Queues with waiting time dependent service

R. Bekker[†], G.M. Koole[†], B.F. Nielsen[⋆], T.B. Nielsen[⋆]

[†]Dept. Mathematics
VU University Amsterdam
De Boelelaan 1081, 1081 HV, the Netherlands.

[⋆]Dept. Informatics and Mathematical Modelling
Technical University of Denmark
Richard Petersens Plads, 2800 Kgs. Lyngby, Denmark.

Abstract

Motivated by service levels in terms of the waiting time distribution seen in e.g. call centers, we consider two models for systems with a service disciple that depends on the waiting time. The first model deals with a single server that continuously adapts its service rate based on the waiting time of the first customer in line. In the second model, one queue is served by a primary server which is supplemented by a secondary server when the waiting of the first customer in line exceeds a threshold. Using level crossings for the waiting time process of the first customer in line, we derive steady-state waiting time distributions for both models. The results are illustrated with numerical examples.

Keywords: Waiting-time distribution; Adaptive service rate; Call centers; Contact centers; Queues; Deterministic threshold; Overflow; Level crossing.

1 Introduction

In service systems, the tail probability (or distribution function) of the waiting time of customers is one of the main service-level indicators. For example, in call centers the service level is generally characterized by the telephone service factor (TSF), i.e., the fraction of calls whose delay fall below a prespecified target. Typically, call centers use a 80-20 TSF meaning that 80% of the calls should be taken into service within 20 seconds, see [9]. Motivated by performance measures in terms of tail probabilities of waiting times, we consider queueing systems where the service mechanism is based on waiting times of customers. This differs from the traditional queueing literature, where routing and control are commonly based on the number of customers present.
The main goal of this paper is to find the steady-state waiting time distribution for queueing systems where the service characteristics depend on the waiting time of the first customer in line. This type of service control seems to be new in the queueing literature, despite its widespread use in the industry. In the sequel we use FIL as an abbreviation of first customer in line.
We consider two Markovian queueing models: (i) single-server queues with FIL waiting-time dependent service speed and (ii) a queue with two heterogeneous servers, where the

secondary server is only activated as soon as the FIL waiting time exceeds some target level. For both models, the analysis is based on the waiting process of the first customer in line (FIL-process). Using level crossings, we find the steady-state distribution of the FIL-process and derive the waiting-time distribution as a corollary.

First, in Section 2, we study the single-server model, where the service speed can be continuously adapted based on the waiting time of the first customer in line. This model is related to the study of dams and queueing systems with workload-dependent service rates, see e.g. [3], [4], [13] or [20]. The difference is that the service speed here depends on the waiting time instead of the amount of work present.

Second, in Section 3, we consider a system with a single queue and two heterogeneous servers, where the secondary server takes the first customer in line into service as soon as his waiting time exceeds some threshold. The primary motivation for this model stems from routing mechanisms in call centers with operators in front and back offices. Typically, the only task of operators in the front office would be to answer calls whereas operators in the back office would have other assignments and only answer calls under high load. A common problem is then how to meet the service level agreements while keeping the disturbance of the back office operators to a minimum, see [9] and references therein. Overflow problems are in general difficult to analyze, see [15], because the overflow traffic is not Poisson; the deterministic threshold of this model only adds to this. We believe though that the model is of independent interest and has its applications in other areas where the service level involves the (tail) distribution of the waiting time, as in, e.g., telecommunication and production systems.

Related to the heterogeneous-servers model above is the slow-server problem, see [14], [16], [19] and [21]. In the slow-server model, a single queue is served by two heterogeneous servers with service rates μ_1 and μ_2, where $\mu_1 > \mu_2$. In [19], the author gives qualitative and explicit quantitative results on when to maintain or discard the slow server. In the models of [14] and [16], customers can be assigned to one of the servers depending on the number of customers present. There it was shown that the fast server should always be used and that the slow server should only be used if the number of customers exceeds some threshold. This result was derived for an infinite waiting space. We note that in case of a finite queue length, the optimal policy is not necessarily of a threshold type, see [21].

The literature on queueing models where the service time process depends on the waiting time is limited. There are some studies of single-server queues where the service time depends on the waiting time experienced by the customer in service (instead of the first customer in line), see [5], [18] and [22]. Furthermore, in [6] the authors consider an M/M/2 queue where non-waiting customers receive a different rate of service than customers who first wait in line. Their analysis is based on the "system point method" [7], which is closely related to the level crossing equations of Section 3.

Some numerical results are presented in Section 4. Conclusions and topics for further research can be found in Section 5.

2 Single-server queue

In this section we consider a single-server queue where the service speed depends on the waiting time of the first customer in line. In particular, we assume that customers arrive according to a Poisson process with rate λ and have exponentially distributed service requirements with mean $1/\mu$. The service discipline is assumed to be FIFO. Denote by W_t the waiting time of the first customer in the queue at time t, with the convention that $W_t = 0$ if the queue is empty. Also, let Y_t denote the number of customers in service at time t (thus $Y_t \in \{0,1\}$). The service speed depends on the waiting time of the first customer in line and the service speed function is denoted by $r(\cdot)$. Let $r(0)$ be the service speed for state $(W_t, Y_t) = (0,1)$ and 0 be the speed for state $(0,0)$. For convenience, define $\rho_0 = \lambda/(\mu r(0))$. We assume that $r(\cdot)$ is strictly positive, left-continuous, and has a strictly positive right limit on $(0,\infty)$.

The process $\{(W_t, Y_t), t \geq 0\}$ can now be described as follows. Given that $W_{t_0} = w > 0$ and the next service completion is at time $t_1 > t_0$, the waiting time process of the first customer in line during (t_0, t_1) behaves as $W_{t_0+t} = w + t$ and $Y_{t_0+t} = 1$. If S_w denotes the time until the next service completion, conditioned on the initial waiting time $w > 0$, then $\mathbb{P}(S_w > t) = \exp\left(-\mu \int_w^{w+t} r(y)\mathrm{d}y\right)$. At the moment of a service completion, the second customer in line (if there is any) becomes the first customer in line. Since the interarrival times between customers are exponentially distributed, we have

$$W_{t_1^+} = \left(W_{t_1^-} - A_\lambda\right)^+, \tag{1}$$

where $(x)^+ = \max\{x, 0\}$ and A_λ denotes an exponential random variable of rate λ.

