p(v)\,dv\right),$$

where, for $u \geq 0$,

$$C_p(u) = \mathbb{E}_\nu\left[p(X(0))p(X(u))\right] - \mathbb{E}_\nu\left[p(X(0))\right]^2$$

is, up to the factor ε^2, the covariance function of the extra-capacity of the perturbed queue.

It is straightforward to conclude from the expression in Proposition 11 that $\mathbb{E}(\widehat{B}-\widetilde{B}^\varepsilon)$ is negative when ε is small.

Corollary 12 (Negative impact of the variation of the service rate)**.** *When the environment is positively correlated i.e. when the function $u \to C_p(u)$ is non-negative, then the first term of the expansion of $\mathbb{E}(\widehat{B}-\widetilde{B}^\varepsilon)$ in ε is of order 2 and is negative.*

The following expression gives a closed form expression of the second term of the expansion when the environment has an exponential decay.

Proposition 13. *When the correlation function of the environment is exponentially decreasing, i.e. when, for some $\alpha > 0$,*

$$C_p(x) = \mathrm{Var}[p(X(0))]\,e^{-\alpha x}, \qquad x \geq 0,$$

then the difference between reduced and variable service rates satisfies the relation

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon^2}\,\mathbb{E}\left(\widehat{B}-\widetilde{B}^\varepsilon\right) \overset{def.}{=} \Delta_2(\alpha) = -\,\frac{\mathrm{Var}[p(X(0))]}{(\mu-\lambda)^3}\,\mathbb{E}\left(e^{-\alpha Z}\right) \leq 0, \tag{19}$$

where, Z is the random variable whose density function on $\mathbb{R}_+$ is given by

$$x \to \frac{1}{\mu(1-\rho)^2} \int_x^{+\infty} \mathbb{P}\left(B \geq u\right)\, du.$$

In particular, the function $\alpha \to \Delta_2(\alpha)$ is non-decreasing and concave.

Proof. For a square integrable random variable A on $\mathbb{R}^+$, A^* denotes the random variable with density $x \to \mathbb{P}(A \geq u)/\mathbb{E}(A)$ on $\mathbb{R}_+$. Note that for $\alpha \geq 0$,

$$\mathbb{E}\left(e^{-\alpha A^*}\right) = \frac{1 - \mathbb{E}(e^{-\alpha A})}{\alpha \mathbb{E}(A)} \tag{20}$$

and $\mathbb{E}(A^*) = \mathbb{E}(A^2)/(2\mathbb{E}(A))$.

To simplify notations, it is assumed that $\mathrm{Var}[p(X(0))] = 1$. Proposition 11 gives that the coefficient $\Delta_2(\alpha)$ of ε^2 is in this case

$$\begin{aligned}\Delta_2(\alpha) &= -\frac{1}{\mu}\mathbb{E}\left(\int_0^B (B-v)\, e^{-\alpha v}\, dv\right) = -\frac{1}{\mu}\mathbb{E}\left(\int_0^B v\, e^{-\alpha(B-v)}\, dv\right) \\ &= -\frac{1}{\mu}\mathbb{E}\left(\frac{B}{\alpha} - \frac{1}{\alpha^2} + \frac{e^{-\alpha B}}{\alpha^2}\right) = -\frac{\mathbb{E}(B)\mathbb{E}(B^*)}{\mu}\frac{1 - \mathbb{E}\left(e^{-\alpha B^*}\right)}{\alpha \mathbb{E}(B^*)}.\end{aligned}$$

The Proposition is proved by using Relation (20). □

5.2. **Non-Positive Perturbation Functions.** It is assumed in this section that the perturbation function is non-positive so that the environment uses a part of the capacity of the $M/M/1$ queue with constant service rate μ. This application is motivated by the following practical situation: Coming back to the coexistence of elastic and streaming traffic in the Internet, assume that priority is given to streaming traffic in a buffer of a router. The bandwidth available for non-priority traffic is the transmission link reduced by the bit rate of streaming traffic. Denoting by $\varepsilon d(X_t)$ the bit rate of streaming traffic at time t (for instance ε may represent the peak rate of a streaming flow and $d(X_t)$ the number of such flows active at time t), the service rate available for non-priority traffic is $\mu - \varepsilon d(x)$. Setting $p(x) = -d(x)$, the function $p(x)$ is non-positive. We are then in the framework when the environment gives a reduced bandwidth to a non-priority $M/M/1$ queue. The same notation as in the previous section is used extensively.

