Queueing Systems (2020) 95:145–171
https://doi.org/10.1007/s11134-020-09647-x

Ticket queues with regular and strategic customers

Gabi Hanukov[1,3] · **Shoshana Anily**[2] · **Uri Yechiali**[3]

Received: 9 March 2019 / Revised: 8 October 2019 / Published online: 15 February 2020

Abstract
We study a Markovian single-server ticket queue where, upon arrival, each customer can draw a number from a take-a-number machine, while the number of the customer currently being served is displayed on a panel. The difference between the above two numbers is called the "virtual queue length." We consider a nonhomogeneous population of customers comprised of two types: "regular" and "strategic." Upon arrival, a regular customer, regardless of the value of the virtual queue length, draws a number from the machine, joins the queue and waits in the system until being served. A strategic customer, depending on the virtual queue length, may either join, leave, or go to "orbit" for a random duration. If, upon return from orbit, a strategic customer realizes that s/he missed her/his turn, s/he balks. Otherwise, s/he joins the queue and waits to be served. We analyze this intricate stochastic system, calculate its steady-state probabilities, derive the sojourn time's Laplace–Stieltjes transform of a regular and of a strategic customer and calculate the system's performance measures. Finally, an economic analysis is performed to determine the optimal mean orbiting time of strategic customers for two types of objective functions. Numerical examples are presented.

Keywords Ticket queues · Strategic customers · Orbit · Sojourn times · Matrix geometric

✉ Gabi Hanukov
german.kanukov@biu.ac.il

Shoshana Anily
anily@tauex.tau.ac.il
https://en-coller.tau.ac.il/profile/anily

Uri Yechiali
uriy@tauex.tau.ac.il

1 Department of Management, Bar-Ilan University, 5290002 Ramat Gan, Israel

2 Coller School of Management, Tel Aviv University, 6997801 Tel Aviv, Israel

3 Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, 6997801 Tel Aviv, Israel

Mathematics Subject Classification 60K25 · 68M20 · 90B22

1 Introduction

Queueing problems, known as *Ticket Queues*, are considered highly intricate for analytic solution. A recent extensive survey of queueing publications [8] mentions a single paper on ticket queues [19]. That paper considers an infinite capacity Markovian queue that serves a homogeneous population, where, upon arrival, each customer is issued a ticket number, while the ticket number of the last customer currently being served is displayed on a panel. Based on this information, customers have to decide whether to join or to balk, while reneging after joining is not considered an option. If the difference between the above two numbers, called the *virtual queue length*, exceeds a certain pre-specified common patience-queue level, the customer balks never to return. This threshold-type decision is made with incomplete information on the *actual queue length*, as the virtual queue length includes, in addition to the actual queue length, customers that have balked upon arrival. In a recent paper, Kerner et al. [10] show that no threshold strategy can be a Nash equilibrium strategy for the model analyzed in [19]. Furthermore, they show that if all customers adopt any threshold strategy, the best response of any individual is to use a double-threshold strategy of the following type: join if and only if the virtual queue length is either (i) smaller than one threshold, or (ii) larger than a second threshold. Jennings and Pender [9] compare between standard queues and ticket queues of the type described above and prove heavy traffic limit theorems for both ticket and standard queueing processes, discovering that if the customer population is relatively patient, then the two processes converge to the same limit. Further, the latter paper heuristically estimates several performance metrics of ticket queues and tests them by using simulation.

Comparisons of perceptions of, and behavior in ticket queues and physical queues (stand-in-line systems) are discussed in Kuzu [11]. The paper conducts a series of surveys to understand customer preferences and patience in such systems. The paper also investigates whether the assumptions used in analytical queuing models for customers abandonment behavior are realistic. The results of the paper indicate that customers generally prefer ticket queues over physical queues and are more patient in ticket queues. Ding [2] develops an approximation procedure that calculates the percentage of reneging customers in single-server ticket queues and illustrates numerically that the percentage of reneging customers can be significantly reduced by offering customers extra information on the actual queue length. Many other papers on *strategic queues* assume that the balking/waiting decision rule used by customers is based on a complete information of the actual queue length; see, for example, [1, 8, 16, 21].

A recent paper by Kuzu et al. [12] takes a holistic empirical approach to examine how realistic customer behaviors drive ticket queue performance. Their empirical results reveal that customers are capable of adapting their patience to the waiting context and that they dynamically improve their forecast and decision making over time. The conclusion is that the inefficiency of ticket queues is much smaller than that predicted in the literature.

In our paper, we propose a model where customers in ticket queues adapt their behavior strategically to their forecast of the waiting time. More specifically, we assume that a fraction of the customers' population acts strategically by quitting the queue temporarily in order to get an extra reward by using the waiting time for completing some other duties (unrelated to the service offered by the system). Such customers take the risk of missing their turn when returning to the system. If the type of service provided by the system is not a onetime service, the strategic customers are expected to improve their forecast over the waiting time duration, in order to maximize their total expected reward, as suggested by Kuzu et al. [12].

In this paper, we follow the assumption of the above-mentioned papers [19] and [10] that the decisions taken by customers are based on the virtual queue length, which is an upper bound on the unobserved actual queue length. However, our model deviates from these papers by allowing a nonhomogeneous population of customers, in addition to more involved decision rules described in the sequel.

In practice, there are two types of models for ticket queues: (i) models where customers must first draw a ticket in order to view their number, as is the case in [19], and (ii) models where customers see the number that they can draw before drawing the ticket, as we consider here. Thus, a customer that arrives to such a system can see both the ticket number that s/he can draw and the displayed number of the customer currently being served. Based on the difference between the two numbers (the virtual queue length), s/he decides whether to balk or to draw the number and join the queue.

The first model, where customers must draw a ticket in order to see their number, is operated by *electronic queueing management systems* and fits high volume businesses that provide a range of services, where each type of service has its own queue. The second type of model is applied in *take-a-number* systems that fit small service facilities where a single type of service is provided, like pharmacies, cell phone repair stores, post offices, etc. Technically speaking, the main difference between the two models is that in the second type, customers who decide to balk upon arrival, based on the virtual queue length, leave the system without drawing a number. Thus, under the second type of ticket queues, the virtual queue length better reflects the actual queue length, as customers who leave upon arrival, do not draw a ticket and therefore, prevent an artificial increase of the virtual queue length. For this reason, from the point of view of customers, take-a-number systems are preferable to electronic queueing management systems. We note that in reality both types of ticket queues are common. Moreover, as electronic queueing management systems become more popular, the ticket number that is next to be drawn will probably be displayed on the panel. That is, modern systems will probably behave similarly to *take-a-number* systems.

In both types of systems, namely electronic queueing management systems and take-a-number systems, customers can take advantage of the fact that their turn is kept while going to wander around (*orbit* hereafter), as long as they return to the waiting room on time, i.e., before missing their turn. Thus, customers can try to decrease the burden of the waiting time by utilizing this time efficiently, as, for example, for completing some errands (like shopping) at a nearby location, or by relaxing at some coffee shop.

There is a vast literature on increasing the servers' efficiency by utilizing them during their idle time for doing some ancillary duties, so-called "vacations;" see, for

example, [3, 13–16, 20, 22]. Recently, Hanukov et al. [5, 6, 7] have proposed an innovative idea by which the servers' idle time is utilized for preparing and storing some preliminary services for future incoming customers. In this current study, we adopt a similar idea, but now it is the strategic customers who utilize their waiting time efficiently by going to "orbit."

This paper considers a ticket queue operated by take-a-number-system, and a non-homogeneous population of customers that consists of two types: the first type, called *regular customers*, arrives at a Poisson rate and each individual draws a ticket regardless of the queue length and stays in the waiting room until being served. In marketing terminology these customers are called *captive customers*. In our case, the population of such customers consists of, for example, elderly or handicapped people or ones that had to travel a long way to reach the service facility. The customers of the other type, called *strategic customers*, also arrive at a Poisson rate, but immediately after observing the virtual queue size, they exercise a two-threshold policy defined by a pair of positive integers (m, n), where $m < n$, as follows: if the virtual queue length does not exceed m, they draw a ticket and act like regular customers, i.e., they join the queue and wait until being served. If the virtual queue length is n or above, the customer leaves the system never to return. If the virtual queue length is greater than m but is less than n, a strategic customer draws a ticket and becomes an *orbit* customer, namely s/he goes out temporarily for having a cup of coffee or to run some errands, and after a random time s/he returns back to the waiting room. When an orbit customer returns to the system, s/he observes the display panel that shows the number of the customer currently being served, and if that number is larger than her/his own ticket number (i.e., the customer missed her/his turn), the customer leaves the system. Otherwise, the customer joins the queue and waits there until being served.

An interesting issue related to the conclusions of Kuzu et al. [12], in the context of strategic customers in our model, is the improvement over time of their forecast on the optimal orbiting duration. However, since the orbit time depends on many accidental factors that are not under control, we look here for the *mean* orbit time, taking into account the following components associated with the orbiting customers: (i) the service reward obtained by those that do not miss their turn; (ii) the reward rate (per unit time) of orbiting; and (iii) the opportunity cost for the total time that elapses from the moment of drawing a ticket until the final departure from the system. This last cost refers to a time duration that contains the orbiting time, and it usually represents the loss of income due to taking some time off from work. Note that the longer the orbiting time, the higher the reward from orbiting but, simultaneously, the higher the risk of missing the turn for service. We also analyze an alternative extended model, where we consider a penalty to be incurred if a certain deadline that the customer needs to meet is missed.

The system's complexity increases rapidly with the parameters m and n. We present a complete analysis of the case where $m = 1$ and $n = 3$. The generalization procedure for larger values of the parameters (m, n), $0 \leq m < n$, is detailed in the Appendix.

The outline of the paper is as follows: Sect. 2 presents the exact model formulation. Section 3 consists of a full steady-state probabilistic analysis of the system. In Sect. 4, various performance measures are derived regarding the expected number of regular and of strategic customers in the waiting room and in the system, where the latter

includes the orbiting customers. In Sects. 5 and 7 the Laplace–Stieltjes transforms of the sojourn time of a regular and of a strategic customer, respectively, are derived. The probability that a strategic customer is actually served is computed in Sect. 6. Section 8 deals with the question of the optimal orbit time, as mentioned above. For this sake, we provide several numerical examples that demonstrate how the economic optimal mean orbit time can be obtained. Section 9 concludes the paper.

2 Notation and model formulation

Consider a Markovian queue with a single server that serves two types of customers, called "regular" and "strategic." Regular customers arrive according to a Poisson process with rate λ, while strategic customers arrive according to a Poisson process with rate α. The service time of any customer reaching the server is exponentially distributed with mean $1/\mu$. All processes are mutually independent. A regular customer draws a number upon arrival and waits in the queue until being served. A strategic customer first observes the displayed running number of the current customer being served, and then looks at the number of the ticket that s/he can draw from the take-a-number machine. Let D ($D \geq 1$) be the difference between those two numbers, which is called the *virtual queue length*. Strategic customers are assumed to use a double-threshold policy (m, n) where $1 \leq m < n - 1$: if D is greater than or equal to a pre-specified number, n, the strategic customer balks never to return, without drawing a number. If $D \leq m$, the strategic customer draws a number and joins the queue. However, in the in-between case, where $m < D < n$, the strategic customer becomes a *Real Orbit* (RO) customer, namely the customer draws a ticket and goes out to wander around for an exponentially distributed random time with mean $1/\beta$. In order for the case $m < D < n$ to be nonempty, we assume that $n \geq m + 2$. It is also assumed that the system is such that customers can see if the server is busy or idle. When a RO customer returns to the waiting room, s/he acts as follows: (a) if the displayed number is equal to her/his own ticket number, then, clearly, no one is waiting in queue and the server is free, so the customer that has just returned from orbiting starts her/his service immediately, without waiting. (b) if the displayed number is lower than her/his own number, the customer joins the queue. (c) if the customer returns and finds that s/he missed her/his turn, then there are two options: (c1) if the server is busy then the customer abandons the system. (c2) if the server is free, we assume, that the server will accept the customer only if the last customer that got served had a ticket number which is lower than the number of the RO customer that has just returned. In this last case, where the server accepts the RO, the displayed running number is moved backward to indicate the number of the RO that is currently in service. The assumption described in case (c2) guarantees that the order in which the customers are served is aligned with the order of their ticket numbers, i.e., no customer gets served if a customer with a higher number has already been served. In the sequel, we will use the terminology "in queue" in order to refer to the "waiting room".

