M/G/1 queue with event-dependent arrival rates

Benjamin Legros

Noname manuscript No.
(will be inserted by the editor)

M/G/1 queue with event-dependent arrival rates

Benjamin Legros

the date of receipt and acceptance should be inserted later

Abstract Motivated by experiments on customers' behavior in service systems, we consider a queueing model with event-dependent arrival rates. Customers' arrival rates depend on the last event which may either be a service departure or an arrival. We derive explicitly the performance measures and analyze the impact of the event-dependency. In particular, we show that this queueing model, in which a service completion generates a higher arrival rate than an arrival, performs better than a system in which customers are insensitive to the last event. Moreover, contrary to the M/G/1 queue, we show that the coefficient of variation of the service does not necessarily deteriorate the system performance. Next, we show that this queueing model may be the result of customer's strategic behavior when only the last event is known. Finally, we investigate the historical admission control problem. We show that under certain conditions a deterministic policy with two thresholds may be optimal. This new policy is easy to implement and provides an improvement compared to the classical one-threshold policy.

Keywords queueing systems · performance evaluation · M/G/1 · threshold policy · strategic behavior

Mathematics Subject Classification (2000) 90B22 · 60K25 · 68M20

1 Introduction

Context and motivation. In many service settings, customers encounter queues and have to decide between joining or balking. For instance in a theme park, the decision to join a queue for an attraction may depend on the queue length but also on the possible entertainment that this attraction could offer. In practice, a quick observation of the queue influences customers about the utility to join or not. In particular, the last realized event often has an important impact on their decisions. A too high importance given to the last

Laboratoire Genie Industriel, CentraleSupélec, Université Paris-Saclay, Grande Voie des Vignes, 92290 Chatenay-Malabry, France
E-mail: benjamin.legros@centraliens.net

realized event may however bias customers' decisions by inducing illusory correlation. For instance, if the last event was a service then a reduction of the queue size is observed. This event is either irrelevant or at least insufficient to evaluate the quality of the queueing system. Yet, one can conclude that this queue has a high speed of service since a customer has just left service. Another bias is to believe that a positive event is usually followed by negative ones. In this case, a customer may believe that since a service completion has just occurred, the next service times will be long.

Although the last observed event is not a rational indicator of the quality of a service system, [1] showed that the evolution of the queue size (increasing or decreasing) due to service completions or arrivals impacts strongly the decision of new arriving customers, by using laboratory experiments in which participants experience several observable queues with different characteristics in terms of queue length and service times.

Queues with workload-dependent arrival or service rates have already been widely studied. Therefore, we aim to investigate a queue with another feature for the customers' behavior which is the *event-dependency*. In order to analyze the impact of this new feature, we neglect the other aspects of customer's behavior like the workload sensitivity. To the best of our knowledge, this paper is the first to model this behavior. In what follows we describe the queueing model studied in this article.

Model description. We consider a single-server first-come-first-served queueing model with a general service time distribution. We assume that the iid service times have a density, denoted by g. We denote by $\overline{x}$ the expected service time and by cv the coefficient of variation of the service time distribution; it is the ratio of the standard deviation divided by the expected value. Based on the last event which can either be an arrival or a service completion, the feature of the event-dependency is captured through two exponential arrival rates; λ^+ and λ^-. More precisely, at a given time, if the last event was an arrival (respectively a service), the next customer will arrive in the system after a random time exponentially distributed with parameter λ^+ (respectively λ^-).

Contributions. The key contributions of this paper are the following.

- **Performance Analysis.** First, we consider the embedded Markov chain right after a service completion. This allows us to derive the probability generating function explicitly together with the stability condition and the probability of an empty system at departure instants. Contrary to the M/G/1 queue, the queue length distribution is not identical at arrival and arbitrary instants. Therefore, we also study the relation between these distributions. This leads to explicit expressions of the performance measures at arbitrary instants. Next, we consider three special cases of service time distributions (Exponential, Erlang and Deterministic). We show that our queueing model performs better than a system in which customers are insensitive to the last event if a service generates a higher arrival rate than an arrival ($\lambda^- > \lambda^+$). Contrary

to the M/G/1 queue, we show that the performance of our model can improve when the coefficient of variation of the service increases.

- **Equilibrium Strategy.** We next question if the event-dependency of the arrival rates can be the result of a rational strategy when customers only know the last realized event. For this purpose, we study the remaining service time distribution for an arriving customer given the last event. Next, we derive the expected waiting time of customers who arrive after an arrival and the one of customers who arrive after a service. We prove that the first one is always higher than the second one. Given a non-empty queue, a necessary and sufficient condition which involves the variability of the service time determines the comparison between the expected waiting times. As a conclusion, we show that $\lambda^- \geq \lambda^+$ can be the result of a strategic behavior.
- **Admission Control Problem.** We investigate the historical admission control problem with event-dependency. A controller has to determine at the arrival of a customer whether to let this customer enter the queue or to reject this customer. The decision is based on the number of customers present in the system, the distribution of the remaining service time and the last realized event. The objective is to maximize the throughput of served customers with a service level constraint on the expected number of customers in the system. A two-thresholds policy depending on the last realized event and on the remaining service time of the customer in service is the only possible deterministic policy. Using a Markov decision process approach and approximating the service time distribution by a Coxian distribution allows us to find necessary conditions under which a deterministic threshold type policy is optimal. These conditions are that $\lambda^- \geq \lambda^+$ and that the departure rate out of each service phase of the Coxian distribution is decreasing in the number of remaining service phases. The relation between the optimal thresholds and the remaning service time might make the optimal policy complicate to implement in practice. We then propose a simplified version of this policy where the two thresholds depend only on the number of customers in the system. We finally develop an exact numerical method to obtain the performance measures under this policy at arbitrary instants and show that this policy outperforms the classical one-threshold policy.

The remainder of this paper is structured as follows. We conclude this section with a literature survey. In Section 2, we compute the performance measures and investigate the impact of the event-dependent arrival rates. In Section 3, we prove that the case $\lambda^- \geq \lambda^+$ may be the result of a strategic behavior. In Section 4, we investigate the admission control problem. Finally, Section 5 opens on future research perspectives. The notations used are summarized at the end of the article.

Literature Review. Methodologically, the analysis of this paper is related to (i) queues with general service and state definition based on the residual service time [15, 28, 8], (ii) queueing systems with phase-type service time distributions [24] and (iii) Markov decision process approach [18].

A stream of literature related to this paper is that of queueing systems with state-dependent parameters. Single-server queues with workload-dependent arrival or service rates have been widely analyzed (e.g. see [15, 5, 9, 11]). Historically, Markovian models with general release rule have been considered as dam processes. Later, another application came in packet-switched communication systems; the transmission rate of data connections can also be adapted based on the queue size. Other single-server queueing models have been proposed where the arrival or the service rates are depending on the waiting time of the customer in service [10, 31], the waiting time of the first customer in line [6], or on the remaining service time [21]. Other examples with workload dependent resources can be found via the slow server problem. With two servers (a slow and a fast one) [19], [22], [30], and [17] show that the fast server should be always used, and the slow server should be only used when the fast server is busy and the number of customers waiting in the queue exceeds a given threshold. Extension of these studies for more than two servers can be found in [20], [26] and [23]. Our paper differs from these papers since the arrival-dependency is based on the last event instead of the observed workload or realized waiting time.

2 Performance Analysis

We investigate the impact of an event-dependent arrival process on the performance measures. In Section 2.1 we consider the embedded Markov chain right after a service completion. This approach, in line with the standard analysis of the M/G/1 queue, allows us to obtain explicitly the probability generating function. Contrary to the M/G/1 queue, the queue length distribution is not identical at arrival and arbitrary instants. Therefore, in Section 2.2 we study the relation between these distributions in order to reach the performance measures at arbitrary instants. Finally in Section 2.3, we apply our results to particular service time distributions (Exponential, Erlang and Deterministic) to better understand the effect of the event-dependency of the arrival process. An alternative method to obtain the performance measures with a state definition based on the remaining service time is proposed in Section 1 of the online supplement.

2.1 The embedded Markov chain at service completion instants

Consider the system just after a customer has completed a service. The random variable X_i represents the number of customers remaining in the system as the ith customer departs. We can write that

$$X_{i+1} = \begin{cases} X_i - 1 + A_{i+1}, \text{ if } X_i > 0, \\ A_{i+1}, \text{ if } X_i = 0, \end{cases} \tag{1}$$

where A_{i+1} is the number of customer who arrived during the service time of the $(i+1)$st customer. The service time of the $(i+1)$st customer is independent of previous service times and the length of the queue,

so we can denote by S this random variable without mentioning the index of the $(i+1)$st customer. We now evaluate the distribution of A_{i+1}. Two cases should be considered.

Case 1. If $X_i > 0$, the service initiation time of Customer $i+1$ is the service completion time of Customer i. Therefore, the last event for the first customer who arrives during the service of Customer $i+1$ is a service. It is an arrival for all the other customers. Let us denote by N_t the number of customers who arrive during a service of length t. The distribution of N_t is given by the following set of differential equations:

$$\begin{cases} P(N_0 = 0) = 1, \\ \frac{\mathrm{d}P(N_t=0)}{\mathrm{d}t} = -\lambda^- P(N_t = 0), \\ \frac{\mathrm{d}P(N_t=1)}{\mathrm{d}t} = -\lambda^+ P(N_t = 1) + \lambda^- P(N_t = 0), \\ \frac{\mathrm{d}P(N_t=n)}{\mathrm{d}t} = -\lambda^+ P(N_t = n) + \lambda^+ P(N_t = n-1), \text{ for } n \geq 2. \end{cases} \tag{2}$$

After some algebra, we obtain the solution of this system. It is given by $P(N_t = 0) = e^{-\lambda^- t}$, and

$$P(N_t = n) = \frac{\lambda^-}{\lambda^+ - \lambda^-} \left(\frac{\lambda^+}{\lambda^+ - \lambda^-} \right)^{n-1} \left(e^{-\lambda^- t} - e^{-\lambda^+ t} \sum_{k=0}^{n-1} \frac{((\lambda^+ - \lambda^-)t)^k}{k!} \right),$$

for $n \geq 1$. This arrival process is a modified Poisson process where the first interarrival time follows a different distribution than the other interarrival times. Given that $X_i > 0$, the number of customers who arrive during the service of the $(i+1)$st customer is independent of the index i. We therefore simply denote this random variable by A. We hence have

$$P(A_{i+1} = n | X_i > 0) = P(A = n) = \int_0^\infty P(N_t = n) g(t)\, \mathrm{d}t = \alpha_n,$$

for $i, n \geq 0$.

Case 2. If $X_i = 0$, the service initiation of Customer $i+1$ is the arrival time of Customer $i+1$. Therefore, the last event for all customers who arrive during the service of Customer $i+1$ is an arrival. Thus, the number of customer's arrivals during a service of length t given $X_i = 0$ follows a Poisson process with rate $\lambda^+ t$ and is independent of the index i. We then denote by B the random variable which represents the number of customers who arrive during the service of Customer $i+1$ given that $X_i = 0$. We may write

$$P(A_{i+1} = n | X_i = 0) = P(B = n) = \int_0^\infty e^{-\lambda^+ t} \frac{(\lambda^+ t)^n}{n!} g(t)\, \mathrm{d}t = \beta_n,$$

for $i, n \geq 0$. This corresponds to the transition probabilities in the embedded Markov chain of an M/G/1 queue.

As a conclusion Equation (1) can be simplified into

$$X_{i+1} = \begin{cases} X_i - 1 + A, \text{ if } X_i > 0, \text{ and,} \\ B, \text{ if } X_i = 0. \end{cases} \tag{3}$$

The definition of the process in Equation (3) allows us to define a discrete time Markov chain. The related matrix of transition probabilities is given by

$$M = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \cdots \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \cdots \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \cdots \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

Assuming that steady state is reached, we let p_n represent the stationary probability that n customers are in the system at a service departure instant. We now define the generating functions

$$P(z) = \sum_{n=0}^{\infty} p_n z^n, \quad A(z) = \sum_{n=0}^{\infty} \alpha_n z^n, \quad B(z) = \sum_{n=0}^{\infty} \beta_n z^n, \tag{4}$$

for $|z| \leq 1$. In Theorem 1, we derive $P(z)$ and p_0. In addition, we give the condition of existence of $P(z)$. This condition is also the condition which ensures the stationary regime.

