Applied Mathematical Modelling 105 (2022) 355–374

Contents lists available at ScienceDirect

Applied Mathematical Modelling

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

Individual and social customers' joining strategies in a two-stage service system when discount is offered to users of smartphone application

Gabi Hanukov[a,*], Uri Yechiali[b]

[a] Department of Industrial Engineering, Ariel University, Ariel 40700, Israel
[b] Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 6997801, Israel

ARTICLE INFO

Article history:
Received 20 April 2021
Revised 3 December 2021
Accepted 28 December 2021
Available online 12 January 2022

Keywords:
Queueing models
Omnichannel
Strategic customers
Smartphone application
Servers vacations

ABSTRACT

The use of smartphone applications (APP) for the purchase of services, such as food, flights or household goods, is becoming an increasingly common practice. This practice generates two- phase service systems composed of (i) an ordering phase operated by several servers (under a dynamic vacation policy), and (ii) a following preparation phase. Upon arrival, strategic customers can either join an observable first-stage queue and then continue to a partially unobservable second-stage queue, or balk never to return. In contrast, APP users register their order in advance, skip the first phase and join directly the second-stage queue. Each server stationed at the first service stage takes a 'vacation' when a server-dependent queue size drops below a given value. An individual strategic customer is concerned only with his/her net utility, while a social customer is concerned with the overall social welfare. Each type of customers follows a threshold-joining policy. However, contrary to Naor's seminal model, we show that under such a scenario, the social joining threshold is not always smaller than the individual one, and that several equilibrium thresholds may exist. We analyze the profit optimization problem of the service system's manager when a price discount is offered to potential APP users and show that it may increase manager's profit, reduce customers' sojourn times, reduce the first-stage queue size, and reduce the number of servers required to operate the first-stage queue. Numerical examples are presented.

1. Introduction

Ordering goods or services via a website or smartphone application is becoming an increasingly common practice. Typical on-line purchases include fast food, flight tickets, clothing, household goods and a variety of other products and services. Many retailers have chosen to integrate their conventional service operations with electronic service ordering, thus improving operational efficiency for both the retailer and the customers. Consequently, there is a growing focus on "Omni channel retailing", to such an extent that it has been characterized as the "future of shopping" in recent literature (see, e.g., [1–3]). There is good evidence that the integration of service channels, where customers buy a product online and pick it up in store (BOPS), results in higher sales (see, e.g., [4,5]). For additional related works on BOPS the reader is referred to [6–8].

[*] Corresponding author.
E-mail addresses: gabih@ariel.ac.il (G. Hanukov), uriy@tauex.tau.ac.il (U. Yechiali).

https://doi.org/10.1016/j.apm.2021.12.038
0307-904X/© 2022 Elsevier Inc. All rights reserved.

Many restaurants and coffee shops, such as McDonalds, KFC, Subway, Starbucks, and Aroma, have sought to realize the power of the Internet by providing an online self-service facility. This is usually done by developing a mobile application that allows customers to place an order and pay while using their mobile devices. In this way, customers can avoid the ordering queue and need only wait in the restaurant for their order to be prepared. Baron et al. [9] studied a system that implements such self-ordering technologies for ‘strategic customers’, i.e., customers who can choose between the two ordering channels (online via the APP or offline in the store) and can set their own queue-size threshold for deciding whether or not to join the queue. They showed that adding an online ordering option has the unexpected effect of lowering both the customers’ individual utility and the social welfare outcome (defined as the total utility for all customers from both channels). This ’paradox’ is explained by self-interested channel choices. The authors offered strategies to address the aggravating effects of providing an additional channel. In a similar study, Gao and Su [10] modeled a restaurant service in which orders are processed in two stages (ordering and then food preparation) as above. Customers can place orders through their own digital devices and skip waiting in line for the first service stage. However, the customers are not strategic (i.e., they cannot choose the ordering channel and they always enter the system without considering a queue-size joining threshold). The authors concluded, among other findings, that both customer types (including those who do not use the digital technologies) experience reduced waiting costs and that this generates increased demand.

In practice, not all customers are pleased with such an online self-order service. It requires an installation of an appropriate APP. As remarked by the Facebook’s chief executive officer Mark Zuckerberg, “no one wants to have to install a new app for every business or service that they want to interact with” (quoted in [10,11]). Moreover, there are customers who have no access to the required digital device, e.g., a computer or a smartphone, and cannot use the self-order platform. It follows that customers are generally of two kinds: those who do not use the self-order application (ordinary customers) and those who are willing to use the application (APP customers). Upon arrival, ordinary customers decide whether to join the queue or to balk. Thus, ordinary customers are called *strategic*, whereas APP customers, who place (and normally pay for) their order on their way to the restaurant and always join the second-stage queue, are not. The number of APP customers increased dramatically during the COVID-19 pandemic where in some businesses, especially in fast food shops, it became the only possible way of ordering.

We note that, after placing their order, many customers (of both types) do not physically stand in line waiting for the preparation of their order, but instead disperse and wait in different locations utilizing their time for other purposes. As a result, when strategic customers make the join/balk decision with regard to the first-stage queue, they cannot see the full length of the second-stage queue and make their decision based on: (i) the number of customers they see in the first service stage, and (ii) their estimate of the number of customers in the second service stage (which is a function of the number of customers seen in the first service stage). Moreover, a strategic customer has to take into account the number of APP customers who may arrive after he/she joins the first-stage queue and may overtake him/her by joining the second service-stage queue directly. These considerations have not been taken into account in previous research works reported in the literature.

Another common practice exercised in restaurants is that cashiers do not constantly stand at their cash tills, but rather leave their station when the queue size is relatively short and take a ’vacation’ in order to perform other restaurant duties. Such a queue-dependent vacation policy was suggested by Yadin and Naor [12] and is now known as an N-policy (see [13–21]). In many establishments, each cashier may adopt a different vacation policy, i.e., each cashier returns from the vacation to his/her station only when the queue size reaches some predefined number for that particular cashier. Thus, the larger the queue size, the more servers are assigned to the first service stage. Consequently, more customers are served per hour. This feature of the servers’ behavior makes the customers’ decision of whether to join or balk more complex since they need to take account not only of the customers who are already present in the queue, but also of the possibility that future customers may arrive and trigger an increase in the number of active servers, resulting in a reduction of their waiting time.

An additional consideration is that restaurant managers usually prefer customers to place their orders via the APP so that queue sizes are reduced and servers are released for ancillary duties. To encourage this behavior, it is a common practice of restaurants to offer a discount to those ordering via a digital application. Such price discrimination policies have been widely investigated (see, e.g., [22–27]).

To summarize, when making a join/balk decision, a newly arriving strategic (NAS) customer’s action is based on: (i) the number of customers seen in the first service stage; (ii) the estimated mean number of customers in the second service stage (as a function of the observed queue size in the first stage); (iii) the mean number of future APP customers who will join the second stage queue directly and will ‘overtake’ the NAS customer before he/she can join the second stage queue; (iv) the mean number of strategic customers that will join the first stage queue after the NAS customer and may trigger a faster rate of service; (v) the vacation policy adopted by servers, which is at the discretion of the manager; and (vi) the existence and impact of any APP discount, which is also set by the manager.

These six facets of the join/balk decision are very common features of real-life consumer service systems. However, we are not aware of any investigation that has examined the operation of all the above characteristics simultaneously. Hence, the contribution of the current paper is to achieve the following specific outputs in the context of this multi-faceted two-stage service system:

(i) Constructing the two-stage service system with a dynamic server vacation policy, as a QBD process, and developing corresponding quantitative performance measures.

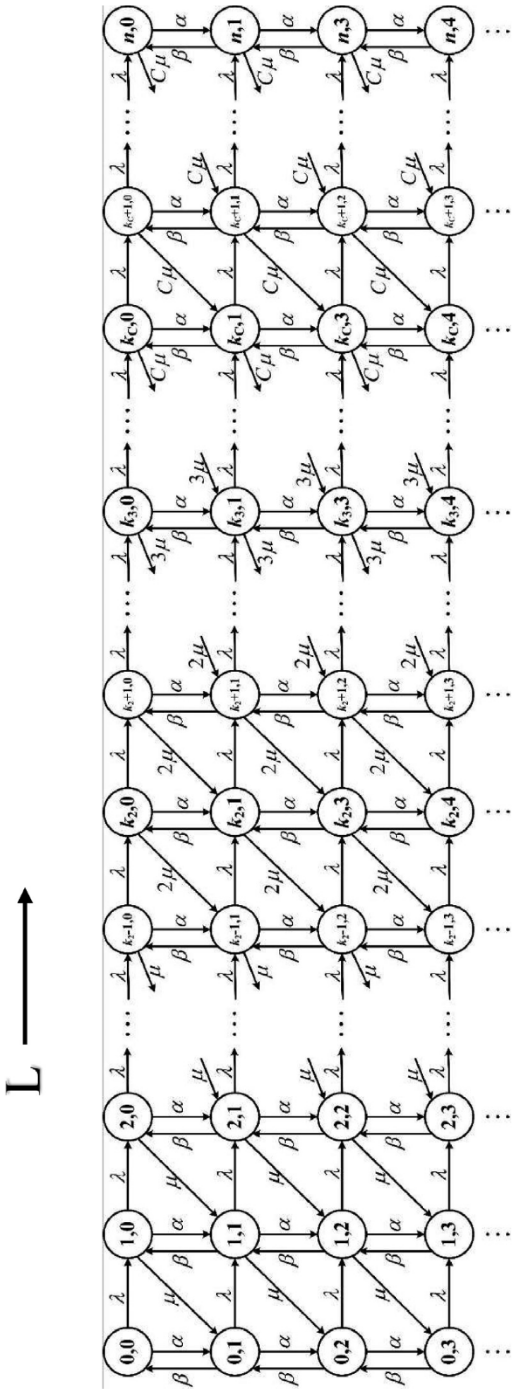

Fig. 1. Transition-rate diagram. The horizontal axis depicts L, while the vertical axis depicts S.

(ii) Developing a methodology for determining equilibrium thresholds for joining/balking at stage one for two categories of strategic customers: (a) those who take into account the sojourn time at the second stage when deciding whether to join the first stage queue (far-sighted customers) and (b) those who do not take account of the second stage sojourn time (myopic customers).

(iii) Examining the monetary increase associated with adopting a far-sighted approach.

(iv) Demonstrating that for individual customers, the number of equilibrium thresholds may exceed one.

(v) Demonstrating that the individual joining threshold is not always larger than the social joining threshold and conducting a comparative study to illustrate cases in which the individual threshold is lower than the social one.

(vi) Developing a method for calculating the optimal values of the APP discount and servers vacation policy.
(vii) Solving the overall service system problem (including the APP discount and servers
(viii) vacation policy) while determining the customers joining threshold.

2. The model and its formulation

We consider a service system in which the service to each individual customer consists of two consecutive stages as discussed above. Each service stage is provided independently by a different team of servers and has a separate queue. Customers go through the first stage (ordering/paying) *either* on-line by using an APP, *or* by physically queueing in the shop to place their order. Whichever route is followed at the first stage, every joining customer joins the second-stage (order preparation) queue. Thus, there are two customer types: (i) Strategic, who physically queue at both stages, and (ii) APP, those who complete the first-stage service online, but physically join the queue at stage 2.

When a strategic customer arrives, s/he joins the stage-1 queue only if fewer than n customers are present there (including those who are being served). Otherwise, a strategic customer balks, leaving the outlet never to return. The joining threshold, n, is determined in accordance with Naor's model [28], as described in Section 3.1 bellow. An APP customer, having completed stage 1 by APP, always joins the stage 2 queue on arrival and waits there until being served. Both strategic and APP customers arrive according to independent Poisson processes, with rates λ and α, respectively.

Note: although each of the variables and parameters used in the paper is defined and described when required, a comprehensive detailed list of notation appears in Appendix D.