The type of systems described above fits many systems where customers that have drawn a number do not need to wait physically in queue in order to be served. The customers just need to make sure not to miss their turn. In a single-server system, like

in a small postal office operated by a single server, where the server and the customers in the waiting room can see each other at any point of time, handling case (c) is simple. However, in large systems, as in medical centers, it is common that the waiting room serves customers of a number of doctors. Thus, any individual doctor cannot know, by inspecting the waiting room, how many customers are waiting for her/him. The assumption that customers can see if the server is busy or idle is needed for handling the RO customers that return to the system, while their number is lower than the one displayed; see case (c) above. Thus, in the example of a medical center (or similar service systems), we assume that an indicator light signals that the doctor is idle. Our model fits systems of small or large size, as described above.

In order to analyze the system in steady state, we use the following notation: let k denote the current number of strategic customers in the system (in orbit, in service or in queue). By definition, $0 \leq k \leq n-1$. For a given k, the system's state is defined by a $(k+1)$-dimensional vector $(c_0, c_1, c_2, \ldots, c_k)$. c_0 is a nonnegative integer that indicates the number of regular customers that are currently in system and have arrived before the first strategic customer. For $1 \leq i \leq k$, c_i is represented by N^q, where $N \geq 0$ is an integer, and q is either {out} or {in}. Specifically, $c_i = N^q$ indicates that (i) N regular customers have arrived after the arrival of the ith strategic customer, and before the arrival of the $(i+1)$st strategic customer. (ii) $q = \text{out}$ indicates that the ith strategic customer is still orbiting, where $q = \text{in}$ indicates that the ith strategic customer is present in the system, as s/he is either an orbiting customer that returned to the queue, or s/he is a strategic customer that has arrived when the virtual queue length did not exceed m.

Next, we consider a few examples that shed light on the above definition: (i) the 2-dimensional vector $(7, 4^{\text{out}})$ describes a state where $k = 1$, i.e., a single strategic customer is in the system, along with $7+4 = 11$ regular customers, 7 of which arrived before the strategic customer who is currently a RO, and 4 arrived after that strategic customer. In particular, we deduce that $n \geq 8$, as the strategic customer found upon arrival a virtual queue length of at least 7, and s/he decided to join the system rather than to balk. (ii) consider a 4-dimensional vector $(7, 4^{\text{out}}, 0^{\text{in}}, 2^{\text{out}})$ describing the state with $k = 3$ strategic customers, in addition to $7+4+0+2 = 13$ regular customers, out of which 7 arrived before the first strategic customer, 4 after the first strategic customer but before the second strategic customer, no regular customer arrived in-between the second and the third strategic customers, and two regular customers arrived after the third strategic customer. The first and third strategic customers are RO (out), while the second one is waiting in queue (in). From this example, we can deduce that $n \geq 14$, as upon arrival of the third strategic customer, the virtual queue length was at least 13, and this customer did not balk. (iii) For the thresholds $m = 1$ and $n = 3$, the state $00^{\text{out}}0^{\text{out}}0^{\text{out}}$ is infeasible, as if it was feasible then there should have been a state from which the state $00^{\text{out}}0^{\text{out}}0^{\text{out}}$ is reachable. We consider the following cases: (a) if $00^{\text{out}}0^{\text{out}}0^{\text{out}}$ was reachable by an end of service of a regular customer that arrived before the third strategic customer, then that regular customer should have arrived before the first strategic customer, as otherwise the strategic customers that arrived before the regular customer, would have missed their turn. Thus, the third strategic customer saw upon arrival a virtual queue length of 3, and as $n = 3$, he should have balked immediately without drawing a ticket. (b) if the state $00^{\text{out}}0^{\text{out}}0^{\text{out}}$ was

reachable from state $00^{\text{out}}0^{\text{out}}$ by an arrival of a third strategic customer, then this last strategic customer would have found upon arrival a virtual queue length of 1, as no one was in queue upon his/her arrival, and by the system's description, the server would be idle and the number on the panel would be the last number that was issued before the new customer drew a ticket. Thus, the virtual queue length upon arrival of the third strategic customer would have been 1, meaning that he would immediately be served, contradicting the fact that this strategic customer went out orbiting.

Before proceeding to the analysis of the system, we show in Proposition 1 that for a given value of n (the threshold for strategic customers to balk upon arrival), the diagram (see Fig. 1) that depicts the transition rates between the states of the system consists of $2\cdot 3^{n-1}$ columns, where each column is infinitely long. Figure 1 is a diagram for the case $m = 1$ and $n = 3$. This diagram indicates the states and the corresponding transition rates between them. In particular, the states are presented by 18 columns of infinite size each (where a diagram for $n = 4$ would have consisted of 54 columns). For the sake of the proof, we define two special types of states:

Definition 1 A state is said to be an *origin state* if (i) it is the empty state 0, or (ii) if it satisfies the following two conditions: it is impossible to reach that state by an arrival of a regular customer and, in addition, there is at least one customer in that state that will eventually be served. Thus, an origin state, if nonempty, contains either a regular customer or a strategic customer that is not orbiting.

Definition 2 A state is called a *semi-empty* state if it consists only of customers that are currently orbiting, i.e., all customers are RO.

In Fig. 1, where $m = 1$ and $n = 3$, there are 18 origin states, namely the states 0, 00^{in}, 10^{in}, 10^{out}, 20^{in}, 20^{out}, $01^{\text{in}}0^{\text{in}}$, $01^{\text{in}}0^{\text{out}}$, $00^{\text{in}}0^{\text{in}}$, $00^{\text{in}}0^{\text{out}}$, $10^{\text{in}}0^{\text{in}}$, $10^{\text{in}}0^{\text{out}}$, $10^{\text{out}}0^{\text{in}}$, $10^{\text{out}}0^{\text{out}}$, $00^{\text{in}}0^{\text{in}}0^{\text{in}}$, $00^{\text{in}}0^{\text{in}}0^{\text{out}}$, $00^{\text{in}}0^{\text{out}}0^{\text{in}}$ and $00^{\text{in}}0^{\text{out}}0^{\text{out}}$. In addition to these states, there are two semi-empty states that consist only of RO customers, namely 00^{out} and $00^{\text{out}}0^{\text{out}}$.

Proposition 1 *For any double-threshold policy $m \geq 1$ and $n \geq m+2$, the transition-rate diagram consists of $2\cdot 3^{n-1}$ columns, where each column is infinitely long.*

Proof Note that any (infinite-length) column in the diagram starts at an origin state, so that the number of columns in the diagram equals the number of origin states. The states in the given column are accessible from the origin state by arrivals of regular customers. In addition, there are $n-1$ semi-empty states, where each one consists of k, $1 \leq k \leq n-1$, RO customers. We assume, without loss of generality, that each of the semi-empty states with k RO customers is at the top of the column whose origin state consists of exactly k strategic customers that are currently in queue or in service (see Fig. 1). It remains to prove that the number of origin states for the double-threshold policy (m, n) is $2\cdot 3^{n-1}$.

For any given $n \geq m+2 \geq 3$, let $\text{OS}(x)$ be the number of origin states that consist of $x \in \{0, 1, 2, \ldots, n\}$ customers. For $x = 0$, $\text{OS}(0) = 1$, as there exists a single origin state with no customers, which is state 0. For $x = 1$, $\text{OS}(1) = 1$, as there exists just one origin state with a single customer, which is 00^{in}. For $x = 2$, $\text{OS}(2) = 4$,

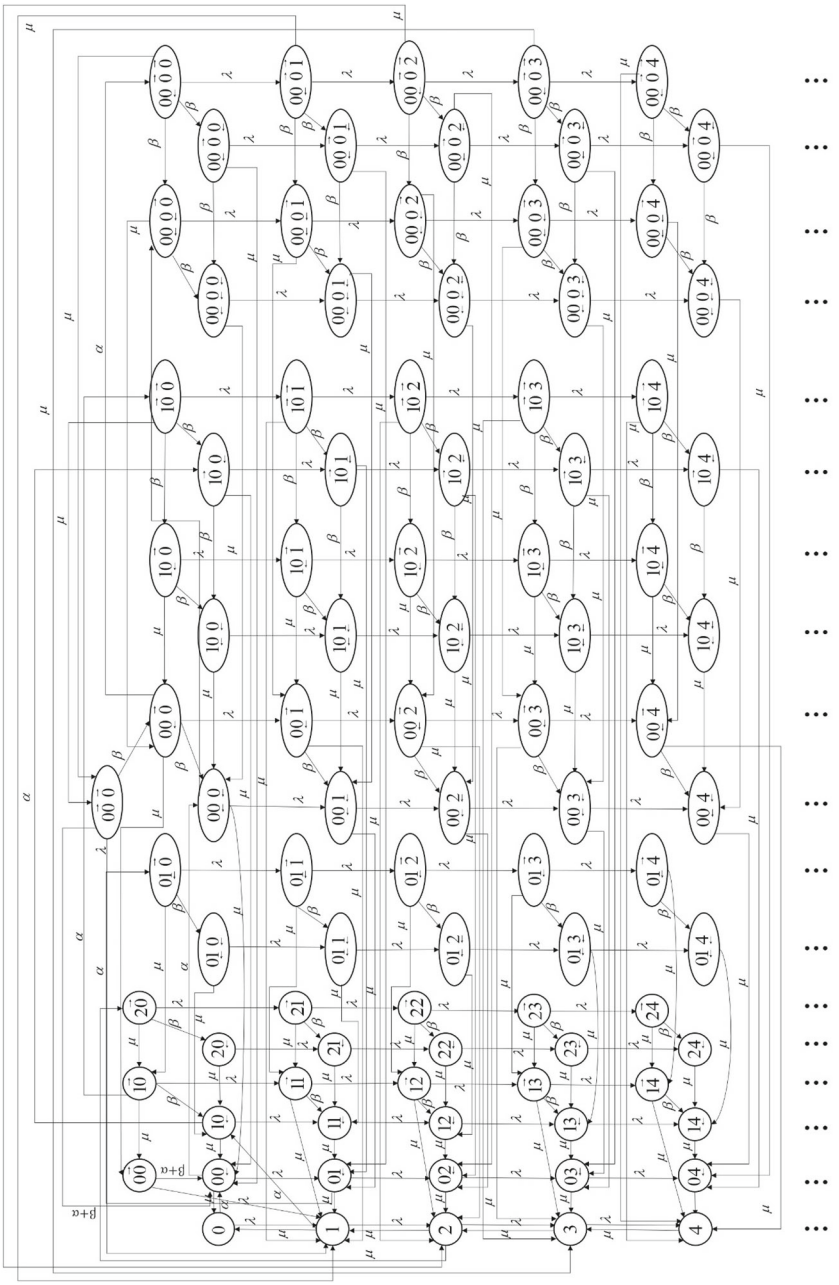

Fig. 1 System's states and transition-rate diagram for the case $n = 3$ and $m = 1$, where N^{in} is denoted by $\underline{N}$ while N^{out} by $\vec{N}$

as there exist 4 origin states each with 2 customers, namely 10^{in}, 10^{out}, $00^{\text{in}}0^{\text{in}}$ and $00^{\text{in}}0^{\text{out}}$. Thereafter, for $n \geq 3$ and $2 < x \leq n$ we prove that $\text{OS}(x) = 3 \cdot \text{OS}(x-1)$. For this sake, we show that any origin state with $x-1$ customers can be extended to 3 origin states with x customers as follows: (i) by concatenating to its end 0^{in} or 0^{out}, we get 2 new origin states with x customers, or (ii) by adding 1 to the number of regular customers that arrived before the first strategic customer, we get a third origin state. For example, the origin state with 2 customers $0^{\text{in}}0^{\text{out}}$ generates 3 origin states with 3 customers, namely $0^{\text{in}}0^{\text{out}}0^{\text{in}}$, $0^{\text{in}}0^{\text{out}}0^{\text{out}}$ and $1^{\text{in}}0^{\text{out}}$. As can be seen, all the new origin states are disjoint.