Theorem 1 *Under the stability condition* $\overline{x} < \frac{1}{\lambda^-}(1 - G^*(\lambda^-)) + \frac{1}{\lambda^+}G^*(\lambda^-)$*, we have*

$$P(z) = p_0 \frac{(1-z)\left((\lambda^+ - \lambda^-)G^*(\lambda^-) - \lambda^+ z G^*(\lambda^+ - \lambda^+ z)\right)}{(1-z)\left((\lambda^+ - \lambda^-)G^*(\lambda^-) - \lambda^+ z\right) + \lambda^- z(1 - G^*(\lambda^+ - \lambda^+ z))},$$

with

$$p_0 = \frac{\lambda^+(1 - \overline{x}\lambda^-) + (\lambda^- - \lambda^+)G^*(\lambda^-)}{\lambda^+ + (\lambda^- - \lambda^+)G^*(\lambda^-)}.$$

Proof. The vector $(p_0, p_1, \cdots)$ is solution of $(p_0, p_1, \cdots) = (p_0, p_1, \cdots) \times M$. Therefore, we have

$$\begin{aligned} p_0 &= \beta_0 p_0 + \alpha_0 p_1, \\ p_1 z &= \beta_1 p_0 z + \alpha_1 p_1 z + \alpha_0 p_2 z, \\ p_2 z^2 &= \beta_2 p_0 z^2 + \alpha_2 p_1 z^2 + \alpha_1 p_2 z^2 + \alpha_0 p_3 z^2, \\ &\vdots \\ p_n z^n &= \beta_n p_0 z^n + \alpha_n p_1 z^n + \alpha_{n-1} p_2 z^n + \cdots + \alpha_0 p_{n+1} z^n, \\ &\vdots \end{aligned}$$

By summing up these equations, we get $P(z) = p_0 B(z) + A(z)(p_1 + p_2 z + p_3 z^2 + \cdots)$. This leads to $P(z) = p_0 B(z) + \frac{A(z)}{z}(P(z) - p_0)$. Finally, we deduce that

$$P(z) = p_0 \frac{A(z) - zB(z)}{A(z) - z}. \tag{5}$$

Using L'Hôpital's rule, we obtain

$$P(1) = 1 = p_0 \frac{A'(1) - B(1) - B'(1)}{A'(1) - 1}. \tag{6}$$

From the results of the M/G/1 queue, we know that $B'(1) = \lambda^+ \overline{x}$ (e.g., see [16], page 185). Moreover, by definition, $B(1) = 1$. From [16] page 184, we know that $B(z) = G^*(\lambda^+ - \lambda^+ z)$.

There remains to compute $A(z)$ and $A'(1)$. For this purpose, let us introduce

$$N(z,t) = \sum_{n=0}^{\infty} P(N_t = n) \cdot z^n,$$

for $t > 0$ and $|z| < 1$. We have

$$A(z) = \int_0^{\infty} N(z,t) \cdot g(t)\, \mathrm{d}t.$$

From Equation (2), we get the differential equation followed by $N(z,t)$;

$$\frac{\mathrm{d}N(z,t)}{\mathrm{d}t} = (1-z)\left(-\lambda^+ N(z,t) + P(N_t = 0) \cdot (\lambda^+ - \lambda^-)\right).$$

Using $P(N_t = 0) = e^{-\lambda^- t}$ and $N(z,0) = 1$, we can solve this differential equation using the method of variation of constants. After some algebra, we obtain

$$N(z,t) = \frac{(1-z)(\lambda^+ - \lambda^-)}{\lambda^+(1-z) - \lambda^-} e^{-\lambda^- t} - \frac{\lambda^- z}{\lambda^+(1-z) - \lambda^-} e^{-\lambda^+(1-z)t}.$$

This leads to

$$A(z) = \frac{(1-z)(\lambda^+ - \lambda^-)}{\lambda^+(1-z) - \lambda^-} G^*(\lambda^-) - \frac{\lambda^- z}{\lambda^+(1-z) - \lambda^-} G^*(\lambda^+(1-z)),$$

where $G^*(.)$ is the Laplace-Stieltjes Transform (LST) of the service time; $G^*(s) = \int_{t=0}^{\infty} g(t) e^{-st}\, \mathrm{d}t$. Moreover,

$$A'(z) = \frac{\lambda^-(\lambda^+ - \lambda^-)}{(\lambda^+(1-z) - \lambda^-)^2}(G^*(\lambda^-) - G^*(\lambda^+(1-z))) + \frac{\lambda^+\lambda^- z}{\lambda^+(1-z) - \lambda^-} G'^*(\lambda^+(1-z)).$$

Using $G^*(0) = 1$ and $G'^*(0) = -\overline{x}$, leads to

$$A'(1) = \frac{\lambda^+ - \lambda^-}{\lambda^-}\left(G^*(\lambda^-) - 1\right) + \lambda^+ \overline{x}.$$

Replacing this expression in Equation (6) gives the expression of p_0 and the stability condition. Finally, replacing the expression of $A(z)$ in Equation (5) allows us to obtain $P(z)$ as in the theorem. □

In Corollary 1, we deduce the expected number of customers in the system at departure instants, $E(Q_d)$. In the case $\lambda^+ = \lambda^-$, we obtain the Pollaczeck-Kinchin formulas for the M/G/1 queue. When $\lambda^+ \neq \lambda^-$, the performance measures presented here not only depend on the first two moment of the service time but also on the LST of the service time at λ^-.

Corollary 1 *We have*

$$E(Q_d) = \frac{\lambda^+\overline{x}(G^*(\lambda^-)((\lambda^-)^2-(\lambda^+)^2)+(\lambda^+)^2(1-\lambda^-\overline{x}))}{(\lambda^+ + (\lambda^- - \lambda^+)G^*(\lambda^-))((\lambda^- - \lambda^+)G^*(\lambda^-)+\lambda^+(1-\lambda^-\overline{x}))} + \frac{(\lambda^+)^2\lambda^-(\overline{x})^2(1+cv^2)}{2((\lambda^- - \lambda^+)G^*(\lambda^-)+\lambda^+(1-\lambda^-\overline{x}))}.$$

Proof. The expected number of customers in the system is equal to $P'(1)$. The expression of $P(z)$ allows us to write $P(z) = N(z)\frac{1-z}{D(z)}$. Therefore, $P'(z) = N'(z)\frac{1-z}{D(z)} - N(z)\frac{D(z)+(1-z)D'(z)}{D^2(z)}$. Using L'Hôpital's rule, one can derive the limit of $\frac{1-z}{D(z)}$ as z tends to 1. For $\frac{D(z)+(1-z)D'(z)}{D^2(z)}$, L'Hôpital's rule should be used twice to obtain the limit of this expression as z tends to 1. This explains why the second moment of the service time is in the expression of $E(Q_d)$. The details of the computation are omitted. □

2.2 Performance analysis at arbitrary instants

We now relate the stationary probabilities at arbitrary instants with those at departure instants in order to use the results of Section 2.1 to obtain the performance of the system. We denote by π_0 the probability of an empty system at arbitrary instants and by $\pi_{n,+}$ and $\pi_{n,-}$ the probability of having n customers in the system after an arrival and after a service respectively ($n \geq 1$) at arbitrary instants.

The queue length distribution is identical at departure instants and arrival instants. The reason is that only one event (an arrival or a service departure) occurs at a time. We therefore have

$$p_0 = \frac{\lambda^-\pi_0}{\lambda^- \sum_{k=0}^{\infty}\pi_{k,-} + \lambda^+\sum_{k=1}^{\infty}\pi_{k,+}}, \tag{7}$$

$$p_n = \frac{\lambda^-\pi_{n,-} + \lambda^+\pi_{n,+}}{\lambda^- \sum_{k=0}^{\infty}\pi_{k,-} + \lambda^+\sum_{k=1}^{\infty}\pi_{k,+}}, \text{ for } n \geq 1. \tag{8}$$

Due to flow conservation, one may write $\lambda^- \sum_{k=0}^{\infty}\pi_{k,-} + \lambda^+\sum_{k=1}^{\infty}\pi_{k,+} = \frac{1}{\overline{x}}(1-\pi_0)$. So, we deduce from Equation (7) that $\pi_0 = \frac{p_0}{p_0+\lambda^-\overline{x}}$. Using now the result of Theorem 1, we deduce that

$$\pi_0 = \frac{\lambda^+(1-\overline{x}\lambda^-) + (\lambda^- - \lambda^+)G^*(\lambda^-)}{\lambda^+ + (\lambda^- - \lambda^+)(1+\overline{x}\lambda^-)G^*(\lambda^-)}. \tag{9}$$

Let us denote by $p_t(n,r,+)$ and $p_t(n,r,-)$ the probability-densities of having n customers in the system after an arrival and after a service respectively, $n \geq 1$ and a remaining service time of r, $r \geq 0$, at time t (given some arbitrary initial distribution). We also define the limit values of these probabilities; $p(n,r,+) = \lim_{t\to\infty} p_t(n,r,+)$ and $p(n,r,-) = \lim_{t\to\infty} p_t(n,r,-)$, for $n \geq 1$. In Lemma 1, we provide the differential equations for $p(n,r,+)$ and $p(n,r,-)$, $r \geq 0$ and $n \geq 1$.

Lemma 1 *For all $r \geq 0$, $p(n,r,+)$ and $p(n,r,-)$ obey the following differential equations*

$$p'(1,r,+) = \lambda^+ p(1,r,+) - \lambda^- \pi_0 g(r), \tag{10}$$

$$p'(n,r,+) = \lambda^+ p(n,r,+) - \lambda^- p(n-1,r,-) - \lambda^+ p(n-1,r,+), \text{ for } n \geq 2, \tag{11}$$

$$p'(n,r,-) = \lambda^- p(n,r,-) - g(r)(p(n+1,0,+) + p(n+1,0,-)), \text{ for } n \geq 1, \tag{12}$$

where $p'(n,r,.) = \frac{\mathrm{d}p(n,r,.)}{\mathrm{d}r}$.

Proof. We will start with the case where $n = 1$ (equation (10)). First, observe that

$$p_{t+\mathrm{d}t}(1,r,+) = (1 - \lambda^+ \mathrm{d}t) p_t(1, r+\mathrm{d}t, +) + \lambda^- \pi_0 g(r) \mathrm{d}t.$$

Taking $t \to \infty$ and dividing by $\mathrm{d}t$ leads to

$$\frac{p(1,r+\mathrm{d}t,+) - p(1,r,+)}{\mathrm{d}t} = \lambda^+ p(1,r+\mathrm{d}t,+) - \lambda^- \pi_0 g(r).$$

Next, taking $\mathrm{d}t \to 0$, we obtain Equation (10). Equation (11) is derived from

$$p_{t+\mathrm{d}t}(n,r,+) = (1-\lambda^+\mathrm{d}t) \cdot p_t(n, r+\mathrm{d}t, +) + \lambda^- \mathrm{d}t \cdot p_t(n-1, r+\mathrm{d}t, -) + \lambda^+ \mathrm{d}t \cdot p_t(n-1, r+\mathrm{d}t, +),$$

and Equation (12) from

$$p_{t+\mathrm{d}t}(n,r,-) = (1-\lambda^-\mathrm{d}t) \cdot p_t(n, r+\mathrm{d}t, -) + g(r)\mathrm{d}t(p(n+1,0,+) + p(n+1,0,-)),$$

with the same approach. □

In Proposition 1, using Lemma 1 we relate $\pi_{n,+}$ and $\pi_{n,-}$ for $n \geq 1$. This proposition proves that the ratio $\pi_{n,+}/\pi_{n,-}$ is constant for $n \geq 1$. This translates that given a non-empty system, the queue length and the last event are independent. This result also holds for the M/G/1 queue.

Proposition 1 *We have $\pi_{n,-} = \frac{\lambda^+}{\lambda^-} \frac{1-G^*(\lambda^-)}{G^*(\lambda^-)} \pi_{n,+}$, for $n \geq 1$.*

Proof. Integrating both sides of Equations (10), (11) and (12) for r from 0 to ∞, we get

$$p(1,0,+) = \lambda^- \pi_0 - \lambda^+ \pi_{1,+}, \tag{13}$$

$$p(n,0,+) = \lambda^- \pi_{n-1,-} + \lambda^+ \pi_{n-1,+} - \lambda^+ \pi_{n,+}, \text{ for } n \geq 2, \tag{14}$$

$$p(n,0,-) = p(n+1,0,+) + p(n+1,0,-) - \lambda^- \pi_{n,-}, \text{ for } n \geq 1. \tag{15}$$

Summing up Equations (14) and (15) yields

$$p(n,0,+) + p(n,0,-) = \lambda^- \pi_{n-1,-} + \lambda^+ \pi_{n-1,+} - \lambda^+ \pi_{n,+} + p(n+1,0,+) + p(n+1,0,-) - \lambda^- \pi_{n,-},$$

for $n \geq 2$. This equation is equivalent to

$$p(n,0,+) + p(n,0,-) - \lambda^- \pi_{n-1,-} - \lambda^+ \pi_{n-1,+} = p(n+1,0,+) + p(n+1,0,-) - \lambda^- \pi_{n,-} - \lambda^+ \pi_{n,+},$$

for $n \geq 2$. Since the difference $p(n,0,+) + p(n,0,-) - \lambda^- \pi_{n-1,-} - \lambda^+ \pi_{n-1,+}$ is not a function of n and tends to zero when n tends to ∞, we get the identity

$$p(n+1,0,+) + p(n+1,0,-) = \lambda^- \pi_{n,-} + \lambda^+ \pi_{n,+}, \tag{16}$$

for $n \geq 1$. Combining Equation (16) with Equation (15) leads to

$$\lambda^+ \pi_{n,+} = p(n,0,-), \tag{17}$$

for $n \geq 1$.