It remains to specify the boundary cases of an empty queue. For $(0, Y_{t_0})$, the time until the next state transition has an exponential distribution with rate $\lambda + \mu r(0) Y_{t_0}$. For $(0,1)$ the next state is $(0,0)$ with probability $\mu r(0)/(\lambda + \mu r(0))$, or W_t starts to increase linearly as described above with probability $\lambda/(\lambda + \mu r(0))$. For $(0,0)$, the next state is $(0,1)$ with probability one.

Since the service requirements and interarrival times are exponentially distributed, the process $\{(W_t, Y_t), t \geq 0\}$ is a Markov process. We assume that this process is has a stationary distribution (we refer to [8, Corollary 4.2] for stability conditions). Below, we determine the steady-state distribution of this process and derive from it the waiting time distribution of an arbitrary customer. For this, we introduce the steady-state distribution of the FIL-process as $W^{\mathrm{FIL}}(x) = \lim_{t\to\infty} \mathbb{P}(W_t \leq x)$ and the corresponding density as $w^{\mathrm{FIL}}(x) = \mathrm{d}W^{\mathrm{FIL}}(x)/\mathrm{d}x$. For the atom in zero, Y_t is included in the notation as $W^{\mathrm{FIL}}(0, y) = \lim_{t\to\infty} \mathbb{P}(W_t = 0, Y_t = y)$.

Theorem 2.1 *We have $W^{\mathrm{FIL}}(0,1) = \rho_0 W^{\mathrm{FIL}}(0,0)$. The density of the FIL-process is*

$$w^{\mathrm{FIL}}(x) = \lambda \rho_0 W^{\mathrm{FIL}}(0,0) \exp\left\{\int_0^x (\lambda - \mu r(y))\mathrm{d}y\right\},$$

where

$$W^{\mathrm{FIL}}(0,0) = \left[1 + \rho_0 + \lambda\rho_0 \int_0^\infty \exp\left\{\int_0^x (\lambda - \mu r(y))\mathrm{d}y\right\}\mathrm{d}x\right]^{-1}.$$

Proof The proof is based on level crossing arguments. For $x > 0$, using (1), the level crossing equations read

$$w^{\mathrm{FIL}}(x) = \int_{y=x}^\infty e^{-\lambda(y-x)} \mu r(y) w^{\mathrm{FIL}}(y)\mathrm{d}y. \tag{2}$$

The left-hand side corresponds to upcrossings of level x and the right-hand side corresponds to the long-run average number of downcrossings through level x. Observe that we have continuous upcrossings of waiting time levels and downcrossings by jumps, where the jump sizes correspond to interarrival times between successive customers (in contrast to workloads in single-server queues). Taking derivatives on both sides of Equation (2) yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}w^{\mathrm{FIL}}(x) &= \lambda\left[\int_{y=x}^{\infty} e^{-\lambda(y-x)}\mu r(y)w^{\mathrm{FIL}}(y)\mathrm{d}y\right] - \mu r(x)w^{\mathrm{FIL}}(x)\\ &= (\lambda-\mu r(x))w^{\mathrm{FIL}}(x),\end{aligned}$$

where the second step follows from (2). The solution of this first-order differential equation can be readily obtained as

$$w^{\mathrm{FIL}}(x) = C\exp\left\{\int_0^x(\lambda-\mu r(y))\mathrm{d}y\right\}. \tag{3}$$

Balancing the transitions between the interior part of the state space and the boundary part, we have

$$\lambda W^{\mathrm{FIL}}(0,1) = \int_0^{\infty} e^{-\lambda y}\mu r(y)w^{\mathrm{FIL}}(y)\mathrm{d}y.$$

Using the above and letting $x\downarrow 0$ in (2) yields $\lim_{x\downarrow 0} w^{\mathrm{FIL}}(x) = \lambda W^{\mathrm{FIL}}(0,1)$. Also, letting $x\downarrow 0$ in (3) determines the constant $C = \lim_{x\downarrow 0} w^{\mathrm{FIL}}(x) = \lambda W^{\mathrm{FIL}}(0,1)$.

Now, balancing the transitions between the two boundary states gives

$$\lambda W^{\mathrm{FIL}}(0,0) = \mu r(0)W^{\mathrm{FIL}}(0,1),$$

which enables us to determine the three constants in terms of $W^{\mathrm{FIL}}(0,0)$. Finally, using normalization, we have

$$W^{\mathrm{FIL}}(0,0)+W^{\mathrm{FIL}}(0,1)+\lambda W^{\mathrm{FIL}}(0,1)\int_0^{\infty}\exp\left\{\int_0^x(\lambda-\mu r(y))\mathrm{d}y\right\}\mathrm{d}x = 1.$$

Expressing $W^{\mathrm{FIL}}(0,1)$ in $W^{\mathrm{FIL}}(0,0)$ and solving for $W^{\mathrm{FIL}}(0,0)$ completes the proof. □

To determine the waiting time, we only need to consider the FIL-process at specific points in time. We introduce the waiting time an arbitrary customer experiences as W and the distribution of this as $W(x) = \mathbb{P}(W \le x)$. Using PASTA, it is easy to see that the atom in zero of the waiting time is given by $\mathbb{P}(W=0) = W^{\mathrm{FIL}}(0,0)$. In case of non-zero waiting times, the waiting times are given by the FIL-process embedded at epochs just before downward jumps.

Let $N_s(u,v)$ denote the number of customers taken into service during the interval $(u,v]$. Consider an infinitesimal interval $(t,t+h]$, $h>0$. Then, $\mathbb{P}(W_t > x; N_s(t,t+h)=1) = \int_x^{\infty}\mu r(y)hw^{\mathrm{FIL}}(y)\mathrm{d}y + o(h)$. Note that $\mathbb{P}(N_s(t,t+h)=1)/h$ (for $h\to 0$) is the rate at which customers are taken into service and, since every customer leaves the queue through the server and the system is stable, equals λ. Combining the above, we have

$$\begin{aligned}\mathbb{P}(W>x) &= \lim_{h\to 0}\mathbb{P}(W_t > x \mid N_s(t,t+h)=1)\\ &= \lim_{h\to 0}\frac{\mathbb{P}(W_t>x; N_s(t,t+h)=1)}{\mathbb{P}(N_s(t,t+h)=1)}\\ &= \frac{1}{\lambda}\int_x^{\infty}\mu r(y)w^{\mathrm{FIL}}(y)\mathrm{d}y.\end{aligned}$$

