European Journal of Operational Research 299 (2022) 566–579

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

An $M/M/c$ queue with queueing-time dependent service rates☆

Bernardo D'Auria[a,*], Ivo J. B. F. Adan[b], René Bekker[c], Vidyadhar Kulkarni[d]

[a] Statistics Department, Madrid University Carlos III Avda Universidad 30, Leganes 28911, Madrid, Spain
[b] Industrial Engineering Department, Technische Universiteit Eindhoven, Postbus 513, Eindhoven 5600 MB, the Netherlands
[c] Department of Mathematics, VU Amsterdam, De Boelelaan 1105, Amsterdam 1081 HV, the Netherlands
[d] Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599, USA

ARTICLE INFO

Article history:
Received 9 July 2021
Accepted 14 December 2021
Available online 21 December 2021

Keywords:
Queueing
M/M/c queue
Variable service rates
Queueing-time dependency

ABSTRACT

Recent studies indicate that in many situations service times are affected by the experienced queueing delay of the particular customer. This effect has been detected in different areas, such as health care, call centers and telecommunication networks. In this paper we present a methodology to analyze a model having this property. The specific model is an $M/M/c$ queue in which any customer may be tagged at her arrival time if her queueing time will be above a certain fixed threshold. All tagged customers are then served at a given rate that may differ from the rate used for the non-tagged customers. We show how it is possible to model the virtual queueing time of this queueing system by a specific Markov chain. Then, solving the corresponding balance equations, we give a recursive solution to compute the stationary distribution, which involves a mixture of exponential terms. Using numerical experiments, we demonstrate that the differences in service rates can have a crucial impact on queueing time performance.

1. Introduction

For classical queueing models, service times are typically assumed to be independent of experienced delay. Such independence assumptions are often crucial for analytical tractability of the queueing system's performance. In practice, however, it has been recognized that the amount of waiting affects service durations, and the assumed independence does therefore not hold. Empirical evidence of this dependence relation primarily stems from the health care domain. The studies (Batt, 0000; Chalfin, Trzeciak, Likourezos, Baumann, & Dellinger, 2007; Chan, Farias, & Escobar, 2017; Chan, Krumholz, Nichol, Nallamothu, & American Heart Association National Registry of Cardiopulmonary Resuscitation Investigators, 2008; Renaud et al., 2009; Richardson, 2002; Siegmeth, Gurusamy, & Parker, 2005; Soltani, Batt, Bavafa, & Patterson, 2019) indicate that delays in admission have adverse effects on patient outcomes and consequently increase the patients length of stay; this is referred to as the slowdown effect in Selen, Adan, Kulkarni, & van Leeuwaarden (2016).

At a more conceptual level, the field of behavioral operations investigates how servers and customers behave in an operational setting. The recent study (Delasay, Ingolfsson, Kolfal, & Schultz, 2019) indicates that in many situations service times are affected by the load. The authors develop a framework for the impact of load on service times, where they distinguish server, network, and customer mechanisms. For the server mechanism, it is observed (and supported by literature) that there is a clear impact of workload on the service speed, and this impact may go in different directions. The authors of Delasay et al. (2019) found far fewer customer mechanisms in the literature, although they expect them to exist. A psychological view of a customers queueing experience during its sojourn time is provided in Carmon, Shanthikumar, & Carmon (1995). Specifically, the authors assume that the dissatisfaction level of a customer increases during waiting, whereas this may be compensated during service. As a consequence, for an acceptable level of dissatisfaction after service, the service time should be longer for a customer experiencing longer delays. Moreover, after excessive waiting customers expect valuable service (Maister et al., 1984), which may also affect the corresponding service time. Similarly, the recent study (Ülkü, Hydock, & Cui, 2020) in a retail environment found that customers waiting longer in fact consume more. Thus, from the customer perspective, it seems conceivable that excessive waits are associated with longer service times.

☆ The research of the first author was partially supported by Spain's Ministry of Science and Innovation [Grants MTM2017-85618-P via FEDER funds and PID2020-116694GB-I00]

* Corresponding author.
E-mail addresses: bernardo.dauria@uc3m.es (B. D'Auria), i.adan@tue.nl (I.J.B.F. Adan), r.bekker@vu.nl (R. Bekker), vkulkarn@email.unc.edu (V. Kulkarni).

https://doi.org/10.1016/j.ejor.2021.12.023

The aim of this paper is to find the steady-state queueing time distribution in multi-server queues where the service time is affected by the experienced queueing time. Despite its apparent practical relevance, such queues have hardly been studied in a service setting with multiple servers. More specifically, we let the service rate of each server depend on whether the experienced queueing time of the customer in service is above or below a given threshold upon service initiation. We envisage that this typically corresponds to a customer mechanism, although the service rate adaptation might also be the consequence of the server adapting to congestion. Despite the inherent model complexity, the steady-state queueing time distribution turns out to be remarkably tractable in this case and can be expressed in terms of a mixture of exponential terms. From our numerical experiments, we see that taking the differences in service rates into account results in crucially different queueing time behavior. As such, ignoring the dependence between experienced waiting and service time might be wholly inadequate. An online implementation of the model is available to further facilitate managerial decision making (D'Auria, 2021a).

Delay thresholds are typically used in empirical health care studies to distinguish delayed and non-delayed patients; if the admission delay is above the delay threshold, a patient is considered to be delayed. For instance, Chalfin et al. (2007), Richardson (2002) investigated the impact of delayed patients at the emergency department on the inpatient length of stay. Based on the patient data, the difference in length of stay is in the order of hours. For cardiac patients, delays are even much more critical; delays in the order of minutes lead to adverse patient outcomes (Chan et al., 2008). Less critical cases, such as surgery of hip fractures, have delay thresholds in the order of days (Siegmeth et al., 2005). For patients with community-acquired pneumonia a similar delay threshold is used (Renaud et al., 2009). For both situations it is shown again that delayed admissions experience extended length of stay. Another example of the impact of physician workload at the emergency department (ED) are Batt (0000), Soltani et al. (2019); amongst others, the authors observe that high physician workload leads to overtesting and generates extra post-ED care.

The health care situations described above have recently inspired the study of multi-server queues, in which the service time (i.e. the length of stay) is affected by delay and congestion at the clinical ward, such as the Intensive Care Unit (Chan et al., 2017; Dong, Feldman, & Yom-Tov, 2015; Selen et al., 2016). The study of Chan et al. (2017) is also supported with data verifying the correlation between delay and length of stay. In Chan et al. (2017), the multi-server queue with delay-dependent service is abbreviated with $M/M(f)/c$; the focus from the queueing perspective is on approximations and bounds for the workload process. The multi-server variant with abandonments in the quality and efficiency driven (QED) regime is considered in Dong et al. (2015). Next to the fact that this involves an asymptotic analysis, the service rate adaptations are also instantaneous instead of the more intricate delay effects on individual customers. Such server mechanisms are referred to as operator slowdown in Selen et al. (2016), as opposed to customer slowdown. The model in Selen et al. (2016) also involves a multi-server queue, where the service rate depends on whether a customer has to wait or not. In terms of the current paper, this means that the waiting threshold is at zero. In addition, Selen et al. (2016) focuses on the number of customers instead of queueing times.

There have been some recent studies on multi-server queues where service times depend on delay. The authors of Wu, Bassamboo, & Perry (2019a) consider a general multi-server queue with abandonments and derive fluid limits as a proxy for expected queueing times. Moreover, Wu, Bassamboo, & Perry (2019b) considers a setting with customer abandonments, where the service time is either endogenously or exogenously determined by the system's dynamics. The focus there is mainly on statistical estimation for both dependency situations. Finally, in Do, Shunko, Lucas, & Novak (2018) the service speed is affected by behavioral factors, such as server speedup due to increased workload and social loafing when multiple workers share the workload. However, the analysis is in terms of queue lengths instead of queueing times.

From the literature discussed above, we observe that almost all studies of multi-server queues with delay-dependent service involve some sort of approximation. This is different for the single-server case, which is much more amenable for analysis. An important observation for the single-server case is that the queueing time then corresponds to the workload a customer finds upon arrival; this is no longer the case for the multi-server setting with delay-dependent service. There is a long tradition of single-server queues with workload-dependent features; we refer to Dshalalow (1997) for an early overview containing many references. Among those early papers are Posner (1973) and Brill & Posner (1981). Interestingly, in 1973 Posner already noted that the server may provide more appropriate service to counter the negative effect of waiting (Posner, 1973); the author then provides a complete analysis for the M/M/1 case in which the service rate is a step function of the queueing time. A little later, Brill & Posner (1981) provides an exact analysis for the M/M/2 queue where non-waiting customers have a different service rate.

For workload-dependent M/G/1 queues, often the service and/or arrival rates are assumed to depend on the workload, but not so often the complete service time. However, generalizations of such systems are Lévy driven queues in which the Lévy exponent depends on the position of the process. The Lévy exponent incorporates the Laplace transform of the service time distribution and, hence, the service time may thus depend on the workload found by a customer entering service. Examples of such Lévy driven queues with state-dependent exponent are Bekker (2009), Bekker, Boxma, & Resing (2009), Palmowski & Vlasiou (2011). Finally, Whitt (1990) and Boxma & Vlasiou (2007) consider G/G/1 queues with service and interarrival times that depend linearly on delays.

Limiting distributions in terms of mixtures of exponentials are also common in Markov-modulated fluid models. In fact, our analysis is along similar lines as such fluid models, although our differential equations differ from the ones found in traditional fluid queues (Anick, Mitra, & Sondhi, 1982), see Kulkarni (1997) for an early overview. Some examples of fluid models with level-dependent features are da Silva Soares & Latouche (2009), Malhotra, Mandjes, Scheinhardt, & Van Den Berg (2009), Scheinhardt, Van Foreest, & Mandjes (2005). A crucial difference with fluid models is the role of the background state. Our state description, where the service time depends on experienced delay, is delicate. In our case, the background state should be interpreted as the server state process; our state description is based on Adan, Hathaway, & Kulkarni (2019).

The paper is organized as follows. In Section 2, a model and state description is provided. The single-server case provides insights in both the approach as well as the results, and is discussed in Section 3. Section 4 presents balance equations that are required to determine the limiting distribution for the multi-server case. The limiting distribution is derived in Section 5, including an illustrative example. Section 6 contains some numerical insights and finally Section 7 draws some conclusions. For readability, most of the technical proofs are deferred to Appendix A. A python algorithm to compute the queueing time distribution is avaiblabe for downloading at the public repository D'Auria (2021b); see D'Auria (2021a) for an online implementation.

2. Model and state description

We consider a queueing system with c identical servers and an infinite waiting room. Customers arrive according to a Poisson process with rate λ. Let $W(t)$ be the virtual queueing time (VQT) at time t. That is, if a customer arrives at time t, his service will start at time $t+W(t)$. Clearly, if at least one server is idle at time t, $W(t)=0$. If all servers are busy at time t, $W(t)>0$. The service times of the customers depend on their queueing time through a critical level $k>0$ as follows: if a customer arrives at time t, and $W(t)\le k$, he is classified as a class 1 customer, and his service time is $\exp(\mu_1)$, otherwise he is classified as class 2 customer and his service time is $\exp(\mu_2)$. In order to describe the dynamics of the VQT process $\{W(t), t\ge 0\}$, we introduce the server state process $S(t)=(S_1(t), S_2(t))$ as follows. We say that $S(t)=(i,j)$ if i servers are serving class 1 customers and j servers are serving class 2 customers at time $t+W(t)$, just before the new service starts at time $t+W(t)$. Clearly, we must have $0\le S_1(t)+S_2(t)\le c-1$ for all $t\ge 0$. Furthermore,

$$W(t)>0 \Rightarrow S_1(t)+S_2(t)=c-1.$$

and

$$0\le S_1(t)+S_2(t)<c-1 \Rightarrow W(t)=0.$$

We discuss the evolution of the $\{(W(t), S_1(t), S_2(t)), t\ge 0\}$ process below. We will use the following notation for the aggregate service rate:

$$\Delta(i,j)=i\mu_1+j\mu_2. \tag{1}$$

Suppose the state at time 0 is $(0,i,j)$ with $0\le i+j<c-1$. If the next event is an arrival, the state jumps to $(0,i+1,j)$; if it is a departure of type 1, it jumps to state $(0,i-1,j)$; and if it is a departure of type 2, it jumps to state $(0,i,j-1)$. Hence, the transition rate from state $(0,i,j)$ to state $(0,i+1,j)$ is λ, to state $(0,i-1,j)$ is $i\mu_1$ and to state $(0,i,j-1)$ is $j\mu_2$.