The first service-stage is provided by $C \geq 1$ potential (statistically identical) independent parallel servers. Servers are listed in the order $m = 1, 2, 3, ..., C$, while each server is associated with a queue size threshold k_m, where $1 = k_1 < k_2 < k_3 < ... < k_C \leq n$. As soon as the queue size falls below k_1, server $m = 1$ goes on 'vacation' and returns immediately when the number of customers increases back to k_1. Similarly, server m goes on vacation when the number of customers falls short of k_m and returns immediately as the queue size reaches k_m again. It readily follows that $k_m \geq m$. This operating procedure describes, for example, the scheduling policy of a shop that employs C cashiers, each of whom may be on vacation, performing other duties when the queue length justifies such action. The second-stage queue is assumed to be serviced by a single server, although in practice the 'server' may comprise a coordinated team that jointly prepares the order. The service duration at each stage is exponentially distributed with parameter μ in the first stage, and parameter β in the second. The two processes are independent.

The system is formulated as a 2-dimensional quasi-birth-and-death (QBD) process. At time t, let L_t and S_t denote the number of customers in the first and in the second stage, respectively. $\{L_t, S_t\}$ defines the state space of the queueing system at time t. Let $L \equiv \lim_{t\to\infty} L_t$ and $S \equiv \lim_{t\to\infty} S_t$. Define the steady-state joint probability distribution function of the two-dimensional Markovian process by $p_{i,j} = \Pr(L = i,\ S = j)$, $i = 0, 1, 2, ..., n$; $j = 0, 1, 2, ...$ The transition-rate diagram for the queueing system's states is depicted in Fig. 1.

In order to construct the infinitesimal generator matrix, Q, of the corresponding QBD process, the system states are arranged in the following lexicographical order: $\{(0, 0), (1, 0), ..., (n, 0); (0, 1), (1, 1), ..., (n, 1); ...; (0, j), (1, j), ..., (n, j); ...\}$, $j = 0, 1, 2, ...$

Then,

$$Q = \begin{pmatrix} B & A_0 & 0 & 0 \cdots \\ A_2 & A_1 & A_0 & 0 \cdots \\ 0 & A_2 & A_1 & A_0 \\ \vdots & & \ddots & \ddots \;\ddots \end{pmatrix},$$

where the matrices B, A_0, A_1 and A_2 are each of order $(n + 1) \times (n + 1)$ and are given below

$$B = \begin{pmatrix} -(\lambda+\alpha) & \lambda & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & -(\lambda+\alpha+\mu) & \lambda & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & -(\lambda+\alpha+\mu) & \ddots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \lambda & & 0 & \cdots & 0 \\ 0 & 0 & 0 & & -(\lambda+\alpha+2\mu) & \ddots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \ddots & \lambda & & 0 \\ 0 & 0 & 0 & \cdots & 0 & & -(\lambda+\alpha+3\mu) & \ddots & 0 \\ \vdots & \vdots & \vdots & & & & & \ddots & \lambda \\ 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & & -(\alpha+C\mu) \end{pmatrix},$$

$$A_0 = \begin{pmatrix} \alpha & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \mu & \alpha & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \mu & \alpha & & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & & & \vdots & & \vdots \\ 0 & 0 & & 2\mu & \alpha & & 0 & \cdots & 0 \\ \vdots & \vdots & & & \ddots & \ddots & & & \vdots \\ 0 & 0 & \cdots & 0 & & 3\mu & \alpha & & 0 \\ \vdots & \vdots & & \vdots & & & \ddots & \ddots & \\ 0 & 0 & \cdots & 0 & \cdots & 0 & & C\mu & \alpha \end{pmatrix},$$

$$A_2 = \begin{pmatrix} \beta & 0 & \cdots & 0 & 0 \\ 0 & \beta & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & \beta & 0 \\ 0 & 0 & \cdots & 0 & \beta \end{pmatrix},$$

$$A_1 = \begin{pmatrix} -(\lambda+\alpha+\beta) & \lambda & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & -(\lambda+\alpha+\mu+\beta) & \lambda & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & -(\lambda+\alpha+\mu+\beta) & \ddots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \lambda & & 0 & \cdots & 0 \\ 0 & 0 & 0 & & -(\lambda+\alpha+2\mu+\beta) & \ddots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \ddots & \lambda & & 0 \\ 0 & 0 & 0 & \cdots & 0 & & -(\lambda+\alpha+3\mu+\beta) & \ddots & 0 \\ \vdots & \vdots & \vdots & & & & & \ddots & \lambda \\ 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & & -(\alpha+C\mu+\beta) \end{pmatrix}.$$

For each row j ($j = 0,1,2,\ldots$) in Fig. 1, define the corresponding $(n + 1)$-dimensional probability vector as $\vec{p}_j \equiv (p_{0,j}, p_{1,j}, p_{2,j}, \ldots, p_{n,j})$. The entire probability vector of all system states is $\vec{p} \equiv (\vec{p}_0, \vec{p}_1, \vec{p}_2, \ldots)$, and let $\vec{e} = (1, 1, 1, \ldots)^T$ be a corresponding infinite dimensional column vector with all its entries equal to one. The system's balance equations are given by

$$\vec{p}Q = \vec{0}, \quad \vec{p} \cdot \vec{e} = 1.$$

Theorem 1. Let $\rho_m = \lambda/m\mu$, $k_0 = 0$ and $k_{C+1} - 1 = n$. Then, the stability condition of the queueing system is given by

$$\alpha + \mu \sum_{m=1}^{C} \frac{m^{k_m}}{\prod_{a=1}^{m-1} a^{k_{a+1}-k_a}} \left(\frac{\rho_m^{k_{m+1}} - \rho_m^{k_m}}{\rho_m - 1} \right) \left(1 + \sum_{m=1}^{C} \frac{m^{k_m - 1}}{\prod_{a=1}^{m-1} a^{k_{a+1}-k_a}} \left(\frac{\rho_m^{k_{m+1}} - \rho_m^{k_m}}{\rho_m - 1} \right) \right)^{-1} < \beta. \tag{1}$$

Proof. See Appendix A.

For example, for $n = 4$, $C = 2$ and $k_2 = 3$, the stability condition is given by

$$\alpha + \mu \frac{2\lambda^4 + 4\lambda^3\mu + 4\lambda^2\mu^2 + 4\lambda\mu^3}{\lambda^4 + 2\lambda^3\mu + 4\lambda^2\mu^2 + 4\lambda\mu^3 + 4\mu^4} < \beta. \tag{2}$$

Note that, in this case, when λ is larger than $C\mu = 2\mu$, the two servers in the first stage are almost always busy so that the effective arrival rate to the unbounded second stage is $\alpha + 2\mu$. Indeed, the left-hand side of the stability condition (2) approaches $\alpha + 2\mu$ as λ approaches infinity, and this rate should be smaller than β, the service rate in that stage.

The steady-state probabilities of the queueing system are calculated (see Neuts, [29]) from $\vec{p}_j = \vec{p}_0 R^j$, $j = 0, 1, 2, \ldots$, where R (the so-called 'rate matrix') is the matrix of size $(n + 1) \times (n + 1)$ that satisfies $A_0 + RA_1 + R^2 A_2 = 0$. In most cases, the matrix R is calculated numerically via successive substitution (see, e.g., Harchol-Balter, [30]). Recently, Hanukov and Yechiali [31] showed that in many cases, when the three matrices A_0, A_1 and A_2 are all upper-triangular, or are all lower-triangular, the entries of R can be calculated *explicitly*, and that the stability condition can be easily obtained. Unfortunately, this is not the case in the current model. The $(n + 1)$-dimensional probability vector $\vec{p}_0$ is calculated as follows: (i) the first vector equation extracted from $\vec{p}Q = \vec{0}$ is $\vec{p}_0 B + \vec{p}_1 A_2 = \vec{p}_0[B + RA_2] = \vec{0}$ (since $\vec{p}_1 = \vec{p}_0 R$). (ii) furthermore, since $\vec{p}_j = \vec{p}_0 R^j$, $j = 0, 1, 2, \ldots$, the normalization equation $\vec{p} \cdot \vec{e} = 1$ translates into $\sum_{j=0}^{\infty} \vec{p}_j \vec{e}_{n+1} = \sum_{j=0}^{\infty} \vec{p}_0 R^j \vec{e}_{n+1} = \vec{p}_0 [I - R]^{-1} \vec{e}_{n+1} = 1$, where here

$\vec{e}_{n+1} = (1, 1, ..., 1)^T$ is an $(n + 1)$-dimensional column vector of ones. Using the last linear equation (involving the $n + 1$ probabilities $p_{0,0}$, $p_{1,0}...p_{n,0}$) together with n equations from $\vec{p}_0[B + RA_2] = \vec{0}$ the vector $\vec{p}_0$ is uniquely calculated.

Let $E[L]$ be the mean number of strategic customers in the first stage (note that all customers in the first stage are strategic), and let $E[S]$ be the mean number of customers in the second stage. Let $E[S^{str}]$ and $E[S^{app}]$, respectively, be the mean number of strategic and of APP customers in the second stage, i.e., $E[S] = E[S^{str}] + E[S^{APP}]$. Let $E[D]$ be the mean sojourn time of a strategic customer in the first stage. Let $E[T^{str}]$ and $E[T^{app}]$, respectively, be the mean sojourn time of a strategic and of an APP customer in the second stage. Let $E[W] = E[D] + E[T^{str}]$ be the mean total sojourn time in the system (in both stages) of a strategic customer. Let $E[V]$ be the mean number of first-stage servers on vacation.

In what follows we use the notation $p_{\bullet,j} \equiv \sum_{i=0}^{n} p_{i,j}$, $j = 0, 1, 2, ...,$ and $p_{i,\bullet} \equiv \sum_{j=0}^{\infty} p_{i,j}$, $i = 0, 1, 2, ..., n$. Let $\vec{v} \equiv (0, 1, 2, ..., n)^T$. Then, the mean number of strategic customers in the first stage is given by

$$E[L] = \sum_{i=1}^{n} i p_{i,\bullet} = \sum_{j=0}^{\infty} (\vec{p}_j \cdot \vec{v}) = \vec{p}_0 \left(\sum_{i=0}^{\infty} R^j \right) \vec{v} = \vec{p}_0 [I - R]^{-1} \vec{v}.$$

By Little's law the mean sojourn time of a strategic customer in the first stage is $E[D] = E[L]/\lambda_{eff}$, where λ_{eff} is the strategic customers' effective arrival rate, calculated as follows. Let $\vec{u} = (1, 1, .., 1, 1, 0)^T$. Then, $\lambda_{eff} = \lambda \sum_{i=0}^{n-1} p_{i,\cdot} = \lambda \sum_{j=0}^{\infty} (\vec{p}_j \cdot \vec{u}) = \lambda \vec{p}_0 [I - R]^{-1} \vec{u}$. The mean number of customers in the second stage is given by

$$E[S] = \sum_{j=1}^{\infty} j p_{\bullet,j} = \sum_{j=1}^{\infty} j \vec{p}_j \cdot \vec{e}_{n+1} = \sum_{j=1}^{\infty} j \left(\vec{p}_0 R^j \right) \vec{e}_{n+1} = \vec{p}_0 \left(\sum_{j=1}^{\infty} j R^j \right) \vec{e}_{n+1} = \vec{p}_0 R [I - R]^{-2} \vec{e}_{n+1}.$$

An APP customer who arrives into state (i, j) stays in the second stage for $j + 1$ service durations. Thus, an APP customer's mean sojourn time in the system is given by

$$E[T^{app}] = \sum_{j=0}^{\infty} (j + 1) \beta^{-1} p_{\bullet,j} = \beta^{-1} \sum_{j=0}^{\infty} (j + 1) \vec{p}_j \cdot \vec{e}_{n+1} = \beta^{-1} \vec{p}_0 [I - R]^{-2} \vec{e}_{n+1}.$$

Using Little's law, the mean number of APP customers in the second stage is calculated as $E[S^{app}] = \alpha E[T^{app}]$. Then, $E[S^{str}] = E[S] - E[S^{app}]$ and $E[T^{str}] = E[S^{str}]/\lambda_{eff}$.