The density of the steady-state waiting time, $w(x)$, can be obtained by differentiating the above:

Corollary 2.1 *For the steady-state waiting time, we have* $\mathbb{P}(W = 0) = W^{\mathrm{FIL}}(0,0)$ *and density*

$$w(x) = \frac{\mu r(x) w^{\mathrm{FIL}}(x)}{\lambda},$$

where $W^{\mathrm{FIL}}(0,0)$ *and* $w^{\mathrm{FIL}}(\cdot)$ *are given in Theorem 2.1.*

Remark 2.1 We note that the steady-state waiting time and FIL distributions take a similar form as the steady-state workload distribution of an M/M/1 queue with workload-dependent arrival and/or service rate, see e.g. [3], [13] or [2], p. 388. Also related is the workload process in an on/off storage system with workload-dependent rates, [4], [20] and the elapsed waiting time process in the M/G/1 queue [17]. $\diamond$

Remark 2.2 For a renewal arrival process, the interior part of the state space can be straightforwardly adapted. In particular, W_t is still a Markov process for positive waiting times and the level crossing equation (2) then reads

$$w^{\mathrm{FIL}}(x) = \int_{y=x}^{\infty} \mu r(y) w^{\mathrm{FIL}}(y)(1 - A(y - x))\mathrm{d}y,$$

where $A(\cdot)$ is the interarrival-time distribution. Note that the above equation can be written as a Volterra integral equation of the second kind, see e.g. [23]. For the FIL process to be a Markov process, a supplementary variable is required to describe the elapsed interarrival time at the boundary of the state space, i.e., in case there is no customer in line. $\diamond$

Example 2.1 The results become even more tractable in various special cases. Here, we consider the case of two service speeds determined by a threshold value of the waiting time of the first customer in the queue. Specifically, we assume that

$$r(x) = \begin{cases} r_1, & \text{for } 0 \leq x \leq K, \\ r_2, & \text{for } x > K. \end{cases}$$

This example may serve as an approximation for the case of two heterogeneous servers in Section 3, where the secondary server is only activated as soon as the FIL-process exceeds K.

Using Theorem 2.1 and Corollary 2.1, we may easily obtain the steady-state distribution of the FIL-process and the waiting time. Here, we present the atom in zero and the density of the waiting time. Let $\rho_i = \lambda/(\mu r_i)$, for $i = 1, 2$. After some straightforward calculations, we obtain

$$w(x) = \begin{cases} r_1 \mu \rho_1 W(0) e^{-r_1 \mu (1-\rho_1) x}, & \text{for } 0 < x \leq K, \\ r_2 \mu \rho_1 W(0) e^{(r_2 - r_1)\mu K} e^{-r_2 \mu (1-\rho_2) x}, & \text{for } x > K, \end{cases}$$

where

$$W(0) = \left[\frac{1}{1-\rho_1} + \rho_1 e^{-r_1 \mu (1-\rho_1) K} \left(\frac{1}{1-\rho_2} - \frac{1}{1-\rho_1} \right) \right]^{-1}.$$

3 Two-server queue

In this section we turn our attention to a system with two heterogeneous servers. As in Section 2 we use the concept of a FIL-process, where W_t denotes the waiting time of the first customer in line at time t. Again customers arrive to the queue according to a Poisson process with rate λ. A primary server handles jobs with exponentially distributed service times with mean $1/\mu_p$. A secondary server starts serving customers when W_t exceeds a threshold K. The service times at the secondary server are exponentially distributed with mean $1/\mu_s$. As in the one-server model of Section 2, the service discipline is FIFO and the servers will always complete a started job, i.e., the secondary server will finish an already started job even if W_t drops below K due to a service completion. In this section Y_t refers to the number of active secondary servers at time t, thus $Y_t \in \{0, 1\}$. For the system to be stable we assume $\lambda < \mu_p + \mu_s$. The described two-server system is depicted in Figure 1.

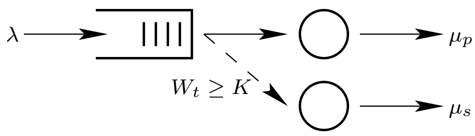


Figure 1: The queue is served by a primary server with rate μ_p which is supplemented by a secondary server with service rate μ_s, when the waiting time of the first in line, W_t, equals or exceeds K.

When dealing with the two-server setup, we introduce the steady-state joint distribution of the FIL-process as $W_i^{\text{FIL}}(x) = \lim_{t\to\infty} \mathbb{P}(W_t \leq x; Y_t = i)$. The joint steady-state density of the FIL-process is denoted $w_i^{\text{FIL}}(x)$.
A sample path of the FIL-process is shown in Figure 2. W_t increases linearly with time whenever a customer is in the queue. When the n'th customer enters service at time t, the waiting time of the first in line decreases with $\min(A_n, W_{t^-})$ from W_{t^-} to $W_{t^+} = \max(W_{t^-} - A_n, 0)$, where A_n is the exponentially distributed interarrival time with rate λ between customers n and $n+1$. Because both service times and interarrival times are exponentially distributed, the FIL-process is Markovian.
The analysis of the system is based on the level crossing equations for the FIL-process. These are more involved, compared to those in Section 2, and are thus presented in Lemma 3.1. From this, the steady state distribution of the FIL-process is determined and given in Theorem 3.1.