In order to complete the proof, it is left to show that any threshold policy (m, n) is associated with $2 \cdot 3^{n-1}$ origin states, each consisting of at most n customers. In other words, we need to prove that $\sum_{x=0}^{n} \text{OS}(x) = 2 \cdot 3^{n-1}$. For this sake, we calculate

$$\sum_{x=0}^{n} \text{OS}(x) = 1 + 1 + 4 + 3 \cdot 4 + 3^2 \cdot 4 + \cdots + 3^{n-2} \cdot 4 = 2 + 4 \sum_{j=0}^{n-2} 3^j = 2 \cdot 3^{n-1}$$

(QED)

In view of the exponential size of the transition-rate diagram as a function of n, see Proposition 1, we perform a complete analysis for the minimal possible values of the thresholds n and m that preserve the structure of the system, namely $m = 1$ and $n = 3$. As will be demonstrated, even this small-size case is quite intricate. Its analysis is basic in understanding the dynamics of similar higher-dimensional systems.

3 The case $m = 1$ and $n = 3$

3.1 Matrix geometric analysis

According to Proposition 1, the system's states in Fig. 1 are arranged in $2 \cdot 3^{n-1} = 18$ columns, each of infinite size. In order to present the associated infinitesimal generator matrix of this Markovian process, we arrange the states of the process in the following rows, indexed by $i \geq 0$.

The case $i = 0$ is depicted by the first double row and includes the 20 origin and semi-origin states

$$(0,\ 00^{\text{in}},\ 00^{\text{out}},\ 10^{\text{in}},\ 10^{\text{out}},\ \ldots,\ 00^{\text{in}}0^{\text{in}},\ 00^{\text{in}}0^{\text{out}},\ 00^{\text{out}}0^{\text{out}},\ \ldots,\ 00^{\text{in}}0^{\text{out}}0^{\text{in}},\ 00^{\text{in}}0^{\text{out}}0^{\text{out}}).$$

Each other case $i \geq 1$ is depicted by double row, which includes 18 states:

$$(i,\ 0i^{\text{in}},\ 1i^{\text{in}},\ 1i^{\text{out}},\ \ldots,\ 00^{\text{in}}0^{\text{out}}i^{\text{in}},\ 00^{\text{in}}0^{\text{out}}i^{\text{out}}),\quad i \geq 1.$$

The above-described vectors are referred to as $\overrightarrow{\text{ST}}_i$ for $i \geq 0$, where, for example, $\overrightarrow{\text{ST}}_0(4) = \left(1, 0^{\text{in}}\right)$, and $\overrightarrow{\text{ST}}_5(4) = \left(1, 5^{\text{out}}\right)$.

Let Q denote the infinitesimal generator matrix of the Markovian process depicted in Fig. 1. The matrix Q is given by

$$Q = \begin{pmatrix} B_{0,0} & B_{0,1} & 0 & 0 & 0 & 0 & \cdots \\ B_{1,0} & B_{1,1} & A_0 & 0 & 0 & 0 & \cdots \\ B_{2,0} & A_2 & B_{2,2} & A_0 & 0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & 0 & A_2 & A_1 & A_0 & \\ \vdots & \vdots & \vdots & & \ddots & \ddots & \ddots \end{pmatrix},$$

where the submatrices of Q are the matrices.
$B_{0,0} = \left[b_{0,0}^{i,j}\right]_{20\times 20}$ $B_{0,1} = \left[b_{0,1}^{i,j}\right]_{20\times 18}$ $B_{1,0} = \left[b_{1,0}^{i,j}\right]_{18\times 20}$ $B_{1,1} = \left[b_{1,1}^{i,j}\right]_{18\times 18}$
$B_{2,0} = \left[b_{2,0}^{i,j}\right]_{18\times 20}$ $B_{2,2} = \left[b_{2,2}^{i,j}\right]_{18\times 18}$ $A_0 = \lambda I_{18\times 18}$ $A_1 = \left[a_1^{i,j}\right]_{18\times 18}$ and $A_2 = \left[a_2^{i,j}\right]_{18\times 18}$ which are specified below:

$$b_{0,1}^{i,j} = \begin{cases} \lambda\ i = 1, 3, 12,\ j = 1;\ i = j = 2;\ j+1 = i = 4, 5, \ldots, 11;\ j+2 = i = 13, 14, \ldots, 20. \\ 0\ \text{otherwise.} \end{cases}$$

$$b_{2,0}^{i,j} = \begin{cases} \alpha\ i = 1,\ j = 7. \\ 0\ \text{otherwise.} \end{cases}$$

$$b_{1,0}^{i,j} = \begin{cases} \mu\ i = j = 1. \\ \alpha\ i = 1,\ j = 4;\ i = 2,\ j = 10. \\ 0\ \text{otherwise.} \end{cases}$$

$$b_{0,0}^{i,j} = \begin{cases} \mu & i = 2, j = 1;\ i = 15, 19, j = 2;\ j+2 = i = 4, \ldots, 7;\ j+3 = i = 13, 14; \\ & j+4 = i = 8, 9, 16;\ j+7 = i = 17, 18;\ j+8 = i = 10, 11, 20. \\ \beta & j+1 = i = 5, 7, \ldots, 11, 12, 14, \ldots, 20;\ j+2 = i = 15, 19, 20. \\ \alpha & i = 1, j = 2;\ i = 4, j = 14;\ i = 5, j = 16;\ i = 10, j = 18;\ j-9 = i = 2, 11. \\ \alpha+\beta & i = 3, 12, j = 2. \\ -(\lambda+\alpha) & i = j = 1. \\ -(\lambda+\alpha+\mu) & i = j = 2, 4, 10. \\ -(\lambda+\alpha+\beta) & i = j = 3. \\ -(\lambda+\alpha+\mu+\beta) & i = j = 5, 11. \\ -(\lambda+\mu) & i = j = 6, 8, 13, 17. \\ -(\lambda+\mu+\beta) & i = j = 7, 9, 14, 15, 18, 19. \\ -(\lambda+\alpha+2\beta) & i = j = 12. \\ -(\lambda+\mu+2\beta) & i = j = 16, 20. \\ 0 & \text{otherwise.} \end{cases}$$

$$b_{1,1}^{i,j} = \begin{cases} \mu & i = 2, 4, 10, 14, 18, j = 1;\ i = 3, 9, 13, 17, j = 2; \\ & j+2 = i = 5, 6, 11, 12;\ j+4 = i = 7, 8;\ j+6 = i = 15, 16. \\ \beta & j+1 = i = 4, 6, \ldots, 18;\ j+2 = i = 13, 14, 17, 18. \\ -(\lambda+\alpha+\mu) & i = j = 1, 2. \\ -(\lambda+\mu) & i = j = 3, 5, \ldots, 11, 15. \\ -(\lambda+\mu+\beta) & i = j = 4, 6, \ldots, 12, 13, 16, 17. \\ -(\lambda+\mu+2\beta) & i = j = 14, 18. \\ 0 & \text{otherwise.} \end{cases}$$

$$b_{2,2}^{i,j} = \begin{cases} -(\lambda+\mu)\ i = j = 2. \\ b_{1,1}^{i,j} \quad \text{otherwise.} \end{cases}$$

$$a_1^{i,j} = \begin{cases} -(\lambda+\mu)\ i=j=1,2. \\ b_{1,1}^{i,j} \qquad \text{otherwise.} \end{cases}$$

$$a_2^{i,j} = \begin{cases} \mu\ i=j=1. \\ 0\ \text{otherwise.} \end{cases}$$

Let $\vec{p}_0 = (p_0 \ p_{0,0^{\text{in}}} \ \cdots \ p_{0,0^{\text{in}},0^{\text{out}},0^{\text{out}}})_{20}$ and $\vec{p}_i = (p_i \ p_{0,i^{\text{in}}} \ \cdots \ p_{0,0^{\text{in}},0^{\text{out}},i^{\text{out}}})_{18}$ for $i = 1, 2, 3, \ldots$ be the probability vectors of the system's states. Further, denote the vector of all probability vectors by $\vec{p} = \left(\vec{p}_0 \ \vec{p}_1 \ \vec{p}_2 \ \cdots\right)$ and let $\vec{e}_y = \left(1\ 1 \cdots 1\right)^{\mathrm{T}}$ be a unit vector of dimension y. Thus, the balance equations can be written as

$$\begin{aligned} &\vec{p}\,Q = \vec{0}, \\ &\vec{p}_0\vec{e}_{20} + \sum_{i=1}^{\infty} \vec{p}_i\vec{e}_{18} = 1. \end{aligned} \tag{1}$$

3.2 Analysis

Using matrix geometric analysis, we first derive the system's stability condition and then calculate the corresponding steady-state probabilities.

3.2.1 The stability condition

According to Hanukov and Yechiali [4], when each of the matrices A_0, A_1 and A_2 is lower triangular (which is the case in our model) the stability condition is given by $a_0^{0,0} < a_2^{0,0}$, which in our model results in $\lambda < \mu$. As can be seen, the stability condition is determined by the regular customers, as the number of strategic customers in the system is bounded from above.

3.2.2 Steady-state probabilities

Let R be the matrix satisfying

$$A_0 + RA_1 + R^2A_2 = 0.$$

In general, the matrix R is calculated via successive substitutions; see [17, 18]. However, in some special cases the matrix R can be obtained directly. One case is when A_2 is of rank 1, satisfying $A_2 = \vec{c} \cdot \vec{r}$, where $\vec{c}$ is a column vector and $\vec{r}$ is a row vector normalized by $\vec{r} \cdot \vec{e} = 1$ (see [18]). In this case, R can be calculated by $R = -A_0(A_1 + A_0\vec{e} \cdot \vec{r})^{-1}$. In our model, $A_2 = \vec{c} \cdot \vec{r}$ with $\vec{c} = (\mu, 0, 0, \cdots, 0)^{\mathrm{T}}$ and $\vec{r} = (1, 0, 0, \cdots, 0)$.