Let $\vec{w}$ be an $(n + 1)$-dimensional column vector defined as follows: for a given m, $m = 0, 1, 2, ..., C$, the i^{th} term of $\vec{w}$, denoted by $\vec{w}[i]$, is $\vec{w}[i] = C - m$ for all $i = k_m, k_m + 1, k_m + 2, ..., k_{m+1} - 1$. $\vec{w}[i]$ indicates the number of servers on vacation in the first stage queue. Thus, the mean number of servers on vacation is given by

$$E[V] = \sum_{j=0}^{\infty} (\vec{p}_j \cdot \vec{w}) = \vec{p}_0 \left(\sum_{i=0}^{\infty} R^j \right) \vec{w} = \vec{p}_0 [I - R]^{-1} \vec{w}.$$

3. The strategic customer problem

Consider, for example, a coffee shop where customers stand in line to place their order at the first stage, and then, instead of physically waiting in line for the completion of their order, they seat at a nearby table or outside the coffee shop. Taking such behavior into account, it is assumed that (i) an NAS customer can see the actual number of customers present in the first-stage queue, but (ii) does not know the number of customers in the second-stage queue (i.e., the second-stage queue is *unobservable*). Thus, based only on the available information on the queue length in the first stage, a strategic customer has to decide whether to join the queue or to balk. Moreover, the sojourn cost-rate in the first stage is not necessarily the same as the one in the second stage.

We distinguish between two types of strategic customers: (i) type *M*, myopic customers, each considers only the mean sojourn time in the first-stage queue; and (ii) type *F*, far-sighted customers, who consider the cumulative mean sojourn time in both stages. We analyze two kinds of threshold joining policies: (i) individual, where each individual customer (*M* or *F*) aims at maximizing his/her own utility (called type *MI* or type *FI*, respectively); or (ii) social threshold policy, where each type (*M* or *F*) aims at maximizing the overall social welfare (type *MS* or type *FS*, respectively). Fig. 2 depicts all possible types of strategic customers. In what follows we derive the equilibrium threshold for each type of strategic customer (see also Section 3.1). The concept of equilibrium follows Naor's model, namely, a threshold joining policy is a number n such that an NAS customer joins the queue only if its size is smaller than n. Thus, we have an infinite number of players (customers), each chooses from an infinite set of strategies ($n = 1, 2, 3, ...$). The players are symmetric regarding their decisions, so that in equilibrium all customers behave according to the same threshold n. The threshold n dictates the servers' dynamic vacation policy which has to be considered as well when a customer makes his/her individual decision. Moreover, a social customer is also affected by other players threshold decision (n) since the overall welfare depends on it. Similar to the basic assumptions listed in Hassin and Haviv's book ([32] page 22) for models similar to ours, the main assumptions in this paper are: (i) a customer's benefit from completed service is r; (ii) the cost rate to a customer from staying in the system is h_1 for stage 1 and h_2 for stage 2; (iii) customers are risk neutral, that is, they maximize the expected value of their net benefit; (iv) utility function of individual customers are identical; (v) a decision to join is irrevocable and reneging is not allowed.

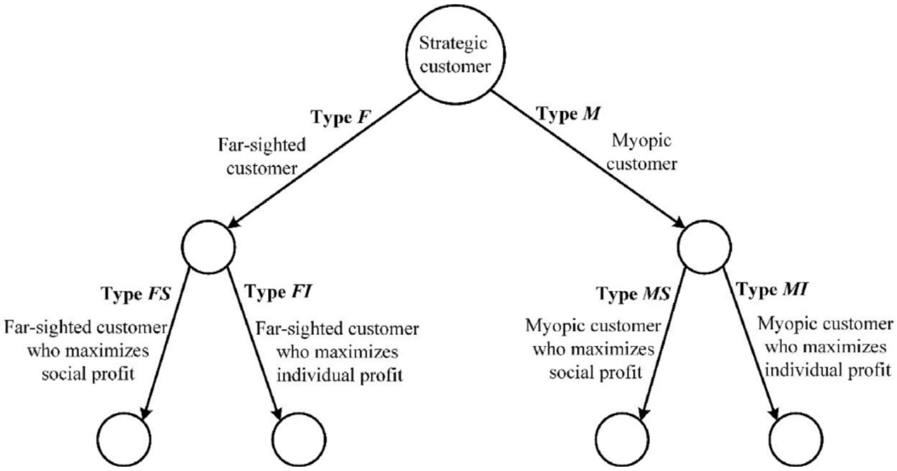
Strategic customer
Type F
Far-sighted customer
Type M
Myopic customer
Type FS
Far-sighted customer who maximizes social profit
Type FI
Far-sighted customer who maximizes individual profit
Type MS
Myopic customer who maximizes social profit
Type MI
Myopic customer who maximizes individual profit

Fig. 2. Summary of strategic customer types.

3.1. Myopic individual (MI) customer

Let r be the customer's reward for obtaining the service and let h_1 be the customer's sojourn cost-rate in the first stage. In the traditional M/M/1 queue, an NAS customer who sees y customers in the queue will spend $y+1$ full-service durations, independent of the number of customers who will arrive after him/her. Thus, the NAS customer will join if $r - h_1(y+1)/\mu > 0$. This allows an equilibrium threshold, n, to be derived using $n = 1 + \max(y|r - h_1(y+1)/\mu > 0) = \lfloor r\mu/h_1 \rfloor$ (Naor, [28]). Furthermore, Yechiali [33,34] considered the GI/M/1 and GI/M/s models and showed that among all randomized joining policies the optimal one is a non-randomized policy (i.e., state-depending joining probabilities are either 0 or 1), and among all non-randomized policies, the threshold one is optimal. Thus, in what follows we consider threshold policies for joining/balking. In the current case, however, it is required to take into consideration the fact that the number of active servers is dynamic, depending on the dynamically changing number of customers in the system. Thus, as noted in the introduction, the mean sojourn time of a customer may be affected by the number of customers who will join the queue after his/her joining instant. Specifically, the larger the queue behind a customer, the more servers will return from their vacation, resulting in a higher overall service rate and a lower customer's sojourn time. Another consideration to be taken is that the arrival rate behind a customer depends on the balking threshold. Thus, let $E[D|y, n]$ be the mean sojourn time in the first stage for an NAS customer who sees y customers in the system and for a given, already known, threshold n. Then, the utility of a strategic type *MI* customer is given by

$$Z_{MI} = r - h_1 E[D|y, n]. \tag{3}$$

A threshold, n, is an equilibrium threshold if, for that n, the following two conditions hold:

$$\text{(i)}\; r - h_1 E[D|y, n] \geq 0 \text{ for all } y < n, \tag{4}$$

$$(ii)\; r - h_1 E[D|y, n] < 0 \text{ for y = n.} \tag{5}$$

In order to calculate $E[D|y, n]$, we formulate a sub-system for an NAS customer. The sub-system is described by states (L, H), where L is the number of customers in the first stage, and H is the position of the customer in the queue. The evolution of the sub-system starts with the customer's arrival and terminates when this customer reaches the service station ($H = 0$). The transition rate diagram of the sub-system is depicted in Fig. 3.

The diagram is constructed as follows: each service completion reduces the number of customers (L) by one, and lower the position of the customer (H) by one (that is, a customer advances one position in the queue). Each customer's arrival increases L by one and has no effect on the customer's position H. The exceptions are the states at which $L = k_m$ or $L = k_{m-1}$. When a service completion occurs at a state $L = k_m$, one server leaves the system for vacation, and thus no change in the customer's position occurs. On the other hand, when a customer arrival occurs at a state where $L = k_{m-1}$, one server returns to the system from vacation, and thus the customer's position in the queue is reduced by one.

While removing the expectation notation, let $D(L, H)$ be the mean total waiting time of a customer starting from state (L, H) until the start of his/her service. For simplicity of presentation, let $m(L)$ be the corresponding m for a given L, according to the servers' vacation policies. Each $D(L, H)$, for all L and H, can be recursively calculated by the following procedure:

Procedure 1. Calculation of *D*(*L, H*) for all *L* and *H*.

(i) $D(L, 1) = \frac{1}{C\mu}, \quad L = k_C + 1, k_C + 2, \ldots, n$

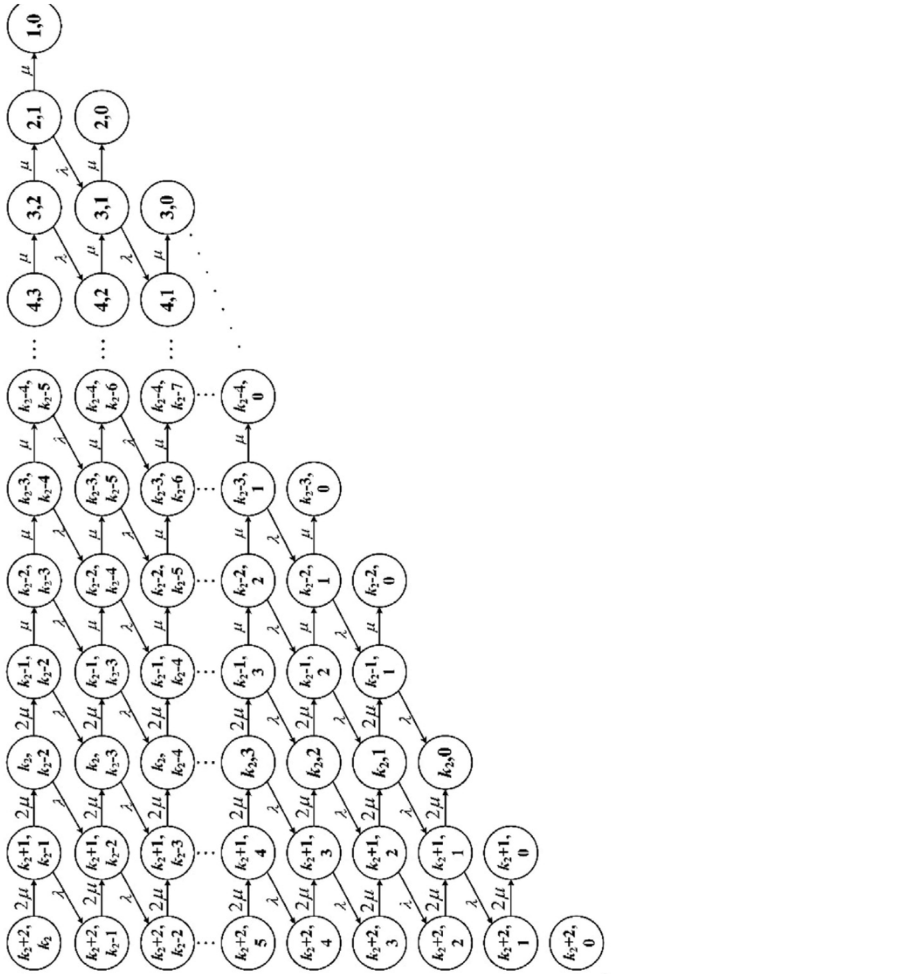

Fig. 3. The sub-system's transition rate diagram.

Table 1
Values of Z_{MI} for different values of y and n.

