Adv. Appl. Prob. **38,** 263–283 (2006)
Printed in Northern Ireland

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

NELSON ANTUNES [*] [**] AND
CHRISTINE FRICKER,[*] [***] *INRIA*
FABRICE GUILLEMIN,[****] *France Télécom*
PHILIPPE ROBERT,[*] [*****] *INRIA*

Abstract

In this paper, motivated by the problem of the coexistence on transmission links of telecommunications networks of elastic and unresponsive traffic, we study the impact on the busy period of an M/M/1 queue of a small perturbation in the service 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 approximation, which consists in comparing the perturbed M/M/1 queue with the M/M/1 queue whose 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.

Keywords: Perturbation analysis; expansion of cycle formula; M/M/1 queue

2000 Mathematics Subject Classification: Primary 60K25
Secondary 60K30

1. Introduction

In this paper we consider an M/M/1 queue with a time-varying service rate. We specifically assume that the service 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 telecommunications network carrying elastic traffic, which is able to adapt to the congestion level of the network, and a small proportion of traffic that is unresponsive to congestion. The problem addressed in this paper is that of deriving 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, 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 (the access link to the network, say) then it is reasonable to assume that bandwidth is distributed among the different competing elastic flows according to the processor-sharing discipline (see, for

Received 16 December 2004; revision received 19 September 2005.
[*] Postal address: INRIA-Rocquencourt, RAP project, Domaine de Voluceau, 78153 Le Chesnay, France.
[**] Email address: nelson.antunes@inria.fr
[***] Email address: christine.fricker@inria.fr
[****] Postal address: France Télécom R&D, CORE/CPN, 22300 Lannion, France.
Email address: fabrice.guillemin@francetelecom.com
[*****] Email address: philippe.robert@inria.fr

instance, [10] and [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 of 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 service rate that 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. In particular, a classical issue is to investigate the validity of the so-called reduced service rate approximation, which states that everything happens as it would if the service rate for long flows were reduced by the mean load of unresponsive traffic. Reduced service rate approximation results (comprising so-called reduced load equivalence) have been shown to hold in a large number of queueing systems in which some distributions are heavy tailed; see [1] and [9], for example.

It is worth noting that queueing systems with time-varying service rates have been studied in the literature in many different situations. In [13] the authors considered a queueing system in which priority is given to some flows driven by Markov-modulated Poisson processes 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 solved the system by means of matrix analysis methods. Similar models have been investigated in [11] and [12] by using the quasi-birth–death process associated with the system, along with 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–death processes, have been studied via a perturbation analysis of a Markov chain, in [2]. Boxma and Kurkova [4] studied the tail distributions of an M/M/1 queue with two service rates.

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

In this paper it is assumed that, at time t, the service rate of the M/M/1 queue is equal to $\mu + \varepsilon p(X(t))$, for some function p, where $(X(t))$ is the process describing the environment affecting the service rate. In [7] it was 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 in ε, which quantifies the magnitude of the perturbation, of the mean busy period duration. As far as the first-order term is concerned, the reduced service rate approximation is valid: the time-varying service rate queue is identical to an equivalent M/M/1 queue with a fixed service rate equal to the average service rate $\mu + \varepsilon \,\mathrm{E}(p(X(0)))$. By combining this observation with the results obtained in [3], we can easily conclude, via a simple regenerative argument, that the reduced service rate approximation holds for the mean number of customers in the queue. The analysis of the second-order term is much more intricate; the

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

correlations of the process $(X(t))$ play a key role and, consequently, the reduced service rate approximation is no longer valid.

The organization of the 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 facts about the M/M/1 queue are recalled in Appendix A.

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 that $L(0) = 1$, $B = \inf\{s \geq 0 \colon L(s) = 0\}$. 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, $B_1 \stackrel{\mathrm{D}}{=} B$, where '$\stackrel{\mathrm{D}}{=}$' denotes equality in distribution. By convention, when the variables B, B_1, and B_1' are used in the same expression, they are assumed to be independent and B_1 and B_1' are assumed to have the same distribution as B. This queue will be referred to as the standard queue, or 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 of each other and independent of the modulating Markov process $(X(t))$. The process $(L(t))$ can be represented as the solution to the stochastic differential equation

$$\begin{aligned} \mathrm{d}L(t) := L(t) - L(t-) &= \mathcal{N}_\lambda([t, t + \mathrm{d}t]) - \mathbf{1}_{\{L(t-)>0\}}\, \mathcal{N}_\mu([t, t + \mathrm{d}t]) \\ &= \mathrm{d}\mathcal{N}_\lambda(t) - \mathbf{1}_{\{L(t-)>0\}}\, \mathrm{d}\mathcal{N}_\mu(t), \end{aligned} \tag{1}$$

where $L(t-)$ is the left limit of $L(s)$ as $s \nearrow t$ and $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function of the set $\{\cdot\}$. For the representation of queueing Markov processes as solutions of stochastic differential equations, see [14].

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

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

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

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 for service in comparison to

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

the s-queue. When $p(x) < 0$, the service rate is slower than in the s-queue. The quantities $p^+(a)$ and $p^-(a)$ are respectively defined as $\max(p(a), 0)$ and $\max(0, -p(a))$. 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 [2].

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

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

For $x \geq 1$, the variable $\tilde{B}^\varepsilon_x$ denotes the duration of a busy period starting with x customers ($\tilde{B}^\varepsilon_1 \stackrel{\mathrm{D}}{=} \tilde{B}^\varepsilon$). In the rest of the paper, we make the following two assumptions.

Assumption 1. *The function $|p(x)|$ is bounded by a constant $M > 0$.*

Assumption 2. $\varepsilon \sup\{|p(x)|,\ x \in \mathcal{S}\} < \mu$.

The queue with time-varying service rate, as just defined, will be referred to as the perturbed queue, or p-queue. The case $\varepsilon = 0$ obviously corresponds to the s-queue.

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 $n \geq 1$,*

$$\sup_{x \in \mathcal{S}} \mathrm{E}(\tilde{B}^\varepsilon_n \mid X(0) = x) \leq Kn.$$

Proof. If we choose ε_0 such that

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

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

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 $(\tilde{L}^\varepsilon(t))$. This is done as follows, provided that for both queues the arrival process is $\mathcal{N}_\lambda$.

