Queues with service speed adaptations

R. Bekker
Department of Mathematics
Vrije Universiteit
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

O.J. Boxma
EURANDOM and Department of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

J.A.C. Resing
Department of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

September 18, 2006

Abstract

In this paper, we consider various queueing models in which the server can work at two different service speeds. The speed of the server depends on either the number of customers, or the workload. Our main interest is in the model in which service speed adaptations can only take place at the arrival instants of an external Poisson observer. Using insightful probabilistic arguments we give the structure of the steady-state queue length and workload distributions in the various models.

Also, in case the service speed can only be adapted right after departure instants based on the number of customers, we provide explicit and intuitively appealing expressions for the steady-state distribution of the number of customers present.

1 Introduction

In this paper we consider queueing models in which the server can work at two different service speeds. The speed at which the server works depends either on the number of customers in the system or on the workload (i.e., the amount of work) in the system.

Models with service speed depending on queue length or workload arise naturally as representation of (congestion) phenomena in, e.g., manufacturing and healthcare processes. Our main interest is in the model in which service speed adaptations can only take place at the arrival instants of an external observer. We assume that these arrival instants occur according to a Poisson process and in the sequel we call this model the model with service speed adaptations at external Poisson instants. However, we also want to compare the performance of this model with the performance of other models, studied before, in which the server can work at two different service speeds. Therefore, we will also discuss the model with instantaneous service speed adaptations (see Cohen [6] and Gaver and Miller [10]) and the model with service speed adaptations at arrival instants (see Bekker [1] and Bekker and Boxma [2]).

The paper is organised as follows. In Section 2, we study a simple $M/M/1$ queue in which the speed of the server depends on the *number of customers* in the system. In this model the steady-state distribution of the number of customers in the system is a mixture of one or more *geometric* distributions. The number of geometric terms depends on the instants at which service speed adaptations can take place and on the part of the state space (few/many customers, high/low service speed).

In Section 3, we will consider the $M/M/1$ model in which the speed of the server depends on the *workload*. It turns out that in this model we get similar results. Now, the steady-state distribution of the workload in the system is a mixture of one or more *exponential* distributions. The number of exponential terms again depends on the variant and on the part of the state space (high/low workload, high/low service speed).

Section 4 is devoted to situations in which service speed adaptations can only take place at departure instants. In this case, the analysis of the system in which the service speed depends on the workload seems very difficult. Therefore, we restrict our attention to the model in which the service speed depends on the number of customers. We will give explicit and intuitively appealing expressions for the steady-state distribution of the number of customers in the system. Instead of restricting our attention only to the $M/M/1$ queue, here we present the results for the $M/G/1$ queue. This model and some extensions have been studied by others, see e.g. [5, 9, 14, 15]. In [5, 14] the authors consider Laplace-Stieltjes transforms, while in [9, 15] the authors focus on computational aspects. We also refer to Dshalalow [8] for an extensive survey, with 277 references, on queueing models with state-dependent parameters.

In the paper, we always try to use probabilistic arguments to find steady-state distributions (instead of, e.g., directly solving balance equations). Although we restrict our attention to models in which the server can work at only two different service speeds, we expect that our probabilistic approach will also be helpful to understand models in which the server can work at more than two service speeds. The ultimate goal of a performance analysis of queueing models like those described above is to optimise and control the system behaviour. While such optimisation and control is not part of the present paper, the analysis presented here does prepare the ground for it.

2 Service speed based on number of customers

In this section we consider an $M/M/1$ queue in which the server can work at two different speeds, r_1 and r_2. The service speed is based on the number of customers. If the number of customers in the system is smaller than or equal to K, then the server should work at speed r_1. If on the other hand the number of customers in the system is bigger than K, the server should work at speed r_2. However, because service adaptations can only take place at special points in time, in some of the models it can happen that the server still works at speed r_2 (resp. r_1) although the number of customers has dropped at or below level K (resp. has increased above level K).

Customers arrive to the system according to a Poisson process with rate λ. Service requirements of customers are independent, exponentially distributed random variables with parameter μ. Define $\rho_i := \lambda/(\mu_i)$ with $\mu_i := \mu r_i,\ i = 1, 2$. In the sequel we assume that $\rho_2 < 1$ to assure that the system is stable. Let $X(t)$ denote the number of customers in the system at time t. Furthermore, $Y(t)$ represents the state of the server at time t. More specifically, $Y(t) = i$ if the speed of the server at time t equals $r_i, i = 1, 2$. With (X, Y) we denote a random vector having the steady-state distribution of the continuous-time Markov process $\{(X(t), Y(t))\}_{t\geq 0}$. With $\pi(k, i)$ we denote the steady-state probability $P((X, Y) = (k, i))$.

2.1 Instantaneous service speed adaptations

In the model with instantaneous service speed adaptations, the speed of the server instantaneously changes from r_1 to r_2 if the number of customers in the system increases from K to $K + 1$. Similarly, the speed of the server instantaneously changes from r_2 to r_1 if the number of customers in the system decreases from $K + 1$ to K. Hence, in this case the state space of the continuous-time Markov process $\{(X(t), Y(t))\}_{t\geq 0}$ equals

$$S = \{(0, 1), \ldots, (K, 1)\} \cup \{(K + 1, 2), (K + 2, 2), \ldots\}.$$

Clearly, at and below level K the system behaves as an $M/M/1/K$ queue with service speed r_1, while above level K the system behaves as an $M/M/1$ queue with service speed r_2. Hence, the steady-state probabilities are given by

$$\begin{aligned} \pi(k, 1) &= C_1\rho_1^k, \quad k = 0, \ldots, K, \\ \pi(K + k + 1, 2) &= C_2\rho_2^k, \quad k = 0, 1, \ldots. \end{aligned}$$

The constants C_1 and C_2 follow from the normalisation equation and the balance equation $\lambda\pi(K, 1) = \mu_2\pi(K + 1, 2)$.

