Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2025, 19(12), 464–477
https://doi.org/10.12913/22998624/210352
ISSN 2299-8624, License CC-BY 4.0

Received: 2025.07.04
Accepted: 2025.09.05
Published: 2025.11.01

Modeling and simulation of heterogeneous multiserver queues with impatient customers

Adam Gregosiewicz[1*], Elżbieta Ratajczyk[1], Łukasz Stępień[1]

[1] Lublin University of Technology, Nadbystrzycka 38A, 20-618 Lublin, Poland
* Corresponding author's e-mail: a.gregosiewicz@pollub.pl

ABSTRACT

The study investigates the dynamics of heterogeneous multiserver queueing systems, where servers operate at different service rates-a scenario often encountered in practice but not captured by classical models. In traditional models, servers are assumed to be homogeneous, serving customers at the same rate, and customers are expected to select servers at random when multiple options are available. However, these assumptions often fail in real-world systems. This research develops a mathematical framework to model and analyze queues where servers have different service speeds. In such systems, a customer at the head of the queue may strategically choose to wait for a faster server, even when slower servers are idle, effectively blocking the queue. This decision-making behavior can improve system efficiency and preserve the "first-come, first-served" principle. However, customers may lose patience over time and opt to be served by the slow server, adding further complexity to the system. In this article, a two-server model that incorporates server preferences and customer hesitation is proposed, and Kolmogorov's forward equations for the corresponding transition probability functions are derived. Simulation experiments are used to illustrate behaviour of the system under various parameter settings. The impact of server heterogeneity and customer decision-making on overall system efficiency is explored.

Keywords: heterogeneus queue, Markov process.

INTRODUCTION

Queueing theory has long been recognized as a powerful mathematical tool for analyzing and optimizing service systems. Originally developed to improve the efficiency of telephone operations, it has since been applied across diverse domains. Recently, in [1] queueing models were used to simulate and analyze fog-computing architectures in order to guide the design of systems that meet Quality of Service requirements; in telecommunications, they are used to analyze complex network architectures [2]. They can also reduce the waiting time of patients in healthcare service [3]. In [4], queueing theory is applied to port operations, modeling ship arrivals, waiting times, and loading processes. Moreover, [5] uses queue-based analysis to evaluate how replica selection algorithms in distributed systems are affected by queue delays. Building on this foundation, a queueing model with heterogeneous servers and strategic customer behavior is explored – conditions often encountered in real systems but rarely captured in classical models (compare to [6]). In the studies of multiserver queueing systems it is customarily assumed that servers are homogeneous, that is, that they serve customers at the same rate (as in [7], where service times are i.i.d random variables, or in [8], where servers are identical and independent of each other). However, in real life queues, this assumption is, more often than not, violated, especially in queueing systems with human servers, but also in systems that are automatic in nature. Nevertheless, as recently observed in [9], modeling the individual characteristics of servers has received little attention. Since it is common to observe servers providing service to identical jobs at different rates, a mathematical study of heterogeneous queueing scenarios is warranted. More specifically, a

two-server system in which service rates differ substantially and customers can choose which server to use is investigated.

In classical models, customers arriving when multiple servers are available typically choose one at random (see [10] for a discussion of queue disciplines in such cases). However, if customers are aware of differences in service rates, they may prefer the faster server. This preference could result in a situation where later-arriving customers, served by the faster server, exit the system before earlier arrivals who chose the slower server. In such cases, if only the slower server is idle, a customer may opt to wait for the faster server rather than accept immediate service – thus deviating from the assumptions of classical queueing disciplines. Such behavior is rational when the performance gap between servers is significant, and may even justify maintaining a first-come, first-served (FCFS) policy in certain scenarios.

Generally, customers involved in queues can exhibit different types of behavior, including balking (a decision not to join the queue if it is too long), reneging (a decision to leave the queue if service waiting time is too long) or jockeying (switching between queues in the hope of receiving the service quickly). Moreover, many queuing situations with discouragement are encountered in real life in which customers get hesitating or impatient. Thus there are numerous cases in which those awaiting service may be allowed to make choices which affect the time spent in the system, as in [11], where a customer who, upon arrival, finds at least two customers in the system may decide not to enter the queue.

In this study, the complex nature of both servers and customers is taken into account, and a specific two-server queueing model for heterogeneous servers is derived that incorporates important behavioral factors. Specifically, the scenario is considered in which a customer at the head of the queue, noticing that a slow server is available, chooses to wait for the faster server to become free, thereby effectively blocking the queue. However, the same customer may eventually lose patience and, after some time, decide to be served by the slow server. Thus model captures this nuanced decision-making process and contributes to a more realistic understanding of queue dynamics in systems with heterogeneous servers.

A key challenge and innovation in formulating our model lies in the careful and appropriate definition of the state space. This step is crucial, as stochastic processes in queueing theory are typically described by N(t), the number of customers in the system, which often results in non-Markovian behavior. To address this, the method of supplementary variables is used, embedding the non-Markovian process into an augmented Markov process – a technique introduced in [12] and applied to single-server systems in [13]. The main idea is to introduce the notion of *configuration*, which extends the state to include indices of active clocks, grouped into an appropriately defined pair. This ensures that the future evolution of the process depends solely on its current state. It is worth emphasizing that our model also provides a foundation for modeling more complex server and customer behaviors. Next, the Kolmogorov forward equations governing the probability densities of the underlying process are derived. This approach is consistent with recent work [14] on transient behavior in multichannel queueing systems, where analytical solutions of Kolmogorov equations have been used to study time-dependent probabilities.

In the third section, the analytical developments are complemented with an extensive computational study. Beyond the classical performance indicators as average queue length, mean sojourn time, and busy-period distributions, fairness measures are also examined such as the proportion of first-come-first-served (FCFS) violations and the utilisation gap between servers. Exploring a wide swath of parameter regimes, behavioural patterns are revealed like impatience-induced load balancing and congestion tipping points, that cannot be replicated by homogeneous-server or non-strategic customer models. This illustrates both the practical relevance and the methodological novelty of our framework.

This study aims to formulate and analyze a two-server queueing model with heterogeneous servers that accounts for customers' strategic waiting, impatience, and switching behavior. Relative to classical heterogeneous-server models, our framework introduces a new state-space construction embedding strategic hesitancy through supplementary clock configurations. This approach allows embedding non-Markovian dynamics into an augmented Markov process, enabling tractable analysis.

MODEL WITH TWO SERVERS

Queue's discipline and the state-space of the related Markov process

A queueing system with two servers X and Y is considered, where Y gives a significantly faster service than X, and customers are aware of this difference. The resulting queue discipline is as follows:

- one customer can be served at a given time in one service point and unserved customers line up in order of arrival keeping in a single queue;
- the customer at the head of the queue is moving forward as soon as the server Y becomes vacant;
- if service point X becomes available, the customer ignores this fact, as he prefers to wait for service at Y,
- as time passes, the customer becomes impatient and may choose to be served by the slow server X anyway.

More precisely, our model is a modification of the standard M/G/2 queue: customers are assumed to arrive according to a Poisson process with rate α, and service times are independently distributed random variables, say, T_1 and T_2, with probability density functions α_1 and α_2 for the servers X and Y, respectively. Moreover, a random variable T_3 is introduced, describing impatience of a customer waiting at the head of the queue when the fast server Y is occupied and the slow server X is free. The probability density function of T_3 is denoted α_3. Functions λ_i defined by

$$\lambda_i(x) = \frac{a_i(x)}{\int_x^\infty a_i(r)\mathrm{d}r}, \qquad x \geq 0,\ i = 1,2,3 \tag{1}$$

formula (1) are thus hazard rate functions for service times at the servers (i = 1,2) and patience time of the customer waiting (i = 3).

Provided that Y is still not free, at T_3 the customer at the head of the queue loses his patience and decides to be served by X.