Equations (5) and (16) give that the expansion

$$\mathbb{E}\left(B - \widetilde{B}^{\varepsilon}\right) = \delta_1 \varepsilon + \delta_2 \varepsilon^2 + o(\varepsilon^2)$$

holds, with $\delta_1 = \mathbb{E}\left[p(X(0))\right]/(\mu - \lambda)^2$ and

$$\delta_2 = -\frac{1}{\mu^3(1-\rho)}\mathbb{E}\left(\sum_{i=1}^{N}\sum_{k=1}^{N'} p(X(0))p(X(B - D_i + D'_k))\right),$$

where, as in (16), $(N, D_1, \ldots, D_N)$ and $(N', D'_1, \ldots, D'_{N'})$ denote the number of departures and the departure times in the busy periods of length B and B_1, respectively. The terms δ_1 and δ_2 are non-positive. Thus, at the first order, the mean of $\widetilde{B}^{\varepsilon}$ is larger than the mean of B. The following proposition which readily follows, compares the length of the busy-period of the P-Queue with the mean of the length of the busy-period $\widehat{B}$ in an $M/M/1$ queue with service rate $\mu + \varepsilon \mathbb{E}_\nu[p(X(0))]$.

Proposition 14 (Comparison with reduced service rate)**.** *If $\widehat{B}$ is the length of a busy period of an $M/M/1$ queue with service rate $\mu + \varepsilon \mathbb{E}_\nu[p(X(0))]$ then*

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} \mathbb{E}\left(\widehat{B} - \widetilde{B}^\varepsilon\right) = -\frac{1}{\mu^3(1-\rho)} \mathbb{E}\left(\sum_{i=1}^{N} \sum_{k=1}^{N'} C_p\left(X(B - D_i + D'_k)\right)\right), \tag{21}$$

where, as in Equation (16)*, $(N, D_1, \ldots, D_N)$ and $(N', D'_1, \ldots, D'_{N'})$ denote the number of departures and the departure times in the busy periods of length B and B_1, respectively and for $u \geq 0$,*

$$C_p(u) = \mathbb{E}_\nu\left[p(X(0))p(X(u))\right] - \mathbb{E}_\nu\left[p(X(0))\right]^2$$

is, up to the factor ε^2, the covariance function of the capacity of the perturbed queue.

This result implies that, as for a non-negative perturbation function, the variation of the service rate has a negative impact on the performance of the system. The following result holds.

Proposition 15 (Negative impact of the variation of the service rate)**.** *When the environment is positively correlated (when the function $u \to C_p(u)$ is non-negative), then the first term of the expansion of $\mathbb{E}\left(\widehat{B} - \widetilde{B}^\varepsilon\right)$ in ε is of order 2 and negative.*

Comparing to the case of a non-negative perturbation function, if the correlation function of the environment is exponentially decreasing, a simple close expression for the right hand side member of Equation (21) seems to be difficult to obtain, though the same qualitative results hold.

Proposition 16 (Exponential decay)**.** *When the correlation function of the environment is exponentially decreasing, i.e. when*

$$C_p(x) = \mathrm{Var}[p(X(0))]\, e^{-\alpha x}, \qquad x \geq 0,$$

and some $\alpha > 0$, the function

$$\alpha \to \lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} \mathbb{E}\left(\widehat{B} - \widetilde{B}^\varepsilon\right)$$

is non-positive, non-decreasing and concave. Moreover when α tends to infinity, this quantity converges to zero.

5.3. **Fast Environments.** A general perturbation function p is considered together with some stationary Markov process $(X(t))$ with invariant probability distribution ν. It is assumed that it verifies a mixing condition such as

$$\lim_{t \to +\infty} |\mathbb{E}_\nu[f(X(0))g(X(t))] - \mathbb{E}_\nu[f(X(0))]\mathbb{E}_\nu[g(X(0))]| = 0, \tag{22}$$

for any Borelian bounded functions f and g on the state space $\mathcal{S}$. Note that this condition is not restrictive in general since it is true for any ergodic Markov process with a countable (or finite) state space or for any ergodic diffusion on $\mathbb{R}^d$, $d \geq 1$.

In this section, the environment is accelerated by a factor $\alpha > 0$, described by the process $(X(\alpha t))$. The behavior when α goes to infinity is investigated. Note that when α goes to 0, the environment is frozen: the service rate remains constant and equal to $\mu + \varepsilon p(X(0))$. Such a situation has also been analyzed by Delcoigne *et al.* [6] through stochastic bounds.

At the intuitive level, when α gets large, for t and $h > 0$ the total service capacity available during t and $t+h$ is given by

$$\mu h + \varepsilon \int_t^{t+h} p(X(\alpha u))\, du \overset{\text{dist.}}{=} \mu h + \varepsilon \frac{1}{\alpha} \int_0^{\alpha h} p(X(u))\, du \sim (\mu + \varepsilon \mathbb{E}_\nu(p(X(0))))h$$

by the ergodic Theorem. Thus, speeding up the environment averages the capacity of the variable queue. This intuitive picture is rigorously established in the following proposition.