So we see that the steady-state distribution has geometric behaviour with parameter ρ_1 on the part of the state space $\{(0, 1), \ldots, (K, 1)\}$ and geometric behaviour with parameter ρ_2 on the part of the state space $\{(K + 1, 2), (K + 2, 2), \ldots\}$.

2.2 Service speed adaptations at arrival instants

In the model with service speed adaptations at arrival instants, the speed of the server is based on the number of customers in the system just after the arrival instant. If the number of customers in the system increases from K to $K+1$, the speed of the server instantaneously becomes r_2 (because there is an arrival). However, if the number of customers in the system decreases from $K+1$ to K, the server continues to work at speed r_2 until the next arrival instant (and even longer if the next arrival brings the system back into state $K+1$). Hence, in this case the state space of the continuous-time Markov process $\{(X(t), Y(t))\}_{t\geq 0}$ equals

$$S = \{(0,1), \ldots, (K,1)\} \cup \{(0,2), \ldots, (K,2)\} \cup \{(K+1,2), (K+2,2), \ldots\}.$$

As before, above level K the system behaves as an ordinary $M/M/1$ queue with service speed r_2, leading to

$$\pi(K+k+1, 2) = C_1 \rho_2^k, \quad k = 0, 1, \ldots.$$

At an arbitrary instant at which the number of customers in the system is at or below level K and the server works at speed r_2, we have

$$(X \,|\, X \leq K, Y = 2) \stackrel{d}{=} \max(K - Z, 0),$$

where Z is geometrically distributed with parameter $\frac{\mu_2}{\lambda+\mu_2}$, leading to

$$\pi(k+1, 2) = C_2 \left(\frac{\lambda+\mu_2}{\mu_2}\right)^k, \quad k = 0, \ldots, K-1.$$

Remark that we only give here the expressions for $\pi(1,2), \ldots, \pi(K,2)$. The expression for $\pi(0,2)$ is slightly different.

At an arbitrary instant at which the number of customers is at or below K and the server works at speed r_1, the system behaves as an $M/M/1/K$ model with service speed r_1, but with the special feature that after each overflow, the number of customers is first instantaneously decreased to the steady-state situation of the system when the number of customers is at or below K and the service speed equals r_2 and, after that, instantaneously increased by 1 (due to the arrival). By using an up- and downcrossing argument for this system we obtain

$$\lambda \pi(k,1) = \mu_1 \pi(k+1,1) + \lambda \left(\frac{\mu_2}{\lambda+\mu_2}\right)^{K-k} \pi(K,1), \quad k = 1, \ldots, K-1,$$

from which we can show that

$$\pi(k+1,1) = C_3 \rho_1^k + C_4 \left(\frac{\lambda+\mu_2}{\mu_2}\right)^k, \quad k = 0, \ldots, K-1.$$

Again, we only give here the expressions for $\pi(1,1), \ldots, \pi(K,1)$. The expression for $\pi(0,1)$ is slightly different.

Hence, the steady-state distribution has geometric behaviour with parameter ρ_2 on the part of the state space $\{(K+1,2),(K+2,2),\ldots\}$, geometric behaviour with parameter $\frac{\lambda+\mu_2}{\mu_2}$ on the part of the state space $\{(1,2),\ldots,(K,2)\}$ and a mixture of geometric behaviour with parameter ρ_1 and geometric behaviour with parameter $\frac{\lambda+\mu_2}{\mu_2}$ on the part of the state space $\{(1,1),\ldots,(K,1)\}$.

2.3 Service speed adaptations at external Poisson instants

In the model with service speed adaptations at external Poisson instants, we assume that these instants occur according to a Poisson process with rate ν. When the number of customers at an external Poisson instant is above level K, then the server works at speed r_2 until the next external Poisson instant. Similarly, if the number of customers at an external Poisson instant is at or below level K, then the server works at speed r_1 until the next external Poisson instant. Because the server can adapt its speed only at external Poisson instants, the state space of the continuous-time Markov process $\{(X(t),Y(t))\}_{t\geq 0}$ in this case equals

$$S=\{(0,1),\ldots,(K,1)\}\cup\{(K+1,1),\ldots\}\cup\{(0,2),\ldots,(K,2)\}\cup\{(K+1,2),\ldots\}.$$

During periods in which the number of customers in the system is above level K and the server works at speed r_1, the system behaves as an ordinary $M/M/1$ queue with service speed r_1 and with disasters occurring with rate ν. For this system it is well-known that the steady-state distribution is geometric with parameter x_1, where x_1 is the root in $(0,1)$ of the equation

$$\mu_1 x^2-(\lambda+\mu_1+\nu)x+\lambda=0. \tag{1}$$

Hence, we have

$$\pi(K+k+1,1)=C_1x_1^k,\quad k=0,1,\ldots.$$

During periods in which the number of customers in the system is above level K and the server works at speed r_2, the system behaves as an ordinary $M/M/1$ queue with service speed r_2 and with the special feature that, whenever the system is empty, additional batches of customers arrive according to a Poisson process with rate μ_2. The batch size distribution of these batches is geometric with parameter x_1. For this system an up- and downcrossing argument yields

$$\lambda\pi(K+k,2)+\mu_2x_1^k\pi(K+1,2)=\mu_2\pi(K+k+1,2),\quad k=1,2,\ldots,$$

from which we can show that

$$\pi(K+k+1,2)=C_2\rho_2^k+C_3x_1^k,\quad k=0,1,\ldots.$$

During periods in which the number of customers in the system is at or below level K and the server works at speed r_2, the system behaves as an ordinary $M/M/1/K$ queue with service speed r_2 and with the special feature that, in every state, additional batches

of customers arrive according to a Poisson process with rate ν. The arrival of such a batch immediately brings the number of customers in the system into state K. For this part of the state space, the balance equations are given by

$$(\lambda + \mu_2 + \nu)\pi(k+1,2) = \lambda\pi(k,2) + \mu_2\pi(k+2,2), \quad k = 0, \ldots, K-1,$$