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

$$\varepsilon \int_a^b p^+(X(s))\,\mathrm{d}s.$$

The points of $\mathcal{N}^+$ are denoted by $t_1^+, t_2^+, \ldots, t_n^+, \ldots$, with $0 < t_1^+ \leq t_2^+ \leq \cdots \leq t_n^+ \leq \cdots$, 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, for $x \geq 0$, by

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

See [8] for an account of inhomogeneous Poisson processes, also referred to as doubly stochastic Poisson processes.

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

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

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

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 $(\tilde{L}^\varepsilon(t))$ has the same distribution as the solution to the stochastic differential equation

$$\mathrm{d}\tilde{L}^\varepsilon(t) = \mathrm{d}\mathcal{N}_\lambda(t) - \mathbf{1}_{\{\tilde{L}^\varepsilon(t-)>0\}}\, \mathrm{d}(\mathcal{N}_\mu + \mathcal{N}^+ - \mathcal{N}^-)(t),$$

which is the analogue of (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 both the s-queue and the p-queue. In this section, we determine the first term of the power series expansion in ε of the expected value of $\tilde{B}^\varepsilon$, namely 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 in which there is either a single additional departure or 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 $x \ge 1$, the stability assumptions ensure that the expected values of the busy periods starting with x customers, namely $\mathrm{E}(B_x)$ and $\mathrm{E}(\tilde{B}_x^\varepsilon)$, are both finite. When the first additional and canceled departures are such that $t_1^+ > \tilde{B}^\varepsilon$ and $t_1^- > \tilde{B}^\varepsilon$, we have $B = \tilde{B}^\varepsilon$. We now consider the different possibilities.

3.1. A single additional departure

If there is only one additional departure and no canceled departure in $(0, \tilde{B}^\varepsilon)$, then at time $\tilde{B}^\varepsilon$ the p-queue is empty and the s-queue has one customer (see Figure 1).

Specifically, we prove the following lemma.

Lemma 1. *In the case of a single departure, we have*

$$\mathrm{E}((B - \tilde{B}^\varepsilon)\, \mathbf{1}_{\{t_1^+ < B\}}) = \varepsilon \frac{\mathrm{E}_\nu(p(X(0))^+)}{(\mu - \lambda)^2} + o(\varepsilon), \tag{4}$$

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

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

$$\mathrm{E}((B - \tilde{B}^\varepsilon)\, \mathbf{1}_{\{t_1^+ < B\}}) = \mathrm{E}((B - \tilde{B}^\varepsilon)\, \mathbf{1}_{\{t_1^+ < \tilde{B}^\varepsilon < t_2^+,\, t_1^- > \tilde{B}^\varepsilon\}}) + \Delta, \tag{5}$$

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

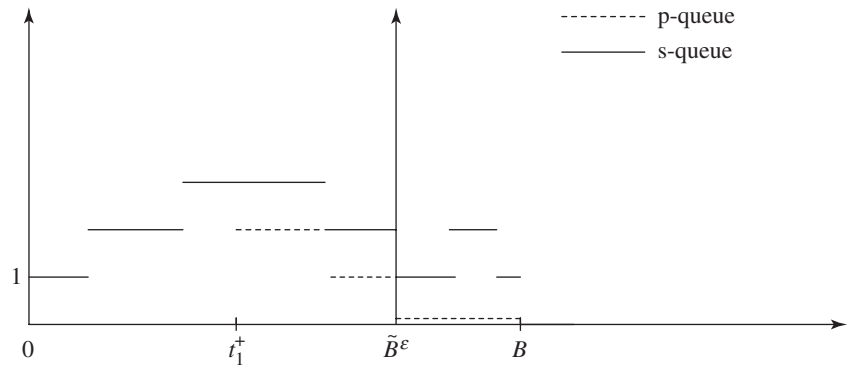

FIGURE 1: A busy period with an additional departure.

where the offset term Δ can be bounded as follows:

$$\Delta \le \mathrm{E}(|B - \tilde{B}^\varepsilon|(\mathbf{1}_{\{t_2^+ < \tilde{B}^\varepsilon,\, t_1^- > \tilde{B}^\varepsilon\}} + \mathbf{1}_{\{t_1^- \le \tilde{B}^\varepsilon,\, t_1^+ \le \tilde{B}^\varepsilon\}})). \tag{6}$$

Let us estimate the first term on the right-hand side of (5). From (2) and the boundedness of p, we have

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

by the independence of B and $(X(t))$ and the stationarity of $(X(t))$. By the strong Markov property at the stopping time $\tilde{B}^\varepsilon$, conditionally on the event $\{t_1^+ < \tilde{B}^\varepsilon < t_2^+,\ \tilde{B}^\varepsilon < t_1^-\}$, at $\tilde{B}^\varepsilon$ the s-queue starts an independent busy period with one customer. Therefore,

$$\begin{aligned}
\mathrm{E}((B - \tilde{B}^\varepsilon)\,\mathbf{1}_{\{t_1^+ < \tilde{B}^\varepsilon < t_2^+,\, t_1^- > \tilde{B}^\varepsilon\}}) &= \mathrm{P}(t_1^+ < \tilde{B}^\varepsilon < t_2^+,\ t_1^- > \tilde{B}^\varepsilon) \\
&\quad \times \mathrm{E}(B - \tilde{B}^\varepsilon \mid t_1^+ < \tilde{B}^\varepsilon < t_2^+,\ t_1^- > \tilde{B}^\varepsilon) \\
&= \mathrm{P}(t_1^+ < \tilde{B}^\varepsilon < t_2^+,\ t_1^- > \tilde{B}^\varepsilon)\,\mathrm{E}(B_1).
\end{aligned}$$

Now, since $\{t_1^+ < \tilde{B}^\varepsilon\} = \{t_1^+ < B\}$ on the event $\{t_1^+ < \tilde{B}^\varepsilon < t_2^+,\ \tilde{B}^\varepsilon < t_1^-\}$, we have

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

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

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

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

Indeed, conditionally on the state of the s-queue, $\mathcal{N}^+([0,B])$ has a Poisson distribution with parameter $\int_0^B \varepsilon p^+(X(s))\,\mathrm{d}s$, which implies that

$$\begin{aligned}\sum_{n\ge 2} n\,\mathrm{P}(\mathcal{N}^+([0,B])=n) &= \mathrm{E}\biggl(\int_0^B \varepsilon p^+(X(s))\,\mathrm{d}s\biggr)\\ &\quad - \mathrm{E}\biggl(\int_0^B \varepsilon p^+(X(u))\,\mathrm{d}u\exp\biggl(-\varepsilon\int_0^B p^+(X(s))\,\mathrm{d}s\biggr)\biggr)\\ &= o(\varepsilon).\end{aligned}$$

The first term on the right-hand side of (6) is thus negligible, to first order in ε.