Another, more general case (see [4]) is when each of the three matrices A_0, A_1 and A_2 is lower triangular, as is the case in the current model. In such a case, the entries of R are given explicitly by

$$r^{v,t} = 0, \quad v < t,$$

$$r^{v,v} = \begin{cases} \frac{-a_1^{v,v} - \sqrt{(a_1^{v,v})^2 - 4a_0^{v,v}a_2^{v,v}}}{2a_2^{v,v}}, & a_2^{v,v} > 0, \quad a_0^{v,v} > 0. \\ 0, & a_2^{v,v} > 0, \quad a_0^{v,v} = 0, \quad \forall v. \\ \frac{-a_0^{v,v}}{a_1^{v,v}}, & a_2^{v,v} = 0. \end{cases}$$

$$r^{v,t} = -\frac{a_0^{v,t} + \sum_{k=t+1}^{v} r^{v,k}a_1^{k,t} + \sum_{\tau=t+1}^{v-1} r^{v,\tau}r^{\tau,t}a_2^{t,t} + \sum_{k=t+1}^{v}\sum_{\tau=k}^{v} r^{v,\tau}r^{\tau,k}a_2^{k,t}}{a_1^{t,t} + a_2^{t,t}(r^{t,t} + r^{v,v})}, \quad v > t.$$

The steady-state probability vectors satisfy

$$\vec{p}_i = \vec{p}_3 R^{i-3}, \quad i = 3, 4, 5, \ldots. \tag{2}$$

In order to calculate those probability vectors, one needs first to obtain the vectors $\vec{p}_i$, $i = 0, 1, 2, 3$. This is achieved by considering the corresponding vector equations from Eq. (1) along with the normalization equation. Thus,

$$\begin{aligned} &\vec{p}_0 B_{0,0} + \vec{p}_1 B_{1,0} + \vec{p}_2 B_{2,0} = \vec{0}, \\ &\vec{p}_0 B_{0,1} + \vec{p}_1 B_{1,1} + \vec{p}_2 A_2 = \vec{0}, \\ &\vec{p}_1 A_0 + \vec{p}_2 B_{2,2} + \vec{p}_3 A_2 = \vec{0}, \\ &\vec{p}_2 A_0 + \vec{p}_3 (A_1 + RA_2) = \vec{0}, \\ &\vec{p}_0 \vec{e}_{20} + \sum_{i=1}^{2} \vec{p}_i \vec{e}_{18} + \vec{p}_3 [I - R]^{-1} \vec{e}_{18} = 1. \end{aligned}$$

Once the vectors $\vec{p}_i$, $i = 0, 1, 2, 3$, are calculated, the rest of the probabilities are given by Eq. (2). Numerical examples are presented in Sect. 8, where the optimal duration of the mean orbiting time is calculated for various system parameters.

4 Performance measures

We consider several performance measures related to the various types of customers. In what follows we derive expressions for (i) $E[D]$—the mean size of the virtual queue, i.e., those customers that are physically waiting in line or being served, augmented with those in orbit; (ii)$E[L]$—the mean number of regular customers in the system; and (iii)$E[S]$—the mean number of strategic customers that are either waiting in line or are being served. With those three measures, one can calculate a few more characteristics of the system: $E[L] + E[S]$ represents the mean total number of customers in queue

or in service, namely it is the number of customers that will eventually be served with probability 1. $E[D] - E[L]$ represents the mean number of strategic customers that are either orbiting, waiting in queue or being served. Finally, the mean number of RO customers is given by $E[D] - E[L] - E[S]$.

In order to calculate $E[D]$, we define the following two vectors: let $\vec{v}_D = (0\ 1\ 1\ 2\ 2\ 3\ 3\ 3\ 3\ 2\ 2\ 2\ 3\ 3\ 3\ 3\ 3\ 3\ 3\ 3)^{\mathrm{T}}$ be a 20-dimensional column vector, associated with the first double row of Fig. 1. The jth entry of the vector $\vec{v}_D$ indicates the total number of customers in the system in the jth state of vector $\overrightarrow{\mathrm{ST}}_0$, for $1 \le j \le 20$; see Sect. 3.1 for definition of the vectors $\overrightarrow{\mathrm{ST}}_i$ for $i \ge 0$. In addition, we define an 18-dimensional column vector, which is associated with each one of the other rows as follows

$$\vec{u}_D = (0\ 1\ 2\ 2\ 3\ 3\ 3\ 3\ 2\ 2\ 3\ 3\ 3\ 3\ 3\ 3\ 3\ 3)^{\mathrm{T}}.$$

In order to explain this last vector, note that the only difference between rows $i \ge 1$ in Fig. 1 is the value of i in c_k, where $c_k = i^q$. In fact, each entry in the vector $\vec{u}_D$ indicates the total number of customers in the system excluding those i regular customers that arrived after the last strategic customer, corresponding to the vector of states $\overrightarrow{\mathrm{ST}}_i$, $i \ge 1$. For example, the total number of customers in the system according to state $\overrightarrow{\mathrm{ST}}_i(13) = (1, 0^{\mathrm{out}}, i^{\mathrm{in}})$ is $3 + i$: $1 + i$ regular and 2 strategic.

For every $i \ge 1$, let $\vec{i}_{18} = (i,\ i,\ \cdots, i)^{\mathrm{T}}$ be an 18-dimensional column vector.Thus, $E[D]$ can be calculated as follows:

$$\begin{aligned}
E[D] &= \vec{p}_0\vec{v}_D + \sum_{i=1}^{2} \vec{p}_i(\vec{u}_D + \vec{i}_{18}) + \sum_{i=3}^{\infty} i\,\vec{p}_i\vec{e}_{18} + \sum_{i=3}^{\infty} \vec{p}_i\vec{u}_D \\
&= \vec{p}_0\vec{v}_D + \sum_{i=1}^{2} \vec{p}_i(\vec{u}_D + \vec{i}_{18}) + \sum_{i=3}^{\infty} i\,\vec{p}_3 R^{i-3}\vec{e}_{18} + \sum_{i=3}^{\infty} \vec{p}_3 R^{i-3}\vec{u}_D \\
&= \vec{p}_0\vec{v}_D + \sum_{i=1}^{2} \vec{p}_i(\vec{u}_D + \vec{i}_{18}) + \vec{p}_3\Big(2[I-R]^{-1} + [I-R]^{-2}\Big)\vec{e}_{18} + \vec{p}_3[I-R]^{-1}\vec{u}_D.
\end{aligned}$$

The following two vectors will be helpful in calculating $E[L]$, the mean number of regular customers in the system. To be consistent, we denote by the letter ν vectors of size 20, and by the letter u vectors of dimension 18. Let

$$\vec{v}_L = (0\ 0\ 0\ 1\ 1\ 2\ 2\ 1\ 1\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0)^{\mathrm{T}}_{20},$$

$$\vec{u}_L = (0\ 0\ 1\ 1\ 2\ 2\ 1\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0)^{\mathrm{T}}_{18}.$$

The vector $\vec{v}_L$ indicates the number of regular customers in the system according to each state in the vector $\overrightarrow{\mathrm{ST}}_0$, while the vector $\vec{u}_L$ corresponds to the vector $\overrightarrow{\mathrm{ST}}_i$, $i \ge 1$, and it indicates the number of regular customers in the system, excluding the i regular customers that arrived after the last strategic customer. Thus, $E[L]$ can be calculated as follows:

$$E[L] = \vec{p}_0\vec{v}_L + \sum_{i=1}^{2}\vec{p}_i(\overrightarrow{u}_L + \overrightarrow{i}_{18}) + \sum_{i=3}^{\infty} i\,\vec{p}_i\vec{e}_{18} + \sum_{i=3}^{\infty}\vec{p}_i\vec{u}_L$$
$$= \vec{p}_0\vec{v}_L + \sum_{i=1}^{2}\vec{p}_i(\overrightarrow{u}_L + \overrightarrow{i}_{18}) + \sum_{i=3}^{\infty} i\,\vec{p}_3R^{i-3}\vec{e}_{18} + \sum_{i=3}^{\infty}\vec{p}_3R^{i-3}\vec{u}_L$$
$$= \vec{p}_0\vec{v}_L + \sum_{i=1}^{2}\vec{p}_i(\overrightarrow{u}_L + \overrightarrow{i}_{18}) + \vec{p}_3\Big(2[I-R]^{-1} + [I-R]^{-2}\Big)\vec{e}_{18} + \vec{p}_3[I-R]^{-1}\vec{u}_L.$$

Finally, in order to calculate $E[S]$, the mean number of strategic customers that are not orbiting, we define the following vectors:

$$\vec{v}_S = (0\,1\,0\,1\,0\,1\,0\,2\,1\,2\,1\,0\,2\,1\,1\,0\,3\,2\,2\,1)^{\mathrm{T}}_{20},$$

$$\vec{u}_S = (0\,1\,1\,0\,1\,0\,2\,1\,2\,1\,2\,1\,1\,0\,3\,2\,2\,1)^{\mathrm{T}}_{18}.$$

The vector $\vec{v}_S$ ($\vec{u}_S$) indicates the number of strategic customers in the system that are not orbiting, according to each state in the vector $\overrightarrow{\mathrm{ST}}_0$ ($\overrightarrow{\mathrm{ST}}_i$, for $i \geq 1$). Thus, $E[S]$ can be calculated as follows:

$$E[S] = \vec{p}_0\vec{v}_S + \sum_{i=1}^{2}\vec{p}_i\overrightarrow{u}_S + \sum_{i=3}^{\infty}\vec{p}_i\vec{u}_S = \vec{p}_0\vec{v}_S + \sum_{i=1}^{2}\vec{p}_i\overrightarrow{u}_S + \sum_{i=3}^{\infty}\vec{p}_3R^{i-3}\vec{u}_S$$
$$= \vec{p}_0\vec{v}_S + \sum_{i=1}^{2}\vec{p}_i\overrightarrow{u}_S + \vec{p}_3[I-R]^{-1}\vec{u}_S.$$

5 Sojourn time of a regular customer

Let $\tilde{\mu}(s) = \mu/(\mu+s)$ be the Laplace–Stieltjes transform (LST) of a single service time (in the sequel we use $\tilde{\mu}$ instead of $\tilde{\mu}(s)$). Let $m = \mu/(\mu+\beta)$ be the probability that a service completion occurs before a RO customer returns to the system. Denote by $\tilde{k}_{a,j}(s) = \tilde{\mu}^a m^j + \tilde{\mu}^{a+1}(1-m^j)$, abbreviated to $\tilde{k}_{a,j}$, the LST of the sojourn time of a regular customer that arrives to the system when there are $(a-1)$ customers waiting in line or in service, and an additional single RO customer. The above follows as the regular customer that has just arrived to the system will either leave the system after a service durations (including his/her own), which will occur if $j \leq a-1$ customers complete their service before the return of the RO (an event that occurs with probability m^j); or, alternatively, this regular customer will leave the system after $(a+1)$ service durations, if the RO returns on time (with the complementary probability $1-m^j$). For the sake of illustration, $\tilde{k}_{2,1}(s)$ describes the case where a regular customer arrives to state $(1, 0^{\mathrm{out}})$, and s/he either completes her/his service after two service durations, which occurs if the customer in service ends her/his service before the return of the RO customer, or, alternatively, s/he completes her/his service after three service durations, which occurs in the case where the RO customer arrived before the start of her/his

service. Note that the LST $\tilde{k}_{2,1}(s)$ applies to other situations that will appear later with respect to the vectors $\vec{v}_Y$ and $\vec{u}_Y$.