Next, suppose the state at time 0 is $(0,i,c-1-i)$ with $0\le i\le c-1$. Again, if the next event is a departure of type 1, it jumps to state $(0,i-1,c-1-i)$; and if it is a departure of type 2, it jumps to state $(0,i,c-1-2)$. Hence, the transition rate from state $(0,i,c-1-i)$ to state $(0,i-1,c-1-i)$ is $i\mu_1$ and to state $(0,i,c-i-2)$ is $(c-1-i)\mu_2$. If the next event is an arrival, all servers become busy, $i+1$ of them serving type 1 customers and $(c-1-i)$ of them serving type 2 customers. The next departure occurs after an $\exp(\Delta(i+1,c-1-i))$ amount of time and the VQT process jumps to level $X\sim\exp(\Delta(i+1,c-1-i))$. Also, the next departure is of type 1 with probability $(i+1)\mu_1/\Delta(i+1,c-1-i)$ and of type 2 with probability $(c-1-i)\mu_2/\Delta(i+1,c-1-i)$. Combining these observations, we see that the transition rate to state $(x,i,c-1-i)$ is

$$\lambda(i+1)\mu_1\exp(-\Delta(i+1,c-1-i)x)dx,$$

and to state $(x,i+1,c-i-2)$ is

$$\lambda(c-1-i)\mu_2\exp(-\Delta(i+1,c-1-i)x)dx.$$

This completes the description of all transitions out of states $(0,i,j)$ with $0\le i+j\le c-1$.

Next, consider states $(w,i,c-1-i)$ with $0<w\le k$ and $0\le i\le c-1$. The state does not change if the next event is a departure. It can change only if the next event is an arrival. An arrival in this state is of type 1. By following the same argument as in the case of state $(0,i,c-1-i)$, we see that the transition rate to state $(w+x,i,c-1-i)$ is

$$(i+1)\mu_1\exp(-\Delta(i+1,c-1-i)x)dx,$$

and to state $(w+x,i+1,c-i-2)$ is

$$(c-1-i)\mu_2\exp(-\Delta(i+1,c-1-i)x)dx.$$

Now consider states $(w,i,c-1-i)$ with $w>k$ and $0\le i\le c-1$. An arrival in this state is of type 2. Hence, following the same argument as above, we see that the transition rate to state $(w+x,i,c-1-i)$ is

$$(c-i)\mu_2\exp(-\Delta(i,c-i)x)dx,$$

and to state $(w+x,i-1,c-i)$ is

$$i\mu_1\exp(-\Delta(i,c-i)x)dx.$$

Finally, if $W(t)>0$, the VQT process changes continuously at rate -1 between arrivals. This completes the description of the evolution of the process $\{(W(t),S(t)), t\ge 0\}$.

3. The single-server queue

Before studying the general multi-server case ($c>1$), in this section we briefly discuss the single-server queue. Since the process $S(t)$ is identically equal to the null vector (0,0), the process $\{W(t), t\ge 0\}$ is sufficient for the state description. In fact, this process corresponds to an M/M/1 queue in which the jump size depends on the state found upon arrival; see Fig. 1 for an illustration of its sample path. In this case, the model is a special case of the M/G/1 variant of Model I in Bekker et al. (2009); see also e.g. Gaver & Miller (1962) for a classical related model with two service speeds.

Defining $F(t)=\lim_{t\to\infty}P(0<W(t)\le x)$ as the stationary workload distribution and by applying a level crossing argument, we obtain the following two integro-differential equations

$$F'(x)=\lambda\pi e^{-\mu_1 x}+\lambda\int_0^x e^{-\mu_1(x-y)}F'(y)dy, \qquad 0<x<k, \tag{2}$$

$$F'(x)=\lambda\pi e^{-\mu_1 x}+\lambda\int_k^x e^{-\mu_2(x-y)}F'(y)dy +\lambda\int_0^k e^{-\mu_1(x-y)}F'(y)dy, \qquad x>k. \tag{3}$$

where $F'(x)$ denotes the derivative of $F(x)$ and $\pi=\lim_{t\to\infty}P(W(t)=0)$. The left-hand side gives the downcrossing probability flow at level x that is only obtained by a continuous decline of the workload in absence of jumps. The right hand side gives the corresponding upcrossing probability flow at the same level x, given by three possible contributions: a jump starting at the origin (the first addendum), a jump starting from a state $y\in(0,k)$ (the second integral addendum) or a jump starting from a state $y\in[k,x)$ (the eventual third integral addendum).

An alternative and more rigorous deduction of the Eqs. (2) and (3) is obtained in Theorem 1 as a special case of the Eqs. (11) and (12).

The integro-differential equations can be readily transformed into ordinary differential equations by taking derivatives, that gives

$$F''(x)+(\mu_1-\lambda)F'(x)=0, \qquad 0<x<k, \tag{4}$$

$$F''(x)+(\mu_2-\lambda)F'(x)=(\mu_2-\mu_1)e^{-\mu_1(x-k)}F'(k), \qquad x>k. \tag{5}$$

Solving these equations, in terms of the density $F'(x)$, we obtain, for $0<x<k$,

$$F'(x)=\lambda\pi e^{-(\mu_1-\lambda)x}$$

whereas, for $x>k$, we have

$$F'(x)=\lambda\pi e^{-(\mu_1-\lambda)k}\left[\frac{(\mu_2-\mu_1)e^{-\mu_1(x-k)}}{\mu_2-\mu_1-\lambda}-\frac{\lambda e^{-(\mu_2-\lambda)(x-k)}}{\mu_2-\mu_1-\lambda}\right]$$

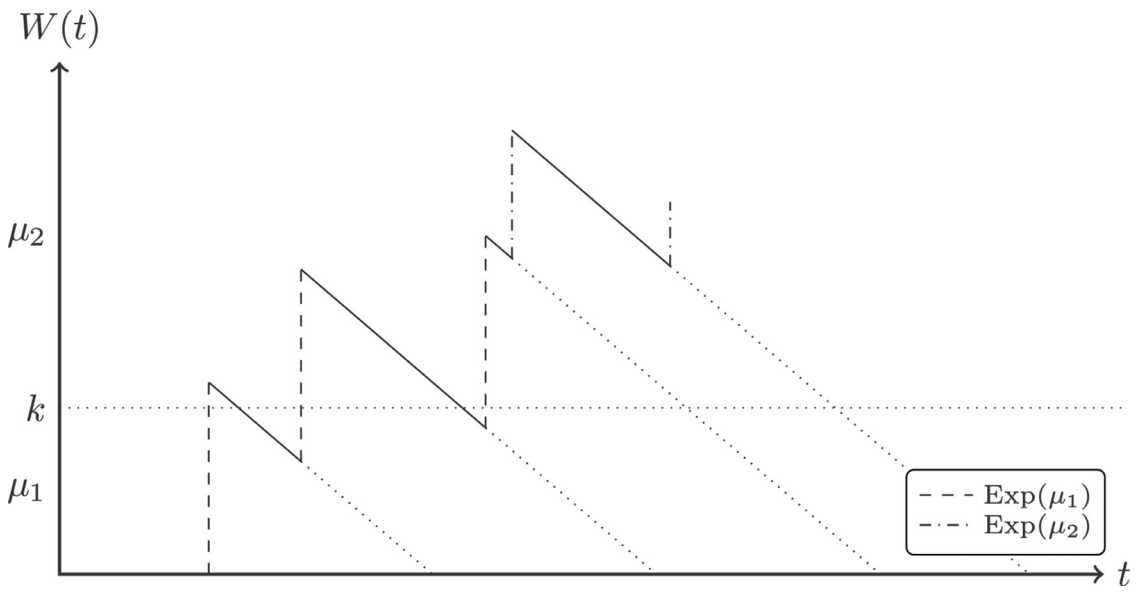

Fig. 1. Sample path of the VQT process $W(t)$ for the case $c = 1$.

with

$$\pi = \frac{(\mu_1/\lambda - 1)(\mu_2/\lambda - 1)}{(\mu_1/\lambda)(\mu_2/\lambda - 1) - (\mu_2/\mu_1 - 1)e^{(\lambda-\mu_1)k}}.$$

The VQT density allows for an intuitive interpretation. Specifically, in the region $(0, k)$ jump sizes are always $\exp(\mu_1)$, which implies that $F'(x)$ is proportional to the limiting workload density in an M/M/1 queue with service rate μ_1 (and finite workload capacity k). Also, observe that sample paths of $W(t)$ in the region (k, ∞) are always initiated by an upcrossing of k with a jump of size $\exp(\mu_1)$, after which all jumps are $\exp(\mu_2)$ until a subsequent downcrossing of k. This implies that $W(t)$ in (k, ∞) behaves as the workload process in an M/M/1 queue with service rate μ_2, but with an exceptional first service time in a busy period that has rate μ_1. This directly explains the mixture of the two exponentials in (k, ∞).

4. Multi-server queue: balance equations

In this section, assuming $c > 1$, we derive the balance equations for the VQT process $(W(t), S(t))$ defined in Section 2 that are satisfied by the limiting distribution. It is straightforward to see that the VQT process is stable if

$$\frac{\lambda}{\mu_2 c} < 1. \tag{6}$$

We shall assume stability from now on and focus on the limiting distribution of $(W(t), S(t))$.

Now let, for $x \geq 0$ and $t \geq 0$,

$$F_i(t, x) = P(0 < W(t) \leq x; S(t) = (i, c - 1 - i)), \quad 0 \leq i \leq c - 1.$$

Define $F_i(x) = \lim_{t\to\infty} F_i(t, x)$, and define the row vector function

$$F(x) = [F_0(x),\ F_1(x),\ \ldots, F_{c-1}(x)],$$

whose first two derivatives are denoted by $F'(x)$ and $F''(x)$. Also, for the case that no customers are waiting, let

$$\pi(i, j) = \lim_{t\to\infty} P(W(t) = 0, S(t) = (i, j)), \quad 0 \leq i + j \leq c - 1,$$

and

$$\delta_i = [\pi(j, i - j),\ 0 \leq j \leq i],\ 0 \leq i \leq c - 1.$$

Using the transition rates derived above, we see that the π's satisfy the following balance equations:

$$\begin{aligned}&(\lambda + i\mu_1 + j\mu_2)\pi(i, j)\\ &= (i + 1)\mu_1\pi(i + 1, j) + (j + 1)\mu_2\pi(i, j + 1)\\ &+ \lambda\mathbf{1}_{(i>0)}\pi(i - 1, j), \quad 0 \leq i + j < c - 1.\end{aligned} \tag{7}$$

with $\mathbf{1}_{(\cdot)}$ denoting the indicator function.

Next we derive the integro-differential equations satisfied by $F(\cdot)$ for the case there is queueing delay.

We denote by I the identity matrix, whose size will be clear from the context, and by $\hat{I}$ a rectangular matrix obtained from I by adding a null column on the left, i.e. $\hat{I} = (\ 0 \quad I\ )$. Throughout the paper, we will use the convention of denoting by $\hat{\cdot}$ a rectangular (instead of square) matrix.

Let B_1 be a $c \times c$ square matrix with entries given by

$$\begin{aligned}B_1(i, i) &= (i + 1)\mu_1, \quad 0 \leq i \leq c - 1,\\ B_1(i, i + 1) &= (c - i - 1)\mu_2, \quad 0 \leq i < c - 1,\\ B_1(i, j) &= 0 \quad \text{for all other } (i, j),\end{aligned}$$

and B_2 be a $c \times c$ square matrix with entries given by

$$\begin{aligned}B_2(i, i) &= (c - i)\mu_2, \quad 0 \leq i \leq c - 1,\\ B_2(i, i - 1) &= i\mu_1, \quad 1 \leq i \leq c - 1,\\ B_2(i, j) &= 0 \quad \text{for all other } (i, j).\end{aligned}$$

We finally define the matrices

$$\Delta_i = \text{diag}(j\mu_1 + (i - j)\mu_2,\ 0 \leq j \leq i), \quad 0 \leq i \leq c - 1, \tag{8}$$

and

$$\tilde{\Delta}_\kappa = \mu_\kappa I + B_\kappa^{-1}(\Delta_{c-1})B_\kappa, \quad \kappa \in \{1, 2\}, \tag{9}$$

$$\tilde{Q}_\kappa(x) = \exp(-\tilde{\Delta}_\kappa x), \quad \kappa \in \{1, 2\}. \tag{10}$$

Theorem 1. *The limiting distribution vector F satisfies the following integro-differential equations:*

$$\begin{aligned}F'(x) = \lambda F(x) - \lambda\int_0^x F(y)B_1\tilde{Q}_1(x - y)dy + F'(0)\\ -\lambda\delta_{c-1}B_1(I - \tilde{Q}_1(x))\tilde{\Delta}_1^{-1}, \quad 0 < x < k\end{aligned} \tag{11}$$

$$F'(x) = \lambda F(x) - \lambda\int_k^x F(y)B_2\tilde{Q}_2(x - y)dy + F'(0)$$

$$-\lambda\delta_{c-1}B_1(I-\tilde{Q}_1(x))\tilde{\Delta}_1^{-1}-\lambda\int_0^k F(y)B_1\tilde{Q}_1(x-y)dy$$
$$-\lambda F(k)\int_k^x (B_1\tilde{Q}_1(x-y)-B_2\tilde{Q}_2(x-y))dy,\quad x>k. \quad (12)$$