Equation (12) can be written as

$$e^{-\lambda^- u}(p'(n,u,-) - \lambda^- p(n,u,-)) = -g(u)(p(n+1,0,+) + p(n+1,0,-))e^{-\lambda^- u}.$$

Since the left hand side here is $(e^{-\lambda^- u} p(n,u,-))'$, integrating both sides for u from 0 to ∞ yields

$$p(n,0,-) = G^*(\lambda^-)(p(n+1,0,+) + p(n+1,0,-)), \text{ for } n \geq 1, \tag{18}$$

Combining Equations (16) and (18) leads to

$$p(n+1,0,+) + p(n+1,0,-) = \frac{p(n,0,-)}{G^*(\lambda^-)} = \lambda^- \pi_{n,-} + \lambda^+ \pi_{n,+}. \tag{19}$$

Since $\lambda^+ \pi_{n,+} = p(n,0,-)$, we obtain $\pi_{n,-} = \frac{\lambda^+}{\lambda^-} \frac{1 - G^*(\lambda^-)}{G^*(\lambda^-)} \pi_{n,+}$, for $n \geq 1$. □

Using the result of Proposition 1 in Equation (8), we obtain

$$\pi_{n,+} = (1-\pi_0)\frac{G^*(\lambda^-)}{\lambda^+ \overline{x}} p_n, \text{ and } \pi_{n,-} = (1-\pi_0)\frac{1 - G^*(\lambda^-)}{\lambda^- \overline{x}} p_n, \tag{20}$$

for $n \geq 1$. This allows us to derive the generating functions

$$P^+(z) = \sum_{n=1}^{\infty} \pi_{n,+} \cdot z^n \text{ and, } P^-(z) = \sum_{n=0}^{\infty} \pi_{n,-} \cdot z^n.$$

After some algebra, we get

$$P^+(z) = \frac{\lambda^- G^*(\lambda^-)((\lambda^- - \lambda^+)G^*(\lambda^-) + \lambda^+(1-\lambda^-\overline{x}))}{\lambda^+(\lambda^+ + G^*(\lambda^-)(\lambda^- - \lambda^+)(1+\lambda^-\overline{x}))} \times \frac{z(G^*(\lambda^+ - \lambda^+ z) - 1)(\lambda^+(1-z) + \lambda^-)}{(1-z)\left((\lambda^+ - \lambda^-)G^*(\lambda^-) - \lambda^+ z\right) + \lambda^- z(1 - G^*(\lambda^+ - \lambda^+ z))},$$

$$P^-(z) = \frac{(1-G^*(\lambda^-))((\lambda^- - \lambda^+)G^*(\lambda^-) + \lambda^+(1-\lambda^-\overline{x}))}{\lambda^+ + G^*(\lambda^-)(\lambda^- - \lambda^+)(1+\lambda^-\overline{x})} \times \frac{(1-z)\left((\lambda^+ - \lambda^-)G^*(\lambda^-) - \lambda^+ z G^*(\lambda^+ - \lambda^+ z)\right)}{(1-z)\left((\lambda^+ - \lambda^-)G^*(\lambda^-) - \lambda^+ z\right) + \lambda^- z(1 - G^*(\lambda^+ - \lambda^+ z))}.$$

In Corollary 2, we deduce the expected number of customers in the system, $E(Q)$, and the expected waiting time, $E(W)$, at arbitrary instants. We also derive the expected remaining service time seen by an arriving customer at a non-empty system, $E(R)$.

Corollary 2 *We have*

$$\frac{E(Q)}{\overline{x}(\lambda^+ + G^*(\lambda^-)(\lambda^- - \lambda^+))} = \frac{\overline{x}\lambda^+\lambda^-\left[(1+cv^2)(\lambda^+ + G^*(\lambda^-)(\lambda^- - \lambda^+)) - 2\lambda^+\right] + 2\left[(\lambda^-)^2 G^*(\lambda^-) + (\lambda^+)^2(1 - G^*(\lambda^-))\right]}{2\left[\lambda^+ + (\lambda^- - \lambda^+)(1+\overline{x}\lambda^-)G^*(\lambda^-)\right]\left[-\overline{x}\lambda^+\lambda^- + \lambda^+ + G^*(\lambda^-)(\lambda^- - \lambda^+)\right]}, \tag{21}$$

$$E(W) = \frac{\overline{x}\lambda^+\lambda^-\left[(1+cv^2)(\lambda^+ + G^*(\lambda^-)(\lambda^- - \lambda^+)) + 2(\lambda^- - \lambda^+)\right] - 2\lambda^+(\lambda^- - \lambda^+)(1 - G^*(\lambda^-))}{2\lambda^-\left[-\overline{x}\lambda^+\lambda^- + \lambda^+ + G^*(\lambda^-)(\lambda^- - \lambda^+)\right]}\overline{x}, \text{ and,} \tag{22}$$

$$E(R) = \frac{(1+cv^2)\overline{x}(\lambda^+ + G^*(\lambda^-)(\lambda^- - \lambda^+))}{2\lambda^-} - \frac{(\lambda^- - \lambda^+)(1 - G^*(\lambda^-) - \lambda^-\overline{x})}{(\lambda^-)^2}. \tag{23}$$

Proof. The expected number of customers in the system is $E(Q) = \sum_{n=1}^{\infty} n(\pi_{n,+} + \pi_{n,-})$. Using Equation (20), we deduce that

$$E(Q) = \frac{1-\pi_0}{\overline{x}}\left(\frac{G^*(\lambda^-)}{\lambda^+} + \frac{1-G^*(\lambda^-)}{\lambda^-}\right)\sum_{n=1}^{\infty} n p_n = \frac{1-\pi_0}{\overline{x}}\left(\frac{G^*(\lambda^-)}{\lambda^+} + \frac{1-G^*(\lambda^-)}{\lambda^-}\right)E(Q_d).$$

This leads to the expression of $E(Q)$. The throughput of served customers is $\frac{1}{\overline{x}}(1-\pi_0)$. Applying Little's law, we get the expected time spent in the system for a given customer by dividing the expected number in the system by $\frac{1}{\overline{x}}(1-\pi_0)$. Finally, subtracting $\overline{x}$ to this expression leads to the expected waiting time in the queue. As mentioned above, the queue length distribution is identical at departure instants and arrival instants. Hence, the expected waiting time can be written as a function of $E(R)$ using the stationary probabilities at arrival instants, p_n, for $n \geq 1$;

$$E(W) = \sum_{n=1}^{\infty} p_n\left((n-1)\overline{x} + E(R)\right). \tag{24}$$

This leads to $E(R) = \overline{x} + \frac{E(W) - \overline{x}E(Q_d)}{1-p_0}$. Using the explicit expressions of $E(W)$, $E(Q_d)$ and p_0, the expression of $E(R)$ can be derived. □

2.3 Special cases

2.3.1 Exponential case

In Proposition 2, we give the performance measures associated to an exponential service time distribution.[1] We denote by a^+ and a^- the products $a^+ = \overline{x} \cdot \lambda^+$ and $a^- = \overline{x} \cdot \lambda^-$.

Proposition 2 *Under the stability condition $a^- a^+ < 1$, we have*

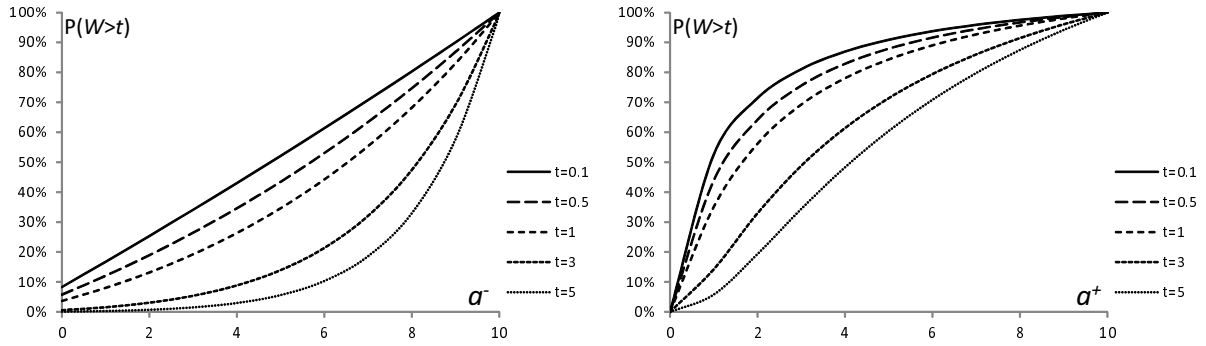
$$\pi_0 = \frac{1 - a^- a^+}{1 + a^-}, \tag{25}$$

$$E(W) = \overline{x} \frac{a^+(1 + a^-)}{1 - a^- a^+}, \text{ and} \tag{26}$$

$$P(W > t) = a^+ \frac{1 + a^-}{1 + a^+} e^{-\frac{t}{\overline{x}} \frac{1 - a^- a^+}{1 + a^+}}, \tag{27}$$

for $t > 0$.

In Figure 1, we illustrate the impact of λ^- and λ^+ on the waiting time distribution. We observe that when the arrival rate after an arrival is higher than the arrival rate after a service then the system performance deteriorates in comparison with an M/M/1 queue. Most of the observed monotonicity results are intuitive since they



(a) Impact of λ^- ($\lambda^+ = 0.1$)

(b) Impact of λ^+ ($\lambda^- = 0.1$)

Fig. 1 Impact of λ^+ and λ^- on $P(W > t)$ ($\overline{x} = 1$)

correspond to the impact of the arrival parameter for an M/M/1 queue. The impact of a^+ is more surprising. The probability $P(W > t)$ can be concave in a^+ for $t > 0$. Note that we have $\lim_{t \to 0} P(W > t) = a^+ \frac{1 + a^-}{1 + a^+}$. The variable a^- is only in the numerator of this expression. This explains the almost linear form of the curves of $P(W > t)$ as functions of a^- when t is small (Figure 1(a)).

[1] The performance measures in the exponential case could also be obtained using a Markov chain analysis.

2.3.2 Erlang case

In Proposition 3, we give the performance measures associated to an Erlang service time distribution with N exponential phases where each phase has an expected duration of $\frac{\overline{x}}{N}$.[2]

Proposition 3 *Under the stability condition* $a^-a^+ < a^-\left(\frac{N}{a^-+N}\right)^N + a^+\left(1-\left(\frac{N}{a^-+N}\right)^N\right)$, *we have*

$$\pi_0 = \frac{a^+(1-a^-)+(a^--a^+)\left(\frac{N}{N+a^-}\right)^N}{a^++(a^--a^+)(1+a^-)\left(\frac{N}{N+a^-}\right)^N}, \text{ and}$$

$$E(W) = \overline{x}\frac{a^+a^-\left(\left(\frac{1}{N}+1\right)\left((a^--a^+)\left(\frac{N}{N+a^-}\right)^N+a^+\right)+2\left(a^--a^+\right)\right)-2a^+\left(a^--a^+\right)\left(1-\left(\frac{N}{N+a^-}\right)^N\right)}{2a^-\left(\left(a^--a^+\right)\left(\frac{N}{N+a^-}\right)^N-a^+a^-+a^+\right)}.$$

The impact of N on the stability of the system depends on the difference $\lambda^+ - \lambda^-$. Consider the function $F(N) = \left(\frac{N}{a^-+N}\right)^N(a^- - a^+) + a^+(1-a^-)$. We have $F(N) > 0$ if and only if $a^-a^+ < a^-\left(\frac{N}{a^-+N}\right)^N + a^+\left(1-\left(\frac{N}{a^-+N}\right)^N\right)$. So, the system is stable if and only if $F(N) > 0$. Since the function $\left(\frac{N}{a^-+N}\right)^N$ is decreasing in N, F is decreasing in N if and only if $\lambda^+ < \lambda^-$. In other words, increasing N stabilizes the system if and only if $\lambda^+ > \lambda^-$.

Finally, we can conclude that the system is stable for all Erlang service time distributions with expected duration $\overline{x}$ if and only if

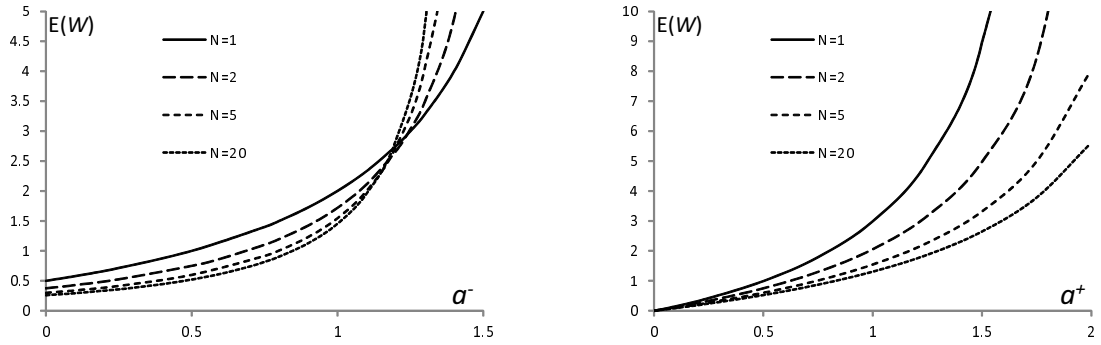
$$\begin{cases} a^-a^+ < 1 \\ a^+ \geq a^- \end{cases} \text{ or } \begin{cases} e^{-a^-} > \frac{(a^--1)a^+}{a^--a^+} \\ a^+ < a^- \end{cases}.$$

These relations are obtained either by choosing $N = 1$ (exponential distribution) or by letting $N \to \infty$ (deterministic distribution).