The problem of choosing a convenient state-space $\mathcal{S}$ for a process describing a queue with this discipline is not as simple as it may appear. In particular, we want $\mathcal{S}$ to contain enough information so that the process gains Markovian nature. To solve the problem, we extend the main idea of Cox [12], and include in $\mathcal{S}$ time-type variables. In the models with single queue and one server one such variable is used and it represents the time that has passed since the last customer started to be served (as it was done in [15]). It might seem that in the case of a number of servers, one should simply deal with several variables, each related to a different server. As it turns out, however, it is better to think in the terms of the time, in what follows denoted by u, that has passed since the last time the system underwent a jump-like change (see below for more details).

The process is determined by the following factors.

- The *counter* that tells us the total number of customers present in the system: including those waiting in the queue and those currently being served on X and Y.
- Three *regular clocks*, say, $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$.
  - $\mathcal{C}_1$ tells us how long ago the last service at X started,
  - $\mathcal{C}_2$ – how long ago the last service at Y started, and
  - $\mathcal{C}_3$ – how long has the hesitating customer been blocking the queue.
- The *special clock*, denoted $\mathcal{C}_0$, that tells us how long ago the last customer arrived in the system.

At a given time not all regular clocks are *active*; for example, when both servers are occupied, $\mathcal{C}_3$ is 'switched off' or, if there is a hesitating customer at the head of the queue, $\mathcal{C}_1$ is inactive. The shortest time shown on the regular active clocks will be denoted by the variable u. Hence, the set indices of active clocks at a given time naturally splits into a pair (K, L) where K is the set of indices of clocks displaying time u, and the set L collects the indices of the remaining active clocks. Of course, not all pairs (K, L) are permissible; for instance, clocks no. 1 and no. 3 are never switched on simultaneously, and no. 3 can be switched only when no. 2. Additionally, we want to keep track of the number of customers in the system. Thus, it seems natural to include $\mathbb{N}_0$ as part of the state space. However, each pair of active clocks corresponds to a specific number of customers, $n \geq 0$. For example, $(\emptyset, \emptyset)$ pairs only with 0, while $(\{1\}, \emptyset)$ and $(\{2\}, \emptyset)$ pair merely with 1. These reasonable triplets

$$E := (K, L, n)$$

will be referred to as *configurations*. The set of all configurations is

$$\begin{aligned}\mathcal{E} := \big\{&(\emptyset,\emptyset,0),(\{1\},\emptyset,1),(\{2\},\emptyset,1),\\ &(\{1\},\{2\},n),(\{2\},\{1\},n),\\ &(\{3\},\{2\},n),(\{2,3\},\emptyset,n),\ n \geq 2\big\}.\end{aligned} \tag{2}$$

Finally, since besides configurations of clocks we want to keep the record of the times shown on these clocks, we define the state-space as

$$\mathcal{S} := \{(\emptyset,\emptyset,0),(\{1\},\emptyset,1,u),(\{2\},\emptyset,1,u),$$
$$(\{1\},\{2\},n,u,x_2),(\{2\},\{1\},n,u,x_1),(3)$$
$$(\{3\},\{2\},n,u,x_2),(\{2,3\},\emptyset,n,u),$$
$$n \geq 2, u \geq 0, x_1, x_2 > 0\}.$$

In principle, the coordinates of a point $(K,L,n,u,x_l) \in \mathcal{S}$ have the following interpretation:

- $K \cup L$ is the set of indices of active clocks,
- n is the number of customers in the system,
- u is the time shown on the clocks $\mathcal{C}_i$ for $i \in K$,
- the time shown on the clock $\mathcal{C}_l$ for $l \in L$ is $u + x_l$.

More specifically, when both servers are idle and the queue is empty, the process is in the state $(\emptyset,\emptyset,0)$. Since $K = \emptyset$, there is no need for a time-type coordinate in this case. The system is in the state $(\{1\},\emptyset,1,u)$ with $u \geq 0$ if there is only one customer, and he is at X with a service time of u. If the only customer is at Y and his service time is u, then the corresponding state is $(\{2\},\emptyset,1,u)$.

Next, the system is in a state $(\{1\},\{2\},n,u,x_2)$ with $n \geq 2, u \geq 0$ and $x_2 > 0$ if there are n customers, the one being served at X has spent the time u there, and another one has been served at Y for the time $u + x_2$, see Figure 1 (left); $x_2 > 0$ is interpreted as the time server Y had already been busy at the moment when the customer started to be served at X. An analogous description holds for the state $(\{2\},\{1\},n,u,x_1)$ with $n \geq 2$, as shown in Figure 1 (middle).

A state $(\{3\},\{2\},n,u,x_2)$ describes the following situation: there is a customer at Y, who has been served for the time $u + x_2$, server X is free, and the hesitating customer at the head of the queue has been waiting for the time u, see Figure 1 (right). If the service time at Y equals the waiting time of the hesitating customer, the process is in the state $(\{2,3\},\emptyset,n,u)$. We note that the states of the form $(\{1,2\},\emptyset,n,u,0)$, $n \geq 2$, are not considered as proper elements of $\mathcal{S}$ because they are never attained with probability 1. In contrast, there is a scenario that leads to the state $(\{2,3\},\emptyset,n,u)$, as shown in line 13 of Table 1.

Description of the process

The resulting stochastic process in $\mathcal{S}$ is a particular example of a piece-wise deterministic process of M.H.A. Davis [16] (see also the more recent [17]), with the following characteristics of its deterministic and random parts. The deterministic part describes what happens between jumps: if started at a (E,u,x), $E \in \mathcal{E}, u \geq 0, x > 0$, the process will be at $(E,u+t,x)$ after time t, provided that no service was completed and no customer arrived in the meantime.

This deterministic motion is interrupted by jumps between copies, and these come in two types. First of all, there are jumps caused by regular clocks. As Table 1 shows, the copy to which the process jumps depends on both the clock that goes off and the current configuration. Since there are $|K \cup L|$ active clocks in a configuration (K,L,n), there are as many possible jumps from the copy indexed by this configuration. A change of configuration is just switching of some of the clocks and switching on some others. It is worth noting that clocks that were switched on before the jump can still be active but reset to zero (as exemplified by $\mathcal{C}_2$ of line 6 or $\mathcal{C}_2$ and $\mathcal{C}_3$ of line 13 in Table 1), or left intact (as is the case, e.g., with $\mathcal{C}_2$ in line 2, and $\mathcal{C}_1$ of line 3 of Figure 1).

Secondly, there are also jumps caused by the special counter $\mathcal{C}_0$; these are described in Table 2. We observe that $\mathcal{C}_0$ always increases the third

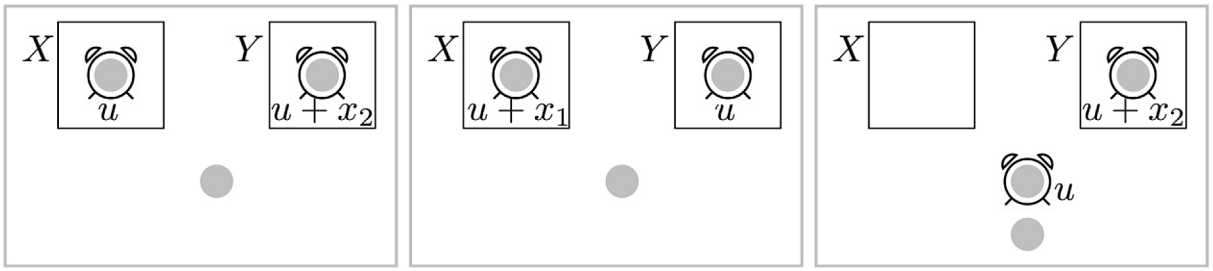

Figure 1. Schematic diagram of the system in states: $(\{1\},\{2\},3,u,x_2)$ on the left, $(\{2\},\{1\},3,u,x_1)$ in the middle and $(\{3\},\{2\},3,u,x_2)$ on the right

