Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 52-13 (2019) 671–676

A multi-server queueing-inventory system with stock-dependent demand

Gabi Hanukov[1], Tal Avinadav[2], Tatyana Chernonog[3], Uri Yechiali[4]

[1]Department of Management, Bar-Ilan University, Ramat Gan, 5290002, Israel, german.kanukov@biu.ac.il
[2]Department of Management, Bar-Ilan University, Ramat Gan, 5290002, Israel, tal.avinadav@biu.ac.il
[3]Department of Management, Bar-Ilan University, Ramat Gan, 5290002, Israel, tatyana.chernonog@biu.ac.il
[4]Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv, 6997801, Israel, uriy@post.tau.ac.il

Abstract: We consider a two-server service system in which idle servers produce and store preliminary services for reducing the sojourn time of incoming customers and increasing customers arrival rate. We model the system as a Markovian process and provide a method, based on matrix geometric (MG) analysis, to obtain closed-form solutions to the steady state probabilities and relevant performance measures. We show the relation of the elements in the rate matrix of the MG to Catalan numbers, and prove that the stability condition remains as is in the traditional M/M/2 queue, although the expected sojourn time of customers has been reduced. We provide an economic analysis for a system in which the PS capacity and the investment to increase customers' arrival rate are decision variables.

Keywords: Markov process, queueing-inventory system, stock-dependent demand, matrix geometric analysis, economic analysis

1. INTRODUCTION

Reducing customers waiting times in service systems is a common objective, which aims to reduce costs associated with customers sojourn time in the system and increase their satisfaction. Usually, it is achieved by using faster serves (e.g., Hwang et al. 2010, Guo and Zhang 2013) or by hiring additional servers. However, these methods are costly. Recent works (Hanukov et al. 2017, 2018a,b) have suggested to use a server's idle time to produce and stock preliminary services, which will be used to serve faster future customers, and thus increase the productivity of the service system without investing in additional or better resources. This innovative approach is modelled using a queueing-inventory system represented by a two-dimensional Markov process (see e.g. Cox 2017), which is solved using a matrix geometric analysis. The main question in these models is how many preliminary services should be produced and stocked when the server is idle. Actually, this approach is in contrast to extensive literature according to which servers' idle times may be used for carrying out ancillary duties (also known as 'vacation models' in e.g. Levy and Yechiali 1975, 1976; Rosenberg and Yechiali 1993; Boxma et al. 2002; Yechiali 2004; Yang and Wu 2015; Mytalas and Zazanis 2015; and Guha et al. 2016) which may increase customers waiting times. So far, this innovative approach has been applied (analytically) only for a single-server service systems and homogenous Poisson (customer) arrival process.

The current work applies the above innovative approach in the framework of a multi-server service system with stock-dependent customers' arrival rate. Multi-server systems are widely spread in real life businesses, but in case of queueing-inventory systems are considerably more complicated to analyze. As for stock-dependent demand, it is known (see, e.g., Urban 2005) that customers buy more units when the level of the presented inventory is higher due to marketing effects (e.g., a wider selection). In our model, the stock of preliminary services has an additional positive effect on customers demand since it implies a shorter sojourn time in the system. To model this effect, we use state-dependent customers' arrival rate (Gupta 1967, Winston 1978, Guo and Hassin 2011, Zhao and Lian 2011).

Our model is mostly motivated by the fast food industry in which food, such as hamburger patties or basic pizza, can be prepared before demand occurs, and only upon the arrival of a customer they are heated up, adjusted according to specific customers requirement, and served to the customer. Another domain in which our model is applicable is bicycle retailing; where mechanics can assemble parts of a bicycle before a customer's arrival, and subsequently assemble the remaining parts in accordance with the customer's specific requirements and preferences. Handmade nameplates for doors are another example in which service can be split up. The server can produce basic nameplates from wood, ceramic or glass (which are not identical) before an order is made, and complete the nameplate upon request (e.g., writing the name, add decoration, etc.).

2. MODEL FORMULATION

Assume a service system with two servers, in which full-service (FS) time is exponentially-distributed with mean $1/\mu$, and customers arrival follows a Poisson process with rate λ. The service can be split into two phases: preliminary service (PS), which can be carried out without the presence of customers, and complementary service (CS), which can be given only when a customer is present. When a server is idle, he/she can produce PSs. In line with prior works in this domain (Benjaafar et al. 2011, Flapper et al. 2012, Iravani et al. 2012, Hanukov et al. 2017), the PS's preparation time is as-

2405-8963 © 2019, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2019.11.124

sumed to be exponentially distributed with parameter *α*. The PSs are stored until the arrival of customers, and, as indicated, are used to reduce the sojourn time of future customers.