We now evaluate the impact of the number of phases on $E(W)$ for different values of λ^+ and λ^- (Figure 2). We observe that the number of phases can deteriorate the expected waiting time when $\lambda^- > \lambda^+$ (Figure 2(a)). This result is surprising since it is in contradiction with the improvement which should result from the reduction of the variability in the service process when the number of phases increases. The reason is related to the presence of the term $G^*(\lambda^-)$ in the expression of $E(W)$ which involves the service time distribution. The improvement related to the decreasing of the coefficient of variability when N increases can be compensated by an increasing number of arrivals during service.

The expected number of customers arriving during a service has already been computed in Section 2.1. It is either equal to $A'(1)$ or to $B'(1)$. The expression of $A'(1)$ explains the position of the curves in Figure 2. Since $G^*(\lambda^-)$ decreases as N increases, the expected number of arrivals during service increases as N increases in the case $\lambda^- > \lambda^+$. This is consistent with the conclusion derived above for the stability region; if $\lambda^- > \lambda^+$, increasing N reduces the stability region.

[2] Another way to compute the performance measures in the Erlang case is to use a Matrix geometric approach.



(a) Impact of λ^- ($\lambda^+ = 0.5$)

(b) Impact of λ^+ ($\lambda^- = 0.5$)

Fig. 2 Impact of λ^+ and λ^- on $E(W)$ ($\overline{x} = 1$)

2.3.3 Deterministic case

By letting $r \to \infty$ in Proposition 3 we obtain the performance measures in the case of a deterministic service with duration $\overline{x}$.

Proposition 4 *Under the stability condition* $a^-a^+ < a^+ + e^{-a^-}(a^- - a^+)$*, we have*

$$\pi_0 = \frac{a^+(1-a^-) + (a^- - a^+)e^{-a^-}}{a^+ + (a^- - a^+)(1+a^-)e^{-a^-}}, \text{ and}$$

$$E(W) = \overline{x}a^+ \frac{(a^-)^2 + (a^- - a^+)(a^- - 2 + (a^- + 2)e^{-a^-})}{2a^- \left((a^- - a^+)\, e^{-a^-} - a^+a^- + a^+\right)}.$$

3 Equilibrium Strategy

An interesting question is to determine if the event-dependency of the arrival process can be the result of an individual rational strategy.[3] For this purpose, we are interested in the service level differentiation after an arrival or after a service. The idea is to determine how does the nature of the last event give an indication on the expected waiting time encountered by an arriving customer. If the last event is a service, an arriving customer may arrive in an empty system. This cannot be the case if the last event is an arrival. This gives an advantage to the customers who arrive after a service. However, given a non-empty system, if the last event is a service, an arriving customer is the first to arrive during a service time. Customers who arrive after an arrival arrive later during the service time. The order of arrival during a service may also influence the expected remaining service time seen by an arriving customer. For instance, with a deterministic service time, the first customer who arrives during a service has a longer expected remaining service time than the next customers. In order to compare the expected waiting time after an arrival or a service, we first evaluate in Section 3.1 the expected remaining service times after an arrival or a after service seen by an arriving customer. In Section 3.2, we derive the expected waiting time of a customer who arrives after an arrival, and the one of a customer who arrives after a service, and compare between them.

[3] We refer the reader to the book of [12] for an overview on equilibrium behavior in queueing systems.

3.1 Expected remaining service time

We are interested in the expected remaining service of an arriving customer. This metric is function of the last realized event and of the number of customers present in the system. We denote by $\overline{r_{n,+}}$ and $\overline{r_{n,-}}$ the expected remaining service time seen by a customer who arrives after an arrival or a service respectively when n customers are present in the system, for $n \geq 1$. We denote by $P^*(n,s,+)$ and $P^*(n,s,-)$ the Laplace-Stieltjes Transform (LST) of $p(n,r,+)$ and $p(n,r,-)$; $P^*(n,s,+) = \int_0^\infty p(n,r,+)e^{-sr}\,\mathrm{d}r$, and $P^*(n,s,+) = \int_0^\infty p(n,r,+)e^{-sr}\,\mathrm{d}r$. We have

$$\overline{r_{n,+}} = \frac{1}{\pi_{n,+}} \int_0^\infty rp(n,r,+)\,\mathrm{d}r = -\lim_{s\to 0} \frac{P'^*(n,s,+)}{\pi_{n,+}},$$

and,

$$\overline{r_{n,-}} = \frac{1}{\pi_{n,-}} \int_0^\infty rp(n,r,-)\,\mathrm{d}r = -\lim_{s\to 0} \frac{P'^*(n,s,-)}{\pi_{n,-}},$$

for $n \geq 1$. In Proposition 5, we provide a recursive formula for the conditional distributions of the residual service times. These relations allows us to compute $\overline{r_{n,+}}$ and $\overline{r_{n,-}}$ for $n \geq 1$. In addition, we show in Section 1 of the online supplement how these relations can be used to obtain the performance measures already found in Section 2.

Proposition 5 *Under the stability condition* $\overline{x} < \frac{1}{\lambda^-}(1-G^*(\lambda^-)) + \frac{1}{\lambda^+}G^*(\lambda^-)$*, we have the following initial relations:*

$$\pi_{1,+} = \frac{\lambda^-}{\lambda^+}(1 - G^*(\lambda^+))\pi_0, \tag{28}$$

$$P^*(1,s,+) = \frac{\lambda^+}{1-G^*(\lambda^+)} \frac{G^*(\lambda^+) - G^*(s)}{s-\lambda^+}\pi_{1,+}. \tag{29}$$

Next, for $n \geq 1$*,*

$$P^*(n,s,-) = \frac{\lambda^-}{1-G^*(\lambda^-)} \frac{G^*(\lambda^-) - G^*(s)}{s-\lambda^-}\pi_{n,-}, \tag{30}$$

$$\pi_{n+1,+} = \frac{\pi_{n,+}}{G^*(\lambda^-)} - \frac{\lambda^-}{\lambda^+}P^*(n,\lambda^+,-) - P^*(n,\lambda^+,+), \tag{31}$$

$$P^*(n+1,s,+) = \lambda^- \frac{P^*(n,\lambda^+,-) - P^*(n,s,-)}{s-\lambda^+} + \lambda^+ \frac{P^*(n,\lambda^+,+) - P^*(n,s,+)}{s-\lambda^+}. \tag{32}$$

Proof. First, Equation (10) can be written as

$$e^{-\lambda^+ u}(p'(1,u,+) - \lambda^+ p(1,u,+)) = -\lambda^- e^{-\lambda^+ u}\pi_0 g(u).$$

Since the left hand side here is $(e^{-\lambda^+ u}p(1,u,+))'$, integrating both sides for u from r to ∞ leads to

$$p(1,r,+) = \lambda^- \pi_0 e^{\lambda^+ r} \int_{u=r}^{\infty} e^{-\lambda^+ u} g(u)\,\mathrm{d}u. \tag{33}$$

Inserting $r = 0$ in (33) leads to

$$p(1,0,+) = \lambda^- \pi_0 G^*(\lambda^+). \tag{34}$$

Combining now Equation (13) of Section 2 and Equation (34), leads to $\pi_{1,+} = \frac{\lambda^-}{\lambda^+}(1 - G^*(\lambda^+))\pi_0$ (Equation (28)).

From Equation (10), one may write

$$\int_{r=0}^{\infty} e^{-sr} p'(1,r,+)\,\mathrm{d}r = \lambda^+ \int_{r=0}^{\infty} e^{-sr} p(1,r,+)\,\mathrm{d}r - \lambda^- \pi_0 \int_{r=0}^{\infty} e^{-sr} g(r)\,\mathrm{d}r.$$

This equation leads to

$$sP^*(1,s,+) - p(1,0,+) = \lambda^+ P^*(1,s,+) - \lambda^- \pi_0 G^*(s).$$

Since $p(1,0,+) = \lambda^- \pi_0 G^*(\lambda^+)$ (Equation (34)), we obtain $P^*(1,s,+) = \lambda^- \frac{G^*(\lambda^+) - G^*(s)}{s - \lambda^+}\pi_0$. Finally, Equation (28) leads to Equation (29).

From Equation (12), we deduce that

$$\int_{r=0}^{\infty} e^{-sr} p'(n,r,-)\,\mathrm{d}r = \lambda^- \int_{r=0}^{\infty} e^{-sr} p(n,r,-)\,\mathrm{d}r - (p(n+1,0,+) + p(n+1,0,-)) \int_{r=0}^{\infty} e^{-sr} g(r)\,\mathrm{d}r,$$

for $n \geq 1$. Thus,

$$sP^*(n,s,-) - p(n,0,-) = \lambda^- P^*(n,s,-) - (p(n+1,0,+) + p(n+1,0,-))G^*(s). \tag{35}$$

Combining Equation (19) and Proposition 1 leads to $p(n+1,0,+) + p(n+1,0,-) = \frac{\lambda^+ \pi_{n,+}}{G^*(\lambda^-)}$. Using now Equation (17) with Proposition 1 in Equation (35), we obtain Equation (30).

With the same approach as for Equations (33) and (34), we obtain

$$p(n,r,+) = e^{\lambda^+ r} \int_{u=r}^{\infty} e^{-\lambda^+ u}(\lambda^- p(n-1,u,-) + \lambda^+ p(n-1,u,+))\,\mathrm{d}u, \text{ for } n \geq 2, \tag{36}$$

$$p(n,0,+) = \lambda^- P^*(n-1,-,\lambda^+) + \lambda^+ P^*(n-1,+,\lambda^+), \text{ for } n \geq 2, \tag{37}$$

using Equation (11). Combining Equation (37) and Equation (14) leads to

$$\lambda^+ \pi_{n,+} = \lambda^- \pi_{n-1,-} + \lambda^+ \pi_{n-1,+} - (\lambda^- P^*(n-1,-,\lambda^+) + \lambda^+ P^*(n-1,+,\lambda^+)),$$

for $n \geq 2$. With Equation (19), $\lambda^- \pi_{n-1,-} + \lambda^+ \pi_{n-1,+} = p(n,0,+) + p(n,0,-)$. Combining Equation (18) with Equation (17) leads to $p(n,0,+) + p(n,0,-) = \frac{p(n-1,0,-)}{G^*(\lambda^-)} = \frac{\lambda^+ \pi_{n-1,+}}{G^*(\lambda^-)}$. The aforementioned relations prove Equation (31).

From Equation (11), we have

$$\begin{aligned}\int_{r=0}^{\infty} e^{-sr} p'(n,r,+)\,\mathrm{d}r =& \lambda^+ \int_{r=0}^{\infty} e^{-sr} p(n,r,+)\,\mathrm{d}r - \lambda^- \int_{r=0}^{\infty} e^{-sr} p(n-1,r,-)\,\mathrm{d}r \\ & - \lambda^+ \int_{r=0}^{\infty} e^{-sr} p(n-1,r,+)\,\mathrm{d}r,\end{aligned}$$

for $n \geq 2$. The equality is equivalent to

$$sP^*(n,s,+) - p(n,0,+) = \lambda^+ P^*(n,s,+) - \lambda^- P^*(n-1,s,-) - \lambda^+ P^*(n-1,s,+).$$

Finally, Equation (37) leads to Equation (32). □

The consequence of Equation (30) is that the expected remaining service time of a customer who arrives in a non-empty system after a service does not depend on the system size. This property allows us to explicitly derive in Corollary 3 the expected remaining service time seen at arrival for a customer who arrives after a service, $E(R^-)$, and after an arrival, $E(R^+)$.