Table 1. Jumps caused by regular clocks

No.	Jump from	Jump to
1.	$\mathcal{C}_1(\{1\},\emptyset,1,u),\ \mathcal{C}_2(\{2\},\emptyset,1,u)$	$(\emptyset,\emptyset,0)$
2.	$\mathcal{C}_1(\{1\},\{2\},2,u,x_2)$	$(\{2\},\emptyset,1,u+x_2)$
3.	$\mathcal{C}_2(\{1\},\{2\},2,u,x_2)$	$(\{1\},\emptyset,1,u)$
4.	$\mathcal{C}_1(\{2\},\{1\},2,u,x_1)$	$(\{2\},\emptyset,1,u)$
5.	$\mathcal{C}_2(\{2\},\{1\},2,u,x_1)$	$(\{1\},\emptyset,1,u+x_1)$
6.	$\mathcal{C}_2(\{2,3\},\emptyset,2,u),\ \mathcal{C}_2(\{3\},\{2\},2,u,x_2)$	$(\{2\},\emptyset,1,0)$
7.	$\mathcal{C}_3(\{2,3\},\emptyset,n,u), n\geq 2$	$(\{1\},\{2\},n,0,u)$
8.	$\mathcal{C}_3(\{3\},\{2\},n,u,x_2), n\geq 2$	$(\{1\},\{2\},n,0,u+x_2)$
9.	$\mathcal{C}_1(\{1\},\{2\},n,u,x_2), n\geq 3$	$(\{3\},\{2\},n-1,0,u+x_2)$
10.	$\mathcal{C}_2(\{1\},\{2\},n,u,x_2), n\geq 3$	$(\{2\},\{1\},n-1,0,u)$
11.	$\mathcal{C}_1(\{2\},\{1\},n,u,x_1), n\geq 3$	$(\{3\},\{2\},n-1,0,u)$
12.	$\mathcal{C}_2(\{2\},\{1\},n,u,x_1), n\geq 3$	$(\{2\},\{1\},n-1,0,u+x_1)$
13.	$\mathcal{C}_2(\{2,3\},\emptyset,n,u),\ \mathcal{C}_2(\{3\},\{2\},n,u,x_2), n\geq 3$	$(\{2,3\},\emptyset,n-1,0)$

Note: Jumps are determined by configurations and indices of clocks going off.

Table 2. Jumps caused by the special clock $\mathcal{C}_0$ that signals the arrival of a new customer

No.	Jump from	Jump to
1.	$(\emptyset,\emptyset,0,u)$	$(\{2\},\emptyset,1,0)$
2.	$(\{1\},\emptyset,1,u)$	$(\{2\},\{1\},2,0,u)$
3.	$(\{2\},\emptyset,1,u)$	$(\{3\},\{2\},2,0,u)$
4.	$(\{1\},\{2\},n,u,x_2), n\geq 2$	$(\{1\},\{2\},n+1,u,x_2)$
5.	$(\{2\},\{1\},n,u,x_1), n\geq 2$	$(\{2\},\{1\},n+1,u,x_1)$
6.	$(\{2,3\},\emptyset,n,u), n\geq 2$	$(\{2,3\},\emptyset,n+1,u)$
7.	$(\{3\},\{2\},n,u,x_2), n\geq 2$	$(\{3\},\{2\},n+1,u,x_2)$

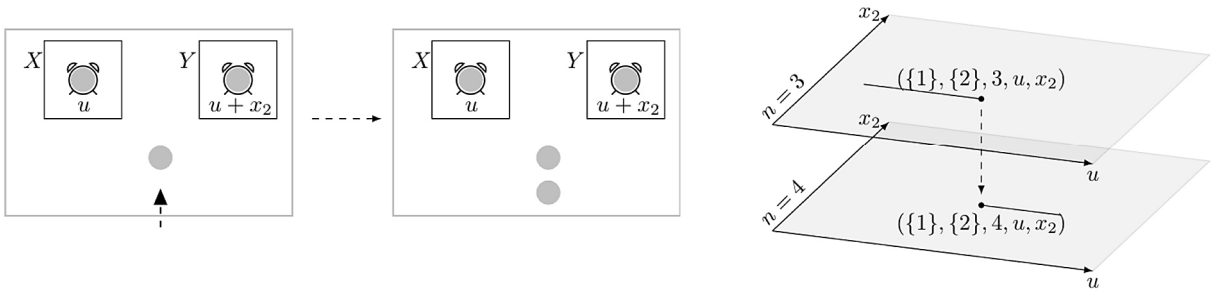

Figure 2. Initially, there are $n=3$ customers, both servers are occupied and service time at Y is greater than at X. Arrival of a new customer results in a process' jump from one copy of the first quadrant in $\mathcal{S}$ to another

coordinate, that is, n, by 1. On the other hand, the regular clocks $\mathcal{C}_1$ and $\mathcal{C}_2$ decrease n by 1, whereas $\mathcal{C}_3$ leaves it intact. Contrary to the previous case, clocks that were switched on before the jump are still active and its time is left intact.

For example, an arrival of a new customer transfers the process from $(\emptyset,\emptyset,0)$ to $(2,\emptyset,1,0)$; that is, the customer immediately chooses server Y and its service time starts from 0. (Customers are aware of the difference in service time.)

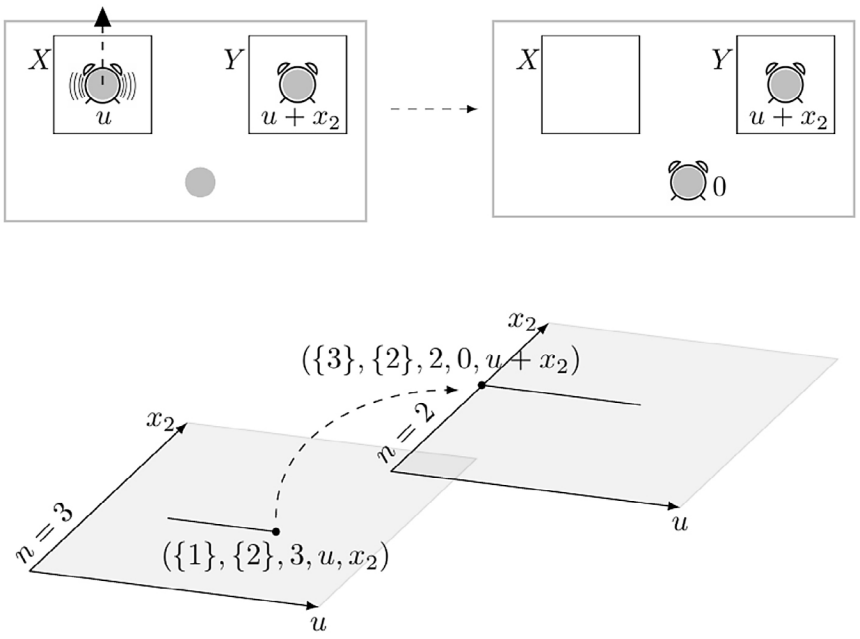

Figure 3. In the situation described in Figure 2, completion of service at X causes the process to jump to the boundary

Figures 2–4 show where the process can jump from a point $(\{1\},\{2\},n,u,x_2)$ with $n \geq 3$ depending on which of the following events takes place first: (i) arrival of a new customer, (ii) completion of service at X, and (iii) completion of service at Y. In the case (i), the process starts anew from $(\{1\},\{2\},n+1,u,x_2)$; an arriving customer takes the place at the end of the queue. In the case (ii), the process recommences at $(\{3\},\{2\},n-1,0,u+x_2)$; the customer at the head of the queue chooses to ignore the fact that X is free and waits for service at Y. In the case (iii), the process starts anew at $(\{2\},\{1\},n-1,0,u)$; the customer at the head of the queue begins to be served at Y, so both servers are again occupied, but now the shortest time is shown on the second clock.