which immediately leads to

$$\pi(k,2) = C_4 x_2^k + C_5 x_3^k, \quad k = 0, \ldots, K,$$

where x_2 and x_3 are the roots of the equation

$$\mu_2 x^2 - (\lambda + \mu_2 + \nu)x + \lambda = 0. \tag{2}$$

Finally, during periods in which the number of customers in the system is at or below level K and the server works at speed r_1, the system behaves as an ordinary $M/M/1/K$ queue with service speed r_1 and with the special feature that, every time an overflow occurs, the number of customers in the system is instantaneously decreased to the steady-state of the system when the number of customers in the system is at or below level K and the server works at speed r_2. An up- and downcrossing argument for this system yields, for some constants D_1 and D_2,

$$\lambda\pi(k,1) = \mu_1\pi(k+1,1) + \lambda\pi(K,1)\left(D_1 x_2^k + D_2 x_3^k\right),$$

which leads to

$$\pi(k,1) = C_6\rho_1^k + C_7 x_2^k + C_8 x_3^k, \quad k = 0, \ldots, K.$$

To summarize, the steady-state distribution has geometric behaviour with parameter x_1 on the part of the state space $S_{2,1} = \{(K+1,1), (K+2,1), \ldots\}$, a mixture of geometric behaviour with parameter ρ_2 and geometric behaviour with parameter x_1 on the part of the state space $S_{2,2} = \{(K+1,2), (K+2,2), \ldots\}$, a mixture of geometric behaviour with parameter x_2 and geometric behaviour with parameter x_3 on the part of the state space $S_{1,2} = \{(0,2), \ldots, (K,2)\}$, and a mixture of geometric behaviour with parameter ρ_1, geometric behaviour with parameter x_2 and geometric behaviour with parameter x_3 on the part of the state space $S_{1,1} = \{(0,1), \ldots, (K,1)\}$.

3 Service speed based on workload

In this section we consider again an $M/M/1$ queue in which the server can work at two different speeds, r_1 and r_2. However, the service speed is now based on the workload instead of on the number of customers. If the workload in the system is smaller than or equal to level K, then the server should work at speed r_1. If on the other hand the workload in the system is bigger than K, the server should work at speed r_2. As before, because service speed adaptations can only take place at special points in time, it can happen that the server still works at r_2 (resp. r_1) although in the meantime, the workload is below (resp.

above) level K. Notation is as in Section 2 and we assume again that $\rho_2 < 1$. Let $V(t)$ denote the workload in the system at time t. As before, $Y(t)$ represents the state of the server at time t, i.e., $Y(t) = i$ if the speed of the server at time t equals $r_i, i = 1, 2$. With (V, Y) we denote a random vector having the steady-state distribution of the continuous-time Markov process $\{(V(t), Y(t))\}_{t\geq 0}$. With $f_i(x)$ we denote, for $i = 1, 2$, the steady-state conditional density of V, given $Y = i$, i.e., $f_i(x)dx = \mathbb{P}(V \in (x, x + dx) \,|\, Y = i)$.

3.1 Instantaneous service speed adaptations

In the model with instantaneous service speed adaptations, the speed of the server instantaneously changes from r_1 to r_2 if an upcrossing of level K of the workload process occurs. Similarly, the speed of the server instantaneously changes from r_2 to r_1 if a downcrossing of level K of the workload process occurs. Hence, in this case the state space of the continuous-time Markov process $\{(V(t), Y(t))\}_{t\geq 0}$ equals

$$S = \{(x, 1) : 0 \leq x \leq K\} \cup \{(x, 2) : x > K\}.$$

Now, above level K the system behaves as an ordinary (non-empty) $M/M/1$ queue with service speed r_2. On the other hand, below level K the system behaves as a finite $M/M/1$ dam with service speed r_1. Using the fact that the steady-state density of the workload process in the finite $M/M/1$ dam is proportional to the steady-state density of the workload process in the ordinary $M/M/1$ queue (see Hooghiemstra [11]), we conclude that

$$\begin{aligned} f_1(x) &= C_1 \cdot e^{-\mu(1-\rho_1)x}, \quad 0 < x \leq K, \\ f_2(x) &= C_2 \cdot e^{-\mu(1-\rho_2)x}, \quad x > K. \end{aligned}$$

So we conclude that the steady-state density has exponential behaviour with exponent $-\mu(1-\rho_2)x$ on the part of the state space $S_{2,2} = \{(x, 2) : x > K\}$ and exponential behaviour with exponent $-\mu(1 - \rho_1)x$ on the part of the state space $S_{1,1} = \{(x, 1) : 0 \leq x \leq K\}$.

3.2 Service speed adaptations at arrival instants

In the model with service speed adaptations at arrival instants, the speed of the server instantaneously changes from r_1 to r_2 if the workload upcrosses level K due to an arrival. However, if the workload downcrosses level K, the server continues to work at speed r_2 until the next arrival instant. Hence, in this case the state space of the continuous-time Markov process $\{(V(t), Y(t))\}_{t\geq 0}$ equals

$$S = \{(x, 1) : 0 \leq x \leq K\} \cup \{(x, 2) : 0 \leq x \leq K\} \cup \{(x, 2) : x > K\}.$$

As before, above level K the system behaves as an ordinary (non-empty) $M/M/1$ queue with service speed r_2, leading to

$$f_2(x) = C_1 \cdot e^{-\mu(1-\rho_2)x}, \quad x > K.$$

However, below level K the system now behaves as a finite $M/M/1$ dam with service speed r_1 but with the special feature that after each overflow the service speed equals r_2 until the next arrival instant. At an arbitrary instant at which the workload is below K and the server works at speed r_2, we have

$$(V \,|\, V < K, Y = 2) \stackrel{d}{=} \max(K - r_2 A, 0),$$

where A is exponentially distributed with parameter λ, leading to

$$f_2(x) = C_2 \cdot e^{+(\lambda/r_2)x}, \quad 0 < x \leq K.$$

At an arbitrary instant at which the workload is below K and the server works at speed r_1, the system behaves as a finite $M/M/1$ dam with service speed r_1, but with the special feature that after each overflow, the workload is first instantaneously decreased to the steady-state situation of the system when the workload is below K and the service speed equals r_2 and, after that, instantaneously increased by an exponentially distributed amount with parameter μ. The instantaneous decrease plus increase is repeated until after the increase no overflow occurs anymore. Let Z be the level of the workload after these successive decreases and increases. We then have

$$Z \stackrel{d}{=} (\max(K - r_2 A, 0) + B \,|\, \max(K - r_2 A, 0) + B \leq K),$$

where A is exponentially distributed with parameter λ and B is exponentially distributed with parameter μ. For the distribution function $Z(\cdot)$ of Z, we have

$$Z(x) = \frac{e^{\frac{\lambda}{r_2}x} - e^{-\mu x}}{e^{\frac{\lambda}{r_2}K} - e^{-\mu K}}, \quad 0 \leq x \leq K.$$