Corollary 3 *We have*

$$E(R^-) = \frac{\overline{x}}{1 - G^*(\lambda^-)} - \frac{1}{\lambda^-}, \text{ and,} \tag{38}$$

$$E(R^+) = \frac{\lambda^- \overline{x}(1+cv^2)(\lambda^+(1-G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)) - 2\lambda^+(\lambda^- \overline{x} + G^*(\lambda^-) - 1)}{2(\lambda^-)^2 G^*(\lambda^-)}. \tag{39}$$

Proof. We have $\overline{r_{n,-}} = -\lim_{s \to 0} \frac{P'^*(n,s,-)}{\pi_{n,-}}$, for $n \geq 1$. Using Equation (30), we get $\overline{r_{n,-}} = \frac{\overline{x}}{1-G^*(\lambda^-)} - \frac{1}{\lambda^-}$, for $n \geq 1$. Therefore, the expected remaining service time seen by an arriving customer after a service does not depend on n. Hence, $E(R^-) = \overline{r_{n,-}}$. The same property does not hold for $\overline{r_{n,+}}$. However, we can compute $E(R^+)$ using the following decomposition:

$$E(W) = \sum_{n=1}^{\infty} ((n-1)\overline{x} + E(R^+)) p_{n,+} + \sum_{n=1}^{\infty} ((n-1)\overline{x} + E(R^-)) p_{n,-}, \tag{40}$$

where $p_{n,+}$ and $p_{n,-}$ are the stationary probabilities to have n customers in the system at arrival instants after an arrival or a service ($p_n = p_{n,+} + p_{n,-}$, for $n \geq 1$). This leads to

$$E(R^+) = \overline{x} + \frac{E(W) - \overline{x}E(Q_d) - (E(R^-) - \overline{x}) \sum_{n=1}^{\infty} p_{n,-}}{\sum_{n=1}^{\infty} p_{n,+}}.$$

All parts of this expression are know except $\sum_{n=1}^{\infty} p_{n,-}$ and $\sum_{n=1}^{\infty} p_{n,+}$. In what follows we derive these metrics. We have $p_{n,-} = \frac{\lambda^- \pi_{n,-}}{\lambda^- \sum_{n=0}^{\infty} \pi_{n,-} + \lambda^+ \sum_{n=1}^{\infty} \pi_{n,+}}$. The denominator of this expression is $\frac{1-\pi_0}{\overline{x}}$ due to flow conservation. Next, from Equation (20), we obtain

$$\sum_{n=1}^{\infty} \pi_{n,-} = (1-\pi_0) \frac{1-G^*(\lambda^-)}{\lambda^- \overline{x}} \sum_{n=1}^{\infty} p_n.$$

Therefore, $\sum_{n=1}^{\infty} p_{n,-} = \frac{\lambda^-(1-\pi_0)\frac{1-G^*(\lambda^-)}{\lambda^-\overline{x}}(1-p_0)}{\frac{1-\pi_0}{\overline{x}}}$. Using the expression of p_0 in Theorem 1, we obtain

$$\sum_{n=1}^{\infty} p_{n,-} = \frac{\lambda^+ \lambda^- \overline{x}(1-G^*(\lambda^-))}{\lambda^- G^*(\lambda^-) + \lambda^+(1-G^*(\lambda^-))}.$$

With the same approach, we obtain $\sum_{n=1}^{\infty} p_{n,+} = \frac{\lambda^+\lambda^-\overline{x}G^*(\lambda^-)}{\lambda^- G^*(\lambda^-)+\lambda^+(1-G^*(\lambda^-))}$. This finishes the proof. □

Remark. The independence between the system size and the remaining service time for a customer who arrives after a service is not surprising. As mentioned above, a customer who arrives after a service in a non-empty system is the first to arrive during a service irrespective of the number of customers already present. For a customer who arrives after an arrival, the number of customers present in the system has an impact. For instance, if a customer arrives after an arrival when one customer is already in the system, then it means that the customer in service arrived in an empty system. So, $\overline{r_{1,+}}$ is the expected remaining service time of the first customer who arrives during a service. As expected, with Equation (29), we obtain $\overline{r_{1,+}} = \frac{\overline{x}}{1-G^*(\lambda^+)} - \frac{1}{\lambda^+}$. This is exactly the expression of $E(R^-)$ by replacing λ^- by λ^+. Consider now the same situation when two customers are already in the system. In this case the arrived customer cannot be the first to arrive during a service, so the expression of $\overline{r_{2,+}}$ may differ from the one of $\overline{r_{1,+}}$.

3.2 Expected waiting time

In Theorem 2, we compare between the expected waiting times after an arrival and after a service. Let us denote by $E(W^+)$, $E(W^-)$, and $E(W^-|\text{non-empty})$ the expected waiting times of a customer who arrives after an arrival, after a service, and after a service given a non-empty system.

Theorem 2 *The following holds.*

1. $E(W^+) \geq E(W^-|\textit{non-empty})$ *if and only if* $\frac{cv^2+1}{2} \geq \frac{E(R^-)}{\overline{x}}$,
2. $E(W^+) \geq E(W^-)$.

Proof. Let us start with the first statement. Proposition 1 proves that the ratio $\pi_{n,+}/\pi_{n,-}$ is constant for $n \geq 1$. Since $p_{n,+} = \frac{\lambda^+ \pi_{n,+}}{\lambda^- \sum_{n=0}^{\infty} \pi_{n,-} + \lambda^+ \sum_{n=1}^{\infty} \pi_{n,+}}$ and $p_{n,-} = \frac{\lambda^- \pi_{n,-}}{\lambda^- \sum_{n=0}^{\infty} \pi_{n,-} + \lambda^+ \sum_{n=1}^{\infty} \pi_{n,+}}$, the ratio $p_{n,+}/p_{n,-}$ is also constant for $n \geq 1$. This translates that given a non-empty system, the queue length at customers' arrival

and the last event are independent. This result is important for the comparison between the expected waiting times after an arrival or after a service given a non-empty system. It means that they only differ in their remaining service times. This can also be shown via the following decomposition:

$$E(W^+) = \frac{\sum_{n=1}^{\infty} p_{n,+}((n-1)\overline{x} + E(R^+))}{\sum_{n=1}^{\infty} p_{n,+}} = E(R^+) - \overline{x} + E(Q_d)\frac{\lambda^+(1-G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)}{\lambda^+\lambda^-}, \text{ and,}$$

$$E(W^-|\text{non-empty}) = \frac{\sum_{n=1}^{\infty} p_{n,-}((n-1)\overline{x} + E(R^-))}{\sum_{n=1}^{\infty} p_{n,-}} = E(R^-) - \overline{x} + E(Q_d)\frac{\lambda^+(1-G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)}{\lambda^+\lambda^-}.$$

Therefore, by comparing $E(R^+)$ and $E(R^-)$ using their expressions in Corollary 3, we get the condition of the first statement.

Let us now consider the second statement. The order of arrival during a service determines the comparison between $E(W^+)$ and $E(W^-|\text{non-empty})$ (first statement). The comparison between $E(W^+)$ and $E(W^-)$ also involves the probability to arrive in an empty system. This makes the comparison more complex since two phenomena are involved. In this case, the explicit expressions of the expected waiting times are required for the comparison. These expressions are computed in a similar way as $E(R^+)$ and $E(R^-)$ in Corollary 3. We have

$$E(W^+)$$
$$= \frac{\left[\lambda^+(1-G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)\right]\left[\lambda^-\overline{x}(1+cv^2)(\lambda^+(1-\lambda^-\overline{x}) + G^*(\lambda^-)(\lambda^- - \lambda^+ + \lambda^+\lambda^-\overline{x})) - 2\lambda^+(1-\lambda^-\overline{x})(\lambda^-\overline{x} + G^*(\lambda^-) - 1)\right]}{2(\lambda^-)^2 G^*(\lambda^-)(\lambda^+(1-\lambda^-\overline{x}) + (\lambda^- - \lambda^+)G^*(\lambda^-))},$$

$$E(W^-|\text{non-empty})$$
$$= \frac{\left[\lambda^+(1-G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)\right]\left[\lambda^-\lambda^+\overline{x}^2(1-G^*(\lambda^-))(1+cv^2) + 2(1-\lambda^+\overline{x})(\lambda^-\overline{x} + G^*(\lambda^-) - 1)\right]}{2\lambda^-(1-G^*(\lambda^-))(\lambda^+(1-\lambda^-\overline{x}) + (\lambda^- - \lambda^+)G^*(\lambda^-))}, \text{ and,}$$

$$E(W^-)$$
$$= \frac{\lambda^+\overline{x}\left[\lambda^+(1-G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)\right]\left[\lambda^-\lambda^+\overline{x}^2(1-G^*(\lambda^-))(1+cv^2) + 2(1-\lambda^+\overline{x})(\lambda^-\overline{x} + G^*(\lambda^-) - 1)\right]}{2(\lambda^+(1-\lambda^-\overline{x}G^*(\lambda^-)) + (\lambda^- - \lambda^+)G^*(\lambda^-))(\lambda^+(1-\lambda^-\overline{x}) + (\lambda^- - \lambda^+)G^*(\lambda^-))}.$$

One then may write

$$E(W^+) - E(W^-)$$
$$= \frac{(\lambda^+(1-G^*(\lambda^-)) + \lambda^- G^*(\lambda^-))(\lambda^-\overline{x}(1+cv^2)(\lambda^+(1-G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)) + 2\lambda^+(1-\lambda^-\overline{x} - G^*(\lambda^-)))}{2(\lambda^-)^2 G^*(\lambda^-)(\lambda^+(1-G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)(1-\lambda^+\overline{x}))}$$

The sign of the denominator depends on the sign of $\lambda^+(1-G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)(1-\lambda^+\overline{x})$. This expression can be rewritten as $\lambda^+\lambda^-\overline{x}(1-G^*(\lambda^-)) + \lambda^+(1-\overline{x}\lambda^- - G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)$. Since $G^*(\lambda^-) \leq 1$, $\lambda^+\lambda^-\overline{x}(1-G^*(\lambda^-)) \geq 0$. Next, we have

$$1 - \overline{x}\lambda^- - G^*(\lambda^-) = \int_0^\infty g(t)(1 - \lambda^- t - e^{-\lambda^- t})\,\mathrm{d}t.$$

Let us define $\phi_1(t) = 1-\lambda^- t - e^{-\lambda^- t}$, for $t \geq 0$. We have $\phi_1'(t) = -\lambda^-(1-e^{-\lambda^- t}) \leq 0$. So, $\phi_1(t)$ is decreasing in t. Moreover, $\phi_1(0) = 0$ then $\phi_1(t) \leq 0$, for $t \geq 0$. This proves that $1 - \overline{x}\lambda^- - G^*(\lambda^-) \leq 0$. Hence, $\lambda^+(1 - \overline{x}\lambda^- - G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)$ is decreasing in λ^+. The stability condition in Theorem 1 is equivalent to

$$\lambda^+ < \frac{\lambda^- G^*(\lambda^-)}{\overline{x}\lambda^- + G^*(\lambda^-) - 1}. \tag{41}$$

As λ^+ tends to $\frac{\lambda^- G^*(\lambda^-)}{\overline{x}\lambda^- + G^*(\lambda^-)-1}$, $\lambda^+(1 - \overline{x}\lambda^- - G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)$ tends to 0. This proves that $\lambda^+(1 - G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)(1 - \lambda^+\overline{x}) \geq 0$.

Consider now the numerator. Since $0 \leq G^*(\lambda^-) \leq 1$, $\lambda^+(1 - G^*(\lambda^-)) + \lambda^- G^*(\lambda^-) \geq 0$. Next, the inequality $\lambda^-\overline{x}(1 + cv^2)(\lambda^+(1 - G^*(\lambda^-)) + \lambda^- G^*(\lambda^-)) + 2\lambda^+(1 - \lambda^-\overline{x} - G^*(\lambda^-)) \geq 0$ is equivalent to

$$\frac{cv^2 + 1}{2} \geq \frac{\lambda^+(G^*(\lambda^-) + \lambda^-\overline{x} - 1)}{\lambda^-\overline{x}(\lambda^- G^*(\lambda^-) + \lambda^+(1 - G^*(\lambda^-))}. \tag{42}$$

We define the function in λ^+, $\psi(\lambda^+) = \frac{\lambda^+(G^*(\lambda^-)+\lambda^-\overline{x}-1)}{\lambda^-\overline{x}(\lambda^- G^*(\lambda^-)+\lambda^+(1-G^*(\lambda^-))}$. We have $\psi'(\lambda^+) = \frac{G^*(\lambda^-)(G^*(\lambda^-)+\lambda^-\overline{x}-1)}{\overline{x}(\lambda^- G^*(\lambda^-)+\lambda^+(1-G^*(\lambda^-))^2} \geq 0$, since we have $G^*(\lambda^-) + \lambda^-\overline{x} - 1 \geq 0$. So ψ is increasing in λ^+. As λ^+ tends to $\frac{\lambda^- G^*(\lambda^-)}{\overline{x}\lambda^- + G^*(\lambda^-)-1}$ (upper bound for λ^+, see Equation (41)), $\psi(\lambda^+)$ tends to $\frac{G^*(\lambda^-)+\lambda^-\overline{x}-1}{(\lambda^-\overline{x})^2}$. Therefore, $\psi(\lambda^+) \leq \frac{G^*(\lambda^-)+\lambda^-\overline{x}-1}{(\lambda^-\overline{x})^2}$.