The total number of PSs in stock is limited to *n* (capacity), and when the inventoried PSs reach this capacity limit, the servers stop producing PSs and return to idle position. When a customer reaches the head of the line and a PS is available, his/her CS phase starts immediately. The CS time is exponentially distributed with mean 1/*β*, where *β* >*μ*. That is, the CS duration is stochastically shorter than an FS duration (first order stochastic dominance). When PSs are available, customers' arrival rate increases to $\delta(>\lambda)$ until the number of customers is equal to the number of PSs (which are both observable by the customers).

The process can be formulated as a three-dimensional continuous-time Markov chain with a state space $\{L_t, S_t, U_t\}$, where $L_t \in \{0,1,2,...,\infty\}$ denotes the number of customers in the system at time *t*, $S_t \in \{0,1,2,..n\}$ denotes the total number of PSs at time *t*, and $U_t \in \{0,1,2\}$ denotes the number of customers at their CS phase at time *t*. Let $L \equiv \lim_{t\to\infty} L_t$, $S \equiv \lim_{t\to\infty} S_t$ and $U \equiv \lim_{t\to\infty} U_t$. Let $p_{i,j,k} = \Pr(L=i, S=j, U=k)$ denote the joint probability distribution of the latter three-dimensional process. $p_{i,j,k}$ represents the long run fraction of time that the system stays in state $(L=i, S=j, U=k)$. As an illustration, Figure 1 bellow depicts the system's states and its corresponding transition rate diagram for the case $n=3$. Each circle in the diagram depicts a state $(L=i, S=j, U=k)$ and the arrows indicate the transition directions and rates.

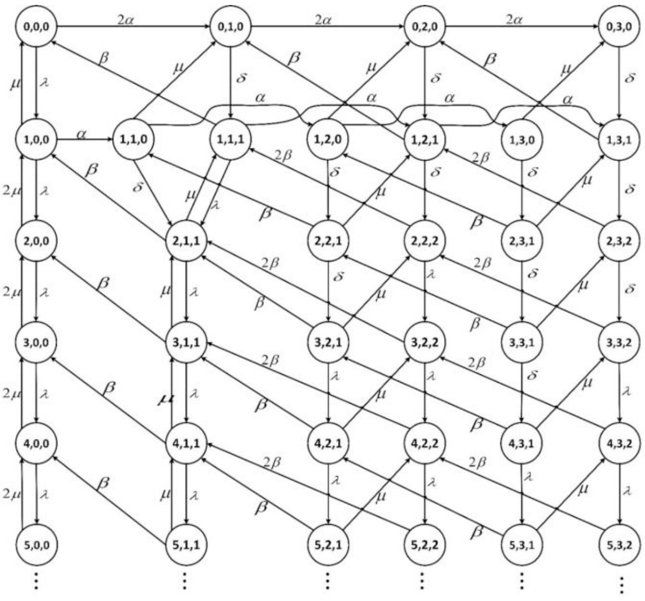

Figure 1. System states and transition-rate diagram for *n* = 3.

Our goal is to obtain the system's stability condition (i.e., the condition on the parameters insuring that the queue size will not grow beyond any bound) and to derive the system's steady-state probabilities (i.e., $p_{i,j,k}$ when the system is stable). Then, various performance measures, such as mean queue size, mean sojourn time, mean inventory level of PSs, mean time a PS spends in inventory, etc. are calculated. To do so, matrix geometric analysis (Neuts 1981) is applied, which entails calculation of the so-called rate matrix *R* (see bellow). The system's states are arranged in the following lexicographic order (*i* = 2,3,…):

$$\{(0,0,0),(0,1,0),...,(0,n,0);(1,0,0),(1,1,0),(1,1,1),...,(1,n,0),(1,n,1);...;$$
$$(i,0,0),(i,1,1),(i,2,1),(i,2,2),...,(i,n,1),(i,n,2);...\},$$

and construct its infinitesimal generator matrix, denoted by *Q*:

$$Q = \begin{pmatrix} B_{1,1} & B_{1,2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ B_{2,1} & B_{2,2} & B_{2,3} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & B_{3,1} & B_{3,2} & B_{3,3} & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & A_2 & B_{4,2} & B_{4,3} & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \ddots & & \vdots & \vdots & \\ 0 & 0 & 0 & 0 & A_2 & B_{n+1,2} & B_{n+1,3} & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & A_2 & A_1 & A_0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \ddots & \ddots & \ddots \end{pmatrix}, \quad (1)$$

where matrices $B_{i,j}$ and A_k ($k=0,1,2$) are given in the Appendix.