By using an up- and downcrossing argument we will show in Lemma A.1 in the appendix that in this situation

$$f_1(x) = C_3 \cdot e^{-\mu(1-\rho_1)x} + C_4 \cdot e^{+(\lambda/r_2)x}, \quad 0 < x \leq K.$$

Alternatively, that result could have been obtained by using, e.g., the Kella-Whitt martingale (see Kella and Whitt [13]).

Hence, the steady-state density has exponential behaviour with exponent $-\mu(1-\rho_2)x$ on the part of the state space $S_{2,2} = \{(x, 2) : x > K\}$, exponential behaviour with exponent $+(\lambda/r_2)x$ on the part of the state space $S_{1,2} = \{(x, 2) : 0 < x \leq K\}$ and a mixture of exponential behaviour with exponent $-\mu(1-\rho_1)x$ and exponential behaviour with exponent $+(\lambda/r_2)x$ on the part of the state space $S_{1,1} = \{(x, 1) : 0 < x \leq K\}$.

3.3 Service speed adaptations at external Poisson instants

In the model with service speed adaptations at external Poisson instants, we again assume that these instants occur according to a Poisson process with rate ν. If the workload in the

system upcrosses level K, the server continues to work at speed r_1 until the next external Poisson instant. Similarly, if the workload in the system downcrosses level K, the server continues to work at speed r_2 until the next external Poisson instant. Hence, in this case the state space of the continuous-time Markov process $\{(V(t), Y(t))\}_{t\geq 0}$ equals

$$S = \{(x,1) : 0 \leq x \leq K\} \cup \{(x,1) : x > K\} \cup \{(x,2) : 0 \leq x \leq K\} \cup \{(x,2) : x > K\}.$$

During periods that the workload is above level K and the service speed is equal to r_1, the system behaves as an $M/M/1$ queue with service speed r_1 and with disasters. Here the instants of the disasters correspond to the external Poisson instants and hence occur with rate ν. For this model it is known (see for example Boucherie and Boxma [4] or Jain and Sigman [12]) that the steady-state workload density has exponential behaviour with exponent $-\mu(1-x_1)x$ where x_1 is the root in $(0,1)$ of $\mu_1 x^2 - (\lambda + \mu_1 + \nu)x + \lambda = 0$, see also Equation (1).

During periods that the workload is above level K and the service speed is equal to r_2, the system behaves as an $M/M/1$ queue with service speed r_2 and with special first service time in a busy period. Here the special first service time is hyperexponentially distributed. With some probability it is exponentially distributed with parameter μ and with some probability it is exponentially distributed with parameter $\mu(1-x_1)$. The first case corresponds to the situation that the set $\{(x,2) : x > K\}$ is entered due to the arrival of a customer. The second case corresponds to the situation that the set $\{(x,2) : x > K\}$ is entered due to the occurrence of an external Poisson instant. For this model it is straightforward to show (using, e.g., Formula (2.35) of Takagi [16] for the LST of the steady-state waiting time in an $M/G/1$ queue with exceptional first service time) that the steady-state workload density has a mixture of exponential behaviour with exponent $-\mu(1-\rho_2)x$ and exponential behaviour with exponent $-\mu(1-x_1)x$.

During periods that the workload is below level K and the service speed is equal to r_2, the system behaves as a finite $M/G/1$ dam with arrival rate $\lambda + \nu$, service speed r_2 and service time

$$B = \begin{cases} \exp(\mu), & \text{with probability } \frac{\lambda}{\lambda+\nu} \\ \infty, & \text{with probability } \frac{\nu}{\lambda+\nu}. \end{cases}$$

Remark that an external Poisson instant is modelled here as an arrival of a customer with infinite service time (implying that after the arrival of such a customer the workload in the dam is always equal to K). For this system we can show that the steady-state workload density has a mixture of exponential behaviour with exponent $-\mu(1-x_2)x$ and exponential behaviour with exponent $-\mu(1-x_3)x$, where x_2 and x_3 are the roots of the equation $\mu_2 x^2 - (\lambda + \mu_2 + \nu)x + \lambda = 0$, see also Equation (2). This will be shown in Lemma A.2 in the appendix by using an up- and downcrossing argument.

During periods that the workload is below level K and the service speed is equal to r_1, the system behaves as a finite $M/M/1$ dam with service speed r_1 in which, after each overflow, the workload is instantaneously decreased to the steady-state situation of the system when the workload is below K and the service speed is equal to r_2. The situation is similar to the situation discussed in Section 3.2. Again by using an up- and downcrossings

argument we will show in Lemma A.3 in the appendix that the steady-state workload density has a mixture of exponential behaviour with exponent $-\mu(1-\rho_1)x$, exponential behaviour with exponent $-\mu(1-x_2)x$ and exponential behavour with exponent $-\mu(1-x_3)x$.

4 Service speed adaptations at departure instants

In this section we consider an $M/G/1$ queue with two different service speeds, r_1 and r_2, which can only be adapted at departure instants. Again, denote the arrival rate by λ and let $B(\cdot)$, $\beta(\cdot)$, and β, be the distribution, LST, and mean, respectively, of a generic service requirement. Also, $\rho_i := \lambda\beta/r_i$ represents the traffic load in case of service speed r_i, $i = 1, 2$.