Let $m_2 = \mu/(\mu + 2\beta)$ be the probability that a single service is completed before any of two RO customers returns to the system. Let $\tilde{x}(s) = \tilde{\mu}^2 m_2 + \tilde{\mu}^3(1 - m_2)m^2 + \tilde{\mu}^4[(1 - m_2)(1 - m) + (1 - m_2)m(1 - m)]$, abbreviated as $\tilde{x}$, be the LST of a regular customer's sojourn time that arrives to the system when there is no one in line, but the server is busy serving a customer, and there are two RO customers. The new customer will spend two, three or four service durations, until completing her/his service. The end of service of the regular customer will be after two service durations if no RO customer returns to the system before the beginning of service of the regular customer (see the first term in the above expression). The end of service of the regular customer will be after three service durations if only one of the two RO customers returns to the system before the beginning of service of the regular customer (see the second term in the above expression). Finally, the case where the regular customer stays in the system during four service durations, occurs if the two RO customers will not miss their turn (see the third term in the above expression).

Using the above notation, let the following two vectors of dimension 20 and 18, respectively, be defined as follows:

$$\vec{v}_Y = (\,\tilde{\mu}\ \tilde{\mu}^2\ \tilde{\mu}\ \tilde{\mu}^3\ \tilde{k}_{2,1}\ \tilde{\mu}^4\ \tilde{k}_{3,2}\ \tilde{\mu}^4\ \tilde{k}_{3,2}\ \tilde{\mu}^3\ \tilde{k}_{2,1}\ \tilde{\mu}\ \tilde{\mu}^4\ \tilde{k}_{3,2}\ \tilde{k}_{3,1}\ \tilde{x}\ \tilde{\mu}^4\ \tilde{k}_{3,2}\ \tilde{k}_{3,1}\ \tilde{x}\,)^{\mathrm{T}},$$

$$\vec{u}_Y = (\,\tilde{\mu}\ \tilde{\mu}^2\ \tilde{\mu}^3\ \tilde{k}_{2,1}\ \tilde{\mu}^4\ \tilde{k}_{3,2}\ \tilde{\mu}^4\ \tilde{k}_{3,2}\ \tilde{\mu}^3\ \tilde{k}_{2,1}\ \tilde{\mu}^4\ \tilde{k}_{3,2}\ \tilde{k}_{3,1}\ \tilde{x}\ \tilde{\mu}^4\ \tilde{k}_{3,2}\ \tilde{k}_{3,1}\ \tilde{x}\,)^{\mathrm{T}}.$$

The vector $\vec{v}_Y$ depicts the corresponding LSTs of a regular customer's sojourn time according to the states in vector $\overrightarrow{\mathrm{ST}}_0$, while $\vec{u}_Y$ depicts the above LSTs, each divided by $\tilde{\mu}^i$, according to vector $\overrightarrow{\mathrm{ST}}_i$, for $i \geq 1$. By using Eq. (2), we conclude that the LST of a regular customer's sojourn time in the system is given by

$$\begin{aligned}\tilde{Y}(s) &= \vec{p}_0\vec{v}_Y + \sum_{i=1}^{\infty}\tilde{\mu}^i\,\vec{p}_i\vec{u}_Y = \vec{p}_0\vec{v}_Y + \sum_{i=1}^{2}\tilde{\mu}^i\,\vec{p}_i\vec{u}_Y + \sum_{i=3}^{\infty}\tilde{\mu}^i\,\vec{p}_3 R^{i-3}\vec{u}_Y \\ &= \vec{p}_0\vec{v}_Y + \sum_{i=1}^{2}\tilde{\mu}^i\,\vec{p}_i\vec{u}_Y + \tilde{\mu}^3\vec{p}_3\sum_{i=3}^{\infty}(\tilde{\mu}R)^{i-3}\vec{u}_Y = \vec{p}_0\vec{v}_Y + \sum_{i=1}^{2}\tilde{\mu}^i\,\vec{p}_i\vec{u}_Y + \tilde{\mu}^3\vec{p}_3[I - \tilde{\mu}R]^{-1}\vec{u}_Y.\end{aligned}$$

Note that $[I - \tilde{\mu}R]^{-1}$ exists as $\tilde{\mu} \leq 1$.

6 The probability that a strategic customer will be served

In this section, we calculate the probability that a strategic customer will be served. Based on this derivation, the sojourn time's LST of a strategic customer will be calculated in Sect. 7. Recall that a strategic customer who becomes a RO and returns to the waiting room after missing her/his turn abandons the system without being served. Recall also that for a system with $m = 1$ and $n = 3$, a strategic customer draws a ticket if s/he arrives to the system when the virtual queue length is at most two. We consider three cases: (i) the strategic customer, upon arrival, finds the system in one

of the following states: (0), (1), (0, 0^{in}), (0, 0^{out}) or (0, 0^{out}, 0^{out}). Under any of these states, the strategic customer joins the system upon arrival as s/he finds a virtual queue length of one, and is served with probability one. (ii) The strategic customer arrives when there are two customers in the system but none of them is in orbit, as is the case with the states (2), (0, 1^{in}), (1, 0^{in}) and (0, 0^{in}, 0^{in}). In such cases, our customer will go out orbiting and will be served if one of the following cases happens: (a) the customer returns before any termination of service; (b) the customer returns after the first service has been completed but before the end of service of the second; (c) the customer returns after both customers completed their service, but before a new arrival occurs. (iii) Our strategic customer arrives to the system when there is one customer in service, hereafter called the first customer, and another in orbit, hereafter called RO, which can happen in the following two states: (1, 0^{out}) and (0, 0^{in}, 0^{out}). Our strategic customer will then go out orbiting and will finally be served if one of the following cases happens: (a) our customer finds out, upon her/his return, that no customer has completed her/his service, i.e., the orbiting duration of our customer was shorter than the service duration; (b) our customer returns from orbiting after the service completion of the first customer, but before the return of the RO customer that precedes her/his, and also before any arrival of a new customer, implying that the server is idle upon her/his return; (c) the RO customer that precedes our customer returns to the system before the first customer has completed her/his service, and our customer returns before the service completion of the RO; (d) the service of the first customer was completed, then the RO customer that precedes our customer, returns, and then, our customer returns before the service of the RO is completed; (e) the service of the first customer was completed, the RO customer that precedes our customer, returns and completes her/his service, and then our customer returns, but all this happens before any new arrival of customers; (f) the RO that precedes our customer returns while the first customer is still being served, then the services of these two customers are completed, and then our customer returns, where all of that occurs before a new customer arrives. Taking into account all the above possibilities, we now calculate the probability that a strategic customer ends up getting service. Let

$$
\begin{aligned}
p_{s1} = {} & p_0 + p_{0,0^{\text{out}}} + p_{0,0^{\text{out}},0^{\text{out}}} + p_1 + p_{0,0^{\text{in}}} \\
& + (p_2 + p_{0,1^{\text{in}}} + p_{1,0^{\text{in}}} + p_{0,0^{\text{in}},0^{\text{in}}}) \left(\frac{\beta}{\beta+\mu} + \frac{\mu}{\beta+\mu}\frac{\beta}{\beta+\mu} + \frac{\mu}{\beta+\mu+\lambda}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha} \right) \\
& + (p_{1,0^{\text{out}}} + p_{0,0^{\text{in}},0^{\text{out}}}) \left(\begin{array}{r} \frac{\beta}{\beta+\mu} + \frac{\mu}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda+\alpha} + \frac{\beta}{2\beta+\mu}\frac{\mu}{\beta+\mu}\frac{\beta}{\beta+\mu} \\ + \frac{\mu}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda+\alpha}\frac{\beta}{\beta+\mu} \\ + \frac{\mu}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda+\alpha}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha} \\ + \frac{\beta}{2\beta+\mu+\lambda}\frac{\mu}{\beta+\mu+\lambda}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha} \end{array} \right)
\end{aligned}
$$

be the probability that an arriving strategic customer obtains service. Also, let $p_{s2} = p_0 + p_{0,0^{\text{out}}} + p_{0,0^{\text{out}},0^{\text{out}}} + p_1 + p_{0,0^{\text{in}}} + p_2 + p_{0,1^{\text{in}}} + p_{1,0^{\text{in}}} + p_{0,0^{\text{in}},0^{\text{in}}} + p_{1,0^{\text{out}}} + p_{0,0^{\text{in}},0^{\text{out}}}$ be the probability that an arriving strategic customer joins the system, i.e., there are

at most two customers in the system. Then, the probability that a strategic customer who joins the system gets service is

$$p_s = \frac{p_{s1}}{p_{s2}}. \tag{3}$$

7 Sojourn time of a strategic customer

We first derive the LST of a strategic customer's sojourn time while waiting in line or while being served, and then the LST of her/his total sojourn time. For a system with $m = 1$ and $n = 3$, a strategic customer spends a single service duration if s/he arrives while the server is idle, and otherwise s/he spends two or three service durations. The calculation of the LST of the sojourn time of a strategic customer, excluding the orbiting duration, denoted by $\tilde{W}_1(s)$, is based on the cases detailed in Sect. 6, while case (iii) (a) is divided into two sub-cases.

$$\begin{aligned}
\tilde{W}_1(s) = {} & (p_0 + p_{0,0^{\text{out}}} + p_{0,0^{\text{out}},0^{\text{out}}})\tilde{\mu} + (p_1 + p_{0,0^{\text{in}}})\tilde{\mu}^2 \\
& + (p_2 + p_{0,1^{\text{in}}} + p_{1,0^{\text{in}}} + p_{0,0^{\text{in}},0^{\text{in}}})\left(\frac{\beta\tilde{\mu}^3}{\beta+\mu} + \frac{\mu\tilde{\mu}^2}{\beta+\mu}\frac{\beta}{\beta+\mu} + \frac{\mu\tilde{\mu}}{\beta+\mu+\lambda}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha}\right) \\
& + (p_{1,0^{\text{out}}} + p_{0,0^{\text{in}},0^{\text{out}}})\left(\begin{array}{l}
\frac{\beta\tilde{\mu}^2}{2\beta+\mu}\frac{\mu}{\beta+\mu} + \frac{2\beta\tilde{\mu}^3}{2\beta+\mu}\frac{\beta}{\beta+\mu} \\
\quad + \frac{\mu\tilde{\mu}}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda+\alpha} + \frac{\beta\tilde{\mu}^2}{2\beta+\mu}\frac{\mu}{\beta+\mu}\frac{\beta}{\beta+\mu} \\
\quad + \frac{\mu\tilde{\mu}^2}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda+\alpha}\frac{\beta}{\beta+\mu} \\
\quad + \frac{\mu\tilde{\mu}}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha} \\
\quad + \frac{\beta\tilde{\mu}}{2\beta+\mu+\lambda}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha}
\end{array}\right).
\end{aligned}$$

Consequently, the LST of a strategic customer that has joined the system is

$$\tilde{W}(s) = \frac{\tilde{W}_1(s)}{p_{s1}}.$$

Let $\tilde{\beta}(s) = \beta/(\beta + s)$, abbreviated as $\tilde{\beta}$, be the LST of the time a strategic customer spends in orbit. In order to simplify the presentation, we use the following two constants:

$$\Theta = \left(\frac{\beta}{\beta+\mu} + \frac{\mu}{\beta+\mu}\frac{\beta}{\beta+\mu} + \frac{\mu}{\beta+\mu+\lambda}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha}\right)$$

and

$$\Psi = \left(\begin{array}{l} \frac{\beta}{\beta+\mu} + \frac{\mu}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda+\alpha} + \frac{\beta}{2\beta+\mu}\frac{\mu}{\beta+\mu}\frac{\beta}{\beta+\mu} \\ \quad + \frac{\mu}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda+\alpha}\frac{\beta}{\beta+\mu} \\ \quad + \frac{\mu}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda+\alpha}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha} \\ \quad + \frac{\beta}{2\beta+\mu+\lambda}\frac{\mu}{\beta+\mu+\lambda}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha} \end{array} \right).$$