Derivation of Kolmogorov equations

Let $p_{E_0}(t)$ be the probability that the process at time t is in the state

$$E_0 := (\emptyset,\emptyset,0) \tag{4}$$

corresponding to the empty queue, and denote

$$E_1^1 := (\{1\},\emptyset,1), \qquad E_1^2 := (\{2\},\emptyset,1). \tag{5}$$

Let $p_{E_1^1}(t,\cdot)$ be the probability density function of the process in $\{E_1^1\}\times\mathbb{R}_+$ and let $p_{E_1^2}(t,\cdot)$ be such a function for the process in $\{E_1^2\}\times\mathbb{R}_+$ at this time. Kolmogorov equations for these functions are derived in the usual manner, that is, by specifying the changes that can occur in a small time interval Δt. The system is empty at time t if there was no customer present in it at time $t-\Delta t$ and nobody arrived in the meantime or there was one customer and its service just ended. The total probability mass transfer from E_1^1 to E_0 at time t is $\int_{\mathbb{R}_+}\lambda_1(u)p_{E_1^1}(t,u)\mathrm{d}u$. Combining this with a similar quantity related to E_1^2, we obtain the equation

$$\partial_t p_{E_0}(t) = -\alpha p_{E_0}(t) + \int_{\mathbb{R}_+}\lambda_1(u)p_{E_1^1}(t,u)u + \\ + \int_{\mathbb{R}_+}\lambda_2(u)p_{E_1^2}(t,u)\mathrm{d}u. \tag{6}$$

Next, let

$$E_n^1 := (\{1\},\{2\},n), \\ E_n^2 := (\{2\},\{1\},n), \qquad n \geq 2. \tag{7}$$

Accordingly, let $p_{E_n^1}(t,\cdot,\cdot)$, $p_{E_n^2}(t,\cdot,\cdot)$ be the corresponding probability density functions. The rules listed in Tables 1 and 2 translate into relations

$$\partial_t p_{E_1^1}(t,u) = -\partial_u p_{E_1^1}(t,u) - (\alpha+\lambda_1(u))p_{E_1^1}(t,u) \\ + \int_{\mathbb{R}_+}\lambda_2(u+x_2)p_{E_2^1}(t,u,x_2)\mathrm{d}u + \int_0^u \lambda_2(v)p_{E_2^2}(t,v,u-v)\mathrm{d}v \tag{8}$$

$$\partial_t p_{E_1^2}(t,u) = -\partial_u p_{E_1^2}(t,u) - (\alpha+\lambda_2(u))p_{E_1^2}(t,u) \\ + \int_{\mathbb{R}_+}\lambda_1(u+x_1)p_{E_2^2}(t,u,x_1)\mathrm{d}u + \int_0^u \lambda_1(v)p_{E_2^1}(t,v,u-v)\mathrm{d}v \tag{9}$$

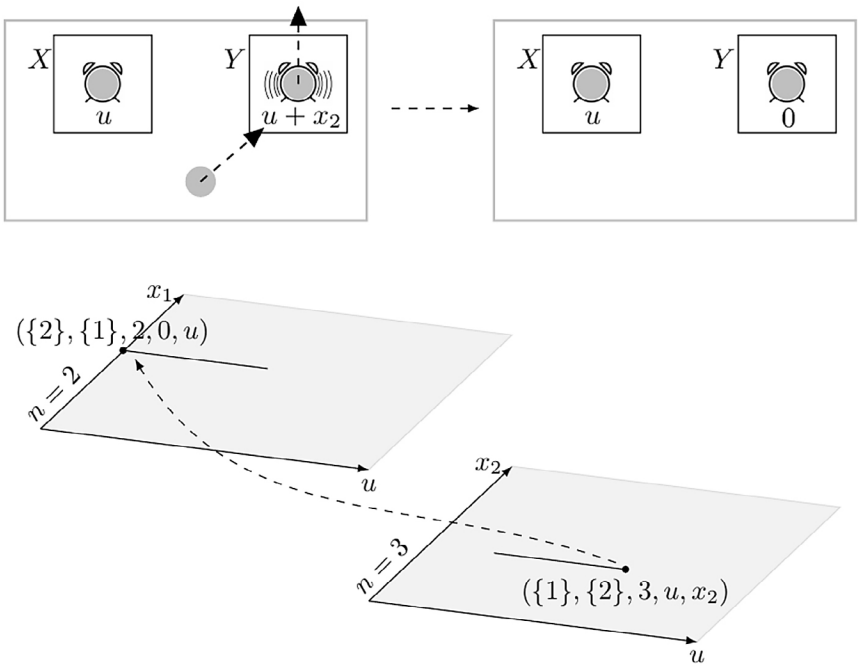

Figure 4. In the situation described in Figure 2, completion of service at Y causes the process to jump from the interior of one copy of the first quadrant to the boundary of another copy of this quadrant

and

$$\partial_t p_{E_n^1}(t,u,x_2) = -\partial_u p_{E_n^1}(t,u,x_2) - \\ -(\alpha+\lambda_1(u)+\lambda_2(u+x_2))p_{E_n^1}(t,u,x_2) \\ +\alpha p_{E_{n-1}^1}(t,u,x_2)[n\geq 3] \tag{10}$$

$$\partial_t p_{E_n^2}(t,u,x_1) = -\partial_u p_{E_n^2}(t,u,x_1) - \\ -(\alpha+\lambda_1(u+x_1)+\lambda_2(u))p_{E_n^2}(t,u,x_1) \\ +\alpha p_{E_{n-1}^2}(t,u,x_1)[n\geq 3], \tag{11}$$

where the Iverson brackets are used for notational simplicity. For example, equation (10) says that the process is in a state (E_n^1,u,x_2) in one of the two following cases:

a) Δt units of time ago it was in $(E_n^1,n,u-\Delta t,x_2)$ and nothing has happened in the meantime (i.e., no service was completed, and no customer arrived), or
b) if $n\geq 3$, Δt units of time ago the process was at $(E_{n-1}^1,u-\Delta t,x_2)$, and a new customer arrived in the meantime. Let

$$E_n^3 := (\{3\},\{2\},n), \\ E_n^4 := (\{2,3\},\emptyset,n), n\geq 2. \tag{12}$$

Given $n\geq 2$, we define $p_{E_n^3}(t,\cdot,\cdot)$ to be the probability density function for the process in the quadrant $\{E_n^3\}\times\mathbb{R}_+^2$ and $p_{E_n^4}(t,\cdot)$ be such a function for the process in $\{E_n^4\}\times\mathbb{R}_+^2$ at this time. The mechanism governing this quantities is rather similar to that visible in Equations 10–11, and consequently

$$\partial_t p_{E_n^3}(t,u,x_2) = -\partial_u p_{E_n^3}(t,u,x_2) - \\ -(\alpha+\lambda_2(u+x_2)+\lambda_3(u))p_{E_n^3}(t,u,x_2) \\ +\alpha p_{E_{n-1}^3}(t,u,x_2)[n\geq 3] \tag{13}$$

$$\partial_t p_{E_n^4}(t,u) = -\partial_u p_{E_n^4}(t,u) - \\ -(\alpha+\lambda_2(u)+\lambda_3(u))p_{E_n^4}(t,u) \\ +\alpha p_{E_{n-1}^4}(t,u)[n\geq 3]. \tag{14}$$

Derivation of boundary conditions for Kolmogorov equations

Equations 8–11 and 13–14 express the rules pertaining to the states in $\mathcal{S}$ with $u>0$, where u is interpreted to be 'the time from the last significant change' in the system. To explain, if, for example, there are already 2 customers in the system, an arrival of a new customer, even though it lengthens the queue, does not change the nature of the situation: the arrival does not affect the crucial parameter u.