To estimate the second term on the right-hand side of (6), we need to consider the different possibilities for the locations of the points t_1^+ and t_1^-. In the case that t_1^+ and t_1^- occur during $[0,B]$ and $\tilde{B}^{\varepsilon}\ge B$, at time B the p-queue has at most $p\ge 0$ customers (where $p+1$ is the number of canceled departures). If $\mathcal{D}([0,B])$ is the number of customers during the busy period of the s-queue, then clearly

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

by Proposition 1. However, we also have

$$\mathrm{E}(|\tilde{B}^{\varepsilon}-B|\,\mathbf{1}_{\{\tilde{B}^{\varepsilon}<B,\,t_1^-\le B,\,t_1^+\le B\}}) \le \mathrm{E}(B\,\mathbf{1}_{\{t_1^-\le B,\,t_1^+\le B\}}) = o(\varepsilon).$$

Finally,

$$\mathrm{E}(|\tilde{B}^{\varepsilon}-B|\,\mathbf{1}_{\{t_1^-\le\tilde{B}^{\varepsilon},\,t_1^+\le\tilde{B}^{\varepsilon}\}}) \le \mathrm{E}(|\tilde{B}^{\varepsilon}-B|\,\mathbf{1}_{\{t_1^-\le B,\,t_1^+\le B\}}) + \mathrm{E}(B\,\mathbf{1}_{\{t_1^-\le B,\ B\le t_1^+\le\tilde{B}^{\varepsilon}\}}),$$

where it can be shown, as above, that the last term is $o(\varepsilon)$. We conclude that the term Δ is $o(\varepsilon)$ as ε goes to 0. By using (5), we obtain the desired result.

The estimation of the right-hand side of (5) may appear quite cumbersome. However, it is worth noting that the environment $(X(t))$ of the p-queue introduces delicate dependencies, which have to be handled with care. This is why we have chosen to explicitly present the precise method by which the strong Markov property is used to obtain the first-order term. In the following, similar arguments will not be explicitly formulated.

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

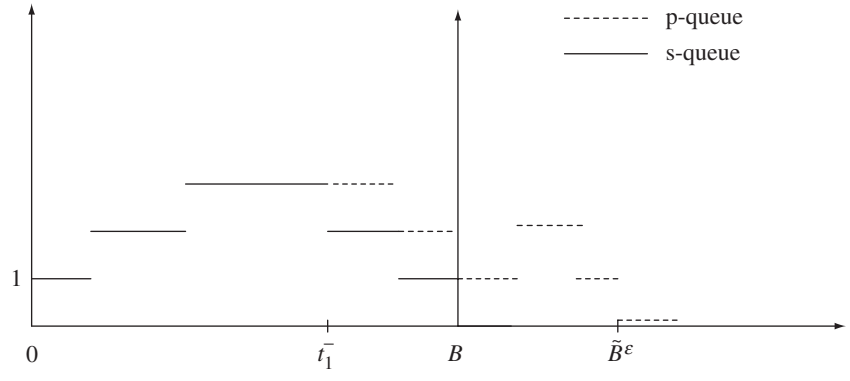

FIGURE 2: A busy period with a canceled departure.

3.2. A single canceled departure

Now suppose that there is only one canceled departure, i.e. a departure of the s-queue is canceled in the p-queue, and no additional jumps are made 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 or additional departures during $(B, \tilde{B}^\varepsilon)$ in the p-queue, the difference between the busy periods has the same distribution as the length B_1 of a standard busy period (see Figure 2).

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

$$\mathrm{E}((\tilde{B}^\varepsilon - B)\,\mathbf{1}_{\{t_1^- \le B\}}) = \varepsilon \frac{\mathrm{E}_\nu(p^-(X(0)))}{(\mu-\lambda)^2} + o(\varepsilon). \tag{7}$$

Proof. By using the same arguments as before, we obtain the relation

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

To estimate $\mathrm{P}(t_1^- \le B)$, denote by (D_i) the sequence of departures times in the s-queue and by N the number of customers served during the busy period of length B. Equation (3) then gives the identity

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

by stationarity of $(X(t))$ and Wald's formula. Since $\mathrm{E}(N) = \mu/(\mu-\lambda)$ (see Appendix A), (7) follows.

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

In the expansion of the busy period of the p-queue, the term in ε depends on either the event that there is only one canceled departure during the busy period of the s-queue, or the event that there is only one additional departure during the busy period of the s-queue. The next proposition follows from (4) and (7).

Proposition 2. (First-order expansion.) *We have*

$$\mathrm{E}(\tilde{B}^{\varepsilon}) = \frac{1}{\mu - \lambda} - \varepsilon \frac{\mathrm{E}_{\nu}(p(X(0)))}{(\mu - \lambda)^2} + o(\varepsilon). \tag{8}$$

Equation (8) is consistent with the reduced service rate approximation. As a matter of fact, as indicated in the introduction, everything happens as if we had a classical M/M/1 queue with service rate $\mu + \varepsilon \,\mathrm{E}_{\nu}(p(X(0)))$ and arrival rate λ. In such a queue, the mean length of the busy period is given by

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

which coincides with (8). In the following section we investigate the second-order term and show that the reduced service rate approximation is no longer valid.

4. Busy period: second-order term

In this section, we calculate the coefficient of ε^2 of the mean busy period $\mathrm{E}(\tilde{B}^{\varepsilon})$. Similarly to the first-order coefficient, 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 will be seen, this coefficient stresses the importance of the evolution of the variable 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 appears there.

In order to find the coefficient of ε^2, we must consider the different possibilities for the locations of the points t_1^+, t_2^+, t_1^-, and t_2^-. By using arguments similar to those 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 $\mathrm{E}(\tilde{B}^{\varepsilon} - B)$.