Consider now $\frac{cv^2+1}{2} - \frac{G^*(\lambda^-)+\lambda^-\overline{x}-1}{(\lambda^-\overline{x})^2} = \frac{1-\lambda^-\overline{x}+\frac{(\lambda^-)^2}{2}E(S^2)-G^*(\lambda^-)}{(\lambda^-\overline{x})^2}$, where $E(S^2) = \int_0^\infty t^2 g(t)\,\mathrm{d}t$. We have

$$1 - \lambda^-\overline{x} + \frac{(\lambda^-)^2}{2}E(S^2) - G^*(\lambda^-) = \int_0^\infty \left(1 - \lambda^- t + \frac{(\lambda^-)^2}{2}t^2 - e^{-\lambda^- t}\right) g(t)\,\mathrm{d}t.$$

We define $\phi_2(t) = 1 - \lambda^- t + \frac{(\lambda^-)^2}{2}t^2 - e^{-\lambda^- t}$, for $t \geq 0$. We have $\phi_2'(t) = -\lambda^-\phi_1(t) \geq 0$. So $\phi_2(t)$ is increasing in t. Moreover, $\phi_2(0) = 0$. So, $\phi_2(t) \geq 0$, for $t \geq 0$ and $1 - \lambda^-\overline{x} + \frac{(\lambda^-)^2}{2}E(S^2) - G^*(\lambda^-) \geq 0$. This proves that $0 \leq \frac{cv^2+1}{2} - \frac{G^*(\lambda^-)+\lambda^-\overline{x}-1}{(\lambda^-\overline{x})^2} \leq \frac{cv^2+1}{2} - \psi(\lambda^+)$. Hence, Inequality (42) holds in all cases. This finishes the proof of the second statement. □

For the first statement, the condition $\frac{cv^2+1}{2} \geq \frac{E(R^-)}{\overline{x}}$ reveals the importance of the service variability in the performance comparison. With high variability, the indication that the last event was an arrival is a signal that the expected waiting time may be longer than if the last event was a service.

Examples.

1. The service time follows an exponential distribution. Then, $E(R^-) = \overline{x}$ and $cv = 1$. So $E(W^+) = E(W^-|\text{non-empty})$.
2. The service time follows a deterministic distribution. Then $\frac{E(R^-)}{\overline{x}} = \frac{1}{1-e^{-\lambda^-\overline{x}}} - \frac{1}{\lambda^-\overline{x}}$ and $cv = 0$. We can show that $\frac{1}{1-e^{-\lambda^-\overline{x}}} - \frac{1}{\lambda^-\overline{x}} \geq 1/2$. This proves that $E(W^+) \leq E(W^-|\text{non-empty})$.
3. The service time follows a particular hyperexponential distribution for which a customer is either served with an exponential duration with rate μ_0 with probability q or is instantaneously served with probability

$1-q$. We have $G^*(\lambda^-) = \frac{q\mu_0}{\mu_0+\lambda^-}$, $\overline{x} = \frac{q}{\mu_0}$ and $cv^2 = \frac{2}{q} - 1$. Then, $\frac{cv^2+1}{2} - \frac{E(R^-)}{\overline{x}} = \frac{(1-q)\left(\frac{\mu_0}{\lambda^-}+1\right)^2}{q\left(1+\frac{\mu_0}{\lambda^-}(1-q)\right)} \geq 0$. So, $E(W^+) \geq E(W^-|\text{non-empty})$.

The second statement indicates that in all cases, the expected waiting time of an arriving customer after an arrival is longer than the expected time after a service. This might be surprising. Given the first statement, one could imagine that some counterexamples could be found for instance in cases with a low service variability and a high workload situation.

We now explain how the model studied in the article can be the result of a strategic behavior. Consider an initial unobservable system with a potential arrival rate λ. At arrival, a customer can decide to join the queue or not to join. Although the system is unobservable, an arriving customer is informed about the last realized event; an arrival or a service completion. We define the net benefit for a customer who joins by the value of service, B, minus the cost of waiting proportional to a waiting cost per time unit, C. Given that the expected waiting time of an arriving customer is different whether the last event was an arrival or a service, a strategy after an arrival or after a service can be described by two different probabilities of joining $\widetilde{p}^+$ and $\widetilde{p}^-$, such that $\lambda^+ = \widetilde{p}^+ \cdot \lambda$ and $\lambda^- = \widetilde{p}^- \cdot \lambda$. The possible values for $\widetilde{p}^+$ and $\widetilde{p}^-$ are such that the stability condition in Theorem 1 is satisfied. The net benefit for a customer who arrives after and arrival is therefore $B - C \cdot E(W^+)$ and it is $B - C \cdot E(W^-)$ for a customer who arrives after a service completion. We have the following cases:

- Case 1: $E(W^+) \leq \frac{B}{C}$ for $\widetilde{p}^+ = \widetilde{p}^- = 1$. In this case, since $E(W^-) \leq E(W^+)$, we also have $E(W^-) \leq \frac{B}{C}$ for $\widetilde{p}^+ = \widetilde{p}^- = 1$. In this case, even if all potential customers join after an arrival or a service, they all enjoy a non-negative benefit. Therefore, the strategy of joining with probability $\widetilde{p}^+ = \widetilde{p}^- = 1$ is an equilibrium strategy. [4]
- Case 2: $E(W^+) \geq \frac{B}{C}$ as $\widetilde{p}^+$ tends to 0. As $\widetilde{p}^+$ tends to 0, $E(W^+)$ tends to $\overline{x}\frac{1+cv^2}{2}$ and $E(W^-)$ tends to 0.[5] Even if no other customer joins after an arrival, the net benefit of a customer who joins after an arrival is non-positive. Therefore, the strategy of joining after an arrival with probability $\widetilde{p}^+ = 0$ is an equilibrium strategy and no other equilibrium is possible after an arrival. If all customers join after a service completion, they all enjoy a non-negative benefit since $E(W^-) = 0$. Therefore the strategy of joining with $\widetilde{p}^+ = 0$ and $\widetilde{p}^- = 1$ is an equilibrium strategy.
- Case 3: $E(W^+) > \frac{B}{C}$ for $\widetilde{p}^+ = \widetilde{p}^- = 1$ and $E(W^+) < \frac{B}{C}$ as $\widetilde{p}^+$ tends to 0. In this case if $\widetilde{p}^+ = 1$ then a customer who joins after an arrival suffers a negative benefit. This cannot be an equilibrium strategy. If $\widetilde{p}^+ = 0$, then all customers balk after an arrival. Yet, a customer who joins after an arrival would get a positive benefit. This contradiction shows that $\widetilde{p}^+ = 0$ cannot be an equilibrium strategy. There exists a unique equilibrium where λ^+ solves $E(W^+) = \frac{B}{C}$ for $\widetilde{p}^- = 1$. In this case $E(W^-) \leq \frac{B}{C}$, therefore if

[4] In this case $E(W^+) = \overline{x}\frac{1+cv^2}{2}\left(\frac{1}{G^*(\lambda)} + \frac{\lambda\overline{x}}{1-\lambda\overline{x}}\right) + \frac{1-\lambda\overline{x}}{\lambda G^*(\lambda)} - \frac{1}{\lambda}$, and $E(W^-) = \overline{x}\left(\frac{1+cv^2}{2}\frac{(\lambda\overline{x})^2(1-G^*(\lambda))}{(1-\lambda\overline{x})(1-\lambda\overline{x}G^*(\lambda))} + \frac{G^*(\lambda)+\lambda\overline{x}-1}{1-\lambda\overline{x}G^*(\lambda)}\right)$.

[5] The limit of $E(W^+)$ corresponds to the expected remaining service time in an M/G/1 queue.

all customers join after a service, they all enjoy a non-negative benefit. Therefore the strategy of joining with $0 < \widetilde{p}^+ < 1$ and $\widetilde{p}^- = 1$ is also an equilibrium strategy.

As a conclusion, the case $\lambda^- \geq \lambda^+$ can be seen as the result of a strategic behavior.

4 The Admission Control Problem

We question here the possibility to take different decisions after a service or after an arrival to answer a classical routing problem in the queueing theory. This problem is referred in the literature as the admission control problem (Section 1 in [18]).

4.1 The optimisation problem

We propose to solve this problem under event-dependency. As mentioned in the previous section, the event-dependency may be the result of a customer strategic behavior when only the information of the last event is given. This leads to $\lambda^- \geq \lambda^+$. However, it might be interesting for the system to better control the arrival process by accepting or rejecting customers based on the system size. Let us specify that the arrival of a customer who was rejected is not seen as an arrival for the purpose of the rate changing.

A controller has to determine at arrival of a new customer whether we allow this new customer to enter the system or whether we reject this customer from the system. The optimization problem may be written as

$$\begin{cases} \text{Maximize } T_S, \\ \text{subject to } E(Q) \leq \overline{E(Q)}, \end{cases} \tag{43}$$

where T_S is the throughput of served customers, $E(Q)$ is the expected number of customers in the queue and, $\overline{E(Q)}$ is the service level constraint on $E(Q)$. We are restricting the class of admissible policies to the class of deterministic policies. In real system, for instance in a shop, deterministic policies are easier to implement than non deterministic ones which may require randomization. Yet, deterministic policies are not necessarily optimal. In order to saturate the constraint, it may be useful to randomize between two or more deterministic policies. This however may only improve the class of deterministic policies. It does not lead to the optimal policy.[6] In Section 4.2, we give conditions under which deterministic policies (or randomization between a number of them) are optimal.

Let us now specify the nature of a deterministic policy. Any policy within the class of deterministic stationary policies is equivalent to a two-thresholds policy for a given remaining service time. More precisely, for a given remaining service time r and k customers in the system, a *two-thresholds policy* is defined by two thresholds k_r^+ and k_r^- $(k_r^+, k_r^- \geq -1)$ such that,

[6] [3] shows an example where deterministic policies are not optimal for the admission control in an M/G/1 queue.

– if an arrival occurs after an arrival, then this customer is rejected if $k > k_r^+$, otherwise this customer is accepted,
– if an arrival occurs after a service, then this customer is rejected if $k > k_r^-$, otherwise this customer is accepted.

The two-thresholds policy can be used to improve the classical one-threshold policy where the same decision is taken after an arrival or a service. By allowing different decisions after an arrival or after a service, a larger range of values may be reachable for $E(Q)$. This may lead to a better solution for the optimization problem.

4.2 Optimal policy

We propose to formulate the routing problem as a Markov decision process (MDP) and next use the value iteration technique to prove the threshold structure of the optimal policy. We choose to approximate the service time duration by a *Coxian distribution*. Since Coxian distributions are dense in the field of all non-negative distributions [27], the obtained results apply for a general distribution. Consider a Coxian random variable which represents the service duration. It is defined by the parameters μ_j ($\mu_j > 0$, $1 \leq j \leq N$), and r_j ($r_j \in [0,1]$, $0 \leq j \leq N$) with $r_1 = 0$. The quantity r_j is the probability to enter the remaining phase $j-1$ after leaving remaining phase j and the parameter μ_j is the rate of the exponential distribution describing the random duration spent at remaining phase j.

However, our system is not a standard Markov decision process (MDP). The form of the problem makes it a constrained MDP; maximize the throughput of served customers with a constraint on the expected number of customers in the system. Constrained MDP's can be solved using various techniques. Here we use one that introduces the constraint in the objective using a Lagrange multiplier. Under weak conditions it can be seen that the optimal policy for a certain Lagrange multiplier is optimal for the constrained problem if the value of the constraint under this policy attains exactly $\overline{E(Q)}$. From the theory on constrained MDP's it follows that this policy is stationary and randomizes in at most 1 state. For this and other results on constrained MDP's, see the book of [2]. The optimization problem may then be rewritten as $\min(E(Q) - c \cdot T_S)$, where the coefficient c ($c \geq 0$) is the Lagrange multiplier which translates the relative importance given, by the system manager, to the throughput of served customers (T_S) compared to expected number of customers in the system ($E(Q)$).

Let us denote by (x, y) a state of the system where x is the number of potential remaining phases of work for the server, $x \geq 0$ and y is the nature of the last event; an arrival ($y = +$) or a service ($y = -$). We denote the transition rate from state (x, y) to state (x', y') by $q_{(x,y),(x',y')}$. Hence for $x, x' \geq 0$ and $y, y' \in \{+, -\}$,

we have

$$q_{(x,y),(x',y')} = \begin{cases} \lambda^+, \text{ if } x' = x + N, y = y' = + \text{ for } x \geq 0, \\ \lambda^-, \text{ if } x' = x + N, y = -, y' = + \text{ for } x \geq 0, \\ r_j \mu_j, \text{ if } x = kN + j, x' = x - 1, y' = y \text{ for } k \geq 0, 1 \leq j \leq N, \text{ and } y \in \{+, -\}, \\ (1 - r_j)\mu_j, \text{ if } x = kN + j, x' = kN, y' = - \text{ for } k \geq 0, 1 \leq j \leq N, \text{ and } y \in \{+, -\}, \\ 0, \text{ otherwise}, \end{cases}$$

which corresponds to arrivals and service departures.