	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$	$\mathbf{y = 7}$	$\mathbf{y = 8}$	$y = 9$
$n = 4$	5.94	4.44	3.32	2.19					
$n = 5$	5.94	4.6	3.8	2.67	1.55				
$n = 6$	5.94	4.6	3.87	3.03	1.9	0.78			
$\mathbf{n = 7}$	5.94	4.6	3.87	3.08	2.2	1.07	**−0.05**		
$\mathbf{n = 8}$	5.94	4.6	3.87	3.08	2.23	1.31	0.18	**−0.94**	
$n = 9$	5.94	4.6	3.87	3.08	2.23	1.33	0.37	−0.76	−1.88

(ii) $D(L,1) = \frac{1}{\lambda+m(L)\mu} + \frac{\lambda}{\lambda+m(L)\mu}D(L+1,1)$,
$L = 2, 3, \dots, k_C - 1 \neq k_m - 1, k_m, m = 2, 3, \dots, C$

(iii) $D(L,1) = \frac{1}{\lambda+(m-1)\mu}$, $L = k_m - 1, m = 2, 3, \dots, C$

(iv) $D(L,1) = \frac{1}{\lambda+m(L)\mu} + \frac{m(L)\mu}{\lambda+m(L)\mu}D(L-1,1) + \frac{\lambda}{\lambda+m(L)\mu}D(L+1,1)$,
$L = k_m, m = 2, 3, \dots, C$

(v) $D(L, L-m(L)-l) = \frac{1}{\lambda+m(L)\mu} + \frac{m(L)\mu}{\lambda+m(L)\mu}D(L-1, L-1-m(L-1)-l)$
$+\frac{\lambda}{\lambda+m(L)\mu}D(L+1, L-m(L+1)-l)$,
$L = 3, 4, \dots, n-1, l = 0, 1, 2, \dots, L-m(L)-2$

(vi) $D(L,H) = \frac{1}{C\mu} + D(L-1, H-1)$, $L = n, H = 2, 3, 4, \dots, n - C.$

When an arriving customer joins the system, the number of customers increases from y to $y + 1$ ($y = 1, 2, 3, \dots, n - 1$). The position H of the customer in the queue depends on the number of active servers, m, at the instant of the new arrival. Specifically, the position of the customer is $y + 1 - m$. Thus, a new customer changes the system state to $(y + 1, y + 1 - m)$, and consequently, the mean waiting time of the customer in the queue is $D(y + 1, y + 1 - m)$. Note that all states $(y + 1, y + 1 - m)$ are arranged in the upper row in Fig. 3. Finally, the mean sojourn time of the customer in the first stage, for all y and its corresponding m, is given by

$$E[D|y,n] = D(y+1, y+1-m) + 1/\mu. \tag{6}$$

In order to illustrate the derivation of the equilibrium threshold for *MI* customers, denoted by n_{MI}, we present in Table 1 values of Z_{MI} Eq. (3)) for different values of y and n for the following numerical example: $C = 2$, $k_2 = 4$, $\lambda = 16$, $\mu = 20$, $r = 10$ and $h_1 = 45$. As $n \geq k_C = 4$, the table begins with row $n = 4$. The values of Z_{MI} in Table 1 are calculated while using the recursive calculation given in Procedure 1 for $D(\bullet, \bullet)$, and Eq. (6) for $E[D|y, n]$. It is shown that two values of n, $n = 7$ and $n = 8$, satisfy the conditions for an equilibrium threshold given in Eqs. (4),((5). Thus, two equilibrium thresholds are established in that case, $n_{MI} = 7$ and $n_{MI} = 8$.

It is seen that for each y, the utilities are equal for all $n \geq y + k_2 - 1 = y + 3$. This is explained as follows: a joining customer sees y customers in the system. Once $k_2 - 2$ customers arrive after him/her (so that the number of customers in the system becomes $y + 1 + k_2 - 2 = y + 3$), the customer's service rate will remain 2μ until s/he starts service. Thus, once $k_2 - 2$ new customers have arrived, the waiting time of the customer already in the queue no longer depends on the number of additional arrivals.

3.2. Myopic social (MS) customer

For a given threshold policy n the overall rate of utility for type *MS* customers is given by

$$Z_{MS} = r\lambda_{eff} - h_1 E[L]. \tag{7}$$

Let n_{MS} be the equilibrium threshold of a type *MS* customer. Then, n_{MS} is calculated by maximizing Z_{MS}. Using the parameter values from the above example, the equilibrium threshold for a type *MS* customer is $n_{MS} = 6$. It is classically putative that a social threshold should be lower than the individual one (see, e.g., Naor [28], and Hassin, [35]). This phenomenon is explained by the *negative externalities* that a joining customer imposes on *future customers*, but are ignored by a customer who maximizes his/her individual utility. In our model, however, a joining customer also imposes *positive externalities* on customers already in the system, since by joining s/he increases the system's overall service rate. Thus, it is of interest to investigate these two kinds of thresholds. For this purpose, we consider the values used to construct Table 1, and calculate the equilibrium thresholds while systematically changing the value of one parameter at a time, keeping the others constant. The results are presented in Figs. 4–7. The figures do indeed show that, in some cases, the individual threshold is lower than the social one, i.e., $n_{MI} < n_{MS}$. Fig. 4 shows that when the sojourn cost h_1 is relatively small, the individual threshold is higher than the social threshold; but as h_1 increases, the individual threshold falls below the social one. Similarly, Fig. 5 shows that the individual threshold is lower than the social one when the reward gained from the service, r, is small, but becomes higher than the social threshold when r increases. Fig. 6 shows that the individual threshold increases slightly with customers arrival rate, λ. This result can be explained as follows: a higher arrival rate increases the overall service rate and consequently reduces the waiting time. For small arrival rates, the social threshold is higher than the individual one. As

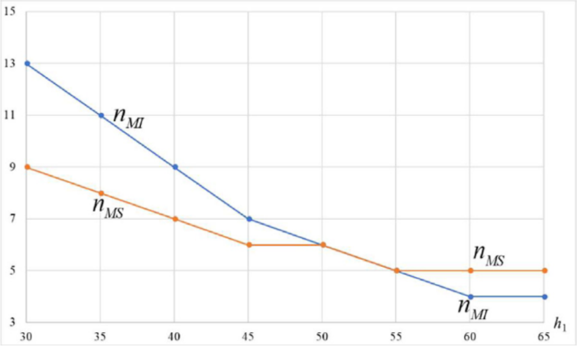

Fig. 4. Social and individual thresholds as a function of h_1 for $C = 2$, $k_2 = 4$, $\lambda = 16$, $\mu = 20$ and $r = 10$.

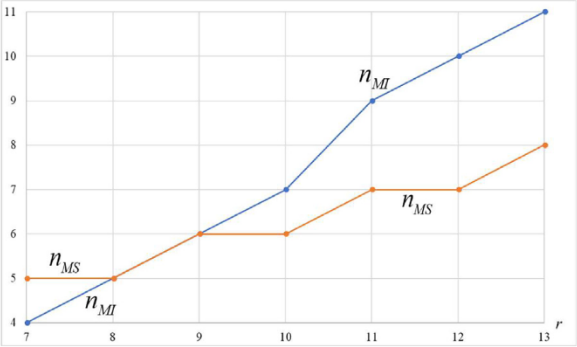

Fig. 5. Social and individual thresholds as a function of r for $C = 2$, $k_2 = 4$, $\lambda = 16$, $\mu = 20$ and $h_1 = 45$.

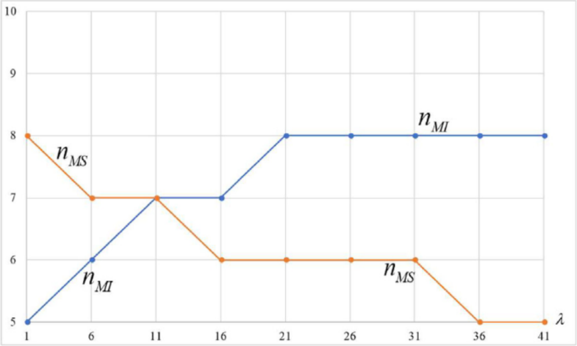

Fig. 6. Social and individual thresholds as a function of λ for $C = 2$, $k_2 = 4$, $\mu = 20$, $r = 10$and $h_1 = 45$.

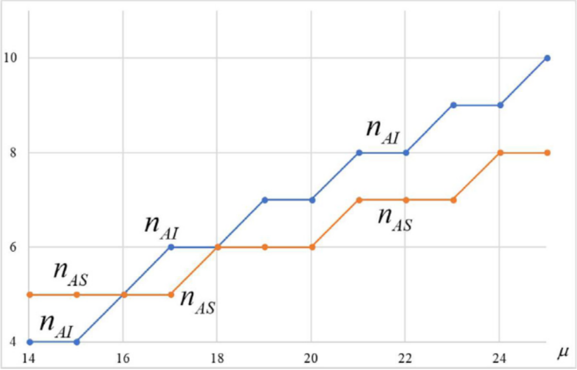

Fig. 7. Social and individual thresholds as a function of μ for $C = 2$, $k_2 = 4$, $\lambda = 16$, $r = 10$and $h_1 = 45$.

Table 2
Values of Z_{FI} for different values of y and n.

	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$\mathbf{y = 5}$	$\mathbf{y = 6}$	$y = 7$
$n = 4$	4.926	3.337	2.130	0.969			
$\mathbf{n = 5}$	4.844	3.332	2.355	1.087	**−0.129**		
$\mathbf{n = 6}$	4.814	3.280	2.347	1.314	0.033	**−1.179**	
$n = 7$	4.800	3.262	2.324	1.336	0.265	−1.007	−2.207

the arrival rate increases, the social threshold decreases. This is explained as follows: the higher the arrival rate, the higher the probability that the number of customers in the system is larger than k_2; this, in turn, diminishes the positive externalities effect. Thus, at some value of the arrival rate, the individual threshold becomes higher than the social one. Fig. 7 shows that the individual threshold is lower than the social threshold when the service rate, μ, is small, but that it increases at a slightly greater rate with μ than the social threshold. Thus, the individual threshold is higher than the social one when the service rate is high.

3.3. Far-sighted individual (FI) customer

In this section we derive the equilibrium threshold for far-sighted customers (i.e., those who consider the cumulative mean sojourn time at both stages of the process) and seek to maximize their individual utility (type *FI*). The calculation of the sojourn time in the first stage is the same as for a type *MI* customer. However, it is also required to derive, for a given threshold n, a methodology to calculate the sojourn time in the second stage for an NAS customer who sees y customers in the first stage. For this purpose, let $p_{j|i}$, $i = 0, 1, 2, ..., n$, $j = 0, 1, 2, ...$, be the probability of j customers in the second stage given that there are i customers in the first stage, let $\psi_{\gamma,j}(t)$ be the probability that γ customers are in the second stage t time units from the present time, given that j customers are currently in the second stage, let $E[S(t)|y]$ be the mean number of customers who will occupy the second stage, given that y customers are currently in the first stage. Let $\vec{s_i}$ be a column vector with $n+1$ entries, the i^{th} of which equals one, and all other entries equal zero. Then, $p_{j|i}$ is given by

$$p_{j|i} = \frac{p_{i,j}}{\sum_{k=0}^{\infty} p_{i,k}} = \frac{p_{i,j}}{\sum_{k=0}^{\infty} \vec{p}_k \cdot \vec{s_i}} = \frac{p_{i,j}}{\sum_{k=0}^{\infty}\left(\vec{p}_0 R^k\right)\vec{s_i}} = \frac{p_{i,j}}{\vec{p}_0[I-R]^{-1}\vec{s_i}}, \ i = 0, 1, 2, ..., n,\ j = 0, 1, 2, ... \tag{8}$$

According to Kleinrock [36] (p. 77) the value of $\psi_{\gamma,j}(t)$ is given by

$$\psi_{\gamma,j}(t) = e^{-(\alpha_{eff}+\beta)t}\left[\rho^{(\gamma-j)/2} I_{\gamma-j}(at) + \rho^{(\gamma-j-1)/2} I_{\gamma+j+1}(at) + (1-\rho)\rho^{\gamma}\sum_{k=\gamma+j+2}^{\infty}\rho^{-k/2} I_k(at)\right], \tag{9}$$

where (i) α_{eff} is the total arrival rate at the second stage (including strategic and APP customers), the calculation of which is described below, (ii) $\rho = \alpha_{eff}/\beta$, (iii) $a = 2\beta\sqrt{\rho}$, and (iv) $I_\gamma(x) = \sum_{\upsilon=0}^{\infty}(x/2)^{\gamma+2\upsilon}/((\gamma+\upsilon)!\ \upsilon!)$ is Bessel function. Then, $E[S(t)|y]$ is given by

$$E[S(t)|y] = \sum_{\gamma=0}^{\infty}\sum_{j=0}^{\infty}\int_0^{\infty}\gamma\psi_{\gamma,j}(t)p_{j|y}f_{D|y}(t)dt, \tag{10}$$

where $f_{D|y}(t)$ is a density function of a customer's sojourn time in the first stage starting from his/her arrival until the start of his/her service. It follows that this customer's mean sojourn time in the second stage, for a given y and n, is given by

$$E[T|y, n] = (E[S(t)|y] + 1)/\beta. \tag{11}$$

The calculation of α_{eff} and $f_{D|y}(t)$ requires an involved procedure, which is given in Appendices B and C, respectively.