Lemma 3.1 *We consider the level crossing equations with upcrossings of level x on the left-hand side and downcrossings on the right-hand for three different cases.*

(*i*) *For $W_t < K$ and an active secondary server we have*

$$
\begin{aligned}
w_1^{\text{FIL}}(x) + \mu_s W_1^{\text{FIL}}(x) = \mu_p \int_{y=x}^{\infty} e^{-\lambda(y-x)} w_1^{\text{FIL}}(y)\mathrm{d}y \\
+ \mu_s \int_{y=K}^{\infty} e^{-\lambda(y-x)} w_1^{\text{FIL}}(y)\mathrm{d}y \\
+ w_0^{\text{FIL}}(K) e^{-\lambda(K-x)}.
\end{aligned}
$$

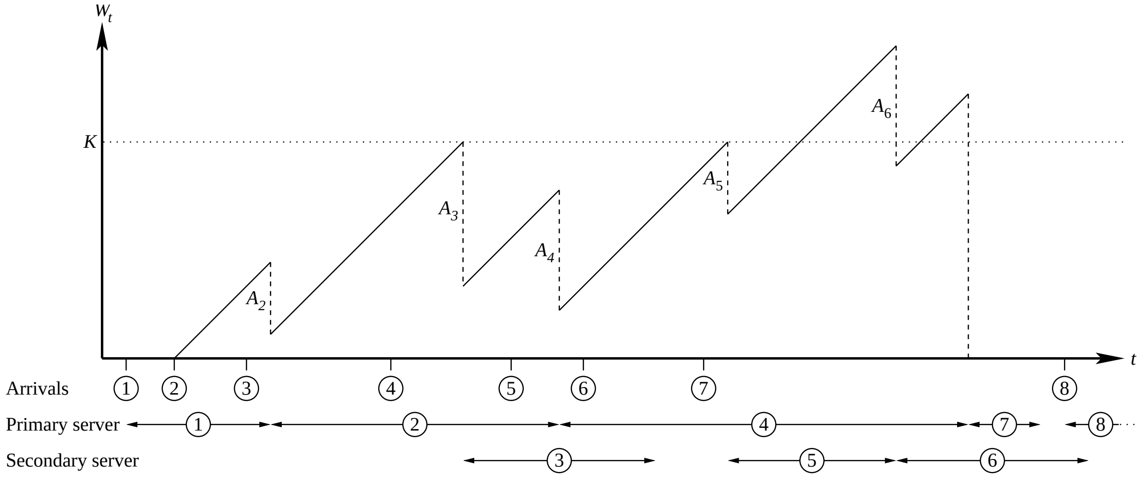


Figure 2: Elapsed waiting time of the first customer in line, W_t. The occupation of the servers are shown beneath the graph. Notice how W_t keeps increasing after customer #3 finishes service as the secondary server is not allowed to start a new service until the level K is reached.

(ii) For $W_t < K$ and an inactive secondary server

$$w_0^{\text{FIL}}(x) = \mu_p \int_{y=x}^{K} e^{-\lambda(y-x)} w_0^{\text{FIL}}(y)\mathrm{d}y + \mu_s W_1^{\text{FIL}}(x).$$

(iii) For $W_t \geq K$ the secondary server will always be active

$$w_1^{\text{FIL}}(x) = (\mu_p + \mu_s) \int_{y=x}^{\infty} e^{-\lambda(y-x)} w_1^{\text{FIL}}(y)\mathrm{d}y.$$

Proof Only case (i) is dealt with in detail as it is the most complicated. The level crossing equations are obtained from setting up forward Kolmogorov equations. For case (i) this becomes

$$\begin{aligned} &\mathbb{P}(W_{t+\mathrm{d}t} \leq x + \mathrm{d}t; Y_{t+\mathrm{d}t} = 1) \\ =& (1 - \mu_p\mathrm{d}t - \mu_s\mathrm{d}t)\mathbb{P}(W_t \leq x; Y_t = 1) \\ &+ \mu_p\mathrm{d}t\mathbb{P}(W_t \leq x + A_n; Y_t = 1) \\ &+ \mu_s\mathrm{d}t\mathbb{P}(K < W_t \leq x + A_n; Y_t = 1) \\ &+ (1 - \mu_p\mathrm{d}t)\mathbb{P}(W_t \in [K - \mathrm{d}t, K]; W_t \leq x + A_n; Y_t = 0) + o(\mathrm{d}t). \end{aligned}$$

Subtracting $\mathbb{P}(W_t \leq x + \mathrm{d}t; Y_t = 1)$ from both sides, dividing by $\mathrm{d}t$ and letting $\mathrm{d}t \to 0$ allows us to rewrite the term on the left side and the first term on the right side as derivatives with regard to t and x respectively. Moreover $\mathrm{d}t$ cancels from the rest of the terms except the last. Note that $\mu_p\mathbb{P}(W_t \in [K - \mathrm{d}t, K]; W_t \leq x + A_n; Y_t = 0) \to 0$ for

$\mathrm{d}t \to 0$. Hence,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(W_t \leq x; Y_t = 1)\\
=& -\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(W_t \leq x; Y_t = 1) - (\mu_p + \mu_s)\mathbb{P}(W_t \leq x; Y_t = 1)\\
&+ \mu_p\mathbb{P}(W_t \leq x + A_n; Y_t = 1) + \mu_s\mathbb{P}(K < W_t \leq x + A_n; Y_t = 1)\\
&+ \lim_{\mathrm{d}t\to 0}\frac{\mathbb{P}(W_t \leq K; Y_t = 0) - \mathbb{P}(W_t \leq K - \mathrm{d}t; Y_t = 0)}{\mathrm{d}t}\cdot\mathbb{P}(A_n > K - x).
\end{aligned}$$

By letting $t \to \infty$, the left side of the expression tends to zero. The probabilities can be written in form of density and distribution functions, using convolution for the probabilities involving A_n; e.g. $\mathbb{P}(W_t \leq x + A_n; Y_t = 1) = W_1^{\mathrm{FIL}}(x) + \mathbb{P}(x < W \leq x + A_n, Y = 1) = W_1^{\mathrm{FIL}}(x) + \int_{y=x}^{\infty} e^{-\lambda(y-x)} w_1^{\mathrm{FIL}}(y)\mathrm{d}y$. Using $\lim_{\mathrm{d}t\to 0, t\to\infty}\left(\frac{\mathbb{P}(W_t\leq K; Y_t=0)-\mathbb{P}(W_t\leq K-\mathrm{d}t; Y_t=0)}{\mathrm{d}t}\right) = w_0^{\mathrm{FIL}}(K)$, then leads to:

$$\begin{aligned}
0 =& -w_1^{\mathrm{FIL}}(x) - (\mu_p + \mu_s)W_1^{\mathrm{FIL}}(x)\\
&+ \mu_p\left(W_1^{\mathrm{FIL}}(x) + \int_{y=x}^{\infty} e^{-\lambda(y-x)} w_1^{\mathrm{FIL}}(y)\mathrm{d}y\right) + \mu_s\int_{y=K}^{\infty} e^{-\lambda(y-x)} w_1^{\mathrm{FIL}}(y)\mathrm{d}y\\
&+ w_0^{\mathrm{FIL}}(K)e^{-\lambda(K-x)}.
\end{aligned}$$