Proof. The proof follows standard probabilistic arguments and uses an infinitesimal approach, which we defer to the Appendix A. □

Eqs. (11) and (12) are integro-differential equations. These equations are related to level crossings principles; see Section 3 for the single-server case providing additional intuitive insight. To find the limiting distribution, we first convert them to second order linear non-homogeneous differential equations. They are given in the following theorem. To do so, first define the differential operators $\mathcal{L}_1$ and $\mathcal{L}_2$ as follows:

$$\mathcal{L}_\kappa G(x)=G''(x)-G'(x)(\lambda I-\tilde{\Delta}_\kappa)+\lambda G(x)(B_\kappa-\tilde{\Delta}_\kappa)\quad \kappa\in\{1,2\}.$$

Theorem 2. *The limiting distribution vector F satisfies the following second order differential equations:*

$$\mathcal{L}_1F(x)=\alpha_0,\quad 0<x<k \quad (13)$$

$$\mathcal{L}_2F(x)=\alpha_1+\alpha_2\tilde{Q}_1(x-k)(\tilde{\Delta}_1-\tilde{\Delta}_2),\quad x>k \quad (14)$$

where

$$\alpha_0=F'(0)\tilde{\Delta}_1-\delta_{c-1}\lambda B_1 \quad (15)$$

$$\alpha_1=\alpha_0\tilde{\Delta}_1^{-1}\tilde{\Delta}_2-\lambda F(k+)(B_1\tilde{\Delta}_1^{-1}\tilde{\Delta}_2-B_2) \quad (16)$$

$$\alpha_2=\alpha_1\tilde{\Delta}_2^{-1}-F'(k+)+\lambda F(k+)(I-B_2\tilde{\Delta}_2^{-1}). \quad (17)$$

Proof. This follows from rewriting the integro-differential Eqs. (11) and (12), see Appendix A. □

The next corollary gives the boundary conditions for F. Here and later we use the notation $g(x\pm):=\lim_{y\to x^\pm}g(y)$ to denote the one-sided limits of the function g at x.

Corollary 1. *The limiting distribution vector F satisfies the following boundary conditions:*

$$F(0)=0, \quad (18)$$

$$F(k-)=F(k+), \quad (19)$$

$$F'(k-)=F'(k+), \quad (20)$$

$$F'(0)=\delta_{c-1}(\lambda I+\Delta_{c-1})-\delta_{c-2}\lambda\hat{I}. \quad (21)$$

Proof. See Appendix A for details. □

Remark 1. In the single-server case, Eq. (21) would reduce to $F'(0)=\lambda\pi$ that combined with Eqs. (11) and (12) would give Eqs. (2) and (3), respectively. Similary, Eqs. (13) and (14) respectively reduce to Eqs. (4) and (5).

5. Solution of the balance equations

In this section we determine the limiting distribution by developing the analytical solution of the differential equations in Theorem 2 and the boundary conditions in Corollary 1. In fact, we present two different ways to express the limiting distribution of the VQT process. The first is based on a scalar representation and clearly reveals that $F(x)$ can be written as a mixture of exponentials. This representation gives insight in the probabilistic interpretation of the queueing delay and is presented in Section 5.1. The second concerns a matrix representation and is more compact. This representation is more amenable for numerical computations and can be found in Section 5.2. Finally, in Section 5.3 a detailed example is explicitly solved.

5.1. Limiting distribution

For the solution of Eq. (13) in Theorem 2, we first need to solve the homogeneous equation. Hence, we first need to define several backgrounds quantities. Consider the following quadratic eigenvalue equation:

$$\phi[\theta^2I-\theta(\lambda I-\tilde{\Delta}_1)+\lambda(B_1-\tilde{\Delta}_1)]=0. \quad (22)$$

There are $2c$ solutions $\{(\theta_i,\phi_i),\ 0\le i\le 2c-1\}$ to the above system. Since the matrices involved in the above equation are all upper triangular, it is easy to see that these $2c$ solutions are given by the solutions to the following quadratic equations:

$$\theta^2-\theta(\lambda-(i+1)\mu_1-(c-1-i)\mu_2)$$
$$-(c-1-i)\lambda\mu_2=0,\quad 0\le i\le c-1. \quad (23)$$

If both conditions $\lambda\neq c\mu_1$ and $\lambda\neq c(\mu_1-\mu_2)$ are satisfied, all these eigenvalues are real and distinct, see also Remark 3 for the other cases. For the rest of the paper we implicitly assume that these conditions hold. The solutions $\{\theta_i, 0\le i\le c-1\}$ are given by

$$\theta_i=\frac{1}{2}(\lambda-(i+1)\mu_1-(c-1-i)\mu_2)$$
$$-\frac{1}{2}\sqrt{(\lambda-(i+1)\mu_1-(c-1-i)\mu_2)^2+4(c-1-i)\lambda\mu_2}. \quad (24)$$

Note that $\{\theta_i, 0\le i\le c-2\}$ are negative and $\theta_{c-1}=\min\{0,\lambda-c\mu_1\}$. The solutions $\{\theta_{i+c}, 0\le i\le c-1\}$ are positive and are given by

$$\theta_{i+c}=\frac{1}{2}(\lambda-(i+1)\mu_1-(c-1-i)\mu_2)$$
$$+\frac{1}{2}\sqrt{(\lambda-(i+1)\mu_1-(c-1-i)\mu_2)^2+4(c-1-i)\lambda\mu_2}. \quad (25)$$

Note that $\theta_{2c-1}=\max\{0,\lambda-c\mu_1\}$. The corresponding eigenvectors $\{\phi_i, 0\le i\le 2c-1\}$ are easy to compute. In particular, the eigenvector corresponding to the null eigenvalue is denoted by

$$\phi_*=[0,\ 0,\ldots,0,\ 1], \quad (26)$$

which is a row vector of length c.

Next, we turn to the homogeneous equation based on Eq. (14). For this, consider the following quadratic eigenvalue equation:

$$\psi[\beta^2I-\beta(\lambda I-\tilde{\Delta}_2)+\lambda(B_2-\tilde{\Delta}_2)]=0. \quad (27)$$

There are $2c$ solutions $\{(\beta_i,\psi_i),\ 0\le i\le 2c-1\}$ to the above system. The β_i's for $0\le i\le c-1$ are given by

$$\beta_i=\frac{1}{2}\left(\lambda-i\mu_1-(c-i)\mu_2-\sqrt{(\lambda-i\mu_1-(c-i)\mu_2)^2+4i\lambda\mu_1}\right), \quad (28)$$

and the β_{i+c}'s, for $0\le i\le c-1$ are given by

$$\beta_{i+c}=\frac{1}{2}\left(\lambda-i\mu_1-(c-i)\mu_2+\sqrt{(\lambda-i\mu_1-(c-i)\mu_2)^2+4i\lambda\mu_1}\right). \quad (29)$$

Similarly to Eqs. (24), and (25), all these eigenvalues are real and distinct, assuming that $\lambda \neq c(\mu_2 - \mu_1)$. The eigenvalues $\{\beta_i, 0 \leq i \leq c-1\}$ are negative, $\beta_c = 0$ and $\{\beta_{i+c}, 1 \leq i \leq c-1\}$ are positive. The corresponding eigenvectors $\{\psi_i, 0 \leq i \leq 2c-1\}$ are easy to compute. In particular,

$$\psi_c = [1,\ 0, \ldots, 0,\ 0], \tag{30}$$

which is a row vector of length c.

Remark 2. The parameters θ_i and β_i may be related to the virtual waiting in regular M/M/c queues. For instance, $\exp(\theta_{c-1}) = \exp(\lambda - c\mu_1)$ and $\exp(\beta_0) = \exp(\lambda - c\mu_2)$ are proportional to the stationary densities of the VQT in M/M/c queues with only service rates μ_1 and μ_2, respectively. Moreover, consider the process $W(t) \in (0, k)$ and fix the server state process $S(t) = (i, c-1-i)$; then the VQT process is decreasing with rate 1 and makes jumps with rate λ of size $\exp(\Delta(i+1, c-1-i))$. Upon a jump, the server state process $S(t)$ changes with probability $(c-1-i)\mu_2/\Delta(i+1, c-1-i)$, which may be interpreted as a type of clearing (Boxma, Perry, & Stadje, 2001). It may be verified that the stationary density of such a 'clearing system' is a mixture of $\exp(\theta_i)$ and $\exp(\theta_{i+c})$. A similar argument applies to $W(t) > k$ in terms of β_i.

As mentioned, to get the solution of the differential equations of Theorem 2 we first need to find the solution of the homogeneous differential equations using the above, and then we look for a particular solution. However, in order to construct a particular solution, since both Eqs. (13) and (14) admit zero as eigenvalue, we would need together with the left eigenvectors ϕ_* and ψ_c, respectively defined in Eqs. (26) and (30), the corresponding right eigenvectors that we denote by $\tilde{\phi}_*$ and $\tilde{\psi}_c$. The following result shows an important relation between those eigenvectors that will be used later in the proof of Theorem 3 to show that $F(x)$ does not have a linear term.

Lemma 1. *Let $\tilde{\phi}_*$ and $\tilde{\psi}_c$ be the right eigenvectors corresponding to the left eigenvectors ϕ_* and ψ_c, i.e. satisfing the following relations*

$$(B_1 - \tilde{\Delta}_1)\tilde{\phi}_* = 0, \tag{31}$$

$$(B_2 - \tilde{\Delta}_2)\tilde{\psi}_c = 0. \tag{32}$$

It follows that

$$\alpha_1 \cdot \tilde{\psi}_c = 0 \Rightarrow \alpha_0 \cdot \tilde{\phi}_* = 0. \tag{33}$$

Proof. The proof uses linear algebra techniques and is included in the Appendix A. □

The next result gives the solution of the differential equations of Theorem 2 in terms of $4c$ unknowns $\{a_i, 0 \leq i \leq 2c-1\}$ and $\{b_i, 0 \leq i \leq c\}$.

Theorem 3. *For $0 < x < k$, the solution is given by*

$$F(x) = \sum_{i=0}^{2c-1} a_i e^{\theta_i x}\phi_i + \alpha_0 M_0, \quad 0 \leq x \leq k, \tag{34}$$

with

$$M_0 = (\lambda(B_1 - \tilde{\Delta}_1) + \mathrm{diag}(\phi_*))^{-1}. \tag{35}$$

For $x > k$, the solution is given by

$$F(x) = \sum_{i=0}^{c-1} b_i e^{\beta_i(x-k)}\psi_i + b_c\psi_c + \alpha_1 M_1 + \alpha_2 \tilde{Q}_1(x-k)(\tilde{\Delta}_1 - \tilde{\Delta}_2)M_2, \tag{36}$$

with

$$M_1 = (\lambda(B_2 - \tilde{\Delta}_2) + \mathrm{diag}(\psi_c))^{-1}, \tag{37}$$

$$M_2 = ((c\mu_1 + \lambda)(c\mu_1 I - \tilde{\Delta}_2) + \lambda B_2)^{-1}. \tag{38}$$

Proof. This follows from solving the systems of second order linear differential equations below and above level k in Theorem 2, thereby also utilizing Lemma 1; see Appendix A. □

Remark 3. We note that in the special case that $\lambda = c\mu_1$, we have two identical eigenvalues $\theta_{c-1} = \theta_{2c-1} = 0$. Furthermore, in case $\lambda = c(\mu_1 - \mu_2)$, it holds that $\theta_i = \theta_{c-1} = \lambda - c\mu_1$ for all $i = 0, \ldots, c-1$. In that case, $F(x)$ for $0 < x < k$, contains terms of the form $\sum_{i=0}^{c-1} a_i x^i e^{(\lambda - c\mu_1)x}$. Similarly, if $\lambda = c(\mu_2 - \mu_1)$, then the β_i for $i = 0, \ldots, c-1$ are identical to β_0 and the mixture of exponentials in $F(x)$ for $x > k$ needs to be replaced.

Remark 4. In case $\mu_1 = \mu_2$ the model reduces to a simple M/M/c queue. Therefore, in this case the VQT distribution is given by

$$P(W \leq x) = 1 - C(c, \lambda/\mu_2)\frac{1}{1-\rho}\exp\{-c\mu_2(1-\rho)x\},$$

where $C(c, \lambda/\mu_2) = 1/(1 + (1-\rho)(\frac{c!}{c\rho^c})\sum_{k=0}^{c-1}\frac{(c\rho)^k}{k!})$ is the Erlang's C formula and $\rho = \lambda/(c\mu_2)$ is the *server utilization*.