On the other hand, suppose there are at least three customers in the system and two of them are being served: one at X, the other at Y. For definiteness, suppose the process is at (E_n^1,u,x_2) with $n\geq 2$. If the service at X is completed, the counter u is reset to $u=0$, and the process starts

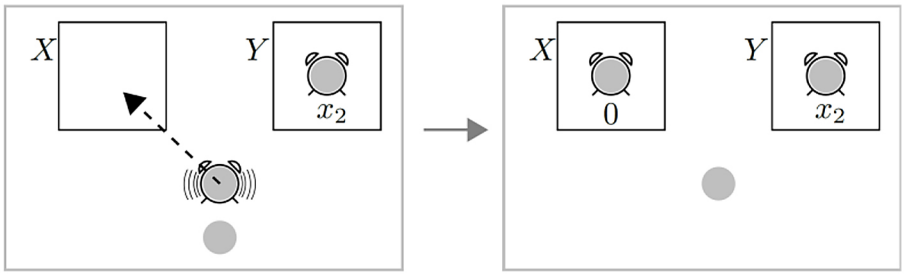

Figure 5. Equation 17 for $n = 3$: the hesitating customer runs out of his patience, decides not to wait any longer and chooses the slow server X

anew from $(E^3_{n-1}, 0, u + x_2)$, that is, from a point at the boundary. Similarly, if the service at Y is completed, the process jumps to $(E^2_{n-1}, 0, u)$. The jumps that involve resetting u are seen as 'significant', and, more importantly, lead to the following boundary conditions that accompany (8)–(11):

$$p_{E^1_1}(t, 0) = 0 \tag{15}$$

$$p_{E^2_1}(t, 0) = \alpha p_{E_0}(t) + \\ + \iint_{\mathbb{R}^2_+} \lambda_2(u + x_2) p_{E^3_2}(t, u, x_2) \mathrm{d}u \mathrm{d}x_2 \\ + \int_{\mathbb{R}_+} \lambda_2(u) p_{E^4_2}(t, u) \mathrm{d}u \tag{16}$$

and

$$p_{E^1_n}(t, 0, x_2) = \int_0^{x_2} \lambda_3(u) p_{E^3_n}(t, u, x_2 - u) \mathrm{d}u + \\ + \lambda_3(x_2) p_{E^4_n}(t, x_2) \tag{17}$$

$$p_{E^2_n}(t, 0, x_1) = \alpha p_{E^1_1}(t, x_1)[n = 2] + \\ + \int_{\mathbb{R}_+} \lambda_2(u + x_1) p_{E^1_{n+1}}(t, x_1, u) \mathrm{d}u \\ + \int_0^{x_1} \lambda_2(u) p_{E^2_{n+1}}(t, u, x_1 - u) \mathrm{d}u \tag{18}$$

for $n \geq 2$ and $t \geq 0$. To explain Equation 16: there are two possibilities for a process to start anew at $(E^2_1, 0)$. Either there were two customers in the system: one at X and one hestitaning, and the service at Y was completed or the system was empty and a new customer arrived. The Equation 17 says that the process starts anew at $(E^1_n, 0, x_2)$ with $n \geq 2$, when a hesitating customer lost his patience and decided to be served at X anyway, see Figure 5. Finally, the three terms on the right-hand side of Equation 18 come from the following three possibilities for the process to start anew at $(E^2_n, 0, x_1)$ (Figure 6):

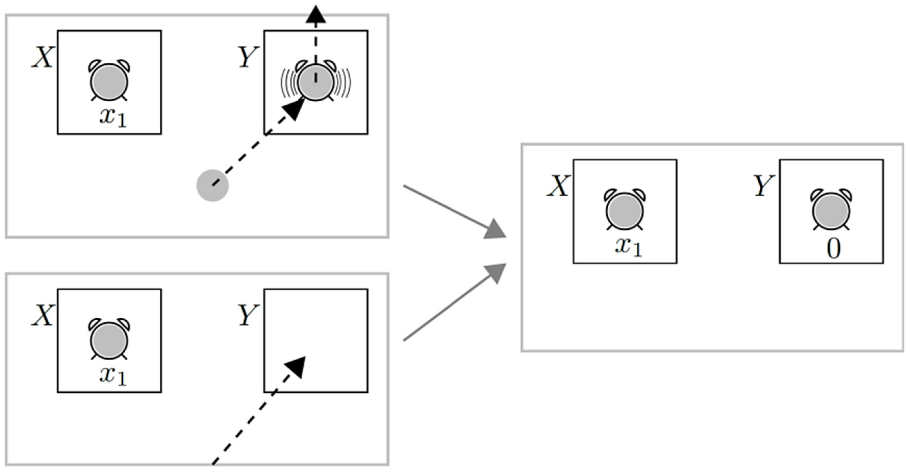

Figure 6. Explanation of Equation 18 for $n = 2$. The left server is busy, and meanwhile: either the service on the fast server Y ends and the only customer waiting in the queue takes it over, or a new customer comes to the system, finds the fast server idle and immediately takes it. After each of these events the process starts anew at $(E^2_2, 0, x_1)$.

1) the process was at (E^1_{n+1}, x_1, u) for some $u \geq 0$, and the service at Y was completed,
2) the process was at $(E^2_{n+1}, u, x_1 - u)$ for some u between 0 and x_1 and the service at Y was completed, and
3) this case is possible only if $n = 2$, the process was at (E^1_1, x_1) and then a new customer arrived.

Finally, an analogous analysis of Figures 7 and 8 leads to the boundary conditions

$$p_{E^3_n}(t, 0, x_2) = \alpha p_{E^2_1}(t, x_2)[n = 2] + \\ + \int_0^{x_2} \lambda_1(u) p_{E^1_{n+1}}(t, u, x_2 - u)\mathrm{d}u \\ + \int_{\mathbb{R}_+} \lambda_1(u + x_2) p_{E^2_{n+1}}(t, x_2, u)\mathrm{d}u \quad (19)$$

$$p_{E^4_n}(t, 0) = \iint_{\mathbb{R}^2_+} \lambda_2(u + x_2) p_{E^3_{n+1}}(t, u, x_2)\mathrm{d}u\mathrm{d}x_2 \\ + \int_{\mathbb{R}_+} \lambda_2(u) p_{E^4_{n+1}}(t, u)\mathrm{d}u \quad (20)$$

which are to be satisfied for $n \geq 2$ and $t \geq 0$.

TWO APPROACHES TO SIMULATING THE QUEUING SYSTEM

Two complementary computational approaches are employed to characterize the queueing system: first, an event-driven Monte–Carlo simulation [18], and second, a finite-difference, operator-splitting solver for the age-structured PDE system [19]. All of our numerical experiments and figures are fully reproducible using the open-source code available at GitHub [20].