Proposition 17. *When the environment is given by $(X(\alpha t))$ and Equation (22) holds then if $\delta_2(\alpha)$ is the ε^2 coefficient of the expansion of with respect to ε,*

$$\mathbb{E}(\widetilde{B}^\varepsilon - B) = \frac{\mathbb{E}_\nu\left(p(X(0))\right)}{(\mu-\lambda)^2}\,\varepsilon + \delta_2(\alpha)\,\varepsilon^2 + o(\varepsilon^2),$$

the following relation holds,

$$\lim_{\alpha \to +\infty} \delta_2(\alpha) = \frac{\mathbb{E}_\nu\left(p(X(0))\right)^2}{(\mu-\lambda)^3}.$$

Proof. The quantity $\delta_2(\alpha)$ is equal to $a_- - a_+$ where a_- and a_+ are given by Equations (16) and (15), respectively. We shall deal only with the first term of a_- in Equation (16). Let

$$F(\alpha) \overset{\text{def.}}{=} -\mathbb{E}\left(\sum_{i=1}^{N} \int_0^{B+B_1} p^-(X(\alpha D_i)) p^+(X(\alpha s))\, ds\right)$$

where N is the number of customers in the busy period of length B and their departure times are denoted by $(D_i, 1 \le i \le N)$. We have

$$F(\alpha) = -\mathbb{E}\left(\sum_{i=1}^{N} \int_0^{B+B_1} \mathbb{E}\left(p^-(X(\alpha D_i)) p^+(X(\alpha s)) \mid B, N\right)\, ds\right).$$

Relation (22) and the boundedness of p (Assumption (H_1)) show that, almost surely,

$$\lim_{\alpha \to +\infty} \mathbb{E}\left(p^-(X(\alpha D_i)) p^+(X(\alpha s)) \mid B, N\right) = \mathbb{E}_\nu\left(p^-(X(0))\right)\, \mathbb{E}_\nu\left(p^+(X(0))\right),$$

therefore Lebesgue's theorem gives

$$\begin{aligned}\lim_{\alpha \to +\infty} \frac{-F(\alpha)}{\mathbb{E}_\nu(p^-(X(0)))\,\mathbb{E}_\nu(p^+(X(0)))} &= \mathbb{E}\left(NB\right) + \mathbb{E}\left(B_1\right)\mathbb{E}\left(N\right) \\ &= \frac{1+\rho}{\mu(1-\rho)^3} + \frac{1}{\mu-\lambda}\frac{1}{1-\rho} = \frac{2}{\mu(1-\rho)^3},\end{aligned}$$

by using the expressions of $\mathbb{E}(N)$ and $\mathbb{E}(NB)$ in the Appendix. Similar calculations can be conducted for all the other terms to finally give the Proposition. □

6. Appendix: Some useful quantities for the $M/M/1$ queue

Let (A_k) (resp. (D_k)) denotes the arrival times (resp. departure times) in a busy period of an $M/M/1$ queue with arrival rate λ and service rate μ. A busy period denoted by B that starts at time 0 will last a time t and will consist of N services if, and only if,

(i) there are $(N-1)$ arrivals in $(0,t)$;
(ii) $D_N = t$;
(iii) $A_{k+1} \le D_k$, $k = 1, \ldots, N-1$.

If conditions (i) and (ii) are satisfied then $(A_2, \ldots, A_N)$ and $(D_1, \ldots, D_{N-1})$ are independent and represent the ordered values of two sets of $N-1$ uniform $(0,t)$ random variables. Hence,

$$b_n(t) = d\mathbb{P}(B < t, N = n)/dt = \frac{e^{-\lambda t}(\lambda t)^{(n-1)}}{(n-1)!} \frac{\mu e^{-\mu t}(\mu t)^{(n-1)}}{(n-1)!} \times \mathbb{P}(A_2 \leq D_1, \ldots, A_n < D_{n-1}).$$

The first two moments of the stationary busy period are given by

$$\mathbb{E}(B_1) = \frac{1}{\mu - \lambda}, \qquad \mathbb{E}(B_1^2) = \frac{2}{\mu^2(1-\rho)^3}.$$

Expression (2.40) p.190 of Cohen [5] shows that

$$\varphi(z,\xi) = \sum_{n=1}^{+\infty} z^n \int_0^{+\infty} e^{-\xi t} b_n(t)\, dt,$$

is given by

$$\varphi(z,\xi) = \frac{1}{2\rho}\left(1 + \rho + \mu^{-1}\xi - \sqrt{(1+\rho+\mu^{-1}\xi)^2 - 4\rho z}\right)$$

for $|z| \leq 1$, $\Re(\xi) \geq 0$. It is easy to derive

$$\mathbb{E}(N) = \int_0^{+\infty} dt \sum_{n=1}^{+\infty} n b_n(t) = \frac{1}{1-\rho},$$

$$\mathbb{E}(NB) = \int_0^{+\infty} t dt \sum_{n=1}^{+\infty} n b_n(t) = -\frac{d^2\varphi}{dz d\xi}(1,0) = \frac{1+\rho}{\mu(1-\rho)^3},$$

$$\mathbb{E}[N(N-1)] = \int_0^{+\infty} dt \sum_{n=1}^{+\infty} n(n-1) b_n(t) = \frac{d^2\varphi}{dz^2}(1,0) = \frac{2\mu^2\lambda}{(\mu-\lambda)^3}.$$