Now, let Z_{FI} denote the utility enjoyed by a strategic type *FI* customer:

$$Z_{FI} = r - h_1 E[D|y, n] - h_2 E[T|y, n], \tag{12}$$

where h_2 is a customer's sojourn cost in the second stage. The threshold, n, is an equilibrium threshold if, for that n, the following two conditions hold:

$$\text{(i) } r - h_1 E[D|y, n] - h_2 E[T|y, n] \geq 0 \text{ for all } y < n, \tag{13}$$

$$\text{(ii) } r - h_1 E[D|y, n] - h_2 E[T|y, n] < 0 \text{ for } y = n. \tag{14}$$

In order to illustrate the derivation of the equilibrium threshold for a type *FI* customer, denoted n_{FI}, we use the example from previous sections and add the following parameter values: $\beta = 35$ and $h_2 = 25$. Table 2 presents values of Z_{FI} for different values of y and n for the above parameter values. Note that $I_\gamma(x)$ is calculated via hypergeometric functions, while the calculation of $E[S(t)|j]$ is terminated when the summation from $\gamma = 0$ to $\gamma = \gamma_0$ differs from the summation up to $\gamma = \gamma_0 + 1$ by less than 10^{-10}. A similar procedure is used when calculating $\psi_{\gamma,j}(t)$. It is shown that two values of n, $n = 5$ and $n = 6$, satisfy the conditions for an equilibrium threshold given in Eqs. (13),(14). Thus, two equilibrium thresholds are established in that case, $n_{MI} = 5$ and $n_{MI} = 6$.

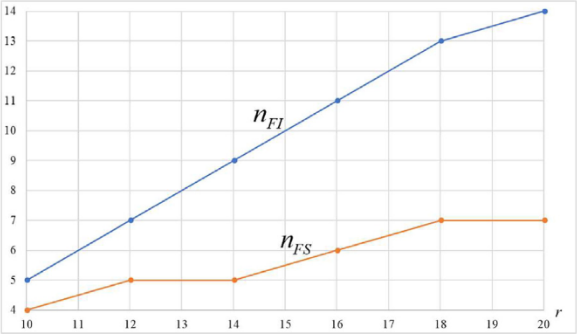

Fig. 8. Social and individual thresholds as a function of r for $C = 2$, $k_2 = 4$, $\lambda = 16$, $\mu = 20$, $h_1 = 45$, $\beta = 35$ and $h_2 = 25$.

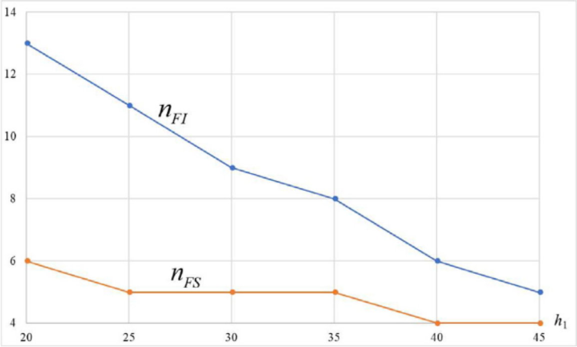

Fig. 9. Social and individual thresholds as a function of h_1 for $C = 2$, $k_2 = 4$, $\lambda = 16$, $\mu = 20$, $r = 10$, $\beta = 35$ and $h_2 = 25$.

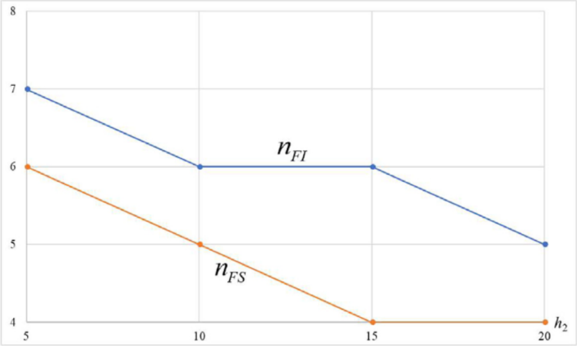

Fig. 10. Social and individual thresholds as a function of h_2 for $C = 2$, $k_2 = 4$, $\lambda = 16$, $\mu = 20$, $r = 10$, $h_1 = 45$ and $\beta = 35$.

3.4. Far-sighted social (FS) customer

Type *FS* customers' utility is given by

$$Z_{FS} = r\lambda_{eff} - h_1 E[L] - h_2 E[S^{reg}]. \tag{15}$$

Let n_{FS} be the equilibrium threshold of a type *FS* customer. Its value is calculated by maximizing Z_{FS}. Using the parameter values above, the equilibrium threshold is $n_{FS} = 4$. In line with the case of myopic customers, we wish to analyze far-sighted customers decisions with regard to social and individual thresholds. Using the above example as a baseline case, Figs. 8–10 present the two kinds of threshold as a function of various parameter values. The Figures show that the individual threshold is always higher than the social one.

3.5. Comparison between myopic and far-sighted customer's utilities

A myopic customer does not take into account the sojourn costs in the second stage, and, if such a customer also maximizes social utility, his/her behavior will be determined by the equilibrium threshold n_{MS}. However, in reality an MS cus-

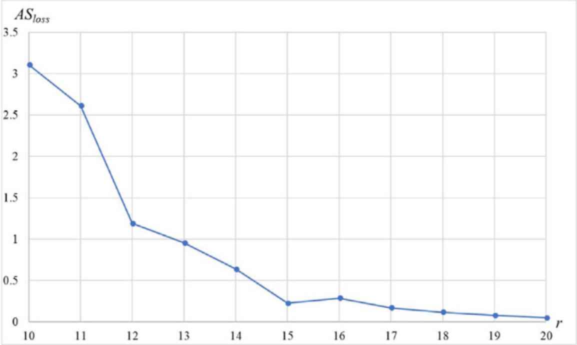

Fig. 11. Percentage of return for an FS customer as a function of r for $C = 2$, $k_2 = 4$, $\lambda = 16$, $\mu = 20$, $h_1 = 45$, $\beta = 35$ and $h_2 = 25$.

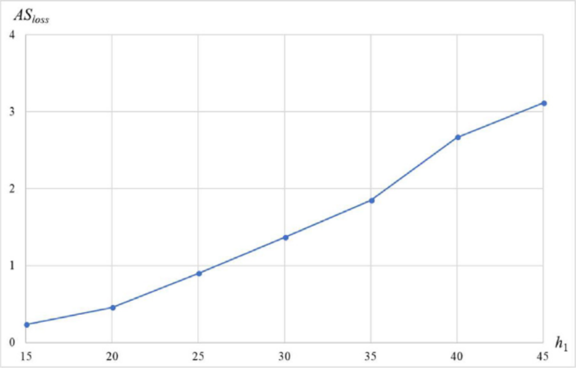

Fig. 12. Percentage of return for an FS customer as a function of h_1 for $C = 2$, $k_2 = 4$, $\lambda = 16$, $\mu = 20$, $r = 10$, $\beta = 35$ and $h_2 = 25$.

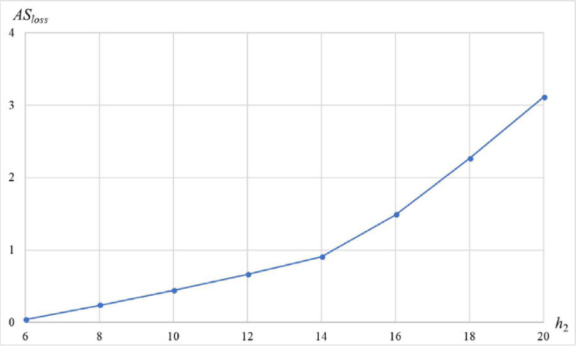

Fig. 13. Percentage of return for an FS customer as a function of h_2 for $C = 2$, $k_2 = 4$, $\lambda = 16$, $\mu = 20$, $r = 10$, $h_1 = 45$, and $\beta = 35$.

tomer still has to spend time in the second stage, and thus his/her real utility will be affected by the sojourn costs experienced at that second stage; this feature is captured by substituting the n_{MS} threshold into the Z_{FS} utility function.

Let $Z_{FS}(n_{MS})$ be the MS customer's actual utility, and let $Z_{FS}(n_{FS})$ be the FS customer's utility. Since n_{FS} maximizes Z_{FS}, it follows that $Z_{FS}(n_{FS}) \geq Z_{FS}(n_{MS})$ and that $[Z_{FS}(n_{FS})$ - $Z_{FS}(n_{MS})]$ represents the increase in utility due to a social utility maximiser being far-sighted. To examine the percentage increase in utility due to a far-sighted approach, we calculate: $FS_{rew} = (\frac{Z_{FS}(n_{FS})-Z_{FS}(n_{MS})}{Z_{FS}(n_{MS})}) \cdot 100$.

Fig. 11 shows that the percentage increase in utility decreases (almost monotonically) with r, and approaches zero when r becomes large. Fig. 12 shows that the return is small for small values of h_1 and then increases with h_1. Fig. 13 shows that the return is small for small values of h_2 and then increases monotonically.

4. The service system manager's problem

There is an implied benefit for the service system for offering customers the opportunity to complete stage 1 by APP for three reasons: (i) it reduces the balking rate since APP customers do not balk; (ii) it reduces customers' sojourn times,

Table 3
Values of n_{MS} for different values of d and k_2.

	$k_2 = 2$	$k_2 = 3$	$k_2 = 4$	$k_2 = 5$	$k_2 = 6$	$k_2 = 7$
$d = 0.00$	6	6	6	7	7	7
$d = 0.01$	6	6	7	7	7	7
$d = 0.02$	6	6	7	7	7	7
$d = 0.03$	6	7	7	7	7	7
$d = 0.04$	6	7	7	7	7	7
$d = 0.05$	7	7	7	7	7	7
$d = 0.06$	7	7	7	7	7	7
$d = 0.07$	7	7	7	7	7	7
$d = 0.08$	7	7	7	7	7	7
$d = 0.09$	7	7	7	7	7	7
$d = 0.10$	7	7	7	7	7	7

since APP customers skip the first service stage; and (iii) it increases the available vacation time for servers and hence their capacity to perform other tasks. Thus, there is an incentive for managers of the system to encourage strategic customers to use the application. To this end the manager can - and many do - set a price discount that is given only to customers who order via the APP.

Let θ be the manager's mean revenue per customer (strategic or APP) and let d be the discount expressed as a fraction of θ. Then, for $d \in [0, 1]$ and $\omega > 0$, let $\eta(d) = \eta(1 - e^{-\frac{\omega d}{1-d}})$ be the arrival rate of customers who switched from strategic to APP mode due to the discount, where $\eta \leq \lambda$ is the full potential arrival rate of customers who may be encouraged to use the application as $d \to 1$. Then, following the introduction of a discount, the arrival rate of strategic customers as a function of the discount is reduced to $\lambda(d) = \lambda - \eta(d)$, and similarly, the arrival rate of the APP customers is increased to $\alpha(d) = \alpha + \eta(d)$.