Monte Carlo simulation

The stochastic model of the queue is simulated by an event-driven Monte Carlo algorithm. In each run, customers arrive according to a Poisson process of rate α, and two servers operate at exponential rates λ_1 (slow) and λ_2 (fast). Waiting customers facing an occupied fast server may abandon their wait after an exponential patience time of rate λ_3 and enter service at the slow server. Time is advanced by scheduling the next event (arrival, service completion, or

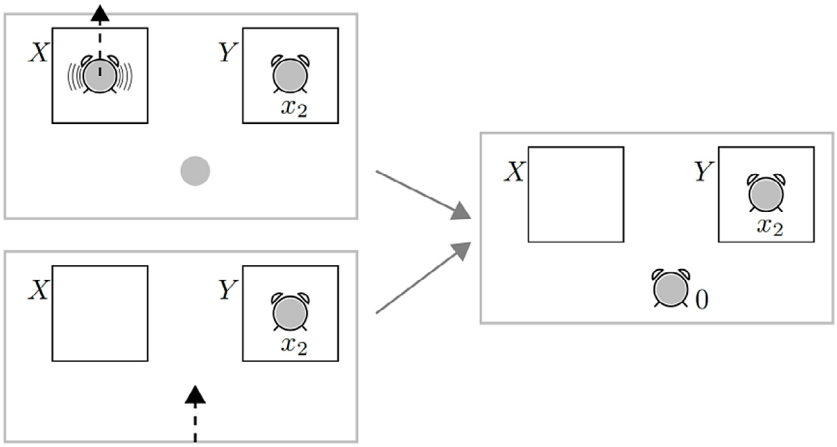

Figure 7. Equation 19 for $n = 2$. For the process to start anew at $(E^3_2, 0, x_2)$, either the server X becomes vacant but the customer at the front of the queue waits to be served at Y, or there was only one customer present in the system (being served at Y) and a new client arrives.

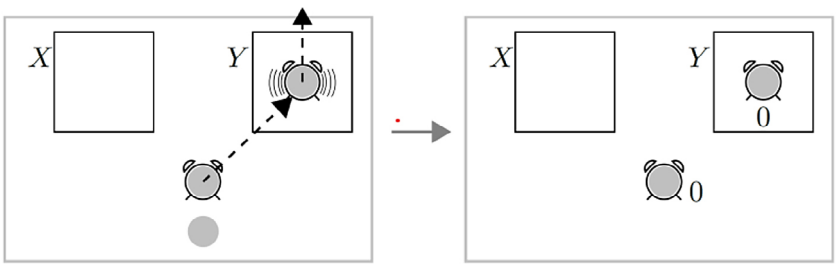

Figure 8. Equation 20 for $n = 2$. After the fast server Y is released, the hesitating customer starts to be served there

patience expiry) via a min-heap priority queue. After each event the system state is updated and any newly enabled events are scheduled. We collect, over 10^6 independent runs up to a fixed time t, the empirical distribution of the number of customers in system.

PDE with boundary conditions numerical solution

The age-structured PDE system (Equations 6, 8–11 and 13–14) with boundary conditions (Equations 15–20) is numerically solved by a finite-difference, operator-splitting scheme combined with composite Simpson quadrature. The aim is to compute the probability distribution of the number of customers in the system at a given time. More specifically, for $t \geq 0$, $N(t)$ is defined as the number of customers in the system at time t and

$$P_n(t) := P(N(t) = n), \qquad n \geq 0. \quad (21)$$

is calculated. To this end, we fix positive reals $U_{\max}$ and $X_{\max}$ as upper bounds for age variables u and x, and we express the probability $P_n(t)$ as marginals (Eq. 22).

Here, we truncate the customer-count at fixed positive integer $N_{\max}$, that is $n = 0, 1, \ldots, N_{\max}$. We introduce a „tail bucket" at $n = N_{\max} + 1$, so that any probability flux into $n > N_{\max}$ is absorbed without spurious feedback. The variables $u \in [0, U_{\max}]$ and $x \in [0, X_{\max}]$ are discretized on uniform grids of N_u and N_x points, with spacings

$$\Delta u = \frac{U_{\max}}{N_u - 1}, \quad \Delta x = \frac{X_{\max}}{N_x - 1}. \quad (23)$$

Time is discretized into N_t steps of size $\Delta t = 0.5\Delta u$, ensuring the advective Courant-Friedrichs-Lewy (CFL) condition $\Delta t/\Delta u \leq 0.5$.

The computations are split into *advect* and *react* parts. In each time-step from t_k to $t_{k+1} = t_k + \Delta t$, first we solve

$$\partial_t p = -\partial_u p \quad (24)$$

by an explicit upwind discretization. At each grid-point, indexed by i, we let

$$p_i^{\text{adv}} = p_i^k - \frac{\Delta t}{\Delta u}(p_i^k - p_{i-1}^k), \quad (25)$$

with inflow boundary values p_0^{adv} given by the boundary conditions (Equations 15–20). Here, p_i^k is the value at t_k and i-th grid-point. Then, the remaining ordinary differential equation of the form

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -(\alpha + \Lambda)p + S \quad (26)$$

is solved by implicit-Euler at each (u, x)-grid point, as

$$p^{k+1} = \frac{p^{\text{adv}} + S\Delta t}{1 + (\alpha + \Lambda)\Delta t}. \quad (27)$$

Here, Λ is the local total departure rate (e.g. $\lambda_1(u) + \lambda_2(u + x)$ for $p_{E_n^1}$), and S is the source term (either the arrival from $n - 1$ or the service-completion integrals from $n + 1$). All one- and two-dimensional integrals are approximated by composite Simpson's rule.

At each grid point, the advective step is followed by a reaction update governed by a local ordinary differential equation. This ODE arises naturally from the decomposition of the Kolmogorov forward equations into transport and reaction parts. The ODE solution corresponds to the probability mass transfer due to arrivals, service completions, or reneging. The initial condition is provided by the empirical distribution of the system at time t=0, where all probability mass is concentrated at the empty state.

Finally, after each full step the total probability mass $\sum_{n=0}^{N_{\max}} P_n$ is computed and all densities are renormalized to enforce mass conservation.

This scheme combines the simplicity of explicit upwind transport with the robustness of implicit reaction updates and higher-order quadrature, yielding numerical solution in reasonable CPU time.

Such age-structured PDEs, after reduction to ODEs, can in be solved by other high-accuracy techniques for initial and boundary value problems [21], or by optimization with respect

$$\begin{cases} p_{E_0}(t), & n = 0, \\ \displaystyle\int_0^{U_{\max}} \left[p_{E_1^1} + p_{E_1^2}\right](t, u)\,\mathrm{d}u, & n = 1, \\ \displaystyle\int_0^{X_{\max}} \int_0^{U_{\max}} \left[p_{E_n^1} + p_{E_n^2} + p_{E_n^3}\right](t, u, x)\,\mathrm{d}u\,\mathrm{d}x + \int_0^{U_{\max}} p_{E_n^4}(t, u)\,\mathrm{d}u, & n \geq 2. \end{cases} \quad (22)$$

to convergence-control parameters in iterative schemes [22]. Our choice of finite-difference operator-splitting is guided by its simplicity, mass conservation, and computational efficiency.

RESULTS

First, the probability distributions obtained by Monte Carlo simulations and numerical solutions are compared. Four sets of parameters are selected (Table 3).

Then the following distances between distributions P_n^{sim} and P_n^{PDE} are calculated:

- Total variation distance:

$$\text{TV} = \frac{1}{2}\sum_{n=0}^{N_{\max}} \left|P_n^{\text{sim}}(t) - P_n^{\text{PDE}}(t)\right|. \quad (28)$$

- Sup-norm distance:

$$L^{\infty} = \max_{n=0,1,\ldots,N_{\max}} \left|P_n^{\text{sim}}(t) - P_n^{\text{PDE}}(t)\right|. \quad (29)$$

- Kolmogorov-Smirnov distance:

$$\text{KS} = \sup_{t\in\mathbb{R}} \left|F^{\text{sim}}(t) - F^{\text{PDE}}(t)\right|, \quad (30)$$

where: F^{sim} and F^{PDE} are cumulative distribution functions.

- Expected value distance:

$$\Delta E = \left|E^{\text{sim}}(t) - E^{\text{PDE}}(t)\right|, \quad (31)$$

where: E^{sim} and E^{PDE} are expected values.

- Variance distance:

$$\Delta\text{Var} = \left|\text{Var}^{\text{sim}}(t) - \text{Var}^{\text{PDE}}(t)\right|. \quad (32)$$

The results are presented in Table 4. The empirical histograms are shown in Figure 9. It is observed that the Monte Carlo histograms and the PDE solutions align closely across all traffic regimes, confirming consistency between the stochastic simulation and PDE formulation.