We next use the boundary conditions in Corollary 1 to solve for the $3c+1$ unknown constants $\{a_i, 0 \leq i \leq 2c-1\}$ and $\{b_i, 0 \leq i \leq c\}$. We also have $c(c+1)/2$ probabilities $\pi(i, j),\ 0 \leq i + j \leq c-1$ that need to be determined. The result is given in the next theorem.

Theorem 4. *The constants $\{a_i, 0 \leq i \leq 2c-1\}$, $\{b_i, 0 \leq i \leq c\}$ and probabilities $\pi(i, j),\ 0 \leq i + j \leq c-1$ satisfy the following equations:*

$$\sum_{i=0}^{2c-1} a_i\phi_i + \alpha_0 M_0 = 0, \tag{39}$$

$$\sum_{i=0}^{2c-1} a_i e^{\theta_i k}\phi_i + \alpha_0 M_0 = \sum_{i=0}^{c-1} b_i\psi_i + b_c\psi_c + \alpha_1 M_1 + \alpha_2(\tilde{\Delta}_1 - \tilde{\Delta}_2)M_2, \tag{40}$$

$$\sum_{i=0}^{2c-1} a_i\theta_i e^{\theta_i k}\phi_i = \sum_{i=0}^{c-1} b_i\beta_i\psi_i - \alpha_2\tilde{\Delta}_1(\tilde{\Delta}_1 - \tilde{\Delta}_2)M_2, \tag{41}$$

$$\sum_{i=0}^{2c-1} a_i\theta_i\phi_i = \delta_{c-1}(\lambda I + \Delta_{c-1}) - \delta_{c-2}\lambda\hat{I}, \tag{42}$$

$$\begin{aligned}&(\lambda + i\mu_1 + j\mu_2)\pi(i, j)\\ &\quad = (i+1)\mu_1\pi(i+1, j) + (j+1)\mu_2\pi(i, j+1)\\ &\quad + \lambda\mathbf{1}_{(i>0)}\pi(i-1, j), \quad 0 \leq i + j < c-1\end{aligned} \tag{43}$$

and the normalizing equation

$$b_c\psi_c\mathbb{1} + \alpha_1 M_1\mathbb{1} + \sum_{i,j:i+j=c-1}^{i+j=c-1} \pi(i, j) = 1, \tag{44}$$

where $\mathbb{1}$ denotes the all-one vector.

Proof. Eqs. (39)–(42) follow from the boundary conditions in Eqs. (18)–(21), respectively. Eq. (43) presents the balance equations in Eq. (7). Finally, Eq. (44) is the normalizing equation. □

Eqs. (39) –(44) from the above theorem yield $4c + c(c-1)/2 + 1 = 3c + 1 + c(c+1)/2$ linear equations for the $3c+1$ unknown constants $\{a_i, 0 \leq i \leq 2c-1\}$, $\{b_i, 0 \leq i \leq c\}$ and the $c(c+1)/2$ unknown probabilities $\pi(i, j),\ 0 \leq i + j \leq c-1$. Theorem 4 gives necessary conditions for the constants to satisfy, but it does not assure that Eqs. (39)–(44) characterize them. In the following section we show, starting from these equations, how we are able to construct the unique solution.

5.2. Computing the solution

For the limiting distribution in Theorem 3, we need Theorem 4 that expresses all constants as a solution of a quite large system of linear equations. In this section we try to express the solution in simpler terms that are easier to implement in a computer language equipped with basic matrix functions. In addition the contruction of the solution given below shows that the system in Theorem 4 only admits a unique solution.

First of all, it helps to write the function $F(x)$, given in Eqs. (34) and (36), in matrix form. We start by defining the matrices $\Phi^- = [\phi_i^t]^t$, $0 \le i \le c-1$ and $\Phi^+ = [\phi_i^t]^t$, $c \le i \le 2c-1$, whose rows are given by the eigenvectors corresponding to the eigenvalues in Eqs. (24) and (25). We also define the diagonal matrices $\Theta^- = \text{diag}(\theta_i)$, $0 \le i \le c-1$, and $\Theta^+ = \text{diag}(\theta_i)$, $c \le i \le 2c-1$. Then, the matrices $U_1^{\pm} = (\Phi^{\pm})^{-1}\Theta^{\pm}\Phi^{\pm}$, solve the equation

$$U_1^2 - U_1(\lambda I - \tilde{\Delta}_1) + \lambda(B_1 - \tilde{\Delta}_1) = 0,$$

with U_1^- having all non-positive (negative) eigenvalues, U_1^+ having all positive (non-negative) eigenvalues if $\lambda > c\mu_1$ ($\lambda < c\mu_1$).

In a similar way we construct the matrices $U_2^{\pm}$ solving the equation

$$U_2^2 - U_2(\lambda I - \tilde{\Delta}_2) + \lambda(B_2 - \tilde{\Delta}_2) = 0$$

with U_2^- with all negative eigenvalues and U_2^+ with all non-negative eigenvalues. This allows us to rewrite the expression in Eqs. (34) and (36) as

$$\begin{aligned} F(x) &= a^- e^{U_1^- x} + a^+ e^{U_1^+ x} + \alpha_0 M_0, \quad 0 \le x \le k, \\ F(x) &= b^- e^{U_2^- x} + b_c\psi_c + \alpha_1 M_1 + \alpha_2 \tilde{Q}_1(x-k)(\tilde{\Delta}_1 - \tilde{\Delta}_2)M_2, \quad x \ge k, \end{aligned}$$

for unknown constant vectors a^-, a^+ and b^-. We then express these vectors in terms of the unknown vector $F'(0)$ by using the continuity conditions given in Corollary 1. This gives the following easier expressions

$$\begin{aligned} F(x) = {} & (F'(0) + \alpha_0 M_0 U_1^-)(U_1^+ - U_1^-)^{-1}(e^{U_1^+ x} - e^{U_1^- x}) \\ & + \alpha_0 M_0 (I - e^{U_1^- x}), \quad 0 \le x \le k, \end{aligned} \tag{45}$$

$$\begin{aligned} F(x) = {} & F(k-)e^{U_2^-(x-k)} + (b_c\psi_c + \alpha_1 M_1)(I - e^{U_2^-(x-k)}) \\ & - \alpha_2(\tilde{\Delta}_1 - \tilde{\Delta}_2)M_2 e^{U_2^-(x-k)} \\ & + \alpha_2 \tilde{Q}_1(x-k)(\tilde{\Delta}_1 - \tilde{\Delta}_2)M_2, \quad x \ge k. \end{aligned} \tag{46}$$

In the theorem below, $F'(0)$ is expressed in terms of δ_{c-1} and b_c.

Theorem 5. *The function F, given in Eqs. (45) and (46), may be written in terms of the constant b_c and the vector δ_{c-1}, since*

$$F'(0) = \delta_{c-1}H_{16} - b_c\psi_c H_{15}, \tag{47}$$

where H_{15} and H_{16} are defined in Eqs. (A.26) and (A.27), respectively.

In particular it follows that

$$F(\infty) = \delta_{c-1}H_{20} + b_c\psi_c H_{19}, \tag{48}$$

with H_{19} and H_{20} defined in Eqs. (A.29) and (A.30), respectively.

Proof. This follows from Corollary 1 and some tedious rewriting, see Appendix A. □

Having expressed the function $F(x)$ for $x \ge 0$, in terms of δ_{c-1}, it is only left to find the probability of the discrete states, together with the constant b_c that can be found by using the normalizing equation. This is the result of following theorem.

Theorem 6. *The discrete probabilities can be computed as follows*

$$\delta_i = b_c\psi_c\hat{H}_i, \quad 0 \le i \le c-1, \tag{49}$$

where the matrices $\hat{H}_i$ are defined in Eq. (A.35) and with the constant b_c computed as

$$b_c^{-1} = \psi_c\left(H_{19} + \hat{H}_{c-1}H_{20} + \sum_{0 \le n \le c-1}\hat{H}_n\right)\mathbb{1}, \tag{50}$$

where H_{19} and H_{20} are given in Eqs. (A.29) and (A.30) in the Appendix A.

Proof. The linear system of equations may be rewritten to recursively express δ_{c-1} in terms of b_c, whereas b_c follows from normalization; see Appendix A. □

5.3. Numerical example for $c = 2$

As an illustration of the balance equations and the limiting distribution, we consider the 2-server case in this section. A visual representation of the transition diagram of $(W(t), S_1(t), S_2(t))$ can be found in Fig. 2. To avoid excessive expressions, we focus on a numerical example with specific parameters. We fix $k = 0.45$, $\lambda = 2$, $\mu_1 = 0.75$ and $\mu_2 = 1.12$. The second order differential equations of Theorem 2 then looks as follows: for $0 \le x < 0.45$, we have

$$\begin{aligned} F_0''(x)1.13F_0'(x)2.24F_0(x) &= 1.87F_0'(0)1.5\pi(0,1), \\ F_1''(x)0.55F_0'(x)0.5F_1'(x)1.13F_0(x) &= 0.55F_0'(0)1.5F_1'(0) \\ &\quad -2.24\pi(0,1)3.0\pi(1,0), \end{aligned}$$

whereas, in the region $x > 0.45$, we obtain

$$\begin{aligned} & F_0''(x) + 2.24F_0'(x) - 2.25F_1'(x) + 2.0F_1(x) = \\ & \quad +2.24F_0'(0) - 0.25F_1'(0) - 1.8\pi(0,1) \\ & \quad +2.68F_0(0.45+) + 1.5F_1(0.45+) \\ & \quad 0.09e^{-1.87x}(F_0'(0) - F_0'(0.45+)) \\ & \quad 0.07e^{-1.87x}\pi(0,1) \\ & \quad 0.11e^{-1.87x}F_0(0.45+), \\ & F_1''(x) - 0.13F_1'(x) - 1.5F_1(x) = \\ & \quad +1.87F_1'(0) - 3.74\pi(1,0) - 1.5F_1(0.45+) \\ & \quad +(-0.29e^{1.87x} - 0.2e^{-1.5x})(F_0'(0) - F_0'(0.45+)) \\ & \quad 0.13e^{1.5x}(F_1'(0) - F_1'(0.45+)) \\ & \quad 0.24e^{1.87x}\pi(0,1) - 0.26e^{-1.5x}\pi(1,0) \\ & \quad +(0.35e^{-1.87x} - 0.39e^{1.5x})F_0(0.45+). \end{aligned}$$

The boundary conditions in Corollary 1 are rather straightforward. By Theorem 5, the solution of the above system of differential equations is unique depending on a constant b_c and the components $\pi(1,0)$ and $\pi(0,1)$. That is the quantities $F_0'(0)$ and $F_1'(0)$ are determined given those values as follows

$$\begin{aligned} F_0'(0) &= 1.42\pi(0,1)2.34\pi(1,0)0.18286b_c, \\ F_1'(0) &= 0.95\pi(0,1)1.09\pi(1,0)0.16026b_c. \end{aligned}$$

It follows that the remaining unknowns $\pi(0,0)$, $\pi(1,0)$, $\pi(0,1)$ and b_c can be determined by imposing the following constraints

$$\begin{aligned} F_0'(0) &= 3.12\pi(0,1), \\ F_1'(0) &= 2.75\pi(1,0) - 2\pi(0,0), \\ 0 &= -2\pi(0,0) + 0.75\pi(1,0) + 1.12\pi(0,1), \\ 1 &= \pi(0,0) + \pi(1,0) + \pi(0,1) + F_0(\infty) + F_1(\infty), \end{aligned}$$

yielding

$$\begin{aligned} &\pi(0,0) = 3.12, b_c = 0.827051, \\ &\pi(0,1) = 1.08889, \pi(1,0) = 4.35035. \end{aligned}$$

The expressions for $F_0(x)$ and $F_1(x)$, in the interval $0 \le x \le 0.45$, are

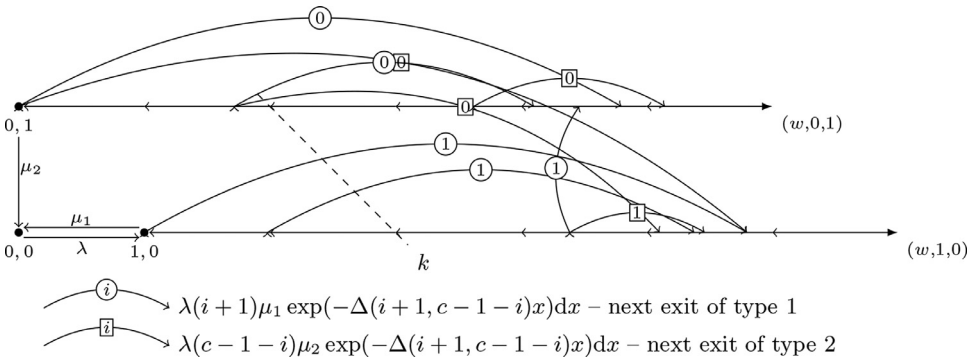

Fig. 2. Sketch of the transition diagram for the case $c = 2$.