We choose to discretize our continuous-time model. This is possible because it is uniformizable (Section 11.5.2. in [25]). We formulate a 2-step value function, in order to separate transitions and actions. We define the dynamic programming value functions $W_n^-(x)$, $W_n^+(x)$, $V_n^-(x)$, and $V_n^+(x)$ over $n \geq 0$ steps, depending on the state of the system. We choose $W_0^{-,+}(x) = 0$ and $V_0^{-,+}(x) = 0$ for $x \geq 0$. We assume without loss of generality that $\lambda^+ + \lambda^- + \sum_{j=1}^N \mu_j = 1$, such that the rate out of each state is equal to 1; thus we can consider the rates to be transition probabilities. We then may write for $1 \leq x \leq N$, and $k \geq 0$,

$$\begin{aligned} V_{n+1}^+(kN + x) &= k + 1 + \lambda^+ W_n^+(kN + x) + r_x \mu_x V_n^+(kN + x - 1) + (1 - r_x)\mu_x V_n^-(kN) \\ &\quad + (1 - \lambda^+ - \mu_x) V_n^+(kN + x), \\ V_{n+1}^-(kN + x) &= k + 1 + \lambda^- W_n^-(kN + x) + r_x \mu_x V_n^-(kN + x - 1) + (1 - r_x)\mu_x V_n^-(kN) \\ &\quad + (1 - \lambda^+ - \mu_x) V_n^-(kN + x), \\ V_{n+1}^-(0) &= \lambda^- W_n^-(0) + (1 - \lambda^-) V_n^-(0). \end{aligned} \tag{44}$$

The operator W_n represents the decision to accept or to reject a new customer from the system. After an arrival we have $W_n^+(kN + x) = \min(V_n^+((k + 1)N + x) - c, V_n^+(kN + x))$, and after a service, we have $W_n^-(kN + x) = \min(V_n^+((k + 1)N + x) - c, V_n^-(kN + x))$.

For each $n > 0$ and every state, there is a minimizing action at customer's arrival: accept this customer or reject this customer. One way of obtaining the long-run average optimal actions is to use the value iteration technique introduced by [7] and [14], by recursively evaluating V_n using Equation (44), for $n \geq 0$. To prove the form of the optimal policy we want to establish structural properties of the value function. In particular, it would be interesting to obtain conditions for which threshold policies based on the number of customers in the system are optimal.

To prove that the optimal policy has a threshold structure, we need the conditions

$$\lambda^- \geq \lambda^+, \text{ and,} \tag{45}$$

$$(1 - r_x)\mu_x \geq (1 - r_{x+1})\mu_{x+1}, \text{ for } 1 \leq x \leq N. \tag{46}$$

In Proposition 6, under Conditions (45) and (46), we prove by induction on the value function that the optimal policy is of threshold type based on the number of customers in the system for a given number of remaining phases. The proof follows a standard MDP methodology where structural properties of the value function are proven by induction. The complete detailed proof is given in Section 2 of the online supplement.

Proposition 6 *Under the condition,* $\lambda^- \geq \lambda^+$*, and* $(1-r_x)\mu_x \geq (1-r_{x+1})\mu_{x+1}$*, for* $1 \leq x \leq N$*, the optimal admission policy has a threshold structure. More precisely, there exists two thresholds* k_x^+ *and* k_x^- *such that, for* $k \geq 0$ *and* $1 \leq x \leq N$*,*

- *if an arrival occurs in state* $(kN + x, +)$*, it is optimal to reject this customer if* $k > k_x^+$*, otherwise this customer is accepted,*
- *if an arrival occurs in state* $(kN + x, -)$*, it is optimal to reject this customer if* $k > k_x^-$*, otherwise this customer is accepted.*

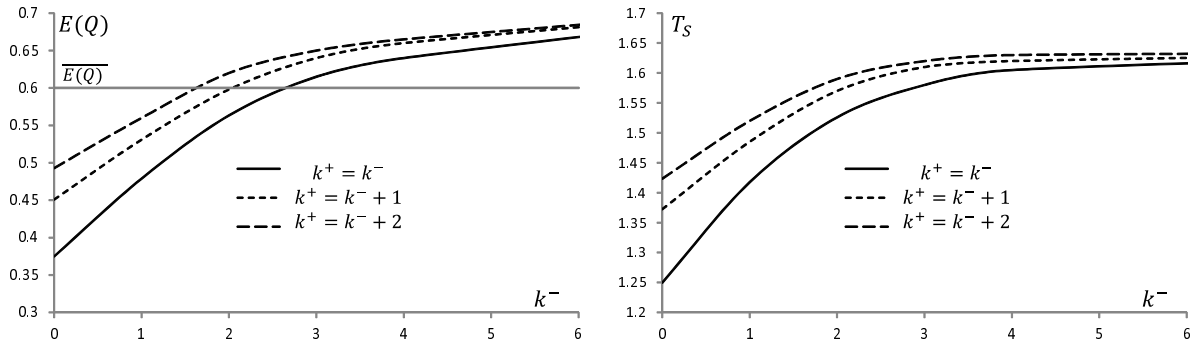
Condition (45) is required to show the convexity and the supermodularity properties of the value function. Condition (46) means that the departure rate out of a given service phase increases with the number of elapsed phases of service. This condition allows the value function to be increasing in the number of remaining service phases. This monotonicity property of the value function is required to prove that the optimal policy has a threshold structure.

Without this condition, the value function can be non-increasing in the number of remaining phases of service. Consider for instance a situation with a Coxian service time distribution with two phases. The first phase of service has an expected duration of 1 and the second one has an expected duration of 100. The probability to end service after the first waiting phase is 90%. So, if at a customer arrival there is 1 customer in the system being in the second phase of service, then the expected waiting time of an arriving customer is 100. If at a customer arrival there are 2 customers in the system and the one in service being in the first phase of service, then the expected waiting time of an arriving customer is $2 \times (1 + 0.1 \times 100) = 22 < 100$. Therefore in this case, the congestion of the system is not positively bound to the number of remaining phases of service and the value function may not be increasing.

Numerical illustration. We consider a particular Coxian distribution for the service time with two exponential phases with rate μ_1 and μ_2 and with probabilities $r_1 = 0$ and $r_2 = 1$.[7] From proposition 6, the optimal policy is determined by the thresholds k_1^+, k_1^-, k_2^+, and k_2^-. The computation of the performance measures can

[7] This particular Coxian distribution is an hypoexponential distribution.

be done using a Matrix geometric approach (e.g., see [24]). In Figures (3(a)) and (3(b)), we represent the performances measures as a function of $k^- = k_1^- + k_2^-$ and $k^+ = k_1^+ + k_2^2$ for three different values for $k^+ - k^-$. The parameters k^- and k^+ represent the thresholds on the number of remaining service phases. In this example, the optimal thresholds to answer the optimization problem are $k^+ = 3$ and $k^- = 2$. This means that an arriving customer should be rejected after an arrival if there is strictly more than 2 customers in the system or if there is two customers in the system and the customer in service still has 2 phases of service to achieve. A customer should be rejected after a service if there is strictly more than one customer in the system. This illustrates a case where the two-thresholds policy achieves a better solution to the optimization problem than the one-threshold policy.



(a) $E(Q)$ (b) T_S

Fig. 3 Performance measures ($\lambda^+ = 1$, $\lambda^- = 2$, $\mu_1 = 5$, $\mu_2 = 10$, $r_1 = 0$, $r_2 = 1$, $\overline{E(Q)} = 0.6$)

4.3 Performance analysis under the two-thresholds policy

The result of Proposition 6 may be difficult to implement in practice. First, it might be difficult to find the appropriate Coxian distribution which well approximates the considered service time distribution. A field of research is dedicated to this problem (e.g., see [29], [4], [13]). Second, the optimal policy depends on the remaining number of service phases or more generally on the remaining service time. This information might also be complicate to obtain in practice.

To overcome these difficulties, we propose a numerical analysis to obtain the performance measures under a two-thresholds policy where the thresholds do not depend on the remaining service time. Although the proposed policy is not optimal, it is simple to implement and may lead to enhanced performance compared to a one-threshold policy based on the number of customers in the system. Moreover, compared to the Coxian approximation followed by a Matrix geometric approach, the method developed here leads to the exact performance measures for any service time distribution.

We denote by k^+ and k^- the thresholds on the system size after an arrival or a service. After an arrival (respectively a service) customers are rejected if there is strictly more than k^+ (respectively k^-) customers

in the system. The approach proceeds in a way very similar to the case with infinite thresholds in Section 2. Yet, the analysis does not lead to explicit expressions. Due to the two thresholds, the stationary probabilities at departure instants can be computed directly since they are in finite number. The transition matrix is given by

$$M = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \cdots\cdots\cdots & 1-\sum_{n=0}^{k^+-1}\beta_n \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \cdots\cdots\cdots & 1-\sum_{n=0}^{k^+-1}\alpha_n \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots\cdots\cdots & 1-\sum_{n=0}^{k^+-2}\alpha_n \\ 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \cdots\cdots\cdots & 1-\sum_{n=0}^{k^+-3}\alpha_n \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots\cdots & \vdots \\ 0 & \cdots\cdots\cdots & \alpha_0 & \alpha_1 & \cdots\cdots & 1-\sum_{n=0}^{k^+-k^-}\alpha_n \\ 0 & \cdots\cdots\cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots\cdots\cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots\cdots\cdots\cdots\cdots & & & 1 & & 0 \end{pmatrix}.$$

The first $k^- + 1$ lines of the matrix are identical to those in the matrix in Section 2.1 except that the number of customers after a service completion is bounded by k^+. From line $k^- + 2$ to line $k^+ + 1$, there cannot be any arrival during a service because the first arrival after the service completion cannot occur. So, the number of customers is forced to be reduced by one at the next service completion. One can then solve the system $P = P \cdot M$, where $P = (p_0, p_1, \cdots, p_{k^+})$ with $\sum_{n=0}^{k^+} p_n = 1$ to obtain the stationary distribution of the system size at departure instants. In Proposition 7, we relate these probabilities to those at arbitrary instants.

Proposition 7 *We have*

$$\pi_{n,+} = (1-\pi_0)\frac{G^*(\lambda^-)}{\lambda^+\overline{x}}p_n, \text{ and } \pi_{n,-} = (1-\pi_0)\frac{1-G^*(\lambda^-)}{\lambda^-\overline{x}}p_n,$$

for $1 \le n \le k^-$, *and*

$$\pi_{n,+} = \frac{1-\pi_0}{\lambda^+\overline{x}}p_n, \pi_{n,-} = (1-\pi_0)p_n, \text{ and } \pi_{k^++1,+} = 1 - \left(\pi_0 + \sum_{n=1}^{k^+}\pi_{n,+} + \pi_{n,-}\right),$$

for $k^- + 1 \le n \le k^+$.

Proof. Contrary to Section 2.2, the system is not identical at arrival instants and departure instants but it is identical at arrival instants of customers who join and departure instants. Therefore, we have

$$p_0 = \frac{\lambda^- \pi_0}{\lambda^- \sum_{k=0}^{k^-} \pi_{k,-} + \lambda^+ \sum_{k=1}^{k^+} \pi_{k,+}}, \tag{47}$$

$$p_n = \frac{\lambda^- \pi_{n,-} + \lambda^+ \pi_{n,+}}{\lambda^- \sum_{k=0}^{k^-} \pi_{k,-} + \lambda^+ \sum_{k=1}^{k^+} \pi_{k,+}}, \text{ for } 1 \leq n \leq k^-, \text{ and} \tag{48}$$

$$p_n = \frac{\lambda^+ \pi_{n,+}}{\lambda^- \sum_{k=0}^{k^-} \pi_{k,-} + \lambda^+ \sum_{k=1}^{k^+} \pi_{k,+}}, \text{ for } k^- + 1 \leq n \leq k^+. \tag{49}$$

Due to flow conservation, one may write $\lambda^- \sum_{k=0}^{k^-} \pi_{k,-} + \lambda^+ \sum_{k=1}^{k^+} \pi_{k,+} = \frac{1}{\overline{x}}(1 - \pi_0)$. So, we deduce from Equation (47) that $\pi_0 = \frac{p_0}{p_0 + \lambda^- \overline{x}}$. The throughput of served customers is hence $\frac{\lambda^-}{p_0 + \lambda^- \overline{x}}$.

Similarly to the proof of Lemma 1, $p(n, r, +)$ and $p(n, r, -)$ obey the following differential equations:

$$p'(1, r, +) = \lambda^+ p(1, r, +) - \lambda^- \pi_0 g(r), \tag{50}$$

$$p'(n, r, +) = \lambda^+ p(n, r, +) - \lambda^- p(n-1, r, -) - \lambda^+ p(n-1, r, +), \text{ for } 2 \leq n \leq k^- + 1, \tag{51}$$

$$p'(n, r, -) = \lambda^- p(n, r, -) - g(r)(p(n+1, 0, +) + p(n+1, 0, -)), \text{ for } 1 \leq n \leq k^-, \tag{52}$$

$$p'(n, r, +) = \lambda^+ p(n, r, +) - \lambda^+ p(n-1, r, +), \text{ for } k^- + 2 \leq n \leq k^+, \tag{53}$$

$$p'(n, r, -) = -g(r)(p(n+1, 0, +) + p(n+1, 0, -)), \text{ for } k^- + 1 \leq n \leq k^+ - 1, \tag{54}$$

$$p'(k^+ + 1, r, +) = -\lambda^+ p(k^+, r, +), \tag{55}$$

$$p'(k^+, r, -) = -g(r) p(k^+ + 1, 0, +). \tag{56}$$

By integrating Equations (51) and (52) for r from 0 to ∞ and summing up the two obtained equations, we deduce that $p(n, 0, +) + p(n, 0, -) - \lambda^+ \pi_{n-1,+} - \lambda^- \pi_{n-1,-}$ is a constant for $2 \leq n \leq k^- + 1$. With the same approach with Equations (53) and (54), we deduce that $p(n, 0, +) + p(n, 0, -) - \lambda^+ \pi_{n-1,+}$ is equal to the same constant for $k^- + 2 \leq n \leq k^+$. Using Equations (53) and (56), proves that this constant is also equal to $p(k^+ + 1, 0, +) - \lambda^+ \pi_{k^+,+}$. Finally, Equation (55), leads to $p(k^+ + 1, 0, +) = \lambda^+ \pi_{k^+,+}$. So, the aforementioned constant is 0. As a conclusion, we may write

$$p(n, 0, +) + p(n, 0, -) = \lambda^+ \pi_{n-1,+} + \lambda^- \pi_{n-1,-}, \text{ for } 2 \leq n \leq k^- + 1,$$

$$p(n, 0, +) + p(n, 0, -) = \lambda^+ \pi_{n-1,+}, \text{ for } k^- + 2 \leq n \leq k^+,$$

$$p(k^+ + 1, 0, +) = \lambda^+ \pi_{k^+,+}.$$

Combining this last set of equations with Equations (52) and (54), we get $p(n,0,-) = \lambda^+ \pi_{n,+}$, for $1 \leq n \leq k^-$.