Regarding the management objective, the system benefits from servers' vacations that allows them to perform other functions. However, increasing the k_m values has a diminishing return effect on the system revenue $U(k_2,k_3,..., k_C)$ from server vacations. This effect is expressed by $U(k_2, k_3, ..., k_C) = a + \sum_{m=2}^{C} \frac{\delta_m (k_m - m)}{k_m - m + 1}$ where a is a constant representing the basic revenue if $k_m = m$ for all m, δ_m is a coefficient, and the term $\frac{(k_m - m)}{k_m - m + 1}$ represents the diminishing return phenomenon. On the other hand, increasing k_m increases customer sojourn costs. Thus, k_m for $m = 2, 3, ..., C$, are considered as system decision variables. Let c_1 and c_2 be the system cost-rate incurred as a result of a customer's sojourn time in the first and second service stage, respectively. Considering the performance measures λ_{eff}, $E[L]$ and $E[S]$ as functions of d and $\vec{k} = (k_2, k_3, ..., k_C)$, the system's profit function is given by

$$Z(d, \vec{k}) = \lambda_{eff}(d, \vec{k})\theta + \alpha(d)(1 - d)\theta + U(\vec{k}) - c_1 E\Big[L(d, \vec{k})\Big] - c_2 E\Big[S(d, \vec{k})\Big]. \tag{16}$$

The values of d and k_m ($m = 2, 3, ..., C$) affect the customers' choice of thresholds for balking, which will differ for each customer type. This effect should be taken into consideration by the system manager when determining the values of d and $\vec{k}$. Thus, the optimal solution is given by the triple $\{d, \vec{k}, n_{type}(d, \vec{k})\}$, $type \in \{MI, MS, FI, FS\}$, that maximizes $Z(d, \vec{k})$.

Using the parameter values from the preceding section, together with $\theta = 20$, $c_1 = 40$, $c_2 = 20$, and $\delta_2 = 35$, extensive numerical calculations result in the following optimal values:

(i) For the case where the customers are myopic and maximize individual utility (MI type) - $\{d = 0.04, \; k_2 = 3, n_{MI}(0.04{,}3) = 7\}$.
(ii) For the case where the customers are myopic and maximize social utility (MS type) - $\{d = 0.04, \; k_2 = 3, n_{MS}(0.04{,}3) = 7\}$.
(iii) For the case where the customers are far-sighted and maximize individual utility (FI type) - $\{d = 0.04, \; k_2 = 3, n_{FI}(0.04{,}3) = 6\}$.
(iv) For the case where the customers are far-sighted and maximize social utility (FS type) - $\{d = 0.04, \; k_2 = 3, n_{FS}(0.04{,}3) = 4\}$.

It is seen that the differences between the optimal values of d, k_2 and n for the four customer types are confined mainly to the level of the threshold values, n. However, since these threshold levels have a limited effect on the optimal value of $Z(d, \vec{k})$, the system's profit is only minorly affected (as it is comparatively insensitive to the optimal discount fraction d, and the server's vacation policy k_2). Thus, we present the calculation of optimal $Z(d, \vec{k})$ for the case where all customers are of type *MS* (see Tables 3 and 4), and the calculations for other customer types are performed in the same manner with slightly different values of $Z(d, \vec{k})$. Table 3 gives the values of the *MS* customer's equilibrium threshold, n_{MS}, for different values of d and k_2. Table 4 shows the values of the system's profit, Z_{MS}, for the same values of d and k_2 using the values of n_{MS} corresponding to each combination of d and k_2. Table 4 shows that the optimal value of $Z(d, \vec{k})$ is achieved when $d = 0.04$ and $k_2 = 3$ (indicated in bold) and its corresponding customer equilibrium threshold is $n_{MS} = 7$ as presented in Table 3.

Table 5 gives the values of the *MI* customer's equilibrium threshold, n_{MI}, for different values of d and k_2. It is seen that n_{MI} decreases with k_2, unlike the findings in Table 3 for an MS customer, where n_{MS} increases with k_2. This result can be explained as follows: an increase in k_2 results in a lower total service rate, which decreases the social reward. To

Table 4
Values of Z_{MS} for different values of d and k_2.

	$k_2 = 2$	$k_2 = 3$	$k_2 = 4$	$k_2 = 5$	$k_2 = 6$	$k_2 = 7$
$d = 0.00$	398.39	403.13	396.46	387.98	379.04	370.09
$d = 0.01$	401.04	407.28	402.8	396.55	390.59	384.96
$d = 0.02$	402.16	409.68	406.86	402.63	398.63	395.18
$d = 0.03$	402.17	410.75	409.25	406.46	403.92	401.89
$d = 0.04$	401.34	**410.78**	410.33	408.62	407.10	406.00
$d = 0.05$	399.88	410.03	410.39	409.49	408.68	408.18
$d = 0.06$	397.95	408.67	409.67	409.36	409.06	408.95
$d = 0.07$	395.66	406.84	408.33	408.47	408.53	408.69
$d = 0.08$	393.10	404.66	406.53	407.00	407.31	407.66
$d = 0.09$	390.35	402.19	404.37	405.09	405.59	406.06
$d = 0.10$	387.45	399.53	401.94	402.85	403.49	404.05

Table 5
Values of n_{MI} for different values of d and k_2.

	$k_2 = 2$	$k_2 = 3$	$k_2 = 4$	$k_2 = 5$
$d = 0.00$	8	8	7, 8	7
$d = 0.01$	8	8	7, 8	6
$d = 0.02$	8	8	7	6
$d = 0.03$	8	8	7	6
$d = 0.04$	8	8	7	6
$d = 0.05$	8	8	7	6
$d = 0.06$	8	8	7	6
$d = 0.07$	8	8	7	5
$d = 0.08$	8	8	7	5
$d = 0.09$	8	8	7	5
$d = 0.10$	8	8	7	5

compensate for this reduction in total service rate, more social customers join the queue (n_{MS} increases), which causes the total service rate to increase. However, MI type customers aim to maximize only their individual reward; thus an increase in mean sojourn time encourages them to balk (n_{MI} decreases). A similar effect occurs in relation to variations in d. Thus, in Table 3 n_{MS} increases with d, whereas in Table 5 n_{MI} decreases with d. This phenomenon is explained as follows: an increase in d causes a decrease in customer arrival rate to the first stage, which in turn decreases the total service rate and increases the customers' mean sojourn time. Thus, for the reasons set out above, n_{MS} increases for social customers, in order to compensate for the reduction in the total service rate, while n_{MI} decreases for MI customers due to the higher mean sojourn time and balking rates.

5. Conclusions and managerial implications

The strategic behavior of two types of customers (strategic and APP users) is studied in a two-stage service system in which servers conducting the first stage may take temporal 'vacation', depending on the actual queue size. Unlike strategic customers, APP customers can skip the first service-stage and join the second-stage queue directly. On the other hand, strategic customers who decide to join rather than to balk, have to pass sequentially through both service stages and incur waiting-time losses in both queues. Since a greater number of servers go on vacation when the first-stage queue size is small, we show that, contrary to Naor's well-known finding, the threshold of social customers is not always lower than the corresponding threshold of individual customers. Furthermore, there can be more than one equilibrium threshold for joining in the case of individual customers. Regarding the overall operation of the service station, offering a price discount to potential APP users may increase the service system manager's profit, reduce customers sojourn times, reduce the queue size in the first service-stage, and reduce the number of servers required to operate at that stage. A possible extension of the model is to assume that first-stage vacationing servers may dynamically (queue size dependent) join the second-stage team of servers to increase the latter's service rate.

We note that the analysis is for Markovian systems, which allow the use of a matrix geometric analysis. If the service times are general rather than Markovian, the analysis will become significantly more complex and elaborate.

Yet, the results of the analysis clearly indicate that the introduction of electronic ordering devices is beneficial for all customers – regular and app users – as well as for the facility management. It implies that it is beneficial for the management to invest in encouraging customers to order via apps. We propose the use of a price discount as a tool for that encouragement, and develop a scheme to obtain the optimal discount level which can be easily implemented. The results further imply that additional encouraging tools should be considered (e.g., offering discount points) to improve overall efficiency. Furthermore, it is shown that shifting servers to perform ancillary duties can increase the facility profit (although it may increase customers waiting time). From customers perspective, the waiting time increase from being myopic may be small in certain occasions. We emphasize the interesting result that is that in some cases, social equilibrium is higher than the individual one. This raises a controversial question: should a toll be imposed on balking customers (in order to increase the optimal individual joining threshold)?

Appendix A

Let

$$A = A_0 + A_1 + A_2 = \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \mu & -(\lambda+\mu) & \lambda & & 0 & 0 & \cdots & 0 \\ 0 & \mu & -(\lambda+\mu) & \ddots & \vdots & \vdots & & \vdots \\ \vdots & & \ddots & \ddots & \lambda & 0 & \cdots & 0 \\ 0 & 0 & & & -(\lambda+\mu) & \lambda & & 0 \\ 0 & 0 & 0 & & 2\mu & -(\lambda+2\mu) & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 2\mu & \ddots & 0 \\ \vdots & \vdots & \vdots & & \vdots & & \ddots & \lambda \\ 0 & 0 & 0 & \cdots & 0 & 0 & & -(\lambda+C\mu) \end{pmatrix}.$$

According to Neuts [29], the system's stability condition is

$$\vec{\pi} A_0 \vec{e}_{n+1} < \vec{\pi} A_2 \vec{e}_{n+1}, \tag{A.1}$$

where $\vec{\pi} = (\pi_0, \pi_1, ..., \pi_{k_2}, \pi_{k_2+1}, ...\pi_{k_3}, \pi_{k_3+1}, ..., \pi_{k_C}, \pi_{k_C+1}, ..., \pi_n)$ is the unique solution of the linear system

$$\vec{\pi} A = \vec{0}, \tag{A.2}$$

$$\vec{\pi} \cdot \vec{e}_{n+1} = 1. \tag{A.3}$$

Without loss of generality, set $k_0 = 0$. Then, by applying Eq. (A.1), we get $\sum_{m=0}^{C} \sum_{i=k_m}^{k_{m+1}-1} (\alpha + m\mu)\pi_i < \beta$, implying that

$$\alpha + \mu \sum_{m=1}^{C} \sum_{i=k_m}^{k_{m+1}-1} m\pi_i < \beta. \tag{A.4}$$

We first calculate π_i, $i = 0, 1, 2, ..., n$, by solving the set of Eqs. (A.2),(A.3). The set (A.2) can be written as

$$\begin{array}{l} \lambda\pi_0 - \mu\pi_1 = 0 \\ \lambda\pi_{i-1} - (\lambda+\mu)\pi_i + \mu\pi_{i+1} = 0 \quad i = 1, 2, ..., k_2 - 2 \\ \lambda\pi_{i-1} - (\lambda+\mu)\pi_i + 2\mu\pi_{i+1} = 0 \quad i = k_2 - 1 \\ \lambda\pi_{i-1} - (\lambda+2\mu)\pi_i + 2\mu\pi_{i+1} = 0 \quad i = k_2, k_2 + 1, ..., k_3 - 2 \\ \lambda\pi_{i-1} - (\lambda+2\mu)\pi_i + 3\mu\pi_{i+1} = 0 \quad i = k_3 - 1 \\ \lambda\pi_{i-1} - (\lambda+3\mu)\pi_i + 3\mu\pi_{i+1} = 0 \quad i = k_3, k_3 + 1, ..., k_4 - 2 \\ \vdots \\ \lambda\pi_{i-1} - (\lambda+(m-1)\mu)\pi_i + m\mu\pi_{i+1} = 0 \quad i = k_m - 1 \\ \lambda\pi_{i-1} - (\lambda+m\mu)\pi_i + m\mu\pi_{i+1} = 0 \quad i = k_m, k_m + 1, ..., k_{m+1} - 2 \\ \vdots \\ \lambda\pi_{i-1} - (\lambda+(C-1)\mu)\pi_i + C\mu\pi_{i+1} = 0 \quad i = k_C - 1 \\ \lambda\pi_{i-1} - (\lambda+C\mu)\pi_i + C\mu\pi_{i+1} = 0 \quad i = k_C, k_C + 1, ..., n \end{array} . \tag{A.5}$$