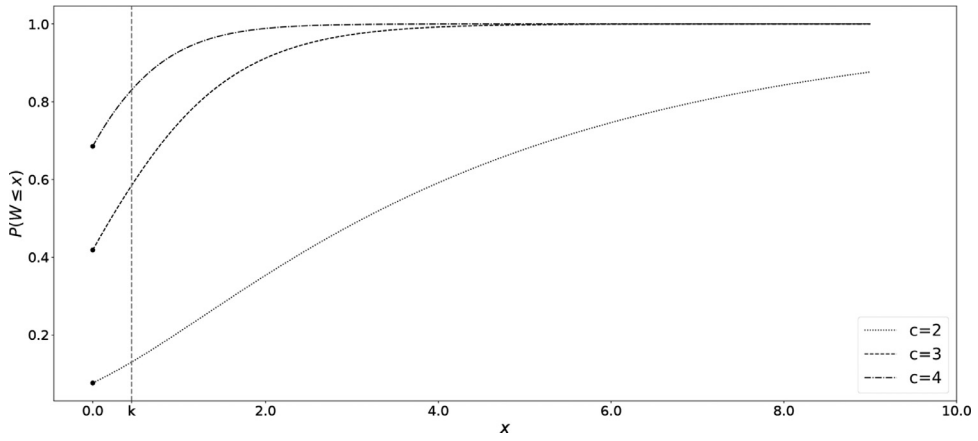

Fig. 3. Stationary VQT cumulative distribution function for $\lambda = 2$, $\mu_1 = 0.75$, $\mu_2 = 1.12$ and $k = 0.45$.[1]

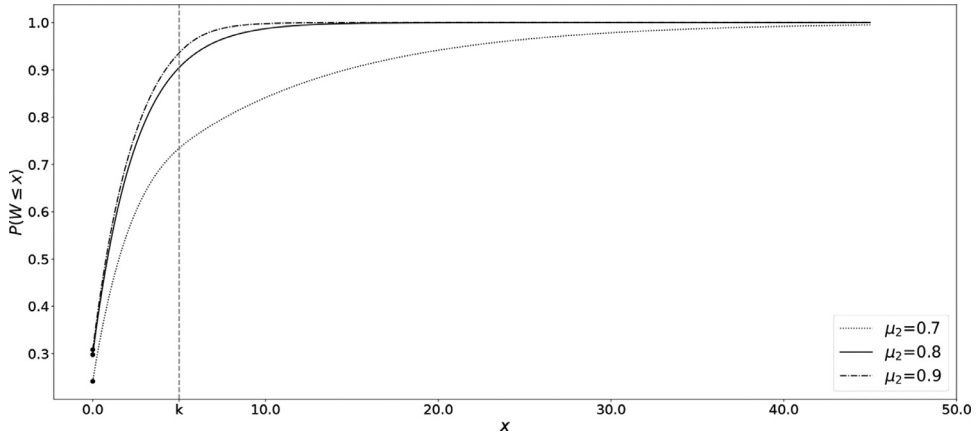

Fig. 4. Cdf of the stationary VQT for $k = 5$, $c = 3$, $\lambda = 2$, $\mu_1 = 0.8$, and $\mu_2 \in \{0.7, 0.8, 0.9\}$.[1]

$$F_0(x) = 0.0214e^{+1.5631x} - 0.0686e^{-1.4331x} - 0.0211,$$
$$F_1(x) = 0.0258e^{+1.5631x} + 0.2303e^{+0.5x} + 0.0085e^{1.4331x} - 0.18910,$$

whereas, in the interval $x > 0.45$, they are given by

$$F_0(x) = -0:9616e^{-0:24(x-0.45)} + 0.2126e^{-1.1615(x-0.45)} + 0.8271 + 0.9281e^{-1.5(x-0.45)}$$
$$F_1(x) = -0.0996e^{-1.1615(x-0.45)} + 0.0961 - 0.5847e^{-1.5(x-0.45)}.$$

Note that $\theta_1 = |\lambda - c\mu_1| = +0.5$ and $\beta_0 = \lambda - c\mu_2 = -0.24$. The distribution of the virtual queueing time is visualized in Fig. 3, along with the cases of 3 and 4 servers.

6. Numerical insights

In this section we focus on numerical insights. Specifically, we consider W, the stationary distribution of the queueing time or VQT. Note that the results in Section 5 contain more information, as they also provide the server state. To obtain VQT, observe that $P(W = 0) = \sum_{i,j:0\le i+j\le c-1} \pi(i, j)$ and $P(W \le x) = P(W = 0) + F(x)\mathbb{1}$. From $F(x)$, we may also directly derive the mean stationary VQT, which we give here in matrix representation.[1]

[1] The python algorithm to generate Figs. 3–7 is avaiblabe for downloading at the public repository (D'Auria, 2021b); see D'Auria (2021a) for an online implementation.

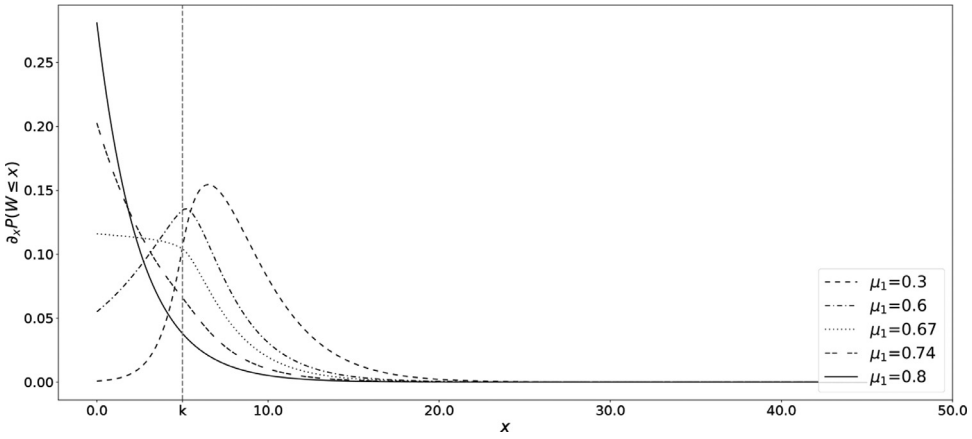

Fig. 5. Stationary VQT density for $k = 5$, $c = 3$, $\lambda = 2$, $\mu_2 = 0.8$, and $\mu_1 \in \{0.3, 0.6, 0.67, 0.74, 0.8\}$.[1]

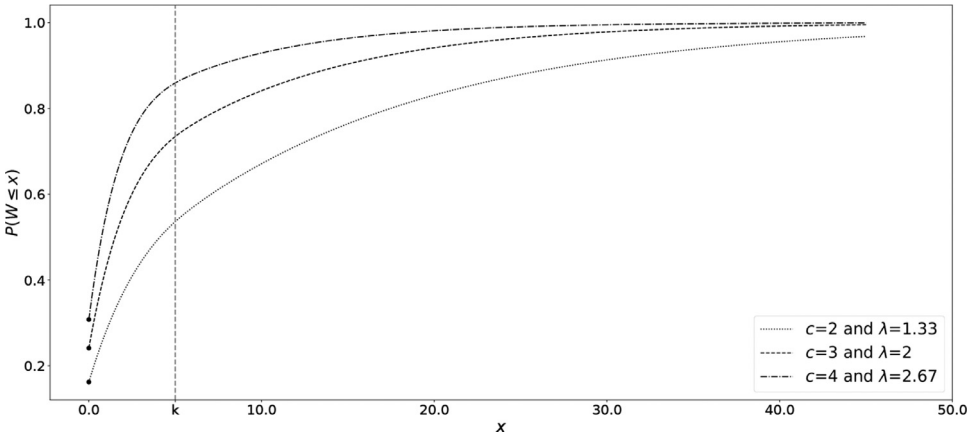

Fig. 6. Cdf of the stationary VQT for $k = 5$, $\mu_1 = 0.8$, $\mu_2 = 0.7$, $(c, \lambda) \in \{(2, 4/3), (3, 2), (4, 8/3)\}$.[1]

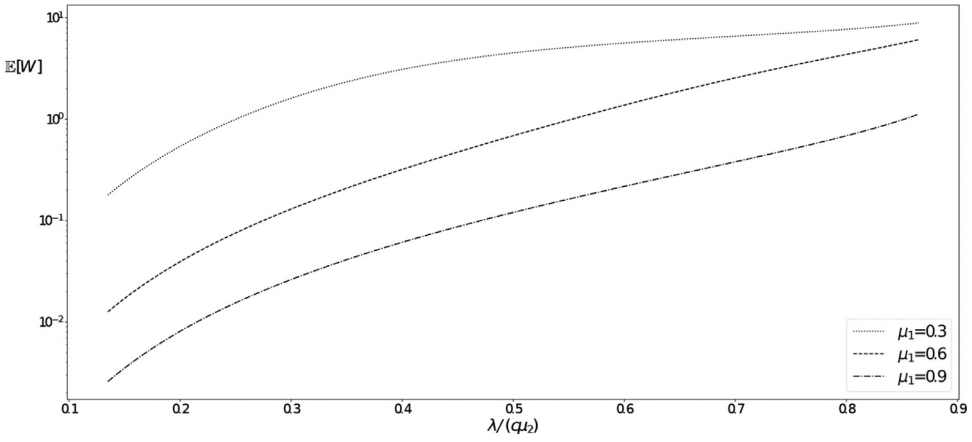

Fig. 7. Expected VQT for $k = 5$, $\mu_2 = 0.8$, $c = 3$ and $\mu_1 \in \{0.3, 0.6, 0.9\}$.[1]

Lemma 2. *The mean stationary VQT is computed as follows*

$$\begin{aligned}\mathbb{E}[W] = {} & (F'(0) + \alpha_0 M_0 U_1^-)(U_1^+ - U_1^-)^{-1} I(0, k; U_1^+)\mathbb{1} \\ & - ((F'(0) + \alpha_0 M_0 U_1^-)(U_1^+ - U_1^-)^{-1} + \alpha_0 M_0) I(0, k; U_1^-)\mathbb{1} \\ & + (F(k) - b_c \psi_c - \alpha_1 M_1 - \alpha_2 (\tilde{\Delta}_1 - \tilde{\Delta}_2) M_2)((U_2^-)^{-1} - kI)\mathbb{1} \\ & - \alpha_2 (\tilde{\Delta}_1^{-1} + kI)(\tilde{\Delta}_1 - \tilde{\Delta}_2) M_2 \mathbb{1}, \end{aligned} \tag{51}$$

where $I(a, b; D)$ is defined as in Eq. (A.38).

Proof. The results follow from the density $F'(x)$ and integration by parts, see Appendix A. □

First, we consider the impact of having different μ_1 and μ_2. In case $\mu_1 = \mu_2$, the system corresponds to the classical M/M/c queue, whereas there is a slowdown (speedup) effect when $\mu_2 < \mu_1$ ($\mu_2 > \mu_1$). As a basic example, we take $k = 5$, $c = 3$, $\lambda = 2$, and $\mu_1 = 0.8$, whereas $\mu_2 \in \{0.7, 0.8, 0.9\}$. The cumulative distribution function (cdf) of W is visualized in Fig. 4. Clearly, in case of a slowdown ($\mu_2 = 0.7$), the queueing time strongly deteriorates compared to the standard situation where $\mu_2 = \mu_1 = 0.8$. In fact, if $\mu_2 \le 2/3$ the system would even become unstable. For the current example, the impact of a speedup ($\mu_2 = 0.9$) is relatively small compared to the standard situation, as the basic service rate of 0.8

is already sufficient to provide reasonable queueing times. Hence, taking differences in service rates into account is crucial to provide reliable queueing times, especially in case of slowdowns.

Second, the shape of the VQT density may also be strongly affected by speedups (or slowdowns), i.e., differences in μ_1 and μ_2. The VQT density is strictly decreasing for the standard M/M/c queue, which will also hold in case $\mu_2 < \mu_1$ (slowdown). However, this is no longer necessarily the case for speedups, see Fig. 5 for $k = 5$, $c = 3$, $\lambda = 2$, $\mu_2 = 0.8$, and $\mu_1 \in \{0.3, 0.6, 0.67, 0.74, 0.8\}$. In particular, for more extreme variants of a speedup effect, the peak in the VQT density may be around or above level k.