With the same approach as in Proposition 1 with Equation (52), we then deduce that $\pi_{n,-} = \frac{\lambda^+}{\lambda^-} \frac{1-G^*(\lambda^-)}{G^*(\lambda^-)} \pi_{n,+}$, for $1 \leq n \leq k^-$. Using Equation (48), yields

$$\pi_{n,+} = (1-\pi_0) \frac{G^*(\lambda^-)}{\lambda^+ \overline{x}} p_n, \text{ and } \pi_{n,-} = (1-\pi_0) \frac{1-G^*(\lambda^-)}{\lambda^- \overline{x}} p_n,$$

for $1 \leq n \leq k^-$. From Equation (49), we get

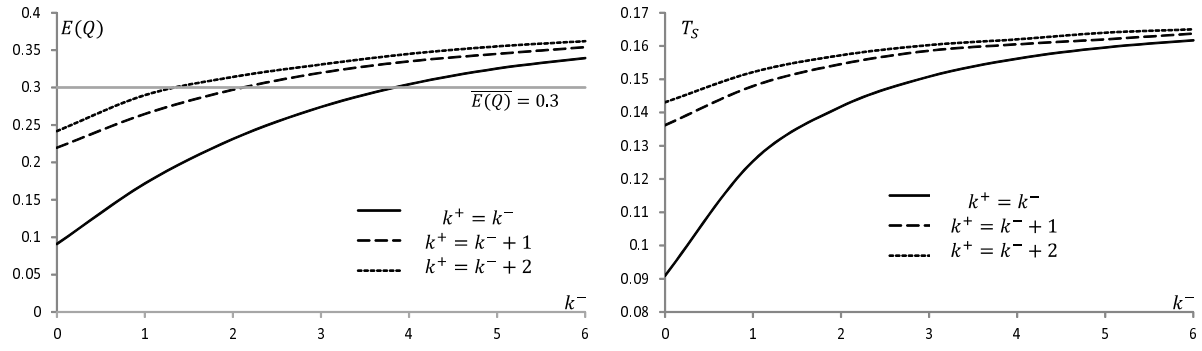
$$\pi_{n,+} = \frac{1-\pi_0}{\lambda^+ \overline{x}} p_n,$$

for $k^- + 1 \leq n \leq k^+$. From Equation (54), we may write $p'(n,u,-) = -g(u)\lambda^+ \pi_{n,+}$, for $k^- + 1 \leq n \leq k^+ - 1$. Integrating this equation for u from r to ∞, we get $p(n,r,-) = \lambda^+ \pi_{n,+} P(S > r)$. We integrate again this equation for r from 0 to ∞. This leads to $\pi_{n,-} = \lambda^+ \overline{x} \pi_{n,+}$, for $k^- + 1 \leq n \leq k^+ - 1$. So,

$$\pi_{n,-} = (1-\pi_0) p_n,$$

for $k^- + 1 \leq n \leq k^+$. The last probability, $\pi_{k^+ +1,+}$, is given by $\pi_{k^+ +1,+} = 1 - \left(\pi_0 + \sum_{n=1}^{k^+} \pi_{n,+} + \pi_{n,-} \right)$. □

Numerical Illustration. In Figures (4(a)) and (4(b)), we represent the performances measures as a function of k^- and k^+ for three different values for $k^+ - k^-$ and an hyperexponential distribution for the service time with 2 rates μ_1 and μ_2 and a probability q to be served with rate μ_1. In this example, the optimal thresholds to answer the optimization problem are $k^+ = 3$ and $k^- = 2$. Again, it illustrates a case where the two-thresholds policy achieves a better solution to the optimization problem than the one-threshold policy.[8]



(a) $E(Q)$ (b) T_S

Fig. 4 Performance measures ($\lambda^+ = 0.8$, $\lambda^- = 0.1$, $q = 1/3$, $\mu_1 = 0.5$, $\mu_2 = 2$, $\overline{E(Q)} = 0.3$)

[8] Yet, we do not claim that the two-thresholds policy is optimal in this case.

5 Future Research

Many questions following this study are open for future research. For instance, it would be interesting to include other features for the customer's behavior like abandonment or workload-dependency. We could also consider a multi-server queue instead of a single-server one. This however would not change the results since the observation of a change in the queue size would only occur when all agents are busy. Another extension of the model is the possibility of a customer's decision not only based on the last event but on a larger finite number of events. Finally, it could also be interesting to consider the symmetrical case where the server adapts its service rates to the last realized event.

References

1. Z. Aksin. The effect of random waits on customer queue joining and reneging behavior: A laboratory experiment. *Proceedings of the EURO XXVII Annual Conference, Glasgow*, 2015.
2. E. Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.
3. E. Altman and R. Hassin. Non-threshold equilibrium for customers joining an M/G/1 queue. In *in Proceedings of 10th International Symposium on Dynamic Game and Applications*. Citeseer, 2002.
4. S. Asmussen, O. Nerman, and M. Olsson. Fitting phase-type distributions via the em algorithm. *Scandinavian Journal of Statistics*, 1:419–441, 1996.
5. R. Bekker, S.C. Borst, O.J. Boxma, and O. Kella. Queues with workload-dependent arrival and service rates. *Queueing Systems*, 46(3-4):537–556, 2004.
6. R. Bekker, G.M. Koole, B.F. Nielsen, and T.B. Nielsen. Queues with waiting time dependent service. *Queueing Systems*, 68(1):61–78, 2011.
7. R.E. Bellman. *Dynamic programming*. Princeton University Press, Princeton, 1957.
8. O.J. Boxma. Joint distribution of sojourn time and queue length in the M/G/1 queue with (in) finite capacity. *European Journal of Operational Research*, 16(2):246–256, 1984.
9. O.J. Boxma, H. Kaspi, O. Kella, and D. Perry. On/off storage systems with state-dependent input, output, and switching rates. *Probability in the Engineering and Informational Sciences*, 19(01):1–14, 2005.
10. O.J. Boxma and M. Vlasiou. On queues with service and interarrival times depending on waiting times. *Queueing Systems*, 56(3-4):121–132, 2007.
11. J.M. Harrison and S.I. Resnick. The stationary distribution and first exit probabilities of a storage process with general release rule. *Mathematics of Operations Research*, 1(4):347–358, 1976.
12. R. Hassin and M. Haviv. *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59. Springer Science & Business Media, 2003.

13. A. Horváth and M. Telek. Phfit: A general phase-type fitting tool. *Computer Performance Evaluation: Modelling Techniques and Tools*, 1:1–14, 2002.
14. R.A. Howard. Dynamic programming and Markov processes. 1960.
15. Y. Kerner. The conditional distribution of the residual service time in the Mn/G/1 queue. *Stochastic Models*, 24(3):364–375, 2008.
16. L. Kleinrock. *Queueing Systems, Theory*, volume I. A Wiley-Interscience Publication, 1975.
17. G. Koole. A simple proof of the optimality of a threshold policy in a two-server queueing system. *Systems & Control Letters*, 26(5):301–303, 1995.
18. G. Koole. *Monotonicity in Markov reward and decision chains: Theory and applications*, volume 1. Now Publishers Inc, 2007.
19. R.L. Larsen and A.K. Agrawala. Control of a heterogeneous two-server exponential queueing system. *IEEE Transactions on Software Engineering*, (4):522–526, 1983.
20. B. Legros and O. Jouini. Routing in a queueing system with two heterogeneous servers in speed and in quality of resolution. *Stochastic Models*, pages 1–19, 2017.
21. B. Legros and A.D. Sezer. Stationary analysis of a single queue with remaining service time dependent arrivals. *Queueing Systems*, 2017. To appear.
22. W. Lin and P.R. Kumar. Optimal control of a queueing system with two heterogeneous servers. *IEEE Transactions on Automatic Control*, 29(8):696–703, 1984.
23. H.P. Luh and I. Viniotis. Threshold control policies for heterogeneous server systems. *Mathematical Methods of Operations Research*, 55(1):121–142, 2002.
24. M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach*. Johns Hopkins University Press, Mineola, 1981.
25. M.L. Puterman. *Markov Decision Processes*. John Wiley and Sons, 1994.
26. V.V. Rykov. Monotone control of queueing systems with heterogeneous servers. *Queueing systems*, 37(4):391–403, 2001.
27. R. Schassberger. *Warteschlangen*. Springer-Verlag Vienna, 1973.
28. K. Sigman and U. Yechiali. Stationary remaining service time conditional on queue length. *Operations research letters*, 35(5):581–583, 2007.
29. M.C. Van Der Heijden. On the three-moment approximation of a general distribution by a coxian distribution. *Probability in the Engineering and Informational Sciences*, 2(2):257–261, 1988.
30. J. Walrand. A note on optimal control of a queuing system with two heterogeneous servers. *Systems & control letters*, 4(3):131–134, 1984.
31. W. Whitt. Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems*, 6(1):335–351, 1990.

Notations

Table 1 Notations

Exogenous parameters	
λ^+, λ^-	Arrival rates after an arrival and after a service
S	Random variable which represents the service time duration
$\overline{x}$	Expected service time
cv	Coefficient of variation of the service time distribution; it is the ratio of the standard deviation divided by the expected value
$g(.)$	Probability density function of the service time
$G^*(.)$	Laplace-Stieltjes Transform (LST) of the service time; $G^*(s) = \int_0^\infty g(t)e^{-st}\,\mathrm{d}t$
Decision parameters	
k^+, k^-	Thresholds on the system size or on the remaining number of service phases to accept or reject customers at arrival after an arrival or a service
$\widetilde{p}^+$, $\widetilde{p}^-$	Probabilities that an arriving customer accepts to join after an arrival or a service
Probabilities	
$p_t(n,r,+)$, $p_t(n,r,-)$	Probability-density of having n customers in the system, $n \geq 1$ and a remaining service time of r, $r \geq 0$, at time t after an arrival or a service
$p(n,r,+)$, $p(n,r,+)$	$p(n,r,+) = \lim_{t\to\infty} p_t(n,r,+)$, $p(n,r,-) = \lim_{t\to\infty} p_t(n,r,-)$ for $n \geq 1$
$\pi_{n,+}$, $\pi_{n,-}$	Stationary probability to have n customers in the system at arbitrary instants after an arrival or a service ($\pi_{n,+} = \int_{r=0}^\infty p(n,r,+)\,\mathrm{d}r$, $\pi_{n,-} = \int_{r=0}^\infty p(n,r,-)\,\mathrm{d}r$ for $n \geq 1$)
π_n	Stationary probability to have n customers in the system at arbitrary instants,$\pi_n = \pi_{n,+} + \pi_{n,-}$ for $n \geq 0$
$p_{n,+}$, $p_{n,-}$	Stationary probability to have n customers in the system at departure instants after an arrival or a service for $n \geq 0$
p_n	Stationary probability to have n customers in the system at departure instants, $p_n = p_{n,+} + p_{n,-}$ for $n \geq 0$
$\overline{r_n^+}$, $\overline{r_n^-}$	Expected remaining service time seen by a customer who arrives after an arrival or a service with n customers present in the system, $n \geq 1$.
α_n, β_n	Probability that n customers arrive during a service if the service is initiated by a service completion or by an arrival, $n \geq 0$
Performance measures	
$E(Q_d)$, $E(Q)$	Expected number of customers in the system at departure and arbitrary instants
$\overline{E(Q)}$	Service level objective on $E(Q)$
$E(W)$	Expected waiting time at arbitrary instants
T_s	Expected throughput of served customers
$E(W^+)$, $E(W^-)$	Expected waiting time of a customer who arrives after an arrival or a service
$E(R)$	Expected remaining service time seen by an arriving customer at a non-empty system
$E(R^+)$, $E(R^-)$	Expected remaining service time seen by a customer who arrives after an arrival or a service
Markov decision process	
$V_n^+(x)$, $V_n^-(x)$	Value function depending on the state of the system
μ_j	Exponential rate of remaining service phase j in the Coxian distribution ($1 \leq j \leq N$)
r_j	Probability to enter remaining service phase $j-1$ after leaving remaining service phase j ($1 \leq j \leq N$) in the Coxian distribution