The model with service speed adaptations at departure instants is especially relevant when the number of customers in the system is considered. In that case, we may choose a convenient one-dimensional Markov chain $\{X_n\}_{n\geq 0}$, by embedding the system at departure instants, i.e., X_n represents the number of customers immediately after the n-th departure epoch. It is easy to see that the steady-state distribution of this embedded process is equal to the distribution of the number of customers at an arbitrary instant. The analysis of these *distributions* in the $M/G/1$ queue with speed adaptations at departure instants based on the number of customers constitutes the main subject of this section.

We note that the model with service speed adaptations at departure instants based on *workload* seems very difficult to analyse. The reason for this is that, in order to determine future departure instants of customers, it is necessary to keep track of the individual workloads of the customers in the system and not only of the total workload. This will lead to a complicated state description.

Now, consider the embedded process $\{X_n\}_{n\geq 0}$ and let $Y_n^{(i)}$ be the number of arrivals between the n-th and $(n+1)$-th departure instant given that the service speed is r_i, $i = 1, 2$. Then,

$$X_{n+1} = \begin{cases} \max(X_n - 1, 0) + Y_n^{(1)}, & \text{for } 0 \leq X_n \leq K, \\ \max(X_n - 1, 0) + Y_n^{(2)}, & \text{for } X_n > K. \end{cases} \tag{3}$$

Clearly, the distribution of $Y_n^{(i)}$ is independent of n. We thus have, for all $n = 0, 1, \ldots$,

$$\alpha_k^{(i)} := \mathbb{P}(Y_n^{(i)} = k) = \int_0^\infty \frac{(\lambda x/r_i)^k}{k!} e^{-\lambda x/r_i} \mathrm{d}B(x) = \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} \mathrm{d}\tilde{B}_i(x), \tag{4}$$

with $\tilde{B}_i(x) := B(r_i x)$. Also,

$$\beta_i(z) := \mathbb{E}[z^{Y_n^{(i)}}] = \beta((1-z)\lambda/r_i). \tag{5}$$

From (4), we observe that the model with two different service speeds may be equivalently interpreted as a model with two different service requirement distributions $\tilde{B}_i(\cdot)$ (see also the remark lateron in this section). Hence, this model is in fact a special case of the general model considered in, e.g., [5, 14]. These papers focus on results in terms of LST.

(See [9, 15] for computational algorithms and [8] for various related models.) In this section, however, we give explicit and intuitively appealing expressions for the steady-state number of customers present.

Let X be the steady-state random variable of X_n, and denote $\pi(k) := \mathbb{P}(X = k)$. Next, we obtain $\pi(k)$ for $k \in \{0, \ldots, K\}$ and then use that result to determine $\pi(k)$ for $k \in \{K+1, K+2, \ldots\}$. Finally the probability of a customer leaving behind an empty system follows from normalization.

First, we derive $\pi(k)$ for $k = 0, \ldots, K$. Consider the conditional process $\{X_n | X_n \leq K\}$, which is obtained by deleting the periods that the number of customers at departure instants is larger than K and pasting together the remaining parts. It is easy to see that the one-step transition matrix of this conditional process is equal to the transition matrix of the $M/G/1/K+1$ queue with service speed r_1, or of an $M/G/1$ queue with speed r_1 conditioned on the event that the number of customers upon departure does not exceed K. Denote by $X^{(i)}$, $i = 1, 2$, the steady-state number of customers in an $M/G/1$ queue with service speed r_i, i.e., the z-transform of $X^{(i)}$ reads

$$\mathbb{E}[z^{X^{(i)}}] = \frac{(1-\rho_i)\beta_i(z)(1-z)}{\beta_i(z) - z},$$

with $\beta_i(\cdot)$, $i = 1, 2$, given in (5). Hence, for $k = 0, 1, \ldots, K$, we have

$$\pi(k) = \frac{\pi(0)}{1-\rho_1}\mathbb{P}(X^{(1)} = k), \tag{6}$$

where the constant $\pi(0)/(1-\rho_1)$ follows from $k = 0$. In general, we allow for $\rho_1 \geq 1$. In that case, $X^{(1)}$ corresponds to the steady-state number of customers immediately after departure instants of the $M/G/1/K+1$ queue, and $1-\rho_1$ has to be replaced by $\mathbb{P}(X^{(1)} = 0)$. However, in the presentation of the results below we have taken $\rho_1 < 1$.

For the distribution of X on $\{K+1, K+2, \ldots\}$, we use the equivalence with the $M/G/1$ queue with two service requirement distributions $\tilde{B}_i(\cdot)$. In particular, rewriting [5, Equation (4.19)] (where $c = p_0\beta_2(0)$), we obtain

$$\begin{aligned}
\mathbb{E}[z^X] &= \pi(0)\frac{\beta_2(z) - z\beta_1(z)}{\beta_2(z) - z} + \sum_{k=1}^{K} z^{k-1}\pi(k)\frac{z\beta_2(z) - z\beta_1(z)}{\beta_2(z) - z} \\
&= \pi(0)\frac{\beta_2(z) - z\beta_1(z)}{\beta_2(z) - z} + \frac{1}{z}\sum_{k=1}^{K} z^k\pi(k)\left(\frac{\beta_2(z) - z\beta_1(z)}{\beta_2(z) - z} - \frac{(1-z)\beta_2(z)}{\beta_2(z) - z}\right). \qquad (7)
\end{aligned}$$

To obtain the distribution of X on $\{K+1, K+2, \ldots\}$ we invert each of the three terms in the above equation separately. For the first term, we note that, upto a constant, this corresponds to an $M/G/1$ queue with service speed r_2 and an exceptional first service time with distribution $B(xr_2/r_1)$, see [16, Equation (2.38)]. Denoting the steady-state number of customers in such a queue by $X^{(2)}_{\text{exc}}$, the inverse of the first term on the rhs of (7) reads

$$\pi(0)\frac{1+\rho_1-\rho_2}{1-\rho_2}\mathbb{P}(X^{(2)}_{\text{exc}} = k).$$

For the second and third term on the rhs of (7) we note that this involves a product of two z-transforms corresponding to the sum of two independent random variables. It is readily seen that $\sum_1^K z^{k-1}\pi(k)$ is the transform of $X-1$, with X restricted to $\{1,2,\ldots,K\}$. Hence, using (6) and the result for an $M/G/1$ queue with exceptional first service again, the inverse of the second term equals

$$\frac{1+\rho_1-\rho_2}{1-\rho_2}\sum_{j=1}^{K}\frac{\pi(0)}{1-\rho_1}\mathbb{P}(X^{(1)}=j)\mathbb{P}(X^{(2)}_{\rm exc}=k+1-j).$$