Then, the LST of a strategic customer's total sojourn time from his/her arrival instant, until her/his departure, including orbit time, is

$$\begin{aligned} \tilde{T}_1(s) \quad &= (p_0 + p_{0,0^{\text{out}}} + p_{0,0^{\text{out}},0^{\text{out}}})\tilde{\mu} + (p_1 + p_{0,0^{in}})\tilde{\mu}^2 \\ &+ (p_2 + p_{0,1^{\text{out}}} + p_{1,0^{\text{out}}} + p_{0,0^{\text{out}},0^{\text{out}}})\tilde{\beta}\left(\frac{\beta\tilde{\mu}^3}{\beta+\mu} + \frac{\mu\tilde{\mu}^2}{\beta+\mu}\frac{\beta}{\beta+\mu} + \frac{\mu\tilde{\mu}}{\beta+\mu+\lambda}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha} + (1-\Theta)\right) \\ &+ (p_{1,0^{\text{out}}} + p_{0,0^{\text{out}},0^{\text{out}}})\tilde{\beta}\left(\begin{array}{l} \frac{\beta\tilde{\mu}^2}{2\beta+\mu}\frac{\mu}{\beta+\mu} + \frac{2\beta\tilde{\mu}^3}{2\beta+\mu}\frac{\beta}{\beta+\mu} \\ \quad + \frac{\mu\tilde{\mu}}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda+\alpha} + \frac{\beta\tilde{\mu}^2}{2\beta+\mu}\frac{\mu}{\beta+\mu}\frac{\beta}{\beta+\mu} \\ \quad + \frac{\mu\tilde{\mu}^2}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda+\alpha}\frac{\beta}{\beta+\mu} \\ \quad + \frac{\mu\tilde{\mu}}{2\beta+\mu+\lambda}\frac{\beta}{2\beta+\lambda+\alpha}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha} \\ \quad + \frac{\beta\tilde{\mu}}{2\beta+\mu+\lambda}\frac{\mu}{\beta+\mu+\lambda}\frac{\mu}{\beta+\mu+\lambda+\alpha}\frac{\beta}{\beta+\lambda+\alpha} + (1-\Psi) \end{array} \right). \end{aligned}$$

Finally, the LST of the total sojourn time of a strategic customer who joins the system is

$$\tilde{T}(s) = \frac{\tilde{T}_1(s)}{p_{s2}}. \tag{4}$$

8 How long to stay in orbit?

A strategic customer who goes out to orbit does not know for sure when s/he will return to the system. It is of interest to determine the economic optimal mean value of the orbit duration $1/\beta$. An analytical determination of the optimal mean orbit time seems to be intractable even under the assumption of exponential orbit time and a simple economic model. For this reason, we analyze below two numerical objective functions that exhibit how the optimal value β^{-1} can be obtained numerically.

Objective function 1 Let $R_{orb} = g(1 - e^{-\xi X})$ be the accumulated reward for an orbit duration of X units of time ($\xi > 0$), which represents a diminishing marginal reward as a function of the orbit duration. Under the assumption of an exponential orbit time, $X \sim \exp(\beta)$, the mean orbiting reward is given by

$$E[R_{\text{orb}}] = \frac{g\xi}{\xi + \beta}. \tag{5}$$

For the particular case of $m = 1$ and $n = 3$, the probability that a strategic customer who joins the system goes to orbit is given by

$$p_{\text{orb}} = \frac{p_2 + p_{0,1^{\text{in}}} + p_{1,0^{\text{in}}} + p_{0,0^{\text{in}},0^{\text{in}}} + p_{1,0^{\text{out}}} + p_{0,0^{\text{in}},0^{\text{out}}}}{p_{s2}}.$$

Assume that (i) the monetary worth of the service offered by the system is r, and (ii) the opportunity cost rate for the time from the arrival instant until the departure instant of a strategic customer is $c > 0$. This cost applies to the orbiting time, the waiting time and the service time. Hence, the expected total reward function of a strategic customer is defined by

$$G(\beta) = rp_s + E[R_{\text{orb}}]p_{\text{orb}} - cE[T]. \tag{6}$$

The effect of β The graph of $G(\beta)$, depicted in Fig. 2, shows the effect of β on the expected total reward when the other parameters are $\lambda = 8$, $\mu = 10$, $\alpha = 9$, $\xi = 10$, $r = 10$, $g = 10$ and $c = 1$.

It shows that $G(\beta)$ is a uni-modal function with a maximum value $G(\beta^*) = 11.1$ at $\beta^* = 12.1$, it increases rapidly as β goes from 0 to $\beta = 12.1$ and then decreases moderately in a convex fashion. This implies that the optimal mean orbiting time is 1/12.1 h $\approx$ 5 min.

A further sensitivity analysis of β^* with respect to each of the monetary parameters, namely r, g and $\boldsymbol{\xi}$, is given below. Figures 3, 4 and 5 show, respectively, the dependence of β^* on each of the monetary parameters:

(a) $\boldsymbol{\beta^*}$ **as a function of** $\boldsymbol{r}$ Figure 3 shows that β^* is an increasing function (almost linearly) of r. That is, the larger is the value of the service, the shorter is the expected orbit duration.

(b) $\boldsymbol{\beta^*}$ **as a function of** $\boldsymbol{g}$ Figure 4 shows that β^* is a convex decreasing function of g. That is, the higher is the orbiting reward, the higher is the mean orbiting time.

(c) $\boldsymbol{\beta^*}$ **as a function of** $\boldsymbol{\xi}$ Figure 5 is the most interesting among the three figures: for ξ between 5 and 11, β^* is a convex decreasing function, implying that the optimal mean orbiting time is increasing. At $\xi = 11$, β^* starts to be an increasing function of ξ, namely the optimal mean orbiting time is decreasing. This phenomenon suggests that an increase of ξ above 11 implies a too low marginal contribution to the orbit reward, which does not compensate for the increasing risk of missing one's turn. Thus, for $\xi > 11$, the mean orbiting time is slowly decreasing.

Objective function 2 Suppose that, in addition to the opportunity cost considered in objective function 1, a strategic customer has a deadline, denoted by ω, for her/his total sojourn time, so that, if the deadline is missed, a penalty of η monetary units is incurred by the customer. Thus, the expected total reward is now given by

$$Z(\beta) = rp_s + E[R_{\text{orb}}]p_{\text{orb}} - cE[T] - \eta \int_{\omega}^{\infty} f_T(t)\mathrm{d}t, \tag{7}$$

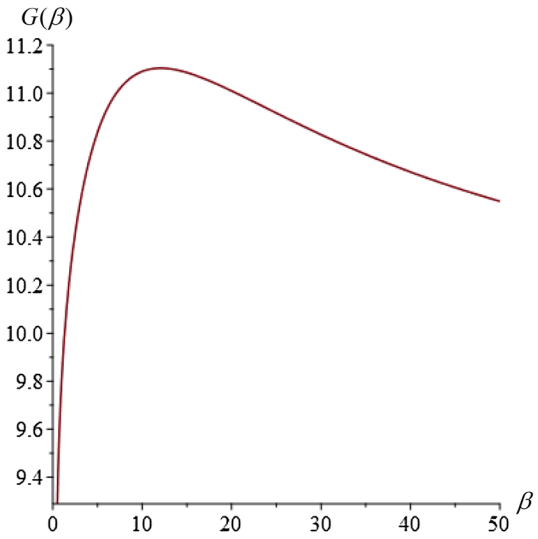

Fig. 2 The expected reward $G(\beta)$, as a function of β for $\lambda = 8$, $\mu = 10$, $\alpha = 9$, $\xi = 10$, $g = 10$ and $c = 1$

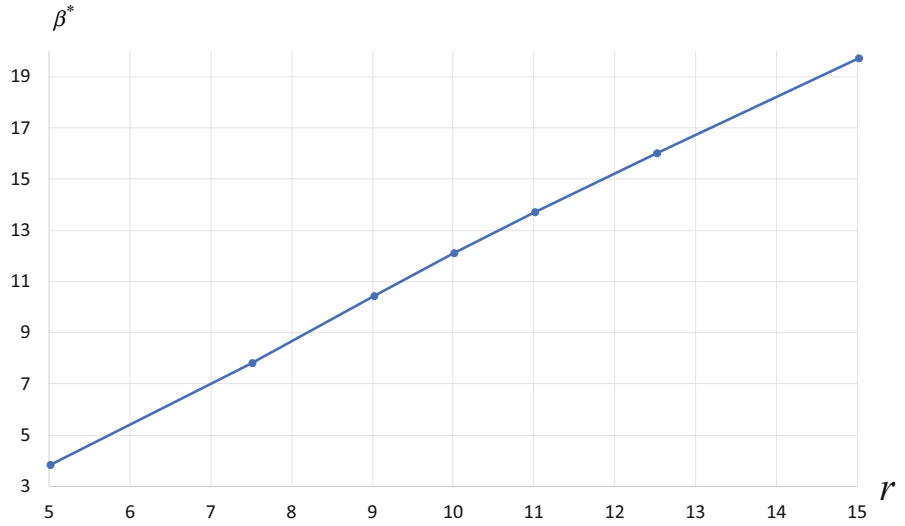

Fig. 3 Optimal β as a function of the worth of service r

where $f_T(t)$ is the probability density function of the total sojourn time of a strategic customer, consisting of orbiting, waiting and service time, up to her/his ultimate departure from the system. This density is calculated as the inverse of the LST $\tilde{T}(s)$ (see Eq. (4)). Figure 6, which corresponds to the parameter values of Fig. 2, depicts the probability that a strategic customer fails to meet her/his deadline, as a function of β, for $\omega = 0.3$. The graph shows the following interesting phenomenon: as β increases beyond $\beta = 5$, or alternatively, the mean orbit time $1/\beta$ decreases below 0.2, the probability of missing the deadline decreases, but this is just until $\beta = 10$, and thereafter, the probability of missing the deadline is weakly increasing. The reason for