To conclude one has to compute $\mathbb{E}(D)$ where $D = D_1 + D_2 + \cdots + D_N$. By using the classical branching argument for the busy-period of the $M/M/1$ queue (see Robert [14] for example), one gets

$$D = \sigma + \sum_{i=1}^{N_\sigma} \left(\left(\sigma + \sum_{j=1}^{i-1} B_j \right) N_i + D_i \right),$$

where σ is the service time of the first customer of the busy-period, N_σ the number of arrivals in the interval $[0,\sigma]$, B_i the busy-period generated by the ith customer arrived during σ, N_i the number of customers in B_i, D_i the sum of the departure times of B_i from the beginning of this busy-period. Taking the expectation, it is easy to derive that

$$\mathbb{E}(D) = \mathbb{E}(\sigma) + \mathbb{E}(\sigma N_\sigma) + \mathbb{E}(B)\mathbb{E}(N_\sigma(N_\sigma - 1)/2)\mathbb{E}(N) + \mathbb{E}(\sigma N_\sigma)\mathbb{E}(D),$$

where N_σ has a geometric distribution with parameter $\lambda/(\lambda+\mu)$. Thus

$$\mathbb{E}(N_\sigma(N_\sigma - 1)) = 2\rho^2.$$

Simple algebra gives $\mathbb{E}(D) = \mu^2/(\mu-\lambda)^3$.

References

[1] Rajeev Agrawal, Armand M. Makowski, and Philippe Nain, *On a reduced load equivalence for fluid queues under subexponentiality*, Queueing Systems. Theory and Applications **33** (1999), no. 1-3, 5–41.

[2] E. Altman, K. Avrachenkov, and R. Núñez Queija, *Perturbation analysis for denumerable markov chains with application to queueing models*, Advances in Applied Probability **36** (2004), no. 3, 839–853.

[3] Nelson Antunes, Christine Fricker, Fabrice Guillemin, and Philippe Robert, *Integration of streaming services and tcp data transmission in the Internet*, Performance'05 (Juan les Pins), IFP WG 7.3, 2005.

[4] O. J. Boxma and I. A. Kurkova, *The $M/M/1$ queue in a heavy-tailed random environment*, Statistica Neerlandica. Journal of the Netherlands Society for Statistics and Operations Research **54** (2000), no. 2, 221–236.

[5] J. W. Cohen, *The single server queue*, 2nd ed., North-Holland, Amsterdam, 1982.

[6] F. Delcoigne, A. Proutière, and G. Régnié, *Modeling integration of streaming and data traffic*, ITC specialist seminar on IP traffic (Würzburg, Germany), July 2002.

[7] C. Fricker, F. Guillemin, and P. Robert, *Perturbation analysis of an $M/M/1$ queue in a diffusion random environment*, preprint, January 2004.

[8] J. Grandell, *Point processes and random measures*, Advances in Applied Probability **9** (1977), 502–526.

[9] Predag Jelenković and Petar Momčilović, *Resource sharing with subexponential distributions*, Infocom'2002 (New York), June 2002.

[10] L. Massoulié and J. Roberts, *Bandwidth sharing: Objectives and algorithms*, INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, 1999, pp. 1395–1403.

[11] R. Núñez-Queija, *Sojourn times in a processor sharing queue with service interruptions*, Queueing Systems **34** (2000), 351–386.

[12] ———, *Sojourn times in non-homogeneous QBD processes with processor sharing*, Stochastic Models (2001), 61–92.

[13] R. Núñez-Queija and O.J. Boxma, *Analysis of a multi-server queueing model of ABR*, J. Appl. Math. Stoch. An. **11** (1998), 339–354.

[14] Philippe Robert, *Stochastic networks and queues*, Stochastic Modelling and Applied Probability Series, vol. 52, Springer, New-York, June 2003.

(Nelson Antunes, Christine Fricker, Philippe Robert) INRIA-Rocquencourt, RAP project, Domaine de Voluceau, 78153 Le Chesnay, France

(Fabrice Guillemin) France Telecom R&D, CORE/CPN, 22300 Lannion, France

E-mail address: Nelson.Antunes@inria.fr

E-mail address: Christine.Fricker@inria.fr

E-mail address: Fabrice.Guillemin@francetelecom.com

E-mail address: Philippe.Robert@inria.fr