Third, we consider the impact of the number of servers, i.e., scale of the system. Let $k = 5$, $\mu_1 = 0.8$, $\mu_2 = 0.7$ (slowdown), and consider systems with 2, 3, and 4 servers. We let λ be 4/3, 2, and 8/3, respectively, such that the loads $\lambda/(c\mu_i)$, for $i = 1, 2$ are identical. The cdf of the stationary VQT is presented in Fig. 6. Clearly, as the number of servers increases the queueing time improves, which is in line with economies of scale for regular M/M/c queues. We like to note that the relative ordering of cdf's below k may change in case $\lambda \geq c\mu_1$ (which we did not visualize here).

Finally, we consider the expected VQT. It is well known that $\mathbb{E}[W]$ is convex and increasing in λ for regular M/M/c queues with c and μ fixed. This property is not necessarily preserved in the current model, see Fig. 7. Specifically, in case μ_1 is relatively small ($\mu_1 = 0.3$ in Fig. 7), a large fraction of the customers will experience a VQT of around k, assuming the system to be stable. This destroys the convexity of $\mathbb{E}[W]$ as a function of λ. Moreover, the impact of μ_1 is also considerable for more heavily loaded systems. For instance, comparing $\mathbb{E}[W]$ for different $\mu_1 \in \{0.3, 0.6, 0.9\}$ with fixed $\lambda/(c\mu_2) = 0.9$, we see that the mean VQT is much smaller when a lot of customers can be served with rate μ_1 (i.e., for $\mu_1 = 0.9$). To conclude this section, we note that neglecting the differences in service rate leads to rather inadequate performance characteristics.

7. Conclusions, implications, and future research

In this paper we analyze the queueing delay in M/M/c types of queues in which the service time is affected by the experienced queueing delay. Such a mechanism may be typical from a customer perspective in which excessive waiting is associated with longer service times. Specifically, the service rate is μ_1 (μ_2) if the queueing time of the customer in service is below (above) a threshold. We show how it is possible to derive the virtual queueing time, using a specific Markov chain. The resulting queueing time can be found in closed form and consists of a mixture of exponentials.

Our key observation is that it is wholy inadequate to ignore differences in service times in the model when they do exist in practice. In case of a slowdown ($\mu_2 < \mu_1$), the performance may strongly deteriorate compared to the standard situation with equal service rates. In case of speedup ($\mu_2 > \mu_1$), situations may arise where the queueing time remains acceptable, but many customers have to wait. In fact, some queueing properties remain valid, such as the economies of scale, whereas some properties are not, such as a decreasing density of the queueing time and the convexity of the expected queueing time as a function of the arrival rate. Assessing the precise queueing time behavior without an appropriate model is difficult, and management should be supported with an online implementation (see e.g. D'Auria, 2021a).

Finally, we mention some topics for further research. First, including abandonments is an interesting topic for further research. Such abandonments can be incorporated in the state description of our specific Markov chain similar to Adan et al. (2019). Second, extending the service times to phase-type distributions provides insight in the impact of the variability in service times. Including the phase of each server in the state description then leads to a higher dimensional state description, see e.g. Ramaswami & Lucantoni (1985) for an algorithmic approach in case of the G/PH/c queue. Third, a challenge is to extend the current model to allow for multiple thresholds.

Appendix A. Proofs

In this appendix, we present the technical proofs of the results presented throughout the paper.

Proof of Theorem 1: Let $0 \leq x < k$, and $0 \leq i \leq c-1$ be fixed. Define $P_i(W(t) \in A) = P(W(t) \in A, S(t) = (i, c-1-i))$, with $A \subset \mathbb{R}$. Conditioning on the jump size being exactly equal to $x - y$, we get

$$\begin{aligned} F_i(t,x) = &\int_0^x P_i(W(t-h) \leq y+h)\lambda h(i+1)\mu_1 e^{-\Delta(i+1,c-1-i)(x-y)}dy \\ &+ \int_0^x P_{i-1}(W(t-h) \leq y+h)\lambda h(c-i)\mu_2 e^{-\Delta(i,c-i)(x-y)}dy \\ &+ (1-\lambda h)P_i(h < W(t-h) \leq x+h) + o(h). \end{aligned} \tag{A.1}$$

Define

$$Q_\kappa(x) = \exp(-(\mu_\kappa I + \Delta_{c-1})x), \quad \kappa \in \{1, 2\},$$

let $t \to \infty$, divide by h, and rearrange terms to get

$$\begin{aligned} \frac{1}{h}(F_i(x) - F_i(x+h)) = &-\lambda F_i(x+h) - (1-\lambda h)\frac{1}{h}(F_i(h) - 0) \\ &+ \lambda \sum_{j=0}^{c-1} \int_0^x (\pi(j, c-1-j) + F_j(y+h)) \sum_{k=0}^{c-1} Q_1(x-y)B_1(k,i)dy \\ &+ o(h)/h. \end{aligned}$$

Letting $h \to 0$ and multiplying both sides by -1, and noting that $F(0) = 0$, we get Eq. (11), after noticing that $Q_1 B_1 = B_1 \tilde{Q}_1$.

For $x > k$, again with $P_i(W(t) \in A) = P(W(t) \in A, S(t) = (i, c-1-i))$ for $A \subset \mathbb{R}$, we have

$$\begin{aligned} F_i(t,x) = &\int_0^k P_i(W(t-h) \leq y+h)\lambda h(i+1)\mu_1 e^{-\Delta(i+1,c-1-i)(x-y)}dy \\ &+ \int_0^k P_{i-1}(W(t-h) \leq y+h)\lambda h(c-i)\mu_2 e^{-\Delta(i,c-i)(x-y)}dy \\ &+ \int_k^x P_i(W(t-h) \leq k)\lambda h(i+1)\mu_1 e^{-\Delta(i+1,c-1-i)(x-y)}dy \\ &+ \int_k^x P_{i-1}(W(t-h) \leq k)\lambda h(c-i)\mu_2 e^{-\Delta(i,c-i)(x-y)}dy \\ &+ \int_k^x P_i(k < W(t-h) \leq y+h)\lambda h(c-i)\mu_2 e^{-\Delta(i,c-i)(x-y)}dy \\ &+ \int_k^x P_{i+1}(k < W(t-h) \leq y+h) \\ &\quad \lambda h(i+1)\mu_1 e^{-\Delta(i+1,c-1-i)(x-y)}dy \\ &+ (1-\lambda h)P_i(h < W(t-h) \leq x+h) + o(h). \end{aligned} \tag{A.2}$$

Following the same steps as above, we get

$$\begin{aligned} -F'(x) = &-\lambda F(x) - F'(0) \\ &+ \lambda \int_0^k (\delta_{c-1} + F(y))Q_1(x-y)B_1 dy + \lambda(\delta_{c-1} + F(k)) \\ &\quad \int_k^x Q_1(x-y)B_1 dy \\ &+ \lambda \int_k^x (\delta_{c-1} + F(y))Q_2(x-y)B_2 dy - \lambda(\delta_{c-1} + F(k)) \\ &\quad \int_k^x Q_2(x-y)B_2 dy. \end{aligned}$$

Substituting $Q_\kappa B_\kappa = B_\kappa \tilde{Q}_\kappa$, $\kappa \in \{1, 2\}$, yields Eq. (12). □

Proof of Theorem 2: First consider Eq. (11). Note that

$$\tilde{Q}_\kappa'(x) = -\tilde{Q}_\kappa(x)\tilde{\Delta}_\kappa, \quad \kappa = 1, 2.$$

Since B_1 is invertible, we can use Eq. (11) to get, for $0 \le x \le k$,

$$\lambda \int_0^x F(y)B_1\tilde{Q}_1(x-y)dy = -F'(x) + F'(0) + \lambda F(x) - \lambda\delta_{c-1}B_1(I - \tilde{Q}_1(x))\tilde{\Delta}_1^{-1}. \tag{A.3}$$

Taking the derivative with respect to x on both sides of Eq. (11), we get

$$F''(x) = \lambda F'(x) - \lambda F(x)B_1 + \lambda \int_0^x F(y)B_1\tilde{Q}_1(x-y)\tilde{\Delta}_1 dy - \lambda\delta_{c-1}B_1\tilde{Q}_1(x).$$

Substituting Eq. (A.3) in the above equation we get Eq. (13) with

$$\begin{aligned}\alpha_0 &= F'(0)\tilde{\Delta}_1 - \delta_{c-1}\lambda B_1,\\ &= [\delta_{c-1}(\lambda I + \Delta_{c-1}) - \delta_{c-2}\lambda\hat{I}]\tilde{\Delta}_1 - \delta_{c-1}\lambda B_1.\end{aligned} \tag{A.4}$$

Here we have used Eq. (21) to eliminate $F'(0)$. This is a second order linear differential equation with constant coefficients and a constant driving function on the right hand side.

Next we consider Eq. (12). First, note that, for $x > k$, we get, by applying Eq. (A.3), that

$$\begin{aligned}\lambda \int_0^k F(y)B_1\tilde{Q}_1(x-y)dy &= \left[\lambda \int_0^k F(y)B_1\tilde{Q}_1(k-y)dy\right]\tilde{Q}_1(x-k)\\ &= [-F'(k) + F'(0) + \lambda F(k)\\ &\quad - \lambda\delta_{c-1}B_1(I - \tilde{Q}_1(k))\tilde{\Delta}_1^{-1}]\tilde{Q}_1(x-k).\end{aligned} \tag{A.5}$$

We also have, for $x > k$,

$$\int_k^x B_\kappa\tilde{Q}_\kappa(x-y)dy = B_\kappa(I - \tilde{Q}_\kappa(x-k))\tilde{\Delta}_\kappa^{-1}, \quad \kappa = 1, 2. \tag{A.6}$$

Substituting in the RHS of Eq. (12) we get, for $x > k$,

$$F'(x) = \lambda F(x) - \lambda \int_k^x F(y)B_2\tilde{Q}_2(x-y)dy + \alpha_1\tilde{\Delta}_2^{-1} - \alpha_2\tilde{Q}_1(x-k) - \lambda F(k)B_2\tilde{Q}_2(x-k)\tilde{\Delta}_2^{-1}, \tag{A.7}$$

where

$$\begin{aligned}\alpha_1\tilde{\Delta}_2^{-1} &= F'(0) - \lambda F(k)(B_1\tilde{\Delta}_1^{-1} - B_2\tilde{\Delta}_2^{-1}) - \delta_{c-1}\lambda B_1\tilde{\Delta}_1^{-1},\\ \alpha_2 &= \lambda F(k)(I - B_1\tilde{\Delta}_1^{-1}) - (F'(k) - F'(0)) - \delta_{c-1}\lambda B_1\tilde{\Delta}_1^{-1}.\end{aligned}$$

Differentiating both sides of Eq. (A.7), we get

$$F''(x) = \lambda F'(x) - \lambda F(x)B_2 + \lambda \int_k^x F(y)B_2\tilde{Q}_2(x-y)\tilde{\Delta}_2 dy \alpha_2\tilde{Q}_1(x-k)\tilde{\Delta}_1 + \lambda F(k)B_2\tilde{Q}_2(x-k). \tag{A.8}$$

Using (A.7) we have

$$\lambda \int_k^x F(y)B_2\tilde{Q}_2(x-y)dy = -F'(x) + \lambda F(x) + \alpha_1\tilde{\Delta}_2^{-1} - \alpha_2\tilde{Q}_1(x-k) - \lambda F(k)B_2\tilde{Q}_2(x-k)\tilde{\Delta}_2^{-1}. \tag{A.9}$$

Substituting Eq. (A.9) in the RHS of Eq. (A.8) we get Eq. (14). This completes the proof. □

Proof of Corollary 1: Eq. (18) follows from the definition of F. Eq. (19) follows by taking the left and right limits at k in Eqs. (A.1) and (A.2), respectively. Eq. (20) follows by taking the left and right limits at k in Eqs. (11) and (12), respectively.