Finally, the level crossing equation for case (i) can be obtained by simply rearranging the above terms.
We now turn to case (ii). Following an approach similar to the one for case (i), the level crossing equation can be found from the initial Kolmogorov equation

$$\begin{aligned}
\mathbb{P}\left(W_{t+\mathrm{d}t} \leq x + \mathrm{d}t; Y_{t+\mathrm{d}t} = 0\right) = &(1 - \mu_p\mathrm{d}t)\mathbb{P}(W_t \leq x; Y_t = 0)\\
&+ \mu_p\mathrm{d}t\mathbb{P}(W_t \leq x + A_n; Y_t = 0)\\
&+ \mu_s\mathrm{d}t\mathbb{P}(W_t \leq x; Y_t = 1) + o(\mathrm{d}t).
\end{aligned}$$

In case (iii) the Kolmogorov equation is of the following form

$$\begin{aligned}
\mathbb{P}(W_{t+\mathrm{d}t} \leq x + \mathrm{d}t; Y_{t+\mathrm{d}t} = 1) = &(1 - \mu_p\mathrm{d}t - \mu_s\mathrm{d}t)\mathbb{P}(W_t \leq x; Y_t = 1)\\
&+ (\mu_p + \mu_s)\mathrm{d}t\mathbb{P}(W_t \leq x + A_n; Y_t = 1) + o(\mathrm{d}t).
\end{aligned}$$

Again, using the same approach as for case (i), the level crossing equation of Lemma 3.1, case (iii), can be obtained. □

Theorem 3.1 *The density of the FIL-process, for $Y_t = 0$, is*

$$w_0^{\mathrm{FIL}}(x) = -c_1e^{(\lambda-\mu_p)x} - r_1c_3e^{r_1x} - r_2c_4e^{r_2x}, \text{ for } 0 < x \leq K,$$

and, for $Y_t = 1$, it is

$$w_1^{\mathrm{FIL}}(x) = \begin{cases} r_1c_3e^{r_1x} + r_2c_4e^{r_2x}, & \text{for } 0 < x \leq K;\\ c_2e^{(\lambda-\mu_p-\mu_s)x}, & \text{for } x > K, \end{cases}$$

with r_1, r_2 given by (6) and (7). The marginal density of the FIL-process for the two-server system becomes

$$w^{\text{FIL}}(x) = \begin{cases} c_1 e^{(\lambda-\mu_p)x}, & \text{for } 0 < x \le K; \\ \\ c_2 e^{(\lambda-\mu_p-\mu_s)x}, & \text{for } x > K. \end{cases}$$

The constants c_i, $i \in \{1,\ 2,\ 3,\ 4\}$, are determined in Subsection 3.1.

Proof The densities of the FIL-process are found from the level crossing equations given in Lemma 3.1. The derivative with respect to x of the level crossing equation in case (i) becomes

$$\begin{aligned} w_1^{\text{FIL}'}(x) + \mu_s W_1^{\text{FIL}'}(x) = \lambda \Bigg[\, \mu_p \int_{y=x}^{\infty} & e^{-\lambda(y-x)} w_1^{\text{FIL}}(y) \mathrm{d}y \\ & + \mu_s \int_{y=K}^{\infty} e^{-\lambda(y-x)} w_1^{\text{FIL}}(y) \mathrm{d}y \\ & + w_0^{\text{FIL}}(K) e^{-\lambda(K-x)} \Bigg] \\ & - \mu_p w_1^{\text{FIL}}(x), \end{aligned}$$

where the first and last term on the right-hand side of the above equation stem from the derivative of $\mu_p \int_{y=x}^{\infty} e^{-\lambda(y-x)} w_1^{\text{FIL}}(y)\mathrm{d}y$. By rearranging and noting that the term inside the brackets equals $w_1^{\text{FIL}}(x) + \mu_s W_1^{\text{FIL}}(x)$, as given in the level crossing equation, we end up with a second-order differential equation:

$$W_1^{\text{FIL}''}(x) + [\mu_p + \mu_s - \lambda] \, W_1^{\text{FIL}'}(x) - \lambda \mu_s W_1^{\text{FIL}}(x) = 0. \tag{4}$$

The general solution of (4) is of the form:

$$W_1^{\text{FIL}}(x) = c_3 e^{r_1 x} + c_4 e^{r_2 x}, \tag{5}$$

where

$$r_1 = \frac{\lambda - (\mu_p + \mu_s) - \sqrt{(\mu_p + \mu_s - \lambda)^2 + 4\lambda\mu_s}}{2}, \tag{6}$$

$$r_2 = \frac{\lambda - (\mu_p + \mu_s) + \sqrt{(\mu_p + \mu_s - \lambda)^2 + 4\lambda\mu_s}}{2} \tag{7}$$

and c_3 and c_4 are constants. The derivative of (5) with respect to x yields the density, $w_1^{\text{FIL}}(x)$, for $0 < x \le K$, as given in Theorem 3.1.
The expressions for $w_0^{\text{FIL}}(x)$ for $x \le K$ and $w_1^{\text{FIL}}(x)$ for $x > K$ can be found in the same way as the solution to the derivative of the level crossing equations in cases (ii) and (iii) of Lemma 3.1 respectively. Finally the marginal density of $w^{\text{FIL}}(x)$ is found as the sum of $w_1^{\text{FIL}}(x)$ and $w_2^{\text{FIL}}(x)$. □

3.1 Constants and atoms

To fully describe the distribution of the FIL-process, the atoms in zero must be determined together with the constants in Theorem 3.1. The atoms, corresponding to the queue being empty, can be divided into four different boundary states; both servers are unoccupied