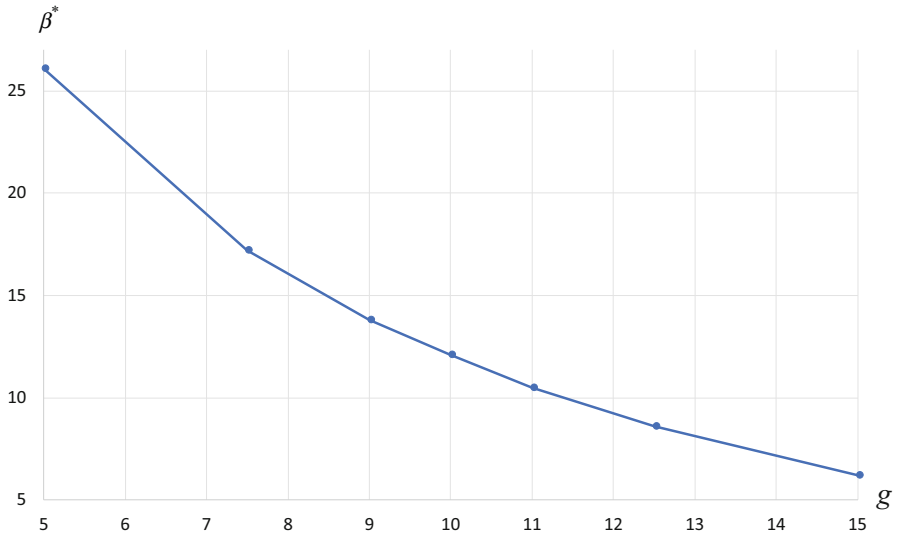

Fig. 4 Optimal β as a function of g

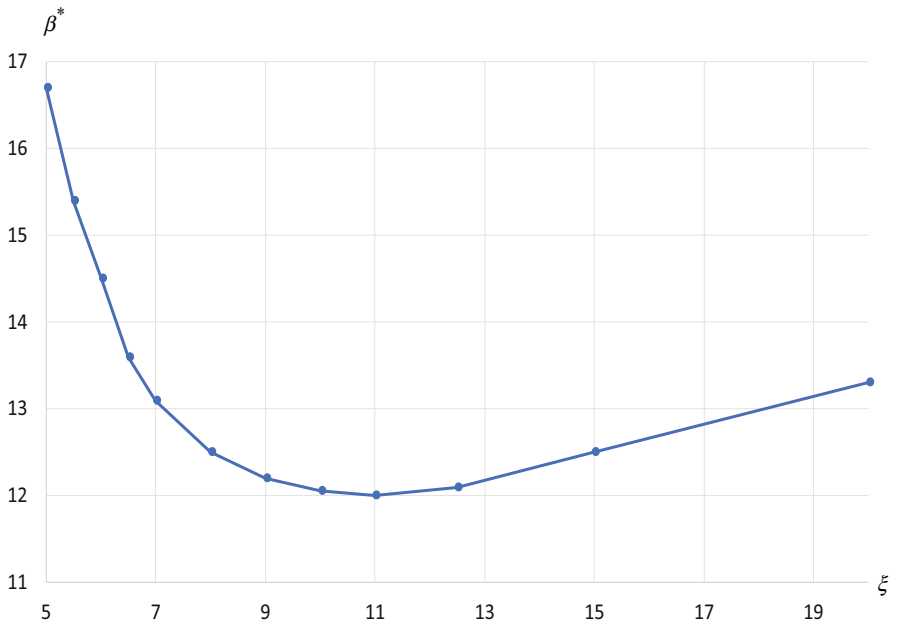

Fig. 5 Optimal β as a function of the diminishing marginal reward $\xi \in (5, 20)$

this peculiar behavior is the fact that initially, when β is small enough, the orbiting time is long, and upon the end of the orbiting duration, it is most probable that the customer misses her/his turn for getting served, so the customer quits the system at the moment s/he returns from orbit, implying that the total sojourn time in the interval

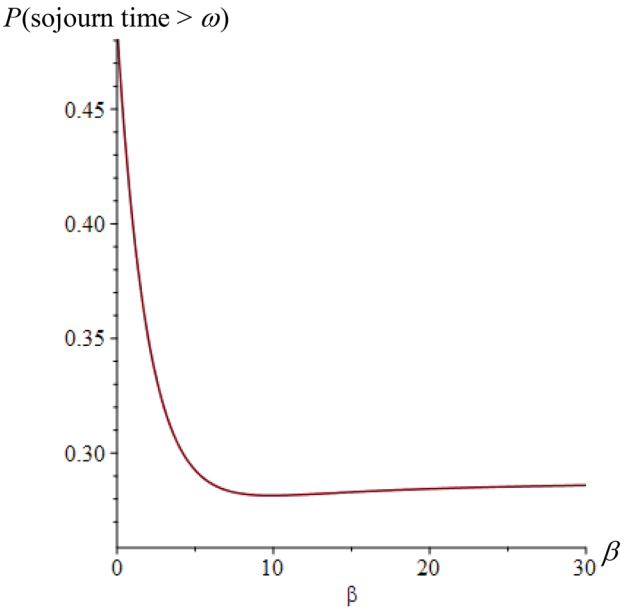

Fig. 6 The probability that a strategic customer misses the deadline $\omega = 0.3$ as a function of β

where $\beta < 10$ consists mainly of the orbit time, implying that the shorter the orbit time in this interval, the smaller the probability of missing the deadline. However, as β continues to increase beyond 10 it becomes more probable that a customer returns to the system on time to get service after orbiting. Thus, as β increases, the total sojourn time of such a customer consists not only of his/her orbit time, but also of the time spent in the waiting room and in service, implying that the probability of missing the deadline is not monotone decreasing in β.

(d) **Effect of the penalty** η The graphs of β^* as function of η for two values of the orbiting rewards, $g = 15$ and $g = 10$, are depicted in Fig. 7.

$$P(\text{sojourn time} > \omega)$$

As can be seen, the lower curve ($g = 15$) is increasing, but always below the upper curve ($g = 10$), which is decreasing. In particular, this implies that the optimal orbit duration for $g = 15$ is strictly higher than that for $g = 10$, independently of η, and that for $g = 15$, $\beta^* < 10$, while for $g = 10$, $\beta^* > 10$.

For $g = 15$, the reward from orbiting is significantly higher than the worth of service r, so a customer is mainly interested in maximizing the orbit reward at the cost of missing his/her turn, as long as the deadline is not missed. This region corresponds to the initial decreasing part of the curve in Fig. 6, i.e., values of β^* below 10.

However, for $g = 10$, the reward from orbit does not compensate for missing one's turn for service, so the customers tend to spend less time in orbit (β^* is above 10) and return on time to get service.

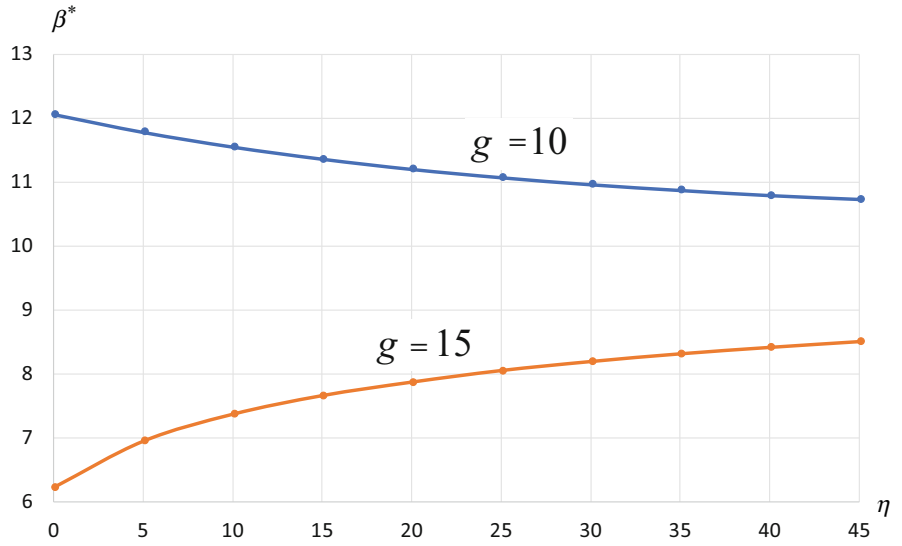

Fig. 7 β^* as a function of η for $g = 10$ and for $g = 15$

The shapes of the curves in Fig. 7 are explained as follows: as the penalty η increases, customers aim at decreasing the probability of missing the deadline, namely decreasing $P(\text{sojourn time} > \omega)$. Figure 6 shows that achieving this goal depends on the value of β. In the range 0 to 10, larger values of β decrease the above probability, while above 10 the opposite occurs. Those insights explain the curves in Fig. 7. When g $= 15$, β^* is smaller than 10, implying that decreasing of $P(\text{sojourn time} > \omega)$ is achieved by increasing β^*. But, when g $= 10$, β^* is higher than 10, and thus, decreasing $P(\text{sojourn time} > \omega)$ is achieved by decreasing β^*.

Figure 8 depicts the strategic customer's expected reward, $Z(\beta^*)$, as a function of η, for $g = 10$. The curve is approximately linear, decreasing in η, and even assumes negative values when the penalty is high ($\eta > 40$), implying that the optimal strategy for such strategic customers is to balk upon arrival to the system.

9 Conclusions

The intriguing ticket queue problem is investigated assuming a nonhomogeneous population of customers that consists of two types: (i) regular customers that draw a ticket regardless of the queue length and stay in line until being served; and (ii) strategic customers that first observe both the displayed running number of the last customer being served, and the number of the ticket that s/he can draw from the take-a-number machine, and then, depending on the difference between the two numbers, namely the virtual queue length, s/he chooses one of the following three decisions: (a) balking; (b) drawing a ticket and joining the line as a regular customer; (c) drawing a ticket, and then going to orbit for a random duration, i.e., leaving the system temporarily. We analyze this involved stochastic system by using matrix geometric methods, derive the

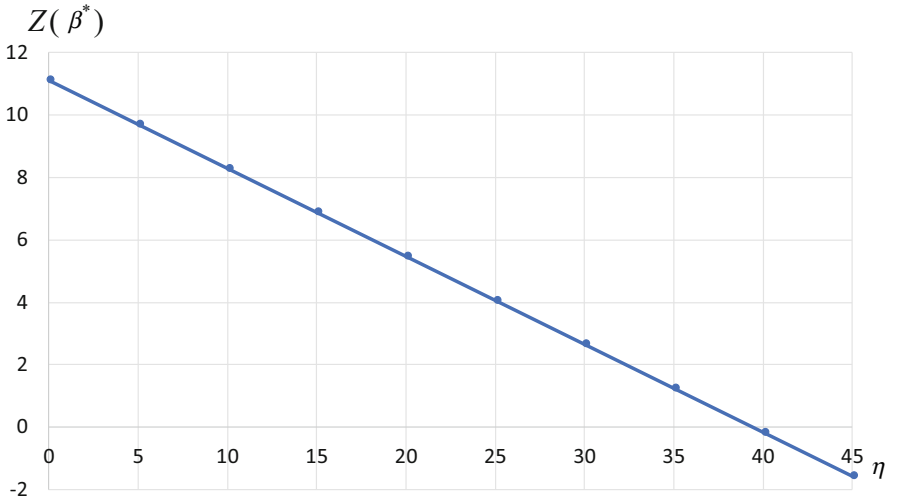

Fig. 8 The optimal expected reward, $Z(\beta^*)$, as a function of η for $g = 10$

system's steady-state probabilities, calculate the LST of a regular customer's sojourn time, the probability that a strategic customer is served, and the LST of the latter's sojourn time. As a consequence, the system's performance measures are constructed. Based on those results, an economic optimal mean orbiting time of a strategic customer is calculated.

Acknowledgements The research of the first and the third authors has been supported by the Israel Science Foundation, Grant No. 1448/17. The research of the first and second author has been supported by the Israel Science Foundation, Grant No. 338/15. The research of the second author was also supported by the Coller Foundation and by the Henry Crown Israeli Institute for Business Research.

Appendix: How to construct diagram $(m, n + 1)$ given diagram (m, n)?

Each node in the diagram is represented by a string. A node having a string of length j represents a state with $j - 1$ strategic customers. Thus, in a state represented by a node of the form (i), $i \geq 0$, there are no strategic customers. Observe from the form of Fig. 1 that all the nodes in any row $i \geq 0$ of diagram (m, n), have i regular customers that have arrived after the arrival of the last strategic customer.

It is quite easy to transform a given diagram (m, n), $m \geq 1$, $n \geq 3$ and $m < n - 1$, to a diagram $(m + 1, n)$ as the number of nodes is independent of m; see Proposition 1. The only changes required are to redirect some of the arcs of the diagram. We skip the details of the changes needed in the diagram for increasing m, as they are straightforward.