The balance equation for state $(0, i, c-1-i)$ yields

$$\begin{aligned}&(\lambda + i\mu_1 + (c-1-i)\mu_2)\pi(i, c-1-i)\\ &= F_i'(0) + \lambda\pi(i-1, c-1-i), \quad 0 \le i \le c-1.\end{aligned}$$

In matrix form this can be written as

$$F'(0) = \delta_{c-1}(\lambda I + \Delta_{c-1}) - \delta_{c-2}\lambda\hat{I},$$

which is Eq. (21). □

Lemma 3. *Let M be a $c \times c$ matrix with entries given by*

$$\begin{aligned}M(0, c-1) &= c\mu_2,\\ M(1, c-1) &= \mu_1,\\ M(i, i-1) &= \frac{i}{c-i}\frac{\mu_1}{\mu_2}, \quad 0 < i < c-1,\\ M(i, j) &= 0 \quad \text{for all other } (i, j).\end{aligned}$$

It is non-singular and satisfies the following equation

$$M(B_1 - \mu_1 I - \Delta_0) = (B_2 - \mu_2 I - \Delta_0). \tag{A.10}$$

Proof. For $0 \le i, j \le c-1$, we write

$$\begin{aligned}B_1(i, j) &= \{i, j\}(i+1)\mu_1 + \{i+1, j\}(c-i-1)\mu_2,\\ B_2(i, j) &= \{i, j\}(c-i)\mu_2 + \{i-1, j\}i\mu_1,\\ \Delta_0(i, j) &= \{i, j\}(i\mu_1 + (c-i-1)\mu_2),\\ M(i, j) &= \{i+c-1, j\}c\mu_2 + \{i+c-1, j+1\}i\mu_1\\ &\quad -\{i, j+1\}i\mu_1/((c-i)\mu_2),\end{aligned}$$

where $\{i, j\}$ is the Kronecker delta function, that is

$$\{i, j\} = \begin{cases}1 & \text{if } i = j,\\ 0 & \text{otherwise.}\end{cases} \tag{A.11}$$

Let $Y_\kappa = B_\kappa - \mu_\kappa I - \Delta_0$, $\kappa \in \{1, 2\}$. It follows that

$$\begin{aligned}Y_1(i, j) &= (\{i+1, j\} - \{i, j\})(c-i-1)\mu_2,\\ Y_2(i, j) &= (\{i-1, j\} - \{i, j\})i\mu_1.\end{aligned}$$

Then by matrix multiplication we have

$$\begin{aligned}&(MY_1)(i, j)\\ &= \sum_{k=0}^{c-1} M(i, k)Y_1(k, j)\\ &= \sum_{k=0}^{c-1}\Big(\{i+c-1, k\}c\mu_2 + \{i+c-1, k+1\}i\mu_1\Big)Y_1(k, j)\\ &\quad - \sum_{k=0}^{c-1}\{i, k+1\}\frac{i}{c-i}\frac{\mu_1}{\mu_2}Y_1(k, j)\\ &= \Big(\{i, 0\}c\mu_2 + \{i, 1\}\mu_1\Big)Y_1(c-1, j) - \frac{i}{c-i}\frac{\mu_1}{\mu_2}Y_1(i-1, j)\\ &= -\frac{i}{c-i}\frac{\mu_1}{\mu_2}Y_1(i-1, j) = (\{i-1, j\} - \{i, j\})i\mu_1 = Y_2(i, j).\end{aligned}$$

□

Proof of Lemma 1: Let us assume that the left equation in Eq. (33) holds, that is $\alpha_1 \cdot \tilde{\psi}_c = 0$. Then using the definition of α_1, given in Eq. (16), we have that

$$\alpha_0\tilde{\Delta}_1^{-1}\tilde{\Delta}_2\tilde{\psi}_c - \lambda F(k+)(B_1\tilde{\Delta}_1^{-1}\tilde{\Delta}_2 - B_2)\tilde{\psi}_c = 0$$

and we are going to show that

$$(B_1 - \tilde{\Delta}_1)\tilde{\Delta}_1^{-1}\tilde{\Delta}_2\tilde{\psi}_c = 0, \tag{A.12}$$

so that the column vector $\tilde{\Delta}_1^{-1}\tilde{\Delta}_2\tilde{\psi}_c$ is parallel to $\tilde{\phi}_*$, because it is a right eigenvector of the matrix $(B_1 - \tilde{\Delta}_1)$ corresponding to the null eigenvalue, whose multiplicity is one.

If Eq. (A.12) holds, we have that

$$(B_1-\tilde{\Delta}_1)\tilde{\Delta}_1^{-1}\tilde{\Delta}_2\tilde{\psi}_c = (B_1\tilde{\Delta}_1^{-1}\tilde{\Delta}_2-\tilde{\Delta}_2)\tilde{\psi}_c$$
$$= (B_1\tilde{\Delta}_1^{-1}\tilde{\Delta}_2-B_2)\tilde{\psi}_c = 0, \quad \text{(A.13)}$$

where in the last equality we used the relation $\tilde{\Delta}_2\tilde{\psi}_c = B_2\tilde{\psi}_c$ given by Eq. (32).

Eq. (A.13), together with Eq. (A.12), implies that $\alpha_0\tilde{\Delta}_1^{-1}\tilde{\Delta}_2\tilde{\psi}_c = 0$ and therefore it also implies the result.

To prove Eq. (A.12), we continue from Eq. (A.13) by rewriting it in the following way

$$\begin{aligned}(B_1\tilde{\Delta}_1^{-1}\tilde{\Delta}_2-B_2)\tilde{\psi}_c &= (B_1\tilde{\Delta}_1^{-1}B_2-B_2)\tilde{\psi}_c\\ &= (B_1\tilde{\Delta}_1^{-1}-I)B_2\tilde{\psi}_c\\ &= ((\mu_1 I+\Delta_{c-1})^{-1}B_1-I)B_2\tilde{\psi}_c\\ &= (\mu_1 I+\Delta_{c-1})^{-1}(B_1-\mu_1 I-\Delta_{c-1})B_2\tilde{\psi}_c\\ &= (\mu_1 I+\Delta_{c-1})^{-1}M^{-1}M(B_1-\mu_1 I-\Delta_{c-1})B_2\tilde{\psi}_c\\ &= (\mu_1 I+\Delta_{c-1})^{-1}M^{-1}(B_2-\mu_2 I-\Delta_{c-1})B_2\tilde{\psi}_c\\ &= (\mu_1 I+\Delta_{c-1})^{-1}M^{-1}B_2^{-1}(B_2-\tilde{\Delta}_2)\tilde{\psi}_c = 0.\end{aligned}$$

In the first equality we used Eq. (32), in the third one that $B_1\tilde{\Delta}_1^{-1} = (\mu_1 I+\Delta_{c-1})^{-1}B_1$, given by the definition in Eq. (9), in the sixth one we used the result of Lemma 3, where the matrix M is defined. Finally in the last two equalities we use again the definition in Eq. (9) and the hypothesis in Eq. (32). □

Proof of Theorem 3: We consider the two regions of x separately. First assume that $0 < x < k$. Eq. (13) is a non-homogeneous linear system of ordinary second order differential equations. Hence, we first try a homogeneous solution of the type

$$F_h(x) = e^{\theta x}\phi,$$

where ϕ is a row vector of length c. Substituting in $\mathcal{L}_1 F_h(x) = 0$ and cancelling $e^{\theta x}$, we get Eq. (22), with $2c$ solutions $\{(\theta_i,\phi_i), 0 \le i \le 2c-1\}$. The eigenvalues θ_i's are given in Eqs. (24) and (25). The homogeneous solution to Eq. (13) is then given by

$$F_h(x) = \sum_{i=0}^{2c-1} a_i e^{\theta_i x}\phi_i, \quad 0 \le x \le k, \quad \text{(A.14)}$$

where the $2c$ constants $\{a_i, 0 \le i \le 2c-1\}$ are to be determined.

For the particular solution we should look for a function of the following type

$$F_p(x) = \eta x + \zeta, \quad \text{(A.15)}$$

because $\theta_{c-1} = 0$ is an eigenvalue for the homogeneous solution. By substitution in Eq. (13) we get that the following equation has to be satisfied

$$-\eta(\lambda I-\tilde{\Delta}_1)+\eta x\lambda(B_1-\tilde{\Delta}_1)+\zeta\lambda(B_1-\tilde{\Delta}_1) = \alpha_0, \quad \text{(A.16)}$$

where the vector ζ can be chosen such that $\zeta\cdot\phi_* = 0$, because for all $a \in \mathbb{R}$,

$$(\zeta + a\,\phi_*)\lambda(B_1-\tilde{\Delta}_1) = \zeta\lambda(B_1-\tilde{\Delta}_1).$$

In addition, in order to have Eq. (A.16) satisfied for any x, the coefficient of the linear term should be null implying that $\eta = a\,\phi_*$.

Taking the scalar product of both sides of Eq. (A.16) by the right eigenvector $\tilde{\phi}_*$ satisfying Eq. (31), we get that

$$-a\,\phi_*(\lambda I-\tilde{\Delta}_1)\cdot\tilde{\phi}_* = \alpha_0\cdot\tilde{\phi}_*,$$

implying that

$$a = -\frac{\alpha_0\cdot\tilde{\phi}_*}{\phi_*(\lambda I-\tilde{\Delta}_1)\cdot\tilde{\phi}_*}. \quad \text{(A.17)}$$

Note that the linear term is missing if $\alpha_0\cdot\tilde{\phi}_* = 0$.

To derive the value of ζ we rewrite the equation $\zeta\cdot\phi_* = 0$ in matrix form as $\zeta\,\mathrm{diag}(\phi_*) = 0$. By adding this equation to Eq. (A.16) we get

$$-\eta(\lambda I-\tilde{\Delta}_1)+\zeta\,(\lambda(B_1-\tilde{\Delta}_1)+\mathrm{diag}(\phi_*)) = \alpha_0$$

and since $(\lambda(B_1-\tilde{\Delta}_1)+\mathrm{diag}(\phi_*))$ is not singular, we have that

$$\zeta = \alpha_0 M_0 + a\,\phi_*(\lambda I-\tilde{\Delta}_1)M_0,$$

where M_0 is as given in Eq. (35).

Finally, we have that the particular solution is equal to

$$F_p(x) = a\,\phi_*\big(xI+(\lambda I-\tilde{\Delta}_1)M_0\big)+\alpha_0 M_0, \quad 0 \le x \le k, \quad \text{(A.18)}$$

where a is defined as in Eq. (A.17). Below we show that $\alpha_0\cdot\tilde{\phi}_* = 0$, implying that $a = 0$ and thereby Eq. (34).

Next consider the region $x > k$, where F satisfies Eq. (14). It is also a non-homogeneous linear system of ordinary second order differential equations. As before we try a homogeneous solution of the type

$$F_h(x) = e^{\beta x}\psi,$$

where ψ is a row vector of length c. Substituting in $\mathcal{L}_2 F_h(x) = 0$ and cancelling $e^{\beta x}$ we get Eq. (27), which has $2c$ solutions (β_i,ψ_i), $0 \le i \le 2c-1$. The eigenvalues β's are given in Eqs. (28) and (29). The homogeneous solution to Eq. (14) is then given by

$$F_h(x) = \sum_{i=0}^{2c-1} b_i e^{\beta_i(x-k)}\psi_i, \quad x \ge k. \quad \text{(A.19)}$$

Since the solution has to be bounded we immediately get that $b_i = 0$ for $c < i \le 2c-1$, since the corresponding β_i's are strictly positive. Moreover, we have $\beta_c = 0$ and ψ_c is as given in Eq. (30). It follows that the homogeneous solution to Eq. (14) can be written as

$$F_h(x) = \sum_{i=0}^{c-1} b_i e^{\beta_i(x-k)}\psi_i + b_c\psi_c, \quad x \ge k, \quad \text{(A.20)}$$

where the $c+1$ constants $\{b_i, 0 \le i \le c\}$ are to be determined.

Next we determine the particular solution. Similarly to what we have done before for the interval $[0,k]$, the particular solution associated with the constant term α_1 in the right hand side of Eq. (14) would be of the form Eq. (A.15), since the associated homogeneous equation $\mathcal{L}_2 F_h(x) = 0$ admits the constant function as solution. However, in this case, the boundary condition, requiring $\lim_{x\to\infty} F(x)$ to be bounded, implies that the vector η is zero and therefore that $\alpha_1\cdot\tilde{\psi}_c = 0$. By applying Lemma 1, this also implies that the linear term in Eq. (A.18) is missing.