(N), only the primary server is occupied (P), only the secondary server is occupied (S), and both servers are occupied (PS). The probabilities of being in these states are referred to as $W_{\mathrm{N}}^{\mathrm{FIL}}(0)$, $W_{\mathrm{P}}^{\mathrm{FIL}}(0)$, $W_{\mathrm{S}}^{\mathrm{FIL}}(0)$ and $W_{\mathrm{PS}}^{\mathrm{FIL}}(0)$, respectively.
Eight independent equations are needed to determine the eight constants; the probability of being in the four boundary states and the c_i's, $i \in \{1,\ 2,\ 3,\ 4\}$. Two equations follow directly from the boundary states in 0, as N and S can only be entered and left from other boundary states. Writing the rate out of the states on the left-hand side and the rate into the states on the right-hand side gives

$$\lambda W_{\mathrm{N}}^{\mathrm{FIL}}(0) = \mu_p W_{\mathrm{P}}^{\mathrm{FIL}}(0) + \mu_s W_{\mathrm{S}}^{\mathrm{FIL}}(0) \tag{8}$$

and

$$(\lambda + \mu_s) W_{\mathrm{S}}^{\mathrm{FIL}}(0) = \mu_p W_{\mathrm{PS}}^{\mathrm{FIL}}(0). \tag{9}$$

The rate out of P is $\lambda + \mu_p$ as this state can only be left by an arrival or a departure from the primary server. The state can be entered by an arrival in state N or a departure from the secondary server in state PS. P can also be entered from the FIL-process for non-zero W_t given that $Y_t = 0$ and $W_t < A_n$. This is represented by the second term on the right-hand side in (10).

$$(\lambda + \mu_p) W_{\mathrm{P}}^{\mathrm{FIL}}(0) = \lambda W_{\mathrm{N}}^{\mathrm{FIL}}(0) + \mu_p \int_{0^+}^{K} e^{-\lambda y} w_0^{\mathrm{FIL}}(y) \mathrm{d}y + \mu_s W_{\mathrm{PS}}^{\mathrm{FIL}}(0). \tag{10}$$

The balance equation for $W_{\mathrm{PS}}^{\mathrm{FIL}}(0)$ is found in the same way:

$$\begin{aligned}(\lambda + \mu_p + \mu_s) W_{\mathrm{PS}}^{\mathrm{FIL}}(0) = \lambda W_{\mathrm{S}}^{\mathrm{FIL}}(0) &+ \mu_p \int_{0^+}^{\infty} e^{-\lambda y} w_1^{\mathrm{FIL}}(y) \mathrm{d}y \\ &+ \mu_s \int_{K}^{\infty} e^{-\lambda y} w_1^{\mathrm{FIL}}(y) \mathrm{d}y + w_0^{\mathrm{FIL}}(K) e^{-\lambda K}. \end{aligned} \tag{11}$$

Three more equations can be obtained by considering boundary conditions. By letting $x \downarrow 0$ in (5) we have

$$W_{\mathrm{S}}^{\mathrm{FIL}}(0) + W_{\mathrm{PS}}^{\mathrm{FIL}}(0) = c_3 + c_4. \tag{12}$$

Letting $x \uparrow K$ in the level crossing equation of case (ii) in Lemma 3.1 gives

$$\begin{aligned} w_0^{\mathrm{FIL}}(K^-) &= \mu_s W_1^{\mathrm{FIL}}(K) \\ &= \mu_s \left[W_{\mathrm{S}}^{\mathrm{FIL}}(0) + W_{\mathrm{PS}}^{\mathrm{FIL}}(0) + \int_0^K w_1^{\mathrm{FIL}}(y) \mathrm{d}y \right], \end{aligned} \tag{13}$$

and the same limit in the level crossing equation of case (i) gives

$$\begin{aligned} w_1^{\mathrm{FIL}}(K^-) + \mu_s W_1^{\mathrm{FIL}}(K) &= w_0^{\mathrm{FIL}}(K) + (\mu_p + \mu_s) \int_{y=K}^{\infty} e^{-\lambda(y-K)} w_1^{\mathrm{FIL}}(y) \mathrm{d}y \\ &= w_0^{\mathrm{FIL}}(K) + c_2 e^{(\lambda - \mu_p - \mu_s)K}. \end{aligned} \tag{14}$$

The final equation is obtained by normalization of the FIL-process:

$$1 = \int_0^K w_0^{\mathrm{FIL}}(y) \mathrm{d}y + \int_0^\infty w_1^{\mathrm{FIL}}(y) \mathrm{d}y + W_{\mathrm{N}}^{\mathrm{FIL}}(0) + W_{\mathrm{S}}^{\mathrm{FIL}}(0) + W_{\mathrm{P}}^{\mathrm{FIL}}(0) + W_{\mathrm{PS}}^{\mathrm{FIL}}(0). \tag{15}$$

Writing out analytical expressions for the constants would be a cumbersome task as the equation system is rather complex. Furthermore, the gained insight would be insignificant. Mathematical software, such as Maple, can be used to determine the eight constants by solving the system of independent equations given in (8)-(15). Proving general independence of the eight equations by direct arguments is inherently difficult but as long as the requirements for stability of the system are fulfilled, a unique solution to the equation array must exist and thus the equations must indeed be independent.

3.2 Waiting-time distribution

We now turn to the waiting-time distribution and use the same definition of this as in Section 2; $W(x) = \mathbb{P}(W \leq x)$, where W is the waiting time an arbitrary customer experiences. Observe that arriving customers are directly taken into service in case the queue is empty and the primary server is available. Using PASTA, it is easy to obtain the atom in zero of the waiting time:

$$\mathbb{P}(W = 0) = W_{\mathrm{N}}^{\mathrm{FIL}}(0) + W_{\mathrm{S}}^{\mathrm{FIL}}(0).$$

In case the waiting time is non-zero, the waiting time corresponds to the FIL-process at epochs right before downward jumps. Here, we again consider an infinitesimal interval $(t, t + h)$ and apply similar arguments as in Section 2. In particular, for $x > K$, we have

$$\mathbb{P}(W_t > x; N_s(t, t + h) = 1) = (\mu_p + \mu_s)h \int_x^\infty w_1^{\mathrm{FIL}}(y)\mathrm{d}y + o(h).$$

For $0 < x \leq K$, we have

$$\begin{aligned}\mathbb{P}(W_t > x; N_s(t, t + h) = 1) = {} & \mu_p h \int_x^{K-h} w_0^{\mathrm{FIL}}(y)\mathrm{d}y + \mu_p h \int_x^K w_1^{\mathrm{FIL}}(y)\mathrm{d}y \\ & + \int_{K-h}^K w_0^{\mathrm{FIL}}(y)\mathrm{d}y + (\mu_p + \mu_s)h \int_K^\infty w_1^{\mathrm{FIL}}(y)\mathrm{d}y + o(h).\end{aligned}$$