A total of 24 distinct parameter configurations were considered. For each configuration, see Table 5, $100\,000$ independent Monte–Carlo runs of length $T = 1000$ time-units were performed. From each run the following performance metrics were extracted and then averaged them over replications:

- $E(N)$ and $\text{Var}(N)$: the time-average mean and variance of the total number of customers in the system;
- $E(W)$: the average waiting time until service begins;
- $\max[N]$: the maximum instantaneous system size observed in any single run;
- ρ_X and ρ_Y: the steady-state utilizations of server X (slow) and server Y (fast);
- λ_{out}: throughput, the long-run departure rate (departures per unit time);
- P_{blk}: blocking probability, the fraction of time that server X is idle while server Y is busy and the queue is non-empty;
- β_Y: fraction of service by Y, the share of all service completions processed on the fast server;
- $P_{\sim\text{FCFS}}$: the proportion of departures that violate first-come-first-served order (i.e. depart before an earlier arrival).

These metrics provide a comprehensive characterization of congestion, delay, server-level load, capacity waste, and fairness under each parameter regime.

Table 3. Parameters sets for comparing distributions

Parameter	α	λ_1	λ_2	λ_3	t
Light traffic	1.0	2.0	3.0	0.8	100
Balanced traffic	2.0	0.8	1.2	2.0	5
Overload traffic	15.0	3.0	4.0	3.0	1
Highly asymmetric servers	4.0	2.0	6.0	3.0	5

Table 4. Comparison of distributions P_n^{sim} and P_n^{PDE}

Parameter	TV	L^{∞}	KS	ΔE	ΔVar
Light traffic	0.0019	0.0019	0.0017	0.0048	0.0224
Balanced traffic	0.0633	0.0328	0.0633	0.3127	1.0735
Overload traffic	0.0471	0.0357	0.0471	0.2313	1.2251
Highly asymmetric servers	0.0806	0.0127	0.0806	0.8226	0.4742

Table 5. Simulation results for selected parameter sets

α	λ_1	λ_2	λ_3	E [N]	Var [N]	Max N	E [W]	ρx	ρy	λout	P_{blk}	βy	$P_{\sim FCFS}$
0.8	1.0	3.0	0.5	0.36	0.47	13	0.10	0.03	0.26	0.80	0.06	0.97	0.03
1.2	1.0	3.0	0.5	0.64	0.97	18	0.17	0.06	0.38	1.20	0.12	0.95	0.06
1.6	1.0	3.0	0.5	1.04	1.90	25	0.28	0.10	0.50	1.60	0.20	0.94	0.09
2.0	1.0	3.0	0.5	1.65	3.87	35	0.44	0.14	0.62	2.00	0.29	0.93	0.13
2.4	1.0	3.0	0.5	2.74	9.04	50	0.75	0.20	0.73	2.40	0.39	0.92	0.16
2.8	1.0	3.0	0.5	5.36	29.95	97	1.52	0.25	0.85	2.79	0.50	0.91	0.20
3.2	1.0	3.0	0.5	19.84	265.68	280	5.78	0.31	0.96	3.18	0.62	0.90	0.24
3.0	1.0	3.0	0.1	30.60	491.17	320	9.82	0.04	0.97	2.96	0.90	0.98	0.04
3.0	1.0	3.0	0.1	23.69	340.72	256	7.51	0.08	0.96	2.97	0.84	0.97	0.07
3.0	1.0	3.0	0.3	12.30	128.75	202	3.71	0.20	0.93	2.99	0.67	0.93	0.16
3.0	1.0	3.0	1.0	6.11	36.93	89	1.61	0.40	0.86	2.99	0.40	0.87	0.30
3.0	1.0	3.0	2.0	4.71	22.31	85	1.12	0.52	0.83	2.99	0.26	0.83	0.38
3.0	1.0	3.0	5.0	3.92	15.60	70	0.83	0.63	0.79	3.00	0.13	0.79	0.45
2.0	0.5	3.0	0.5	1.82	4.30	38	0.47	0.25	0.62	2.00	0.25	0.94	0.23
2.0	2.0	3.0	0.5	1.55	3.56	32	0.43	0.08	0.61	2.00	0.31	0.92	0.08
2.0	1.0	2.0	0.5	6.02	35.86	93	2.44	0.26	0.87	1.99	0.52	0.87	0.20
2.0	1.0	5.0	0.5	0.68	1.05	19	0.11	0.06	0.39	2.00	0.13	0.97	0.06
3.5	1.0	3.0	0.0	246.42	20375.69	808	70.08	0.02	1.00	3.01	0.98	0.99	0.02
3.5	1.0	3.0	4.0	11.36	111.42	171	2.76	0.73	0.92	3.49	0.18	0.79	0.49
4.8	1.0	3.0	0.5	735.77	180573.66	1885	152.96	0.33	1.00	3.33	0.67	0.90	0.25
1.8	0.5	1.2	0.5	179.17	10728.17	586	98.62	0.50	1.00	1.45	0.50	0.83	0.34
2.4	0.5	5.0	1.0	1.04	1.68	23	0.14	0.25	0.45	2.40	0.12	0.95	0.24
5.0	3.0	4.0	2.0	21.30	324.95	286	3.98	0.37	0.96	4.98	0.56	0.77	0.26
10.0	4.0	5.0	5.0	1391.76	646022.87	3392	138.99	0.56	1.00	7.22	0.44	0.69	0.35

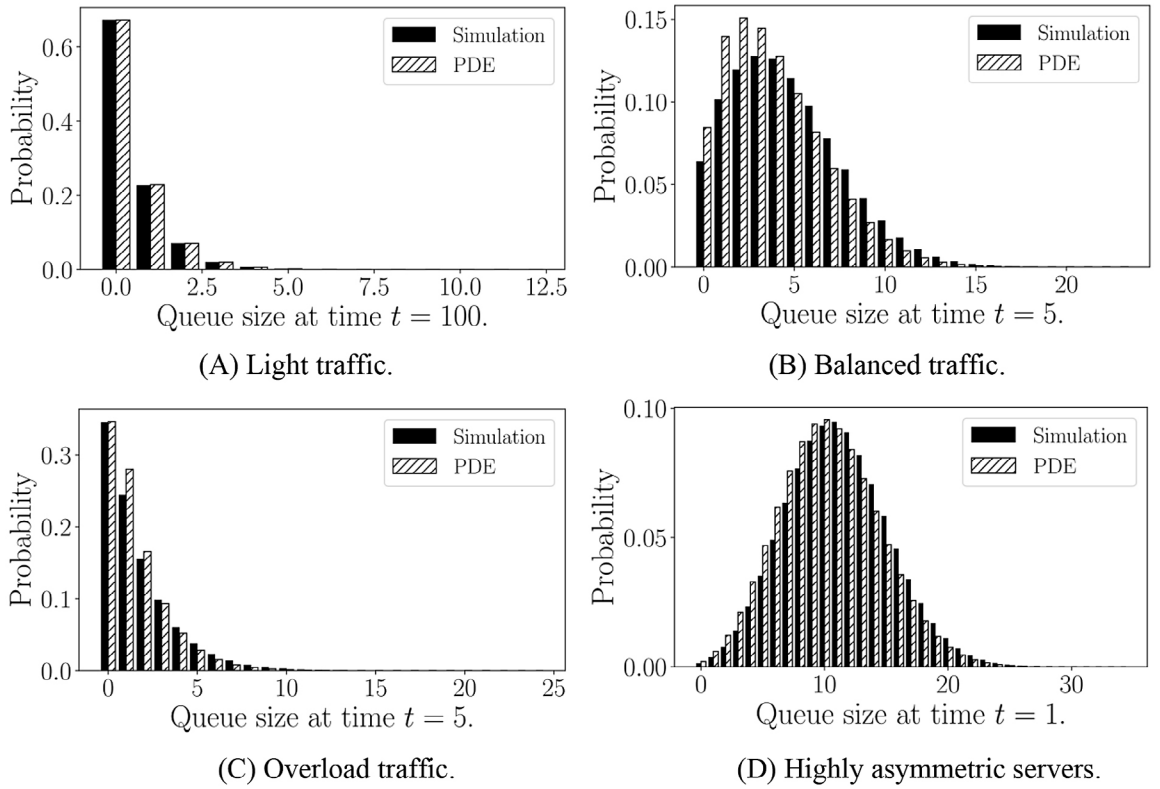

Figure 9. Comparison of the queue-length distribution from Monte Carlo simulations and PDE