Define

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

In 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 the s-queue starting at time t_1 with $L(t_1^+) - 1$ customers). In the event $\mathcal{A}_{\pm} = \{t_1^- \le B,\ B \le t_1^+ \le B + B_1\}$, a canceled departure occurs and another departure is added before the completion of the busy period of the p-queue (B_1 denotes the duration of the additional busy period due to the canceled departure). Finally, in the event $\mathcal{A}_- = \{t_1^- \le B,\ B + B_1 \le t_1^+\}$, at least one canceled departure occurs and no additional departures are added before the completion of the busy period of duration B_1.

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

In a first step, we analyze the case in which there are only additional departures before B; that is, we consider the term $\mathrm{E}((\tilde{B}^{\varepsilon} - B)\,\mathbf{1}_{\mathcal{A}_+})$. When no canceled departure occurs, at most

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

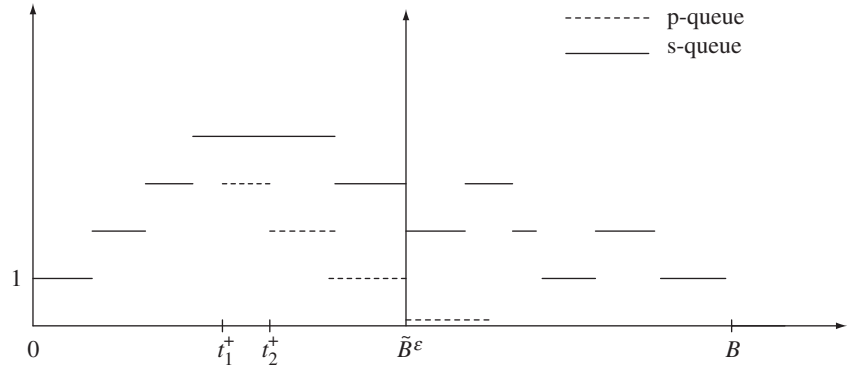

FIGURE 3: A busy period with two additional departures.

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 $\mathrm{E}(B - \tilde{B}^\varepsilon)$. In this case, the difference $B - \tilde{B}^\varepsilon$ is equal to the busy period of an s-queue that starts with either one or two customers, depending on whether or not, in the event $\{t_1^+ \le B\}$, the busy period of the p-queue is already complete at time t_2^+ (see Figure 3).

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

$$\begin{aligned}\mathrm{E}((B - \tilde{B}^\varepsilon)\,\mathbf{1}_{\mathcal{A}_+}) &= \mathrm{E}((B - \tilde{B}^\varepsilon)\,\mathbf{1}_{\{t_1^+ \le B,\, t_1^- \ge t_1^+ + B_{L(t_1^+)-1}\}})\\ &= \mathrm{E}(B_2)\,\mathrm{P}(t_1^+ < B,\, t_2^+ < t_1^+ + B_{L(t_1^+)-1},\, t_1^- \ge t_1^+ + B_{L(t_1^+)-1})\\ &\quad + \mathrm{E}(B_1)\,\mathrm{P}(t_1^+ < B,\, t_2^+ \ge t_1^+ + B_{L(t_1^+)-1}, t_1^- \ge t_1^+ + B_{L(t_1^+)-1})\\ &\quad + o(\varepsilon^2).\end{aligned}$$

This decomposition implies that

$$\begin{aligned}\mathrm{E}((B - \tilde{B}^\varepsilon)\,\mathbf{1}_{\mathcal{A}_+}) &= (\mathrm{E}(B_2) - \mathrm{E}(B_1))\,\mathrm{P}(t_1^+ < B,\, t_2^+ < t_1^+ + B_{L(t_1^+)-1})\\ &\quad + \mathrm{E}(B_1)(\mathrm{P}(t_1^+ < B) - \mathrm{P}(t_1^+ < B,\, t_1^- \le t_1^+ + B_{L(t_1^+)-1})) + o(\varepsilon^2). \quad (9)\end{aligned}$$

From (9), we see that we must expand three expressions with respect to ε. This we do by proving the three following lemmas.

Lemma 3. *The quantity* $\mathrm{P}(t_1^+ < B,\, t_2^+ < t_1^+ + B_{L(t_1^+)-1})$ *can be expanded as*

$$\begin{aligned}&\mathrm{P}(t_1^+ < B,\, t_2^+ < t_1^+ + B_{L(t_1^+)-1})\\ &\quad = \rho\varepsilon^2\,\mathrm{E}\left(\int_0^B (B - v)\,\mathrm{E}_v(p^+(X(0))p^+(X(v)))\right)\mathrm{d}v + o(\varepsilon^2). \quad (10)\end{aligned}$$

Proof. Let us recall the regenerative description of a busy period starting at time 0 with one customer. At time E_1 (which is exponentially distributed with parameter $\lambda + \mu$), the busy period is finished with probability $\mu/(\lambda + \mu)$. Otherwise, with probability $\lambda/(\lambda + \mu)$, a new

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

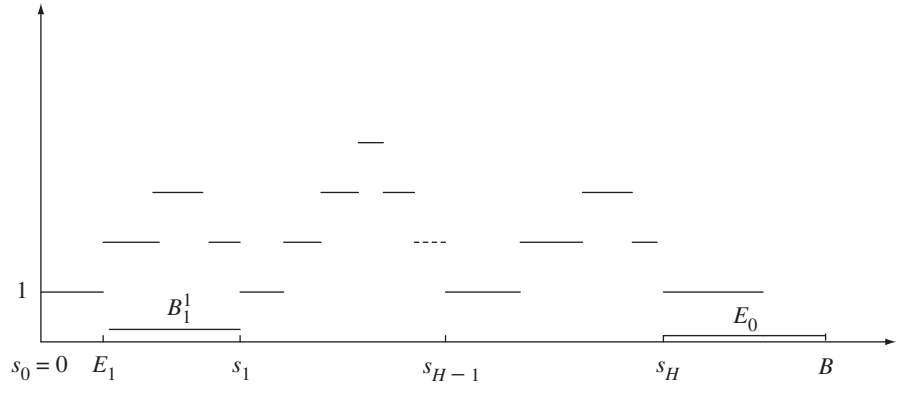

FIGURE 4: Decomposition of a busy period.