Note that $\int_{K-h}^K w_0^{\mathrm{FIL}}(y)\mathrm{d}y/h \to w_0^{\mathrm{FIL}}(K)$, as $h \to 0$. Also, observe that $\mathbb{P}(N_s(t, t + h) = 1)/h$ (for $h \to 0$) is the rate at which customers are taken into service and, since every customer leaves the queue through the server and the system is stable, equals λ. Combining the above and using a similar conditioning as in Section 2, we obtain

$$\mathbb{P}(W > x) = \begin{cases} \frac{1}{\lambda}\Big[\mu_p \int_x^K (w_0^{\mathrm{FIL}}(y) + w_1^{\mathrm{FIL}}(y))\mathrm{d}y \\ \qquad + w_0^{\mathrm{FIL}}(K) + (\mu_p + \mu_s) \int_K^\infty w_1^{\mathrm{FIL}}(y)\mathrm{d}y\Big], & \text{for } 0 < x \leq K, \\ \frac{\mu_p + \mu_s}{\lambda} \int_x^\infty w_1^{\mathrm{FIL}}(y)\mathrm{d}y, & \text{for } x > K. \end{cases} \tag{16}$$

From this, we obtain the density of the steady-state waiting time and the atom at K:

Corollary 3.1 *For the steady-state waiting time, we have two atoms*

$$\begin{aligned}\mathbb{P}(W = 0) &= W_{\mathrm{N}}^{\mathrm{FIL}}(0) + W_{\mathrm{S}}^{\mathrm{FIL}}(0), \\ \mathbb{P}(W = K) &= \frac{w_0^{\mathrm{FIL}}(K)}{\lambda},\end{aligned}$$

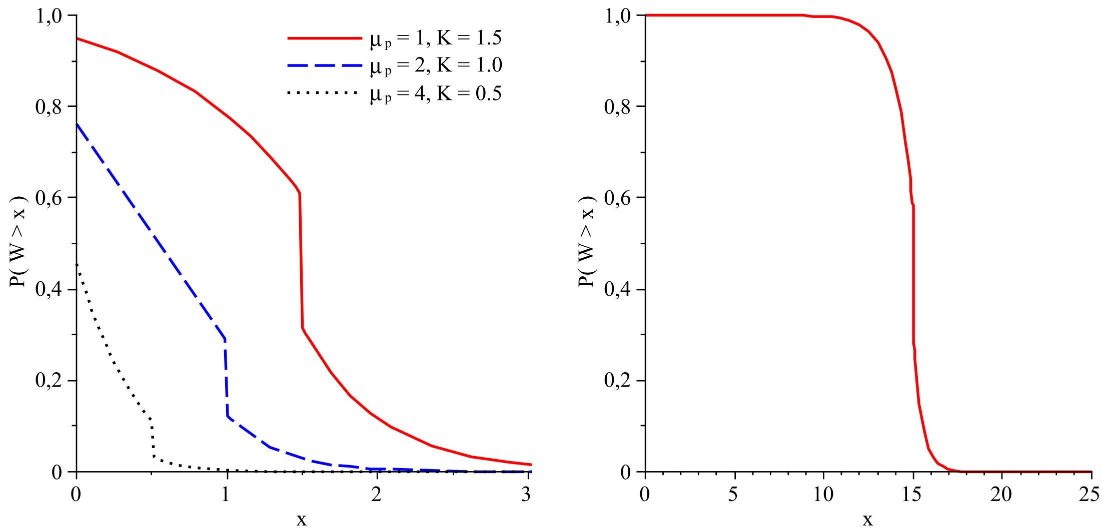
and density

$$w(x) = \begin{cases} \frac{\mu_p}{\lambda} c_1 e^{(\lambda-\mu_p)x}, & \text{for } 0 < x \le K, \\ \frac{\mu_p+\mu_s}{\lambda} c_2 e^{(\lambda-\mu_p-\mu_s)x}, & \text{for } x > K. \end{cases}$$

Remark 3.1 Note that the form of the steady-state waiting time density (and distribution) is closely related to the density in Example 2.1, i.e., the single-server model with two service speeds determined by a threshold on the FIL-process. In particular, the parameters r_i, $i = 1, 2$, and μ should be taken such that $r_1\mu = \mu_p$ and $r_2\mu = \mu_p + \mu_s$ (for instance, let $\mu = \mu_p$, $r_1 = 1$, and $r_2 = 1 + \mu_s/\mu_p$). The main difference between the waiting-time distributions concerns the atom at K. ◇

4 Numerical results

To illustrate the behavior of the waiting-time distribution given in Corollary 3.1, a few numerical results are shown in Figure 3a. The corresponding eight constants, found with Maple, are given in Table 1. It is seen how the relation between λ and μ_p governs the shape of the distribution for $x < K$; it is convex for $\lambda < \mu_p$, concave for $\lambda > \mu_p$ and a straight line for $\lambda = \mu_p$.
Figure 3b illustrates the interesting, but not surprising, phenomenon of how the probability mass gathers around K for $\lambda > \mu_p$ and K large.



(a) Waiting time distribution for $\lambda = 2$, $\mu_s = 3$ and 3 different values of μ_p and K.

(b) Waiting time distribution for large K, ($\lambda = 2$, $\mu_p = 1$, $\mu_s = 3$, $K = 15$).

Figure 3: Numerical examples.

Given the distribution of the waiting time and FIL-process, most of the commonly used performance measures such as TSF are easily found. Other performance measures such

Table 1: Numerical results for common parameters $\lambda = 2$, $\mu_s = 3$.

	$(\mu_p = 1,\ K = 1.5)$	$(\mu_p = 2,\ K = 1.0)$	$(\mu_p = 4,\ K = 0.5)$	$(\mu_p = 1,\ K = 15)$
W_{N}	0.0470	0.2298	0.5318	$6.4968{\cdot}10^{-8}$
W_{P}	0.0860	0.2181	0.2559	$1.2993{\cdot}10^{-7}$
W_{S}	0.0027	0.0078	0.0133	$5.7696{\cdot}10^{-13}$
W_{PS}	0.0135	0.0195	0.0166	$2.8848{\cdot}10^{-12}$
c_1	0.1990	0.4751	0.5451	$2.5987{\cdot}10^{-7}$
c_2	6.3453	2.9956	0.6123	$3.2160{\cdot}10^{12}$
c_3	$-0.6401{\cdot}10^{-4}$	$-0.2673{\cdot}10^{-3}$	$-0.4749{\cdot}10^{-3}$	$-1.3707{\cdot}10^{-14}$
c_4	0.01626	0.0276	0.0304	$3.4754{\cdot}10^{-12}$

as the utilization of the servers can be found as

$$a_p = 1 - W_{\mathrm{N}}(0) - W_{\mathrm{S}}(0),$$
$$a_s = 1 - W_{\mathrm{N}}(0) - W_{\mathrm{P}}(0) - \int_0^K w_0^{\mathrm{FIL}}(y)\mathrm{d}y,$$

where a_p and a_s are the utilization of the primary and secondary server, respectively.