CONCLUSIONS

A heterogeneous two-server queue with an impatient head-of-line customer has been analyzed using both event-driven Monte–Carlo simulation and a finite-difference, operator-splitting PDE solver. It is demonstrated how the interplay between arrival rate, service-rate asymmetry, and reneging behavior governs key performance metrics—mean queue length, sojourn time, blocking probability, and fairness as measured by FCFS violations. It has been shown that increased impatience can both alleviate and exacerbate congestion, depending on system load. In short, larger the impatience (more willingness to take slow service) and higher arrival rate push load onto the slow server, while low arrival rate or impatience leave the slow server barely used. The effect is moderated by service-rate asymmetry (a much faster Y suppresses the utilization of X unless impatience is large).

Relative to the classical heterogeneous-server studies, the model introduces a new state-space construction that embeds strategic hesitancy through supplementary clock configurations. This device permits general service-time distributions while preserving Markovian tractability. Across moderate utilisation levels, the simulations consistently report shorter waiting times and fewer FCFS violations than these benchmark models.

Under pronounced service-rate asymmetry, our policy remains close to the ideal throughput bound while maintaining high fairness, thereby combining efficiency and equity over a broad range of practical conditions.

By coupling a flexible stochastic description with a scalable PDE solver, a tool is provided to practitioners that can be calibrated to empirical service-time distributions and patience profiles. Consequently, our framework offers a practical alternative to existing M/M/2 models for environments ranging from call-centre staffing to cloud-service load balancing, where both heterogeneity and impatience are pervasive.

To summarize, the novelty of our work lies in introducing a flexible stochastic description of heterogeneous two-server queues with a state-space construction that embeds strategic hesitancy and patience dynamics without losing Markovian tractability. This extends existing analyses of strategic or impatient customers in single-server settings to a richer multiserver environment. In practical terms, the model and numerical approach can be calibrated to empirical service-time and patience distributions, offering system designers a tool analogous to recent estimation frameworks for unobserved balking to optimise staffing, load balancing, and policy design in heterogeneous service systems.

REFERENCES

1. Mas, L., Vilaplana, J., Mateo, J. et al. A queuing theory model for fog computing. J Supercomput 2022; 78: 11138–11155. https://doi.org/10.1007/s11227-022-04328-3
2. Giambene G. Queuing theory and telecommunications, networks and applications, Textbooks in Telecommunication Engineering 2021. https://doi.org/10.1007/978-3-030-75973-5
3. Sowndharya K. Preethi, J. Ebenesar Anna Bagyam. Optimal server analysis of M/M/c queueing model to reduce the waiting time of patients in healthcare service. International Journal of Mathematics in Operational Research 2024; 27(2): 254–266. https://doi.org/10.1504/IJMOR.2024.137041
4. Matuszak Z., Bundz S., Jaśkiewicz M., Stokłosa J., Posuniak P. The Application of massing handling theory for evaluation of the application of wharves and loading facilities in the Maritime Port. Advances in Science and Technology Research Journal. 2016; 10(31): 281–288. https://doi.org/10.12913/22998624/64017
5. Jaradat A. Replica selection algorithm in data grids: the best-fit approach. Advances in Science and Technology Research Journal 2021; 15(4): 30–37. https://doi.org/10.12913/22998624/142214
6. Dshalalow J. H. An anthology of classical queueing methods. In: Advances in Queueing Theory, Methods, and Open Problems. CRC Press, 2023; 1–42. https://doi.org/10.1201/9781003418283
7. Boots N. K., Tijms H. A multiserver queueing system with impatient customers, Management Science 1999; 45(3): 444–448. https://doi.org/10.1287/mnsc.45.3.444
8. Dudin A., Jacob V., Krishnamoorthy A. A multi-server queueing system with service interruption, partial protection and repetition of service, Ann. Oper. Res. 2015; 233: 101–121. https://link.springer.com/article/10.1007/s10479-013-1318-3
9. Büke, B. Modelling heterogeneity in many-server queueing systems. Queueing Syst 100, 2022; 401–403. https://doi.org/10.1007/s11134-022-09788-1
10. Ramasamy S., Daman O. A., Sani S. An M/G/2 queue where customers are served subject to a minimum violation of FCFS queue discipline, Eur. J. Oper. Res. 2015; 240(1): 140–146. https://doi.

org/10.1016/j.ejor.2014.06.048

11. Kumar R., Sharma S., Bura G. S. Transient and steady-state behavior of a two-heterogeneous servers' queuing system with balking and retention of reneging customers. In: Deep K., Jain M., Salhi S. (Eds) Performance Prediction and Analytics of Fuzzy, Reliability and Queuing Models. Springer, Singapore 2019; 251–264. https://link.springer.com/chapter/10.1007/978-981-13-0857-4_19
12. Cox D. R. The analysis of non-Markovian stochastic processes by the inclusion of sup-plementary variables, Proc. Cambridge Philos. Soc. 3 1955; 51: 433–441. https://doi.org/10.1017/s0305004100030437
13. Jain M., Kaur S., Singh P. Supplementary variable technique (SVT) for non-Markovian single server queue with service interruption (QSI). Oper Res Int J 2021; 21: 2203–2246. https://doi.org/10.1007/s12351-019-00519-8
14. Vytovtov K. A., Barabanova E. A., Vishnevsky V. M. Modeling and analysis of multi-channel queuing system transient behavior for piecewise-constant rates. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (Eds) Distributed Computer and Communication Networks: Control, Computation, Communications. DCCN 2022. Lecture Notes in Computer Science 2022; 13766: 397–409. Springer, Cham. https://doi.org/10.1007/978-3-031-23207-7_31
15. Bobrowski A. Lord Kelvin and Andrey Andreyevich Markov in a queue with single server, Vestnik YuUrGU. Ser. Mat. Model. Progr. 2018; 3(11): 29–43. https://doi.org/10.14529/mmp180303
16. Davis M. H. A. Markov Processes and Optimization, Chapman and Hall 1993.
17. Rudnicki R., Tyran-Kamińska M. Piecewise Deterministic Processes in Biological Models. Springer Briefs in Applied Sciences and Technology, Springer, Cham, 2017. Springer Briefs in Mathematical Methods.
18. Law A. M. Simulation Modeling and Analysis, McGraw-Hill, 2015.
19. Hundsdorfer W., Verwer J. G. Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations, Springer, 2003.
20. Gregosiewicz A., Stępień Ł., Ratajczyk E. Impatient queues, https://github.com/gregosiewicz/impatient-queues
21. Turkyilmazoglu M. Solution of initial and boundary value problems by an effective accurate method, International Journal of Computational Methods, 2017; 14(6): 1750069. https://doi.org/10.1142/S0219876217500694
22. Turkyilmazoglu M. Optimization by the convergence control parameter in iterative methods, Journal of Applied Mathematics and Computational Mechanics, 2024; 23(2): 105–116. https://doi.org/10.17512/jamcm.2024.2.09