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:

$$B = E_0 + \sum_{i=1}^{H} (E_i + B_1^i).$$

Here H is geometrically distributed with parameter $\lambda/(\lambda+\mu)$, the (E_i) are independent and identically exponentially distributed with parameter $\lambda+\mu$, and the (B_1^i) are also independent and identically distributed. These random variables are all mutually independent. For all i, $0 \le i \le H$,

- s_i denotes the end of the ith sub-busy cycle: $s_0 = 0$ and, for $j \ge 1$,

  $$s_j = s_{j-1} + E_j + B_1^j,$$

  with $B = s_H + E_0$;
- 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 Appendix A. Figure 4 gives an illustration of the above definitions.

It is easy to see that, for the event $\{t_1^+ \le B,\ t_2^+ < t_1^+ + B_{L(t_1^+)-1}\}$ 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

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

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

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

$$\mathrm{E}\left(\int_{0\le u\le v\le B} p^+(X(u))p^+(X(v))\,\mathrm{d}u\,\mathrm{d}v\right) = \mathrm{E}\left(\int_{0\le u\le v\le B} \mathrm{E}_v(p^+(X(0))p^+(X(v-u)))\,\mathrm{d}u\,\mathrm{d}v\right).$$

Finally, since H is geometrically distributed with parameter $\lambda/(\lambda+\mu)$, (10) follows.

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

Lemma 4. *The quantity* $\mathrm{P}(t_1^+ \le B)$ *can be expanded as*

$$\mathrm{P}(t_1^+ \le B) = \varepsilon\frac{\mathrm{E}_v(p^+(X(0)))}{\mu-\lambda} - \varepsilon^2\,\mathrm{E}\left(\int_0^B (B-v)\,\mathrm{E}_v(p^+(X(0))p^+(X(v)))\,\mathrm{d}v\right) + o(\varepsilon^2). \tag{11}$$

Proof. We clearly have

$$\mathrm{P}(t_1^+ \le B) = \mathrm{E}\left(1-\exp\left(-\varepsilon\int_0^B p^+(X(s))\,\mathrm{d}s\right)\right) = \varepsilon\frac{\mathrm{E}_v(p^+(X(0)))}{\mu-\lambda} - \frac{\varepsilon^2}{2}\,\mathrm{E}\left(\left(\int_0^B p^+(X(s))\,\mathrm{d}s\right)^2\right) + o(\varepsilon^2).$$

The second moment of the integral can be expressed as follows, by symmetry and stationarity of the process $(X(t))$:

$$\mathrm{E}\left(\left(\int_0^B p^+(X(s))\,\mathrm{d}s\right)^2\right) = 2\,\mathrm{E}\left(\int_{0\le u\le v\le B} p^+(X(u))p^+(X(v))\,\mathrm{d}u\,\mathrm{d}v\right) = 2\,\mathrm{E}\left(\int_{0\le u\le v\le B} \mathrm{E}_v(p^+(X(0))p^+(X(v-u)))\,\mathrm{d}u\,\mathrm{d}v\right).$$

Equation (11) follows.

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

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

$$\mathrm{P}(t_1^+ < B,\ t_1^- \le t_1^+ + B_{L(t_1^+)-1}) = \frac{\varepsilon^2}{\mu}\,\mathrm{E}\left(\sum_{i=1}^{H}\sum_{j=1}^{N_i}\int_0^{A_i} p^+(X(u))p^-(X(D_j^i))\,\mathrm{d}u\right) + o(\varepsilon^2), \tag{12}$$

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

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

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

Here the (E_i) are independent and identically exponentially distributed with parameter $\mu+\lambda$ and the (B_1^i) are independent and identically distributed with the same distribution as B.

Proof. Using the regenerative description of a standard busy period introduced in the proof of Lemma 3, the variable t_1^- has to occur in some sub-busy period $[s_{i-1}+E_i, s_i]$ of B for some i, $1 \le i \le H$. A little thought shows 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 $\{t_1^+ < B,\ t_1^- \le t_1^+ + B_{L(t_1^+)-1}\}$ 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

$$\mathrm{E}\biggl(\int_{s_{i-1}+E_i}^{B} \varepsilon p^+(X(u)) \exp\biggl(-\varepsilon \int_0^u p^+(X(s))\,\mathrm{d}s\biggr)\,\mathrm{d}u \sum_{j=1}^{N_i} \varepsilon \frac{p^-(X(s_{i-1}+D_j^i))}{\mu} \times \prod_{k=1}^{j-1}\biggl(1-\varepsilon\frac{p^-(X(s_{i-1}+D_k^i))}{\mu}\biggr)\prod_{l=1}^{i-1}\prod_{r=1}^{N_l}\biggl(1-\varepsilon\frac{p^-(X(s_{l-1}+D_r^l))}{\mu}\biggr)\biggr),$$

where the coefficient of ε^2 is

$$\frac{1}{\mu}\mathrm{E}\biggl(\sum_{j=1}^{N_i}\int_{s_{i-1}+E_i}^{B} p^+(X(u))p^-(X(s_{i-1}+D_j^i))\,\mathrm{d}u\biggr).$$

By considering the different subcycles during B and using the stationarity of $(X(t))$, we recover (12).

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

Proposition 3. *The coefficient of ε^2 in the expansion, in terms of $\varepsilon > 0$, of $\mathrm{E}((B-\tilde{B}^\varepsilon)\,\mathbf{1}_{\mathcal{A}_+})$ is given by*

$$a_+ = -\frac{1}{\mu}\mathrm{E}\biggl(\int_0^B (B-v)\,\mathrm{E}_v(p^+(X(0))p^+(X(v)))\,\mathrm{d}v\biggr) - \frac{1}{\mu^2(1-\rho)}\mathrm{E}\biggl(\sum_{i=1}^{H}\sum_{j=1}^{N_i}\int_0^{A_i} p^+(X(u))p^-(X(D_j))\,\mathrm{d}u\biggr). \tag{13}$$

To complete the analysis, we now turn to the expansion of $\mathrm{E}((\tilde{B}^\varepsilon - B)\,\mathbf{1}_{\mathcal{A}_\pm})$ and $\mathrm{E}((\tilde{B}^\varepsilon - B)\,\mathbf{1}_{\mathcal{A}_-})$. In the calculations, it is convenient to consider the sum of these terms; we then have the following result.