5 Conclusions and topics for further research

We have studied queueing systems where the provided service depends on the waiting time of the first customer in line. This has mainly been motivated by a setup often used in call centers, referred to as an "inverted V", see [1]. The main contribution is that we have shown ways to deal with systems where the service changes depending on the waiting time, which can be inherently difficult to deal with in particular in the case of fixed thresholds. The first model of this paper deals with a single server that operates with a service speed depending on the waiting time of the first customer in line. We derived the waiting time distribution of an arbitrary customer entering the system and showed how the model can be used for the threshold case.

The second model of this paper deals with a two-server setup where a secondary server supplements a primary server when the waiting time of the first in line exceeds a threshold. Again the waiting distribution of an arbitrary customer has been derived and numerical examples have been given. The simplicity of the form of the solution for the waiting time given in Corollary 3.1 provides some useful insight.

In the model presented in Section 3, only one primary and one secondary server was considered. This is easily expanded to a more general setup with multiple primary servers by introducing additional states for $W^{\mathrm{FIL}}(0)$ along with the four already used. The extra boundary states should describe the number of unoccupied servers. Analyzing a setup with multiple secondary servers would be much more difficult as the joint distribution of $w_i^{\mathrm{FIL}}(x)$ must be extended to include $i \in \{0,\ 1, ...,\ n\}$, where n is the number of secondary servers.

A related routing setup, often seen in call centers and used as a way to prioritize a group of customers over another, is the "N" design, see [9]. Also related are [10], [11] and [12]. The "N" design is basically an extension to the model of Section 3 where the secondary server

also has a queue of its own, from which it receives jobs. Extending the model presented in this paper to the "N" design, necessitates the use of a 2-dimensional FIL-process in order to keep track of the waiting time of the first customer in line in both queues.
There is still much to be done in relation to analysis of complex queueing systems such as those seen in call centers. Even though simulation may remain the dominant way of modelling these systems, it is indeed worth pursuing analytical approaches to gain insight not obtainable through simulation such as the result in Corollary 3.1.

References

[1] Armony, M. (2005). Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* **51**, 287–329.

[2] Asmussen, S. (2003). *Applied Probability and Queues*, Second Edition. Springer, New York.

[3] Bekker, R., S.C. Borst, O.J. Boxma, O. Kella (2004). Queues with workload-dependent arrival and service rates. *Queueing Systems* **46**, 537–556.

[4] Boxma, O., H. Kaspi, O. Kella, D. Perry (2005). On/off storage systems with state dependent input, output and switching rates. *Probability in the Engineering and Informational Sciences* **19**, 1–14.

[5] Boxma, O.J., M. Vlasiou (2007). On queues with service and interarrival times depending on waiting times. *Queueing Systems* **56**, 121–132.

[6] Brill, P.H., M.J.M. Posner (1981). A two server queue with nonwaiting customers receiving specialized service. *Management Science* **27**, 914–925.

[7] Brill, P.H., M.J.M. Posner (1981). The system point method in exponential queues: a level crossing approach. *Mathematics of Operations Research* **6**, 31–49.

[8] Browne, S., K. Sigman (1992). Work-modulated queues with applications to storage processes. *Journal of Applied Probability* **29**, 699–712.

[9] Gans, N., G.M. Koole, A. Mandelbaum (2003). Telephone call centers: tutorial, review, and research prospects. *Manufacturing and Service Operations Management* **5**, 79–141.

[10] Gans, N., Y.-P. Zhou (2003). A call-routing problem with service-level constraints. *Operations Research* **51**, 255–271.

[11] Gans, N., Y.-P. Zhou (2007). Call-routing schemes for call-center outsourcing. *Manufacturing and Service Operations Management* **9**, 33–50.

[12] Gurvich, I., M. Armony, A. Mandelbaum (2008). Service-level differentiation in call centers with fully flexible servers. *Management Science* **54**, 279–294.

[13] Harrison, J.M., S.I. Resnick (1976). The stationary distribution and first exit probabilities of a storage process with general release rule. *Mathematics of Operations Research* **1**, 347–358.

[14] Koole, G.M. (1995). A simple proof of the optimality of a threshold policy in a two-server queueing system. *Systems & Control Letters* **26**, 301–303.

[15] Koole, G.M., J. Talim (2000). Exponential Approximation of Multi-Skill Call Centers Architecture. *Proc. QNETs 2000* **23**, 1–10.

[16] Lin, W., P.R. Kumar (1984). Optimal control of a queueing system with two heterogeneous servers. *IEEE Trans. Automat. Control* **29**, 696–703.

[17] Perry, D., L. Benny (1989). Continuous Production/Inventory Model with Analogy to Certain Queueing and Dam Models *Advances in Applied Probability* **21**, 123–141.

[18] Posner, M.J.M (1973). Single-server queues with service time dependent on waiting time. *Operations Research* **21**, 610–616.

[19] Rubinovitch, M. (1985). The slow server problem. *Journal of Applied Probability* **22**, 205–213.

[20] Scheinhardt, W.R.W., N. van Foreest, M. Mandjes (2005). Continuous feedback fluid queues. *Operations Research Letters* **33**, 551–559.

[21] Stockbridge, R.H. (1991). A martingale approach to the slow server problem. *Journal of Applied Probability* **28**, 480–486.

[22] Whitt, W. (1990). Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems* **6**, 335–351.

[23] Zabreiko, P.P, A.I. Koshelev, M.A. Krasnosel'skii; transl. and ed. by T.O. Shaposhnikova, R.S. Anderssen and S.G. Mikhlin (1975). *Integral Equations: a Reference Text.* Monographs and textbooks on pure and applied mathematics. Noordhoff, Leiden.