In what follows, we provide a forward recursion that builds diagram $(m, n + 1)$ from any given diagram (m, n), where $m < n - 1$, $m \geq 1$, $n \geq 3$ and $m + n > 4$. The nodes of diagram (m, n) in row $i \geq 0$ that need to be augmented when increasing n by 1 are the nodes for which a new incoming strategic customer has to quit the system

upon arrival as $D = n$, even if the last i regular customers that joined the system are totally removed. Alternatively, in a node that can be augmented, the last strategic customer of its string is preceded by $n - 1$ customers of any type. We call such a node an *augmentable node*. As will be shown below, the augmentation process does not alter the length of the strings that are generated from the augmentable nodes.

In addition to the augmentation process, we need add some new nodes to the diagram $(m, n + 1)$ that are not generated by the augmentable nodes of diagram (m, n). These new nodes are represented by strings of length $n + 2$.

Next, we describe in detail the two steps for generating the diagram for $(m, n + 1)$ from diagram (m, n):

1. New nodes obtained by augmentation: consider the augmentable nodes of row $i \geq 0$ in diagram (m, n) whose string is of length j, $1 < j \leq n + 1$: augment each of these nodes by adding $j - 1$ new nodes of the same length j, where the string that is associated with each augmented node is the same as the string of the node it is augmented from, except that the number in one of the first $j - 1$ locations is increased by 1, without affecting the "in/out" status of the strategic customers. For example, the node $(10^{\text{in}}1^{\text{out}})$ is augmented by two nodes, namely $(20^{\text{in}}1^{\text{out}})$ and $(11^{\text{in}}1^{\text{out}})$. Thus, from each augmentable node in diagram (m, n) that is represented by a string of length $1 < j \leq n + 1$, we get $j - 1$ augmented nodes. For example, in row i of the diagram for $n = 3$, and for $j = 2$, only nodes $(2i^{\text{in}})$ and $(2i^{\text{out}})$ are augmentable. Node $(2i^{\text{in}})$ is augmented by node $(3i^{\text{in}})$, and node $(2i^{\text{out}})$ is augmented by node $(3i^{\text{out}})$.

   Note that, in any row $i \geq 1$, there are i regular customers after the last strategic customer. If the first element of a string in row i of diagram (m, n) is 0, then the second element must be "in", as otherwise the next existing customer (regular or "in") will get served immediately, while throwing away all the "out" strategic customers that were bypassed. Thus, in such a case, we complete the described augmentation of a node by generating one additional augmented string that its first element is 1, the second element is changed to "out", and the remaining elements of the string are unchanged with respect to the augmentable node of diagram (m, n).

We should pay attention that the augmentation process may generate a duplication of augmented nodes as shown in the sequel: first, augment in diagram $(m, n) = (1, 3)$ the augmentable node $(01^{\text{in}}0^{\text{out}})$, which results in the augmented nodes $(11^{\text{in}}0^{\text{out}})$ and $(02^{\text{in}}0^{\text{out}})$. Second, augment the augmentable node $(10^{\text{in}}0^{\text{out}})$, which results in the augmented nodes $(20^{\text{in}}0^{\text{out}})$ and $(11^{\text{in}}0^{\text{out}})$. Thus, the augmented node $(11^{\text{in}}0^{\text{out}})$ is generated twice. More specifically, with respect to diagram (1,3), each of the augmentable nodes $(10^{\text{in}}i^{\text{out}})$, $(10^{\text{in}}i^{\text{in}})$, $(10^{\text{out}}i^{\text{out}})$ and $(10^{\text{out}}i^{\text{in}})$ for $i \geq 0$, is augmented by two nodes, where, for example, node $(10^{\text{in}}i^{\text{out}})$ is augmented by $(20^{\text{in}}i^{\text{out}})$ and by $(11^{\text{in}}i^{\text{out}})$. At the end, eight nodes are augmented by the above four nodes of diagram (1, 3). Consequently, at the end of the process, duplicated nodes should be removed.

By applying the same process to the augmentable nodes $(01^{\text{in}}i^{\text{out}})$, $(01^{\text{in}}i^{\text{in}})$ in diagram (1, 3), we get an additional two new augmentation nodes $(02^{\text{in}}i^{\text{out}})$ and $(02^{\text{in}}i^{\text{in}})$ in addition to two more nodes that are duplicates, namely nodes

$(11^{\text{in}}i^{\text{out}})$ and $(11^{\text{in}}i^{\text{in}})$ that have already been generated. In addition, the strings of the augmentable nodes $(01^{\text{in}}i^{\text{out}})$, $(01^{\text{in}}i^{\text{in}})$ have 0 at the beginning of the string, and therefore they can be also augmented by $(11^{\text{out}}i^{\text{out}})$ and $(11^{\text{out}}i^{\text{in}})$, but these nodes have already been augmented by $\left(10^{\text{out}}i^{\text{out}}\right)$ and $\left(10^{\text{out}}i^{\text{in}}\right)$, respectively. Thus, the augmentation of the nodes $(01^{\text{in}}i^{\text{out}})$ and $(01^{\text{in}}i^{\text{in}})$ results in just two new augmentation nodes.

Finally, the augmentable nodes $(00^{\text{in}}0^{\text{in}}0^{\text{in}})$, $(00^{\text{in}}0^{\text{in}}0^{\text{out}})$, $(00^{in}0^{out}0^{in})$ and $(00^{\text{in}}0^{\text{out}}0^{\text{out}})$ generate four augmentation nodes for each of them. Three are obtained by increasing by 1 the value of one of the first three locations. The last is obtained by increasing the first 0 to 1, and changing the second 0 to an "out". For example, $(00^{\text{in}}0^{\text{in}}0^{\text{in}})$ is augmented by $(10^{\text{in}}0^{\text{in}}0^{\text{in}})$, $(01^{\text{in}}0^{\text{in}}0^{\text{in}})$, $(00^{\text{in}}1^{\text{in}}0^{\text{in}})$ and $(10^{\text{out}}0^{\text{in}}0^{\text{in}})$. Thus, we get 16 such augmented nodes.

In total, we have 28 augmentation nodes added to diagram (1, 3).

2. New nodes. In any diagram (m, n), $m < n-1$, $m \geq 1$, $n \geq 3$, the longest string of nodes is of size $n+1$. Such strings consist of $n+1$ zeros, where the second zero is "in", and the following $n-1$ zeros can be in any combination of "in" or "out". In diagram $(m, n+1)$, the longest string is of length $n+2$. Thus, the total number of new nodes is 2^n.

In conclusion of the presentation of the rows' construction in diagram (1, 4) from that of diagram (1, 3): (i) all nodes in diagram (1, 3) appear in diagram (1, 4), (ii) in each row $i \geq 0$, of diagram (1, 3), 28 augmentation nodes will be added. (iii) 8 new nodes, each having a string 5 zeros, are added to the set of nodes. Thus, each row of diagram (1, 4) consists of 54 nodes, namely $18 + 28 + 8 = 54 = 2 \cdot 3^{4-1}$ nodes.

Construction of the matrices in Q The matrices A_0 and A_2 in the new diagram $(m, n+1)$ will be similar to those corresponding to diagram (m, n). Namely, A_0 and A_2 are each of order $f = 2 \cdot 3^n$: $A_0 = \lambda I_{f \times f}$ and $A_2 = \left[a_2^{i,j}\right]_{f \times f}$, where $a_2^{i,j} = \begin{cases} \mu & i = j = 1. \\ 0 & \text{otherwise.} \end{cases}$. The matrix A_1 is also of order $f \times f$, where $a_1^{i,j} = \begin{cases} -(\lambda + \mu) & i = j = 1, 2. \\ b_{1,1}^{i,j} & \text{otherwise.} \end{cases}$. The entries $\left[b_{1,1}^{i,j}\right]_{f \times f}$ are constructed by adding a column for each augmented node as described above.

References

1. Adiri, I., Yechiali, U.: Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. Oper. Res. **22**(5), 1051–1066 (1974)
2. Ding, D., Ou, J., Ang, J.: Analysis of ticket queues with reneging customers. J. Oper. Res. Soc. **66**(2), 231–246 (2015)
3. Guha, D., Goswami, V., Banik, A.D.: Algorithmic computation of steady-state probabilities in an almost observable GI/M/c queue with or without vacations under state dependent balking and reneging. Appl. Math. Model. **40**(5), 4199–4219 (2016)
4. Hanukov, G., Yechiali, U.: Further relationships between the probability generating functions method and explicit matrix geometric solutions in continuous-time QBD processes. Submitted for publication. (2018)

5. Hanukov, G., Avinadav, T., Chernonog, T., Spiegel, U., Yechiali, U.: A queueing system with decomposed service and inventoried preliminary services. Appl. Math. Model. **47**, 276–293 (2017)
6. Hanukov, G., Avinadav, T., Chernonog, T., Spiegel, U., Yechiali, U.: Improving efficiency in service systems by performing and storing "preliminary services". Int. J. Prod. Econ. **197**, 174–185 (2018)
7. Hanukov, G., Avinadav, T., Chernonog, T., Yechiali, U.: Performance Improvement of a service system via stocking perishable preliminary services. Eur. J. Oper. Res. **274**(3), 1000–1011 (2018)
8. Hassin, R.: Rational queueing. Taylor & Francis Group LLC., Routledge (2016)
9. Jennings, O.B., Pender, J.: Comparisons of ticket and standard queues. Queueing Syst. **84**, 145–202 (2016)
10. Kerner, Y., Sherzer, E., Yanco, M.A.: On non-equilibria threshold strategies in ticket queues. Queueing Syst. **86**, 419–431 (2017)
11. Kuzu, K.: Comparisons of perceptions and behavior in ticket queues and physical queues. Serv. Sci. **7**(4), 294–314 (2015)
12. Kuzu, K., Gao, L., Xu, S.H.: To wait or not to wait: the theory and practice of ticket queues. Manuf. Serv. Oper, Manag (2019)
13. Levy, Y., Yechiali, U.: Utilization of idle time in an M/G/1 queueing system. Manag. Sci. **22**, 202–211 (1975)
14. Levy, Y., Yechiali, U.: An M/M/s queue with servers' vacations. INFOR **14**, 153–163 (1976)
15. Mytalas, G.C., Zazanis, M.A.: An M^X/G/1 queueing system with disasters and repairs under a multiple adapted vacation policy. Nav. Res. Logist. **62**, 171–189 (2015)
16. Naor, P.: The regulation of queue size by levying tolls. Econ J Econ. Soc. **37**, 15–24 (1969)
17. Neuts, M.F.: Matrix-geometric solutions in stochastic models: an algorithmic approach. Johns Hopkins University Press, Baltimore (1981)
18. Ramswami, V., Latouche, G.: A general class of Markov processes with explicit matrix-geometric solutions. Oper. Res. Spektrum **8**(4), 209–218 (1986)
19. Xu, S.H., Gao, L., Ou, J.: Service performance analysis and improvement for a ticket queue with balking customers. Manag. Sci. **53**(6), 971–990 (2007)
20. Yang, D.Y., Wu, C.H.: Cost-minimization analysis of a working vacation queue with N-policy and server breakdowns. Comput. Ind. Eng. **82**, 151–158 (2015)
21. Yechiali, U.: Customers' optimal joining rules for the GI/M/s queue. Manag. Sci. **18**(7), 434–443 (1972)
22. Yechiali, U.: On the M^X/G/1 queue with a waiting server and vacations. Sankhya **66**, 159–174 (2004)