Proposition 4. *The coefficient of ε^2 in the expansion, in terms of $\varepsilon > 0$, of $\mathrm{E}((\tilde{B}^\varepsilon - B)\,\mathbf{1}_{\mathcal{A}_\pm \cup \mathcal{A}_-})$ is given by*

$$a_- = \frac{1}{\mu^2(1-\rho)}\biggl(-\mathrm{E}\biggl(\sum_{i=1}^{N}\int_0^{B+B_1} p^-(X(D_i))p^+(X(s))\,\mathrm{d}s\biggr) + \frac{1}{\mu}\mathrm{E}\biggl(\sum_{i=1}^{N}\sum_{k=1}^{N'} p^-(X(0))p^-(X(B-D_i+D_k'))\biggr)\biggr), \tag{14}$$

where N and $D_1, \dots, D_N$ denote the number of departures and the departure times in the busy period of length B, and N' and $D_1', \dots, D_{N'}'$ respectively denote the same quantities in the busy period of length B_1.

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

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 it into account.

By the strong Markov property, using the same method as in Section 3, we obtain the relation

$$\mathrm{E}((B + B_1 - \tilde{B}^\varepsilon)\,\mathbf{1}_{\{t_1^- \le B,\ B \le t_1^+ \le B+B_1\}}) = \mathrm{E}(B_1')P(t_1^- \le B,\ B \le t_1^+ \le B + B_1) + o(\varepsilon^2),$$

where the random variable B_1' has the same distribution as the random variable B_1. Hence,

$$\begin{aligned}\mathrm{E}((\tilde{B}^\varepsilon - B)\,\mathbf{1}_{\mathcal{A}_\pm}) &= \mathrm{E}((\tilde{B}^\varepsilon - B)\,\mathbf{1}_{\{t_1^- \le B,\ B \le t_1^+ \le B+B_1\}}) \\ &= \mathrm{E}(B_1\,\mathbf{1}_{\{t_1^- \le B,\ B \le t_1^+ \le B+B_1\}}) - \mathrm{E}(B_1')\,\mathrm{P}(t_1^- \le B,\ B \le t_1^+ \le B + B_1) \\ &\quad + o(\varepsilon^2). \end{aligned} \tag{15}$$

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

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

Hence,

$$\begin{aligned}\mathrm{E}((\tilde{B}^\varepsilon - B)\,\mathbf{1}_{\mathcal{A}_-}) &= \mathrm{E}(B_1\,\mathbf{1}_{\{t_1^- \le B,\ t_2^- > B+B_1\}}) - \mathrm{E}(B_1\,\mathbf{1}_{\{t_1^- \le B,\ t_1^+ \le B+B_1\}}) \\ &\quad + \mathrm{E}(B_1\,\mathbf{1}_{\{t_1^- \le B,\ B \le t_2^- \le B+B_1\}}) + \mathrm{E}(B_1')\,\mathrm{E}(\mathbf{1}_{\{t_1^- \le B,\ B \le t_2^- \le B+B_1\}}) \\ &\quad + \mathrm{E}(B_2\,\mathbf{1}_{\{t_2^- \le B\}}) + o(\varepsilon^2)\end{aligned}$$

and, finally,

$$\begin{aligned}\mathrm{E}((\tilde{B}^\varepsilon - B)\,\mathbf{1}_{\mathcal{A}_-}) &= \mathrm{E}(B_1)\,\mathrm{P}(t_1^- \le B,\ t_2^- > B) - \mathrm{E}(B_1\,\mathbf{1}_{\{t_1^- \le B,\ t_1^+ \le B+B_1\}}) \\ &\quad + \mathrm{E}(B_1')\,\mathrm{P}(t_1^- \le B,\ B \le t_2^- \le B + B_1) + 2\,\mathrm{E}(B_1)\,\mathrm{P}(t_2^- \le B) + o(\varepsilon^2).\end{aligned} \tag{16}$$

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

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

has no term in ε^2 in its power series expansion. Thus, the first and the final terms on the right-hand side of (16) cancel out of the expansion.

The following expansions are obtained in a similar way:

$$\mathrm{E}(B_1\,\mathbf{1}_{\{t_1^- \le B,\ t_1^+ \le B+B_1\}}) = \frac{\varepsilon^2}{\mu}\,\mathrm{E}\left(B_1 \sum_{i=1}^{N} \int_0^{B+B_1} p^-(X(D_i))\,p^+(X(s))\,\mathrm{d}s\right) + o(\varepsilon^2),$$

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

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

Here N and $D_1, \dots, D_N$ and N' and $D'_1, \dots, D'_{N'}$ denote the numbers of departures and the departure times in two independent busy periods of lengths B and B_1, respectively.

If we sum up the expansions obtained in the case of canceled departures and the case of one canceled departure and one additional departure ((16) and (15), respectively), standard manipulations yield the second term of the expansion in ε of $\mathrm{E}((\tilde{B}^\varepsilon - B)(\mathbf{1}_{\mathcal{A}_\pm} + \mathbf{1}_{\mathcal{A}_-}))$.

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

Theorem 1. *The coefficient of ε^2 in the power series expansion in ε of $\mathrm{E}(\tilde{B}^\varepsilon - B)$ is equal to $a_- - a_+$, where the coefficients a_+ and a_- are given by (13) and (14), respectively.*

It should be noted that the distributions involved in (13) and (14) can be given explicitly using the classical results concerning the M/M/1 queue; see Appendix A, where they are recalled. In the next section, we examine some applications of the above result.