Eventually it follows that the particular solution in the region $x > k$ is given by

$$F_p(x) = \alpha_1 M_1 + \alpha_2\tilde{Q}_1(x-k)(\tilde{\Delta}_1-\tilde{\Delta}_2)M_2, \quad x \ge k, \quad \text{(A.21)}$$

as can be verified by direct substitution, where M_1 and M_2 are as given in Eqs. (37) and (38). The general solution is then as given in Eq. (36). This completes the proof. □

Proof of Theorem 5: According to the results of Corollary 1, we write $F(k)$ to mean $F(k-) = F(k+)$ and similarly for the derivative in k. By defining

$$H_1 = (U_1^+ - U_1^-)^{-1}(e^{U_1^+ k} - e^{U_1^- k}),$$
$$H_2 = M_0(I - e^{U_1^- k} + U_1^- H_1),$$
$$H_3 = H_1 + \tilde{\Delta}_1 H_2,$$
$$H_4 = -\lambda B_1 H_2,$$

we can rewrite Eq. (45) evaluated in k as

$$F(k) = F'(0)H_3 + \delta_{c-1}H_4. \quad \text{(A.22)}$$

By defining

$$H_5 = (U_1^+ - U_1^-)^{-1}(U_1^+ e^{U_1^+ k} - U_1^- e^{U_1^- k}),$$
$$H_6 = M_0(U_1^- e^{U_1^- k} - U_1^- H_5),$$

$$H_7 = H_5 - \tilde{\Delta}_1 H_6,$$
$$H_8 = \lambda B_1 H_6,$$

we can rewrite the derivative of Eq. (45) evaluated in k as

$$F'(k) = F'(0)H_7 + \delta_{c-1}H_8. \tag{A.23}$$

By defining

$$H_9 = (\tilde{\Delta}_1 - \tilde{\Delta}_2)M_2U_2^- + \tilde{\Delta}_1(\tilde{\Delta}_1 - \tilde{\Delta}_2)M_2,$$

we can rewrite the derivative of Eq. (46) evaluated in k as

$$F'(k) = F(k)U_2^- - b_c\psi_c U_2^- - \alpha_1 M_1 U_2^- - \alpha_2 H_9. \tag{A.24}$$

Then substituting the expression of α_2 in Eq. (17), by employing also Eqs. (16) and (15), and defining

$$H_{10} = U_2^- - \lambda(I - B_2\tilde{\Delta}_2^{-1})H_9,$$
$$H_{11} = M_1U_2^- + \tilde{\Delta}_2^{-1}H_9,$$
$$H_{12} = H_{10} + \lambda(B_1\tilde{\Delta}_1^{-1}\tilde{\Delta}_2 - B_2)H_{11},$$
$$H_{13} = \tilde{\Delta}_2 H_{11},$$
$$H_{14} = \lambda B_1\tilde{\Delta}_1^{-1}\tilde{\Delta}_2 H_{11},$$

we get

$$F'(k)(I - H_9) = F(k)H_{12} - b_c\psi_c U_2^- - F'(0)H_{13} + \delta_{c-1}H_{14}. \tag{A.25}$$

By equating Eqs. (A.23) and (A.25) and defining

$$H_{15} = U_2^-(H_7 - H_7H_9 - H_3H_{12} + H_{13})^{-1}, \tag{A.26}$$

$$H_{16} = (H_{14} + H_4H_{12} - H_8 + H_8H_9)(U_2^-)^{-1}H_{15}, \tag{A.27}$$

we get the first result in Eq. (47).

$$F'(0) = \delta_{c-1}H_{16} - b_c\psi_c H_{15} \tag{A.28}$$

Taking the limit in Eq. (46) and defining

$$H_{17} = \tilde{\Delta}_2 M_1 - \lambda H_3(B_1\tilde{\Delta}_1^{-1}\tilde{\Delta}_2 - B_2)M_1,$$
$$H_{18} = \lambda B_1\tilde{\Delta}_1^{-1}\tilde{\Delta}_2 M_1 + \lambda H_4(B_1\tilde{\Delta}_1^{-1}\tilde{\Delta}_2 - B_2)M_1,$$
$$H_{19} = I - H_{15}H_{17}, \tag{A.29}$$

$$H_{20} = H_{16}H_{17} - H_{18}, \tag{A.30}$$

we get the second result in Eq. (48). □

Proof of Theorem 6: We would need the definition of the following rectangular matrices, for $0 \le n \le c-1$:

$$\hat{B}_n(i,i) = (n - i + 1)\mu_2, \quad 0 \le i \le n,$$
$$\hat{B}_n(i,i-1) = i\mu_1, \quad 1 \le i \le n+1,$$
$$\hat{B}_n(i,j) = 0 \quad \text{for all other } (i,j), \quad 0 \le j \le n.$$

Using Eq. (21) together with the balance Eq. (7) and the normalization equation, we finally get

$$\lambda\delta_n = \delta_{n+1}\hat{B}_0, \qquad n = 0, \tag{A.31}$$

$$\delta_n(\lambda I + \Delta_n) = \delta_{n-1}\lambda\hat{I} + \delta_{n+1}\hat{B}_n, \quad 0 < n < c-1, \tag{A.32}$$

$$\delta_n(\lambda I + \Delta_n) = \delta_{n-1}\lambda\hat{I} + F'(0), \qquad n = c-1, \tag{A.33}$$

$$1 = F(\infty)\,\mathbb{1} + \sum_{0\le n\le c-1}\delta_n\,\mathbb{1}, \tag{A.34}$$

where $F'(0)$ and $F(\infty)$ are given in Eqs. (47) and (48), respectively. This system has $1 + (c+1)c/2$ equations and an equal number of unknowns.

Writing $\delta_n = \delta_{n+1}\hat{C}_n$, we have, by Eq. (A.31), $\hat{C}_0 = \hat{B}_0/\lambda$ and by Eq. (A.32), $\hat{C}_n = \hat{B}_n(\lambda(I - \hat{C}_{n-1}\hat{I}) + \Delta_n)^{-1}$, $0 < n < c-1$. By Eq. (A.33) we have $\delta_{c-1} = b_c\psi_c\hat{C}_{c-1}$ with $\hat{C}_{c-1} = -H_{15}(\lambda(I - \hat{C}_{c-2}\hat{I}) + \Delta_{c-1} - H_{16})^{-1}$.

By defining

$$\hat{H}_{c-1} = \hat{C}_{c-1}, \quad \hat{H}_n = \hat{H}_{n+1}\hat{C}_n, \quad 0 < n < c-1, \tag{A.35}$$

and using the normalization constraint Eq. (A.34) we get the result. □

Proof of Lemma 2: By taking derivatives of Eqs. (45) and (46) we can compute the VQT density function as follows

$$F'(x) = (F'(0) + \alpha_0 M_0 U_1^-)(U_1^+ - U_1^-)^{-1}(U_1^+ e^{U_1^+ x} - U_1^- e^{U_1^- x}) - \alpha_0 M_0 U_1^- e^{U_1^- x}, \quad 0 \le x \le k, \tag{A.36}$$

$$F'(x) = (F(k) - b_c\psi_c - \alpha_1 M_1 - \alpha_2(\tilde{\Delta}_1 - \tilde{\Delta}_2)M_2)U_2^- e^{U_2^-(x-k)} - \alpha_2\tilde{\Delta}_1\tilde{Q}_1(x-k)(\tilde{\Delta}_1 - \tilde{\Delta}_2)M_2, \quad x \ge k. \tag{A.37}$$

We define the following matrix function

$$I(a,b;D) = \int_a^b Dx\, e^{Dx}dx = (be^{Db} - ae^{Da}) - D^{-1}(e^{Db} - e^{Da}), \tag{A.38}$$

that is well defined on the set of non-singular matrices and that can be defined on the set of singular matrices by continuity. That is, if $\det(D) = 0$, we set $I(a,b;D) = \lim_{t\to 0} I(a,b;D+tI)$.

Integrating the expressions in Eqs. (A.36) and (A.37) in their corresponding domains multiplied by x, we obtain the result in Eq. (51) after summing up all components. □

References

Adan, I., Hathaway, B., & Kulkarni, V. G. (2019). On first-come, first-served queues with two classes of impatient customers. *Queueing Systems, 91*(1), 113–142.

Anick, D., Mitra, D., & Sondhi, M. M. (1982). Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal, 61*(8), 1871–1894.

Batt, R. J., & Terwiesch, C. (2012) Doctors under load: An empirical study of state-dependent service times in emergency care (pp. 1–32) Wharton School of Business. https://faculty.wharton.upenn.edu/wp-content/uploads/2012/11/DULnew_v6.pdf.

Bekker, R. (2009). Queues with Lévy input and hysteretic control. *Queueing Systems, 63*(1-4), 281–299.

Bekker, R., Boxma, O. J., & Resing, J. A. C. (2009). Lévy processes with adaptable exponent. *Advances in Applied Probability, 41*(1), 177–205.

Boxma, O. J., Perry, D., & Stadje, W. (2001). Clearing models for M/G/1 queues. *Queueing Systems, 38*(3), 287–306.

Boxma, O. J., & Vlasiou, M. (2007). On queues with service and interarrival times depending on waiting times. *Queueing Systems, 56*(3-4), 121–132.

Brill, P., & Posner, M. (1981). A two server queue with nonwaiting customers receiving specialized service. *Management Science, 27*(8), 914–925.

Carmon, Z., Shanthikumar, J. G., & Carmon, T. F. (1995). A psychological perspective on service segmentation models: The significance of accounting for consumers perceptions of waiting and service. *Management Science, 41*(11), 1806–1815.

Chalfin, D. B., Trzeciak, S., Likourezos, A., Baumann, B. M., & Dellinger, R. P. (2007). Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine, 35*(6), 1477–1483.

Chan, C. W., Farias, V. F., & Escobar, G. J. (2017). The impact of delays on service times in the intensive care unit. *Management Science, 63*(7), 2049–2072.

Chan, P. S., Krumholz, H. M., Nichol, G., Nallamothu, B. K., & American Heart Association National Registry of Cardiopulmonary Resuscitation Investigators (2008). Delayed time to defibrillation after in-hospital cardiac arrest. *New England Journal of Medicine, 358*(1), 9–17.

D'Auria, B. (2021a). App for plots of virtual queueing time. https://brdauria.github.io/VQTPlot/.

D'Auria, B. (2021b). Repository for plots of virtual queueing time. https://github.com/brdauria/VQTPlot.git.

Delasay, M., Ingolfsson, A., Kolfal, B., & Schultz, K. (2019). Load effect on service times. *European Journal of Operational Research, 279*(3), 673–686.

Do, H. T., Shunko, M., Lucas, M. T., & Novak, D. C. (2018). Impact of behavioral factors on performance of multi-server queueing systems. *Production and Operations Management, 27*(8), 1553–1573.

Dong, J., Feldman, P., & Yom-Tov, G. B. (2015). Service systems with slowdowns: Potential failures and proposed solutions. *Operations Research, 63*(2), 305–324.

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

Gaver, D., Miller, R., et al. (1962). Limiting distributions for some storage problems. In K. Arrow, et al. (Eds.), *Studies in applied probability and management science* (pp. 110–126). Stanford Univ. Press.

Kulkarni, V. G. (1997). Fluid models for single buffer systems. In J. Dshalalow (Ed.), *Frontiers in queueing: models and applications in science and engineering* (pp. 321–338). CRC Press.

Maister, D. H., et al. (1984). *The psychology of waiting lines*. Citeseer.

Malhotra, R., Mandjes, M., Scheinhardt, W. R., & Van Den Berg, J. (2009). A feedback fluid queue with two congestion control thresholds. *Mathematical Methods of Operations Research, 70*(1), 149–169.
Palmowski, Z., & Vlasiou, M. (2011). A Lévy input model with additional state-dependent services. *Stochastic Processes and their Applications, 121*(7), 1546–1564.
Posner, M. (1973). Single-server queues with service time dependent on waiting time. *Operations Research, 21*(2), 610–616.
Ramaswami, V., & Lucantoni, D. M. (1985). Algorithms for the multi-server queue with phase type service. *Stochastic Models, 1*(3), 393–417.
Renaud, B., Santin, A., Coma, E., Camus, N., Van Pelt, D., Hayon, J., … Fine, M. J., et al. (2009). Association between timing of intensive care unit admission and outcomes for emergency department patients with community-acquired pneumonia. *Critical Care Medicine, 37*(11), 2867–2874.
Richardson, D. B. (2002). The access-block effect: Relationship between delay to reaching an inpatient bed and inpatient length of stay. *Medical Journal of Australia, 177*(9), 492–495.
Scheinhardt, W., Van Foreest, N., & Mandjes, M. (2005). Continuous feedback fluid queues. *Operations Research Letters, 33*(6), 551–559.
Selen, J., Adan, I. J., Kulkarni, V. G., & van Leeuwaarden, J. S. (2016). The snowball effect of customer slowdown in critical many-server systems. *Stochastic Models, 32*(3), 366–391.
Siegmeth, A., Gurusamy, K., & Parker, M. (2005). Delay to surgery prolongs hospital stay in patients with fractures of the proximal femur. *The Journal of Bone and Joint Surgery, British volume, 87*(8), 1123–1126.
da Silva Soares, A., & Latouche, G. (2009). Fluid queues with level dependent evolution. *European Journal of Operational Research, 196*(3), 1041–1048.
Soltani, M., Batt, R., Bavafa, H., & Patterson, B. (2019). Does what happens in the ED stay in the ED? The effects of emergency department physician workload on post-ED care use.
Ülkü, S., Hydock, C., & Cui, S. (2020). Making the wait worthwhile: Experiments on the effect of queueing on consumption. *Management Science, 66*(3), 1149–1171.
Whitt, W. (1990). Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems, 6*(1), 335–351.
Wu, C. A., Bassamboo, A., & Perry, O. (2019a). Service system with dependent service and patience times. *Management Science, 65*(3), 1151–1172.
Wu, C. A., Bassamboo, A., & Perry, O. (2019b). When service times depend on customers' delays: A solution to two empirical challenges.