Let $\xi = \frac{\lambda}{\mu}$. Rearranging the terms in (A.5) leads to

$$\begin{array}{ll} \pi_i = \xi^i \pi_0 & i = 1, 2, ..., k_2 - 1 \\ \pi_i = \frac{\xi^i}{2^{i-k_2+1}} \pi_0 & i = k_2, k_2 + 1, ..., k_3 - 1 \\ \pi_i = \frac{\xi^i}{2^{k_3-k_2} 3^{i-k_3+1}} \pi_0 & i = k_3, k_3 + 1, ..., k_4 - 1 \\ \vdots & \\ \pi_i = \frac{\xi^i}{C^{i-k_C+1} \prod\limits_{a=1}^{C-1} a^{k_{a+1}-k_a}} \pi_0 & i = k_C, k_C + 1, ..., n \end{array} . \tag{A.6}$$

Without loss of generality set $k_{C+1} - 1 = n$. Then (A.6) can be written as

$$\pi_i = \frac{\xi^i}{m^{i-k_m+1} \prod_{a=1}^{m-1} a^{k_{a+1}-k_a}} \pi_0 \quad m = 1, 2, ..., C \quad i = k_m, k_m + 1, ..., k_{m+1} - 1. \tag{A.7}$$

Substituting (A.7) into (A.3) results in

$$\pi_0\left(1+\sum_{m=1}^{C}\sum_{i=k_m}^{k_{m+1}-1}\frac{\xi^i}{m^{i-k_m+1}\prod_{a=1}^{m-1}a^{k_{a+1}-k_a}}\right)=1. \tag{A.8}$$

Let $\rho_m=\frac{\xi}{m}$. Rearranging the terms in (A.8) we get $\pi_0(1+\sum_{m=1}^{C}\frac{1}{m^{-k_m+1}\prod_{a=1}^{m-1}a^{k_{a+1}-k_a}}\sum_{i=k_m}^{k_{m+1}-1}\frac{\xi^i}{m^i})=1$, which leads to $\pi_0(1+\sum_{m=1}^{C}\frac{m^{k_m-1}}{\prod_{a=1}^{m-1}a^{k_{a+1}-k_a}}(\frac{\rho_m^{k_{m+1}}-\rho_m^{k_m}}{\rho_m-1}))=1$, or

$$\pi_0=\left(1+\sum_{m=1}^{C}\frac{m^{k_m-1}}{\prod_{a=1}^{m-1}a^{k_{a+1}-k_a}}\left(\frac{\rho_m^{k_{m+1}}-\rho_m^{k_m}}{\rho_m-1}\right)\right)^{-1}. \tag{A.9}$$

By substituting (A.9) into (A.7), π_i, $i=0, 1, 2, ..., n$, are calculated.

We are now ready to obtain the desired stability condition. By substituting (A.7) into (A.4) we get $\alpha+\mu\pi_0\sum_{m=1}^{C}\sum_{i=k_m}^{k_{m+1}-1}m\frac{\xi^i}{m^{i-k_m+1}\prod_{a=1}^{m-1}a^{k_{a+1}-k_a}}<\beta$, which leads to

$$\alpha+\mu\pi_0\sum_{m=1}^{C}\frac{m^{k_m}}{\prod_{a=1}^{m-1}a^{k_{a+1}-k_a}}\left(\frac{\rho_m^{k_{m+1}}-\rho_m^{k_m}}{\rho_m-1}\right)<\beta. \tag{A.10}$$

Finally, substituting (A.9) into (A.10) leads to

$$\alpha+\mu\sum_{m=1}^{C}\frac{m^{k_m}}{\prod_{a=1}^{m-1}a^{k_{a+1}-k_a}}\left(\frac{\rho_m^{k_{m+1}}-\rho_m^{k_m}}{\rho_m-1}\right)\left(1+\sum_{m=1}^{C}\frac{m^{k_m-1}}{\prod_{a=1}^{m-1}a^{k_{a+1}-k_a}}\left(\frac{\rho_m^{k_{m+1}}-\rho_m^{k_m}}{\rho_m-1}\right)\right)^{-1}<\beta.$$

Appendix B

Let Θ_1 be the number of customers who pass from the first stage to the second stage in the interval between the customer's arrival and his/her service beginning at stage 1. Let Θ_2 be the number of customers who pass from the first stage to the second stage while a customer is being served. Let $q_{L,H}$ be the probability that an NAS customer who sees y customers in the system will visit state (L, H) before entering the service. Each $q_{L,H}$ for each L and H can be recursively calculated by the following procedure.

Procedure 2. Calculation of $q_{L,H}$ for each *L* and *H*.

(i) $q_{L,L-m(L)}=0,\ L=y+2,\ y+3,\ ...,\ n$
(ii) $q_{L,L-m(L)}=1,\ L=y+1$
(iii) $q_{L,L-m(L)}=q_{L+1,L+1-m(L+1)}\frac{m(L+1)\mu}{\lambda+m(L+1)\mu},\ L=2,\ 3,\ ...,\ y$
(iv) $q_{L,L-m(L)-l}=q_{L+1,L+1-m(L+1)-l}\frac{m(L+1)\mu}{\lambda+m(L+1)\mu}+q_{L-1,L-m(L-1)-l}\frac{\lambda}{\lambda+m(L-1)\mu}$,
$L=3,\ 4,\ ...,\ n-2,\ l=1,\ 2,\ ...,\ L-m(L)-1$
(v) $q_{L,L-m(L)-l}=q_{L+1,L+1-m(L+1)-l}+q_{L-1,L-m(L-1)-l}\frac{\lambda}{\lambda+m(L-1)\mu}$,
$L=n-1,\ l=1,\ 2,\ ...,\ L-m(L)-1$
(vi) $q_{L,L-m(L)-l}=q_{L-1,L-m(L-1)-l}\frac{\lambda}{\lambda+m(L-1)\mu},\ L=n,\ l=1,\ 2,\ ...,\ L-m(L)-1$
(vii) $q_{L,0}=q_{L+1,1}\frac{m(L+1)\mu}{\lambda+m(L+1)\mu},\ L=1,\ 2,\ ...,\ n-2\neq k_m-1,\ k_m,\ m=1,\ 2,\ ...,\ C$
(viii) $q_{L,0}=q_{L+1,1}\frac{m(L+1)\mu}{\lambda+m(L+1)\mu}+q_{L-1,1}\frac{\lambda}{\lambda+m(L-1)\mu},\ L=k_m,\ m=1,\ 2,\ ...,\ C$
(ix) $q_{L,0}=q_{L+1,1},\ L=n-1$

A customer reaches the service station only after all y customers present in the queue upon his/her arrival have started their service. However, not all of those y customers will have completed their first stage service and passed to the second stage when s/he starts service. Specifically, assuming that when a customer starts his/her service, there are m active servers, only $y-m+1$ customers will have passed to the second stage by that time. Thus, the mean number of customers who will have passed from first stage to the second in the interval between the customer's arrival to the queue and their start of service is given by

$$\Theta_1=\sum_{\substack{L=1\\ L\neq k_m-1,\forall m}}^{n}(y-m(L)+1)q_{L,0}.$$

During the time a customer is served, s/he occupies exactly one server, while the rest of the servers can serve other customers. Given that one server is occupied and at least k_2-2 customers are present, in addition to our customer, the system can be described as a parallel system where $C' = C - 1$ and $n' = n - k_2 + 2$. When, in the parallel system,$k_2 - 2$ customers are present, the system is idle (analogous to 0 customers in the regular system); with $k_2 - 1$ customers in a parallel system, there is 1 customer in the regular system, etc. Since this parallel system can be idle during the time the customer obtains service, the passing rate from the first stage to the second during this time is λ'_{eff}, and thus the number of customers who pass from the first stage to the second stage during the time the customer is in the service is given by λ'_{eff}/μ.

Suppose that, when a customer starts service, $i < k_2 - 2$ customers are present in the system. In such a case, if the customer's service is completed before $k_2 - 2 - i$ customers have arrived, then zero customers will have proceeded to the second stage during the customer's service. However, if $k_2 - 2 - i$ customers join the system before the customer's service completion, with probability $(\frac{\lambda}{\lambda+\mu})^{k_2-2-i}$, the system starts acting like the parallel system described above. Similarly, if, when a customer starts service, $i \geq k_2 - 2$ customers are present, the system immediately starts acting like a parallel system. The probability that, when the customer starts service, i customers are present in the system is $q_{i+1,0}$. Thus, the number of customers passing to the second stage during the service is

$$\Theta_2 = \left(\sum_{i=0}^{k_2-3} q_{i+1,0} \left(\frac{\lambda}{\lambda+\mu} \right)^{k_2-2-i} + \sum_{i=k_2-1}^{n-2} q_{i+1,0} \right) \lambda'_{eff}/\mu.$$

The above is the probability that the parallel system will be activated multiplied by the number of passing customers in that case. Note that the probability $q_{k_2-1,0}$ does not exist (see Procedure 2), and thus the probability of $k_2 - 1$ other customers being present in the system by the time the customer starts service is zero. Then, the total number of customers who pass from the first stage to the second stage in the interval between a customer's arrival at the stage 1 queue and their completion of the stage 1 service is $\Theta_1 + \Theta_2$, and the passing rate during this time is $(\Theta_1 + \Theta_2)/E[D|y,n]$. Finally, α_{eff} is given by

$$\alpha_{eff} = \alpha + \frac{\Theta_1 + \Theta_2}{E[D|y, n]}.$$

Appendix C: Laplace-Stieltjes transforms

Let $\widetilde{D}(x) = \frac{x}{x+s}$, and let $\widetilde{D}_{L,H}$ be the Laplace Stieltjes transform (LST) of a customer's waiting time in the first stage, starting from state (L, H), until s/he starts service. Each $\widetilde{D}_{L,H}$, for all L and H, can be recursively calculated by the following procedure:

Procedure 3. Calculation of $\widetilde{D}_{L,H}$for all *L* and *H*.

(i) $\widetilde{D}_{L,1} = \widetilde{D}(C\mu), \quad L = k_C + 1, k_C + 2, ..., n$

(ii) $\widetilde{D}_{L,1} = \widetilde{D}(\lambda + m(L)\mu)(\frac{\lambda}{\lambda+m(L)\mu}\widetilde{D}_{L+1,1} + \frac{m(L)\mu}{\lambda+m(L)\mu} \cdot 1),$
$L = 2, 3, ..., k_C - 1 \neq k_m - 1, k_m, m = 2, 3, ..., C$

(iii) $\widetilde{D}_{L,1} = \widetilde{D}(\lambda + m(L)\mu), L = k_m - 1, m = 2, 3, ..., C$

(iv) $\widetilde{D}_{L,1} = \widetilde{D}(\lambda + m(L)\mu)(\frac{m(L)\mu}{\lambda+m(L)\mu}\widetilde{D}_{L-1,1} + \frac{\lambda}{\lambda+m(L)\mu}\widetilde{D}_{L+1,1}),$
$L = k_m, m = 2, 3, ..., C$

(v) $\widetilde{D}_{L,L-m(L)-l} = \widetilde{D}(\lambda + m(L)\mu)(\frac{m(L)\mu}{\lambda+m(L)\mu}\widetilde{D}_{L-1,L-1-m(L-1)-l} + \frac{\lambda}{\lambda+m(L)\mu}\widetilde{D}_{L+1,L-m(L+1)-l}),$
$L = 3, 4, ..., n - 1, l = 0, 1, 2, ..., L - m(L) - 2$

(vi) $\widetilde{D}_{L,H} = \widetilde{D}(C\mu)\widetilde{D}_{L-1,H-1}, L = n, H = 2, 3, 4, ..., n - C.$

Thus, the LST of a waiting time of a newly arriving customer who sees y customers in the first stage is $\widetilde{D}_{y+1,y+1-m}$ (see the explanation under Procedure 1). Then, the LST of the customer's sojourn time is given by

$$\widetilde{D}|y = \widetilde{D}_{y+1,y+1-m}\widetilde{D}(\mu),$$

and finally, the $f_{D|y}(t)$ is calculated as inverse of $\widetilde{D}|y$.