5. Applications

5.1. Nonnegative perturbation functions

Equations (10) and (11) yield the expansion

$$\mathrm{E}(B - \tilde{B}^\varepsilon) = \delta_1 \varepsilon + \delta_2 \varepsilon^2 + o(\varepsilon^2),$$

with $\delta_1 = \mathrm{E}_\nu(p(X(0)))/(\mu - \lambda)^2$ and

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

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

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

and, hence,

$$\mathrm{E}(B - \tilde{B}^\varepsilon) = \varepsilon \frac{\mathrm{E}_\nu(p(X(0)))}{(\mu - \lambda)^2} - \varepsilon^2 \frac{\mathrm{E}_\nu(p(X(0)))^2}{(\mu - \lambda)^3} - \frac{\varepsilon^2}{\mu}\,\mathrm{E}\left(\int_0^B (B - v)C_p(v)\,\mathrm{d}v\right) + o(\varepsilon^2).$$

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

Proposition 5. (Comparison with reduced service rate.) *If $\hat{B}$ is the length of a busy period of an M/M/1 queue with service rate $\mu + \varepsilon\,\mathrm{E}_\nu(p(X(0)))$, then*

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2}\,\mathrm{E}(\hat{B} - \tilde{B}^\varepsilon) = -\frac{1}{\mu}\,\mathrm{E}\left(\int_0^B (B - v)C_p(v)\,\mathrm{d}v\right),$$

where, for $u \geq 0$,

$$C_p(u) = \mathrm{E}_\nu(p(X(0))p(X(u))) - \mathrm{E}_\nu(p(X(0)))^2$$

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

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

It is straightforward to conclude from Proposition 5 that $\mathrm{E}(\hat{B} - \tilde{B}^{\varepsilon})$ is negative when ε is small.

Corollary 1. (Negative impact of the variation of the service rate.) *When the environment is positively correlated, i.e. the function $u \mapsto C_p(u)$ is nonnegative, the first term of the expansion in ε of $\mathrm{E}(\hat{B} - \tilde{B}^{\varepsilon})$ is of order ε^2 and is negative.*

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

Proposition 6. *When the correlation function of the environment is exponentially decreasing, i.e. when, for some $\alpha > 0$,*

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

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

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} \mathrm{E}(\hat{B} - \tilde{B}^{\varepsilon}) =: \Delta_2(\alpha) = -\frac{\mathrm{var}[p(X(0))]}{(\mu - \lambda)^3} \mathrm{E}(\mathrm{e}^{-\alpha Z}) \leq 0,$$

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

$$x \mapsto \frac{1}{\mu(1-\rho)^2} \int_x^{\infty} \mathrm{P}(B \geq u)\, \mathrm{d}u.$$

In particular, the function $\alpha \mapsto \Delta_2(\alpha)$ is nondecreasing and concave.

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

$$\mathrm{E}(\mathrm{e}^{-\alpha A^*}) = \frac{1 - \mathrm{E}(\mathrm{e}^{-\alpha A})}{\alpha\, \mathrm{E}(A)} \tag{17}$$

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

To simplify our notation, we assume that $\mathrm{var}[p(X(0))] = 1$. In this case, Proposition 5 gives the coefficient $\Delta_2(\alpha)$ of ε^2 as

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

The proposition then follows from (17).

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

5.2. Nonpositive perturbation functions

It is assumed in this section that the perturbation function is nonpositive in such a way that the environment uses part of the capacity of the M/M/1 queue with constant service rate μ. This application is motivated by the following practical situation involving 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 nonpriority 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 nonpriority traffic is $\mu - \varepsilon d(x)$. The function $p(x) = -d(x)$ is nonpositive. We are then in a framework in which the environment gives reduced bandwidth to a nonpriority M/M/1 queue. Notation identical to that in the previous section is used extensively.

Equations (4) and (14) imply that the expansion

$$\mathrm{E}(B - \tilde{B}^\varepsilon) = \delta_1 \varepsilon + \delta_2 \varepsilon^2 + o(\varepsilon^2)$$

holds with $\delta_1 = \mathrm{E}(p(X(0)))/(\mu - \lambda)^2$ and

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

where, as in (14), N and $D_1, \ldots, D_N$ and N' and $D'_1, \ldots, D'_{N'}$ denote the numbers of departures and the departure times in the busy periods of length B and B_1, respectively. The quantities δ_1 and δ_2 are nonpositive. Thus, to first order, the mean of $\tilde{B}^\varepsilon$ is larger than the mean of B. The next 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 $\hat{B}$ in an M/M/1 queue with service rate $\mu + \varepsilon \mathrm{E}_\nu(p(X(0)))$.

Proposition 7. (Comparison with reduced service rate.) *If $\hat{B}$ is the length of a busy period of an M/M/1 queue with service rate $\mu + \varepsilon \mathrm{E}_\nu(p(X(0)))$, then*

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} \mathrm{E}(\hat{B} - \tilde{B}^\varepsilon) = -\frac{1}{\mu^3(1-\rho)} \mathrm{E}\Bigg(\sum_{i=1}^{N} \sum_{k=1}^{N'} C_p(X(B - D_i + D'_k)) \Bigg), \tag{18}$$

where, as in (14), N and $D_1, \ldots, D_N$ and N' and $D'_1, \ldots, D'_{N'}$ denote the numbers of departures and the departure times in the busy periods of length B and B_1, respectively; and, for $u \geq 0$,

$$C_p(u) = \mathrm{E}_\nu(p(X(0)) p(X(u))) - \mathrm{E}_\nu(p(X(0)))^2$$

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

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

Proposition 8. (Negative impact of the variation of the service rate.) *When the environment is positively correlated, i.e. the function $u \mapsto C_p(u)$ is nonnegative, the first term of the expansion in ε of $\mathrm{E}(\hat{B} - \tilde{B}^\varepsilon)$ is of order ε^2 and is negative.*

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

In comparison with the case of a nonnegative perturbation function, if the correlation function of the environment is exponentially decreasing, a simple closed-form expression for the right-hand side of (18) seems to be difficult to obtain, though the same qualitative results hold.

Proposition 9. (Exponential decay.) *When the correlation function of the environment is exponentially decreasing, i.e. when, for some* $\alpha > 0$,

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

the function

$$\alpha \mapsto \lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} \mathrm{E}(\hat{B} - \tilde{B}^{\varepsilon})$$

is nonpositive, nondecreasing, and concave. Moreover, when α *tends to* ∞ *this quantity converges to* 0.

5.3. Fast environments

Here, a general perturbation function p is considered together with some stationary Markov process $(X(t))$ with invariant probability distribution ν. We assume that the process obeys a mixing condition such as

$$\lim_{t \to \infty} |\mathrm{E}_\nu(f(X(0))g(X(t))) - \mathrm{E}_\nu(f(X(0)))\,\mathrm{E}_\nu(g(X(0)))| = 0, \tag{19}$$

for any Borelian bounded functions f and g on the state space $\mathcal{S}$. Note that this condition is not in general restrictive, since it is true for any ergodic Markov process with a countable (or finite) state space and 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 ∞ 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], using stochastic bounds.