Similarly, applying (6) and the result for the standard $M/G/1$ queue with service speed r_2 to the third term yields

$$\frac{1}{1-\rho_2}\sum_{j=1}^{K}\frac{\pi(0)}{1-\rho_1}\mathbb{P}(X^{(1)}=j)\mathbb{P}(X^{(2)}=k+1-j).$$

Summarizing, for $k\in\{K+1,K+2,\ldots\}$, we have

$$\begin{aligned}\pi(k) &= \frac{\pi(0)}{1-\rho_2}\Bigg((1+\rho_1-\rho_2)\mathbb{P}(X^{(2)}_{\rm exc}=k)+\frac{1+\rho_1-\rho_2}{1-\rho_1}\sum_{j=1}^{K}\mathbb{P}(X^{(1)}=j)\mathbb{P}(X^{(2)}_{\rm exc}=k+1-j)\\ &\quad -\frac{1}{1-\rho_1}\sum_{j=1}^{K}\mathbb{P}(X^{(1)}=j)\mathbb{P}(X^{(2)}=k+1-j)\Bigg). \qquad (8)\end{aligned}$$

The first term clearly stems from scenarios in which the first customer in a busy cycle leaves behind more than K new customers upon departure, resulting in an ordinary $M/G/1$ queue with exceptional first service (i.e., the service requirement multiplied by the ratio r_2/r_1). The third term on the rhs of (8) is in fact a correction term. To give an intuitively more appealing expression, we rewrite (7) as

$$\mathbb{E}[z^X]=\pi(0)\frac{\beta_2(z)-z\beta_1(z)}{\beta_2(z)-z}+\sum_{k=1}^{K}z^k\pi(k)\frac{\beta_2(z)-z\mathbb{E}[z^{Y^{(1)}-1}]}{\beta_2(z)-z}.$$

Note that we did not rewrite the intuitively appealing first term. The second term corresponds to a convolution of X (restricted to $\{1,\ldots,K\}$) with a special type of $M/G/1$ queue with exceptional first customer. The special feature is the -1 in $\mathbb{E}[z^{Y^{(1)}-1}]$ accounting for the fact that the number of arrivals in between two successive departure instants has to be corrected for the actual departure of the customer in service. In case $X=0$ this correction is not required as represented in the first term. After the arrival of customers during the first exceptional period in which the speed is r_1, the queue continues as an ordinary $M/G/1$ queue with service speed r_2.

Finally, the probability of an empty system $\pi(0)$ can be found from normalization. More specifically, letting $z\to 1$ in (7) and using (6), we obtain

$$\pi(0)=(1-\rho_2)\left[\frac{\rho_1-\rho_2}{1-\rho_1}\mathbb{P}(X^{(1)}\le K)+1\right]^{-1}.$$

Remark *We note that the results can be easily generalised. Consider the case in which, in addition to the service rate, the service requirement distribution also depends on the number of customers upon departure (having distribution $B_i(\cdot)$). Since the distribution of X_n only depends on the ratio between λ and r_i, we may choose λ fixed without loss of generality (however, special care is required for the distribution at arbitrary instants).*

Now, similar to (4), observe that, for $i = 1, 2$,

$$\alpha_k^{(i)} = \int_0^\infty \frac{(\lambda x/r_i)^k}{k!} e^{-\lambda x/r_i} \mathrm{d}B_i(x) = \int_0^\infty \frac{(\lambda x/r_{3-i})^k}{k!} e^{-\lambda x/r_{3-i}} \mathrm{d}\tilde{B}_i(x),$$

with $\tilde{B}_i(x) = B_i(xr_i/r_{3-i})$. Hence, it may be easily verified that the expressions (6) and (8) for the steady-state distribution of X hold by obvious modifications of $\beta_i(\cdot)$ and definition of the exceptional first service.

Finally, as an example, we consider the number of customers at an arbitrary instant in the $M/M/1$ case. In that case, Equations (6) and (8) reduce to nice tractable expressions. In particular, for $k = 0, \ldots, K$, we have ordinary $M/M/1$ behaviour, yielding

$$\pi(k) = \pi(k, 1) = C_1 \rho_1^k.$$

For $k \in \{K+1, K+2, \ldots\}$, we consider the joint distribution of the service speed and number of customers. First, on the part of the state space

$$S_{2,1} = \{(K+1, 1), (K+2, 1), \ldots\},$$

we have geometric behavior with parameter $\lambda/(\lambda + \mu_1)$, since we consider the number of arrivals before the first departure. On the part of the state space

$$S_{2,2} = \{(K+1, 2), (K+2, 2), \ldots\}$$

we have an $M/M/1$ queue with service speed r_2 and an exceptional first service time which is the result of the behavior on $S_{2,1}$. Summarizing, for $k = 0, 1, 2, \ldots$, we have

$$\begin{aligned} \pi(K+k+1, 1) &= C_2 \left(\frac{\lambda}{\lambda + \mu_1} \right)^k \\ \pi(K+k+2, 2) &= C_3 \left(\frac{\lambda}{\lambda + \mu_1} \right)^k + C_4 \rho_2^k. \end{aligned}$$

5 Conclusion and extensions to $M/G/1$

In this paper, we considered various queueing models in which the server can work at two different service speeds. The speed of the server may depend on either the number of customers present, or the workload. In the former case, we analysed the steady-state number of customers, while in the latter case, we determined the workload density in steady state. Our specific aim was to give insight into the structure of the steady-state results by

Table 1: Structure of the steady-state results in various $M/M/1$ queues (without constants).