Appendix D: Notation

n – customer's joining threshold
λ – arrival rate of strategic customers at the first service stage
α – arrival rate of APP customers at the second service stage
C – maximal number of servers in the first stage
k_m – a pre-specified number such that when the first-stage queue size drops below it, server no. m takes a vacation.
μ – service rate in the first stage
β – service rate in the second stage
L – number of customers in the first stage in steady state
S – number of customers in the second stage in steady state

$p_{i,j} = \Pr(L = i,\ S = j)$ – steady state probabilities where $i = 0, 1, 2, \ldots, n;\ j = 0, 1, 2, \ldots$

$E[L]$ – mean number of customers in the first stage (note that all customers in the first stage are strategic)

$E[S^{reg}]$ – mean number of 'regular' (namely strategic) customers in the second stage

$E[S^{app}]$ – mean number of APP customers in the second stage, $E[S] = E[S^{reg}] + E[S^{APP}] =$ mean total number of customers in the second stage

$E[D]$ – mean sojourn time of a strategic customer in the first stage

$E[T^{reg}]$ – mean sojourn time of a strategic customer in the second stage

$E[T^{app}]$ – mean sojourn time of an APP customer in the second stage

$E[W] = E[D] + E[T^{reg}]$ – total mean sojourn time in the system (in both stages) of a strategic customer

$E[V]$ – mean number of servers on vacation

λ_{eff} – effective arrival rate of strategic customers type

MI – myopic strategic customer who maximizes individual utility type

MS – myopic strategic customer who maximizes social utility type

FI – far-sighted strategic customer who maximizes individual utility type

FS – far-sighted strategic customer who maximizes social utility

n_{MI} – equilibrium threshold exercised by type *MI* customers

n_{MS} – equilibrium threshold exercised by type *MS* customers

n_{FI} – equilibrium threshold exercised by type *FI* customers

n_{FS} – equilibrium threshold exercised by type *FS* customers

Z_{MI} – utility of type *MI* customer

Z_{MS} – utility of type *MS* customer

Z_{FI} – utility of type *FI* customer

Z_{FS} – utility of type *FS* customer

r – reward gained by a customer from obtaining service

h_1 – sojourn cost rate of a customer in the first stage

h_2 – sojourn cost rate of a customer in the second stage

y – number of customers that an NAS customer sees in the first stage

$E[D|y, n]$ – mean sojourn time in the first stage of an NAS customer who sees y customers in the first stage when the threshold is n

H – position of the customer in the first stage queue

$m(L)$ – number of active servers corresponding to a given L, based on the servers vacation policy

$D(L, H)$ – mean total waiting time of a customer starting from state (L, H) until the start of his/her service

p_{ser} – probability of a customer to obtain service

$E[S|i]$ – mean number of customers in the second stage given that there are i customers in the first stage $i = 0, 1, 2, \ldots, n$

$E[S(t)|j]$ – mean number of customers who will occupy the second stage t time units from the present time, given that j customers are currently in the second stage

$E[T|y, n]$ – mean sojourn time in the second stage of a customer who sees y customers in the first stage, for a given, already known, threshold n

$\psi_{\gamma,j}(t)$ – probability that there will be γ customers in the second stage t time units from the present time, given that j customers are currently in the second stage

$f_{D|y}(t)$ – density function of the customer's sojourn time in the first stage starting from his/her arrival until the start of his/her service

α_{eff} – overall arrival rate at the second stage (including strategic and APP customers)

FS_{rew} – percentage increase in monetary utility caused by an FS customer

θ – mean revenue received by the management per customer (strategic or APP)

d – discount fraction on θ for APP customers

$\eta(d)$ – arrival rate of customers who switched from strategic to APP customers due to the discount d

$\lambda(d)$ – arrival rate of strategic customers given the discount offer d

$\alpha(d)$ – arrival rate of APP customers given the discount offer d

$U(k_2,k_3,\ldots, k_C)$ – system revenue associated with the work done by servers on vacation under policy $\{k_m\}$

c_1 – system's cost-rate incurred as a result of a customer's sojourn time in the first service-stage

c_2 – system's cost-rate incurred as a result of a customer's sojourn time in the second service-stage

$Z(d, \vec{k})$ – system profit for policy $\vec{k} = (k_2, k_3, \ldots, k_C)$

Θ_1 – number of customers who pass from the first stage to the second stage from the moment a customer arrives until s/he begins to receive first-stage service

Θ_2 – number of customers who pass from the first stage to the second stage during the time a customer is receiving first stage service

$q_{L,H}$ – probability that an NAS customer who sees y customers in the first stage will visit the state (L, H) before starting service.

References

[1] D. Rigby, The future of shopping, Harv. Bus. Rev. 89 (12) (2011) 65–76.

[2] N. Beck, D. Rygl, Categorization of multiple channel retailing in Multi-, Cross-, and Omni-Channel Retailing for retailers and retailing, J. Retail. Consum. Serv. 27 (2015) 170–178.

[3] S. Gallino, A. Moreno, Integration of online and offline channels in retail: the impact of sharing reliable inventory availability information, Manag. Sci. 60 (6) (2014) 1434–1451.

[4] F. Gao, X. Su, Omnichannel retail operations with buy-online-and-pick-up-in-store, Manag. Sci. 63 (8) (2016) 2478–2492.

[5] F. Gao, X. Su, Online and offline information for omnichannel retailing, Manuf. Serv. Oper. Manag. 19 (1) (2016) 84–98.

[6] E. Kim, M.C. Park, J. Lee, Determinants of the intention to use Buy-Online, Pickup In-Store (BOPS): the moderating effects of situational factors and product type, Telemat. Inform. 34 (8) (2017) 1721–1735.

[7] M. Jin, G. Li, T.C.E. Cheng, Buy online and pick up in-store: design of the service area, Eur. J. Oper. Res. 268 (2) (2018) 613–623.

[8] X. Shi, C. Dong, T.C.E. Cheng, Does the buy-online-and-pick-up-in-store strategy with pre-orders benefit a retailer with the consideration of returns? Int. J. Prod. Econ. 206 (2018) 134–145.

[9] O. Baron, X. Chen, Y. Li, The paradox of choice: the false premise of omnichannel services and how to realize it. SSRN Electron. J. (2019) 1–61. Available at SSRN 3444772.

[10] F. Gao, X. Su, Omnichannel service operations with online and offline self-order technologies, Manag. Sci. 64 (8) (2017) 3595–3608.

[11] R. McMillan, Facebook hopes chatbots can solve app overload, Wall Str. J. (2016) April 17 http://www.wsj.com/articles/facebook-hopes-chatbots-can-solve-app-overload-1460930220 .

[12] M. Yadin, P. Naor, Queueing systems with a removable service station, J. Oper. Res. Soc. 14 (4) (1963) 393–405.

[13] O. Kella, The threshold policy in the M/G/1 queue with server vacations, Nav. Res. Logist. 36 (1) (1989) 111–123.

[14] R. Liu, Z. Deng, On the steady-state system size distribution for a discrete-time Geo/G/1 repairable queue, Discret. Dyn. Nat. Soc. (2014), doi:10.1155/2014/924712.

[15] D.Y. Yang, C.H. Wu, Cost-minimization analysis of a working vacation queue with N-policy and server breakdowns, Comput. Ind. Eng. 82 (2015) 151–158.

[16] M. Haridass, R. Arumuganathan, Analysis of a single server batch arrival retrial queueing system with modified vacations and N-policy, RAIRO Oper. Res 49 (2) (2015) 279–296.

[17] D.H. Lee, W.S. Yang, The N-policy of a discrete time Geo/G/1 queue with disasters and its application to wireless sensor networks, Appl. Math. Model. 37 (23) (2013) 9722–9731.

[18] W. Sun, S. Li, E. Cheng-Guo, Equilibrium and optimal balking strategies of customers in Markovian queues with multiple vacations and N-policy, Appl. Math. Model. 40 (1) (2016) 284–301.

[19] D.E. Lim, D.H. Lee, W.S. Yang, K.C. Chae, Analysis of the GI/Geo/1 queue with N-policy, Appl. Math. Model. 37 (7) (2013) 4643–4652.

[20] C. Sreenivasan, S.R. Chakravarthy, A. Krishnamoorthy, MAP/PH/1 queue with working vacations, vacation interruptions and N policy, Appl. Math. Model. 37 (6) (2013) 3879–3893.

[21] J.C. Ke, H.I. Huang, Y.K. Chu, Batch arrival queue with N-policy and at most J vacations, Appl. Math. Model. 34 (2) (2010) 451–466.

[22] D. Bergemann, B. Brooks, S. Morris, The limits of price discrimination, Am. Econ. Rev. 105 (3) (2015) 921–957.

[23] D. Gao, N. Wang, Z. He, T. Jia, The bullwhip effect in an online retail supply chain: a perspective of price-sensitive demand based on the price discount in e-commerce, IEEE Trans. Eng. Manag. 64 (2) (2017) 134–148.

[24] G. Hanukov, T. Avinadav, T. Chernonog, U. Yechiali, A service system with perishable products where customers are either fastidious or strategic, Int. J. Prod. Econ. 228 (2020), doi:10.1016/j.ijpe.2020.107696.

[25] S. Ma, J. Lin, X. Zhao, Online store discount strategy in the presence of consumer loss aversion, Int. J. Prod. Econ. 171 (1) (2016) 1–7.

[26] C.Y. Dye, T.P. Hsieh, Joint pricing and ordering policy for an advance booking system with partial order cancellations, Appl. Math. Model. 37 (6) (2013) 3645–3659.

[27] R.G. Yaghin, Integrated multi-site aggregate production-pricing planning in a two-echelon supply chain with multiple demand classes, Appl. Math. Model. 53 (2018) 276–295.

[28] P. Naor, The regulation of queue size by levying tolls, Econ. J. Econ. Soc. 37 (1) (1969) 15–24.

[29] M.F. Neuts, Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach, Johns Hopkins University Press, Baltimore, MD, 1981.

[30] M. Harchol-Balter, Performance Modeling and Design of Computer Systems: Queueing Theory in Action, Cambridge University Press, 2013.

[31] G. Hanukov, U. Yechiali, Explicit solutions for continuous-time QBD processes by using relations between matrix geometric analysis and the probability generating functions method, Probab. Eng. Inf. Sci. (2020) 1–16, doi:10.1017/S0269964819000470.

[32] R. Hassin, M. Haviv, To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems, Springer Science & Business Media, 2003.

[33] U. Yechiali, On optimal balking rules and toll charges in the GI/M/1 queuing process, Oper. Res. 19 (2) (1971) 349–370.

[34] U. Yechiali, Customers' optimal joining rules for the GI/M/s queue, Manag. Sci. 18 (7) (1972) 434–443.

[35] R. Hassin, Rational Queueing, CRC Press, Taylor & Francis Group LLC, 2016.

[36] L. Kleinrock, Queueing Systems, I, Theory, Wiley, 1975.