At the intuitive level, when α becomes large, the total service capacity available between t and $t + h > t$ is given by

$$\mu h + \varepsilon \int_t^{t+h} p(X(\alpha u))\,\mathrm{d}u \overset{\mathrm{D}}{=} \mu h + \varepsilon \frac{1}{\alpha} \int_0^{\alpha h} p(X(u))\,\mathrm{d}u \sim [\mu + \varepsilon\,\mathrm{E}_\nu(p(X(0)))]h,$$

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

Proposition 10. *When the environment is given by* $(X(\alpha t))$ *and (19) holds, if* $\delta_2(\alpha)$ *is the coefficient of* ε^2 *in the expansion in* ε *of* $\mathrm{E}(\tilde{B}^{\varepsilon} - B)$, *i.e.*

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

then

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

Proof. The quantity $\delta_2(\alpha)$ is equal to $a_- - a_+$, where a_- and a_+ are given by (14) and (13), respectively. We shall deal only with the first term of a_- in (14). Let

$$F(\alpha) := -\mathrm{E}\left(\sum_{i=1}^{N} \int_0^{B+B_1} p^-(X(\alpha D_i))\,p^+(X(\alpha s))\,\mathrm{d}s\right),$$

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

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

$$F(\alpha) = -\mathrm{E}\Big(\sum_{i=1}^{N} \int_0^{B+B_1} \mathrm{E}(p^-(X(\alpha D_i))p^+(X(\alpha s)) \mid B, N)\,\mathrm{d}s\Big).$$

Equation (19) and the boundedness of p (Assumption 1) show that, almost surely,

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

therefore, Lebesgue's theorem gives

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

using the expressions of $\mathrm{E}(N)$ and $\mathrm{E}(NB)$ given in Appendix A. Similar calculations can be performed for all the other terms, to finally prove the proposition.

Appendix A. Some useful quantities for the M/M/1 queue

Let (A_k) and (D_k) respectively denote the arrival times and departure times in a busy period of an M/M/1 queue with arrival rate λ and service rate μ. A busy period 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) $A_{k+1} \le D_k$ for $k = 1, \ldots, N-1$.

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,

$$b_n(t) = \frac{\mathrm{dP}(B < t,\ N = n)}{\mathrm{d}t} = \frac{\mathrm{e}^{-\lambda t}(\lambda t)^{n-1}}{(n-1)!}\,\frac{\mu \mathrm{e}^{-\mu t}(\mu t)^{n-1}}{(n-1)!}\,\mathrm{P}(A_2 \le D_1, \ldots, A_n < D_{n-1}).$$

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

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

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

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

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press

is given by

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

for $|z|\le 1$ and $\mathrm{Re}(\xi)\ge 0$. It is easy to show that

$$\mathrm{E}(N)=\int_0^\infty \mathrm{d}t\sum_{n=1}^\infty nb_n(t)=\frac{1}{1-\rho},$$

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

$$\mathrm{E}(N(N-1))=\int_0^\infty \mathrm{d}t\sum_{n=1}^\infty n(n-1)b_n(t)=\frac{\mathrm{d}^2\varphi}{\mathrm{d}z^2}(1,0)=\frac{2\mu^2\lambda}{(\mu-\lambda)^3}.$$

To conclude, we must compute $\mathrm{E}(D)$, 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 [14], for example), we obtain

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

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 (duration of the) busy period generated by the ith customer that arrives during σ, N_i the number of customers in B_i, and D_i the sum of the departure times of B_i from the beginning of this busy period. By taking the expectation, we can easily derive that

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

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

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

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

References

[1] Agrawal, R., Makowski, A. M. and Nain, P. (1999). On a reduced load equivalence for fluid queues under subexponentiality. *Queueing Systems Theory Appl.* **33,** 5–41.

[2] Altman, E., Avrachenkov, K. and Núñez-Queija, R. (2004). Perturbation analysis for denumerable Markov chains with application to queueing models. *Adv. Appl. Prob.* **36,** 839–853.

[3] Antunes, N., Fricker, C., Guillemin, F. and Robert, P. (2005). Integration of streaming services and TCP data transmission in the Internet. Submitted.

[4] Boxma, O. J. and Kurkova, I. A. (2000). The M/M/1 queue in a heavy-tailed random environment. *Statist. Neerlandica* **54,** 221–236.

[5] Cohen, J. W. (1982). *The Single Server Queue*, 2nd edn. North-Holland, Amsterdam.

[6] Delcoigne, F., Proutière, A. and Régnié, G. (2004). Modeling integration of streaming and data traffic. *Performance Evaluation* **55,** 185–209.

[7] Fricker, C., Guillemin, F. and Robert, P. (2004). Perturbation analysis of an M/M/1 queue in a diffusion random environment. Submitted.

[8] Grandell, J. (1977). Point processes and random measures. *Adv. Appl. Prob.* **9,** 502–526.

[9] Jelenković, P. and Momčilović, P. (2002). Resource sharing with subexponential distributions. In *Proc. IEEE Infocom 2002* (New York, June 2002), IEEE, pp. 1316–1325.

[10] Massoulié, L. and Roberts, J. (1999). Bandwidth sharing: objectives and algorithms. In *Proc. IEEE Infocom 1999* (New York, March 1999), IEEE, pp. 1395–1403.

[11] Núñez-Queija, R. (2000). Sojourn times in a processor sharing queue with service interruptions. *Queueing Systems* **34,** 351–386.
[12] Núñez-Queija, R. (2001). Sojourn times in non-homogeneous QBD processes with processor sharing. *Stoch. Models* **17,** 61–92.
[13] Núñez-Queija, R. and Boxma, O. J. (1998). Analysis of a multi-server queueing model of ABR. *J. Appl. Math. Stoch. Anal.* **11,** 339–354.
[14] Robert, P. (2003). *Stochastic Networks and Queues* (Appl. Math. (New York) **52**). Springer, New York.

https://doi.org/10.1239/aap/1143936150 Published online by Cambridge University Press