Model: adaptations	service speed	Number of customers		Workload	
		$\pi(k)$, $k \leq K$	$\pi(k)$, $k > K$	$f(x)$, $x \leq K$	$f(x)$, $x > K$
Continuous	r_1	$\text{geo}(\rho_1)$	-	$\exp(-\mu(1-\rho_1)x)$	-
	r_2	-	$\text{geo}(\rho_2)$	-	$\exp(-\mu(1-\rho_2)x)$
Arrival instants	r_1	$\text{geo}(\rho_1)$ $+ \text{geo}(\frac{\lambda+\mu_2}{\mu_2})$	-	$\exp(-\mu(1-\rho_1)x)$ $+ \exp(\frac{\lambda}{r_2}x)$	-
	r_2	$\text{geo}(\frac{\lambda+\mu_2}{\mu_2})$	$\text{geo}(\rho_2)$	$\exp(\frac{\lambda}{r_2}x)$	$\exp(-\mu(1-\rho_2)x)$
Poisson instants	r_1	$\text{geo}(\rho_1)$ $+ \text{geo}(x_2)$ $+ \text{geo}(x_3)$	$\text{geo}(x_1)$	$\exp(-\mu(1-\rho_1)x)$ $+ \exp(-\mu(1-x_2)x)$ $+ \exp(-\mu(1-x_3)x)$	$\exp(-\mu(1-x_1)x)$
	r_2	$\text{geo}(x_2)$ $+ \text{geo}(x_3)$	$\text{geo}(\rho_2)$ $+ \text{geo}(x_1)$	$\exp(-\mu(1-x_2)x)$ $+ \exp(-\mu(1-x_3)x)$	$\exp(-\mu(1-\rho_2)x)$ $+ \exp(-\mu(1-x_1)x)$

using probabilistic arguments. The structure of the results, by neglecting the constants, is summarised in Table 1, where x_1 is the root in $(0, 1)$ of Equation (1) and x_2 and x_3 are the roots of (2).

We also considered the $M/G/1$ queue with two service rates depending on the number of customers right after departure instants, yielding an intuitively appealing form for the steady-state distribution of the number of customers present. The LST of the steady-state number of customers has been analysed by other authors, see e.g. [5, 14]. However, to the best of our knowledge, no explicit formula for its distribution has been given. Giving $M/G/1$ analogs for the other models with adaptations based on the number of customers (of Section 2) remains a significant challenge. The difficulty with these models is that a more complex state description is required to obtain a Markov process.

In case the service speed is based on the workload, it is possible to generalise the models of Section 3 to their $M/G/1$ variants. The model with continuous adaptations (based on the workload) is the classical model, which is also often refered to as dam model, see e.g. [6, 10], or [7], p. 556 for the results. In fact, its steady-state workload distribution shows some similarities with the steady-state number of customers of the model with speed adaptations at departure instants. For the $M/G/1$ case of the model with adaptations at arrival instants, see [1]. The model with adaptations at Poisson instants is in fact generalised to the more general case of Lévy processes without negative jumps, see [3].

Appendix

Lemma A.1 *Consider a finite $M/M/1$ dam with capacity K and with the special feature that every time an overflow of the dam occurs, the workload is instantaneously, and independently of the past, decreased to the random level Z, where Z has probability distribution*

$$Z(x) = \frac{e^{\frac{\lambda}{r_2}x} - e^{-\mu x}}{e^{\frac{\lambda}{r_2}K} - e^{-\mu K}}, \quad 0 \le x \le K.$$

The service times are exponentially distributed with parameter μ. Denote with λ the arrival rate, r_1 the service speed and let $\rho_1 := \lambda/(\mu r_1)$. For the steady-state workload density we then have

$$v(x) = C_3 e^{-\mu(1-\rho_1)x} + C_4 e^{\frac{\lambda}{r_2}x}, \quad 0 < x \le K,$$

for some constants C_3 and C_4.

Proof: Denote with $V(x)$ the steady-state workload distribution. An up- and downcrossing argument for level x of the workload then yields

$$\lambda \int_0^x v(y) e^{-\mu(x-y)} dy + \lambda V(0) e^{-\mu x} = r_1 v(x) + \lambda Z(x) \int_0^K v(y) e^{-\mu(K-y)} dy. \tag{9}$$

Now by multiplying (9) with $e^{\mu x}$ and by introducing $f(x) := v(x)e^{\mu x}$ we obtain

$$\lambda \int_0^x f(y) dy + \lambda V(0) = r_1 f(x) + \lambda Z(x) e^{\mu x} \int_0^K f(y) e^{-\mu K} dy. \tag{10}$$

After differentiation of (10) and using the distribution $Z(\cdot)$ we obtain

$$\lambda f(x) = r_1 f'(x) + C e^{\left(\frac{\lambda}{r_2} + \mu\right)x}. \tag{11}$$

The solution of (11) is given by

$$f(x) = C_3 e^{\frac{\lambda}{r_1}x} + C_4 e^{\left(\frac{\lambda}{r_2} + \mu\right)x}.$$

and hence

$$v(x) = f(x)e^{-\mu x} = C_3 e^{\left(\frac{\lambda}{r_1} - \mu\right)x} + C_4 e^{\frac{\lambda}{r_2}x} = C_3 e^{-\mu(1-\rho_1)x} + C_4 e^{\frac{\lambda}{r_2}x}.$$

Lemma A.2 *Consider a finite $M/G/1$ dam with capacity K, arrival rate $\lambda + \nu$, service speed r_2 and service time*

$$B = \begin{cases} exp(\mu), & \text{with probability } \frac{\lambda}{\lambda+\nu} \\ \infty, & \text{with probability } \frac{\nu}{\lambda+\nu}. \end{cases}$$

The steady-state workload distribution is then given by

$$V(x) = D_2 e^{-\mu(1-x_2)x} + D_3 e^{-\mu(1-x_3)x}, \quad 0 \le x \le K,$$

for some constants D_2 and D_3, where x_2 and x_3 are the roots of the equation

$$\mu_2 x^2 - (\lambda + \mu_2 + \nu)x + \lambda = 0. \tag{12}$$

Proof: An up- and downcrossings argument for level x of the buffer content yields

$$r_2 v(x) = \lambda \int_0^x v(y) e^{-\mu(x-y)} dy + \lambda V(0) e^{-\mu x} + \nu V(x).$$

Using that $V'(x) = v(x)$, we obtain after differentiation

$$\begin{aligned} r_2 V''(x) &= (\lambda + \nu)V'(x) - \mu \left(\lambda \int_0^x V'(y) e^{-\mu(x-y)} dy + \lambda V(0) e^{-\mu x} \right) \\ &= (\lambda + \nu)V'(x) - \mu \left(r_2 V'(x) - \nu V(x) \right). \end{aligned}$$

The solution of this second-order differential equation is given by

$$V(x) = D_2 e^{\xi_2 x} + D_3 e^{\xi_3 x}$$

where ξ_2 and ξ_3 are zero's of the characteristic polynomial of the differential equation, i.e.,

$$r_2 \xi^2 - (\lambda + \nu - \mu r_2)\xi - \mu\nu = 0. \tag{13}$$

It is straightforward to show that ξ_i is a solution of (13) if and only if $\xi_i = -\mu(1 - x_i)$, where x_i is a solution of (12).

Remark *We note that the model of Lemma A.2 shows some similarities with a finite $M/M/1$ dam with clearings at exponential times. However, in this case, the clearing instants do not remove all the work present, but instantaneously move the workload to the upper boundary, K.*

Lemma A.3 *Consider the finite $M/M/1$ dam of Lemma A.1, but let the probability distribution of Z now be given by*

$$Z(x) = D_2 e^{-\mu(1-x_2)x} + D_3 e^{-\mu(1-x_3)x}, \quad 0 \le x \le K.$$

For the steady-state workload density we then have

$$v(x) = C_1 e^{-\mu(1-\rho_1)x} + C_2 e^{-\mu(1-x_2)x} + C_3 e^{-\mu(1-x_3)x}, \quad 0 < x < K,$$

for some constants C_1, C_2 and C_3.

Proof: Like in the proof of Lemma A.1 we introduce $f(x) := v(x)e^{\mu x}$, which satisfies (10). After differentiation of (10) we now obtain

$$\lambda f(x) = r_1 f'(x) + \tilde{D}_2 e^{\mu x_2 x} + \tilde{D}_3 e^{\mu x_3 x}. \tag{14}$$

The solution of (14) is given by

$$f(x) = C_1 e^{\frac{\lambda}{r_1}x} + C_2 e^{\mu x_2 x} + C_3 e^{\mu x_3 x},$$

and hence

$$v(x) = f(x)e^{-\mu x} = C_1 e^{-\mu(1-\rho_1)x} + C_2 e^{-\mu(1-x_2)x} + C_3 e^{-\mu(1-x_3)x}.$$

Remark *We note that the model of Lemma's A.1 and A.3 are related to a finite $M/M/1$ dam with exceptional first service time. However, in this case, a special period starts after hitting the upper boundary, instead of an exceptional first service occuring due to an arrival while the system is empty.*

Acknowledgments: The research was done within the framework of the BRICKS project and the European Network of Excellence Euro-NGI. Part of the research was done while the first author was affiliated to CWI, Amsterdam, The Netherlands.

References

[1] Bekker, R. (2005). Queues with state-dependent rates. Ph.D. Thesis, Eindhoven University of Technology, The Netherlands.

[2] Bekker, R., and O.J. Boxma (2005). *Queues with adaptable service speed.* In: B.D. Choi (ed.), Korea-Netherlands Joint Conference on Queueing Theory and its Applications to Telecommunication Systems, 91–100.

[3] Bekker, R., O.J. Boxma, and J.A.C. Resing (2006). Queueing systems with adaptable Lévy input. Preprint.

[4] Boucherie, R.J., and O.J. Boxma (1996). The workload in the $M/G/1$ queue with work removal. *Probability in the Engineering and Informational Sciences* **10**, 1–20.

[5] Boxma, O.J., and V.I. Lotov (1996). On a class of one-dimensional random walks. *Markov Processes and Related Fields* **2**, 349–362.

[6] Cohen, J.W. (1976). On the optimal switching level for an $M/G/1$ queueing system. *Stochastic Processes and Their Applications* **4**, 297–316.

[7] Cohen, J.W. (1982). *The Single Server Queue*, North-Holland, Amsterdam.

[8] Dshalalow, J.H. (1997). Queueing systems with state dependent parameters. In: *Frontiers in Queueing: Models and Applications in Science and Engineering*, 61–116.

[9] Federgruen, A., and H.C. Tijms (1980). Computation of the stationary distribution of the queue size in an $M/G/1$ queueing system with variable service rate. *Journal of Applied Probability* **17**, 515–522.

[10] Gaver, D.P., and R.G. Miller (1962). Limiting distributions for some storage problems. In: *Studies in Applied Probability and Management Science*, 110–126.

[11] Hooghiemstra, G. (1987). A path construction for the virtual waiting time of an $M/G/1$ queue, *Statistica Neerlandica* **41**, 175–181.

[12] Jain, G., and K. Sigman (1996). A Pollaczek-Khintchine formula for $M/G/1$ queues with disasters. *Journal of Applied Probability* **33**, 1191–1200.

[13] Kella, O., and W. Whitt (1992). Useful martingales for stochastic storage processes with Lévy input. *Journal of Applied Probability* **29**, 396–403.

[14] Roughan, M., and C.E.M. Pearce (2002). Martingale methods for analysing single-server queues. *Queueing Systems* **41**, 205–239.

[15] Schellhaas, H. (1986). Computation of the state probabilities in a class of semiregenerative queueing models. In: *Semi-Markov models (Brussels, 1984)*, 111–130.

[16] Takagi, H. (1991). *Queueing Analysis*. Vol. 1, North-Holland, Amsterdam.