3. STEADY STATE ANALYSIS

We now claim that the condition for stability of the system is identical to that of a regular M/M/2 queue without PSs (i.e., a Markovian service system with two parallel identical servers; see, e.g., Harchol-Balter 2013 p. 258):

Theorem 1. The stability condition of the queuing-inventory system is $\lambda < 2\mu$.

Proof. The condition for stability of our system, according to Neuts (1981, p. 83) is

$$\vec{\pi} A_0 \vec{e} < \vec{\pi} A_2 \vec{e}, \quad (2)$$

where A_0 and A_2 (as well as A_1) are given in the Appendix. $\vec{e} = (1,1,...,1)^T$ is a vector of ones, and $\vec{\pi} = (\pi_1, \pi_2, ..., \pi_{2n})$ is the unique solution of the linear system

$$\begin{aligned} \vec{\pi} A &= \vec{0} \\ \vec{\pi}\vec{e} &= 1, \end{aligned} \quad (3)$$

where $A \equiv A_0 + A_1 + A_2$. In our case, $\vec{\pi} = (1,0,...,0)$. Thus, using A_0 and A_2, the stability condition in equation (2) translates into $\lambda < 2\mu$.

Theorem 1 implies that although preparing PSs improves the system's performance, it does not stringent its stability condition. This result can be further explained as follows: due to the stochastic nature of the system, at some point of time both servers will be fully occupied, and will not be able to produce any PS, so the system will become a regular M/M/2 queue. This explanation leads us to state the following conjecture:

Conjecture 1. The condition for stability of the queueing-inventory system as described above having *s* identical servers is $\lambda < s\mu$.

The matrix geometric analysis is based on derivation of a so-called rate-matrix $R \equiv [r_{i,j}]_{2n\times 2n}$, which is obtained by solving the quadratic matrix equation (see Neuts 1981)

$$A_0 + RA_1 + R^2 A_2 = 0_{2n\times 2n}. \qquad (4)$$

Define the $(n+1)$-dimensional vector of probabilities

$\vec{p}_0 \equiv \left(p_{0,0,0} \quad p_{0,1,0} \quad p_{0,2,0} \quad p_{0,3,0} \quad \cdots \quad p_{0,n,0} \right)$,

the $(2n+1)$-dimensional vector of probabilities

$\vec{p}_1 \equiv \left(p_{1,0,0} \quad p_{1,1,0} \quad p_{1,1,1} \quad p_{1,2,0} \quad p_{1,2,1} \quad p_{1,3,0} \quad p_{1,3,1} \quad \cdots \quad p_{1,n,0} \quad p_{1,n,1} \right)$

and the $(2n)$-dimensional vectors of probabilities

$\vec{p}_i \equiv \left(p_{i,0,0} \quad p_{i,1,1} \quad p_{i,2,1} \quad p_{i,2,2} \quad p_{i,3,1} \quad p_{i,3,2} \quad \cdots \quad p_{i,n,1} \quad p_{i,n,2} \right)$

$i = 2,3,4,\ldots$. Following Neuts (1981), the vectors $\vec{p}_i$ satisfy

$$\vec{p}_i = \vec{p}_{n+1} R^{i-(n+1)}, \quad i = n+1, n+2, n+3, \ldots, \infty. \qquad (5)$$

In most cases, the entries $r_{i,j}$ of the matrix R are found by numerical calculations (see Chapter 8 in Latouche and Ramaswami 1999), mostly using *successive substitutions* (Neuts 1981, p. 37). However, for the current model, it is possible to derive closed-form expressions for all $r_{i,j}$, as given in Theorem 2, due to the structure of the stochastic process given in Figure 1.

Let C_m be the m-th Catalan number (i.e., $C_m \equiv \dfrac{(2m)!}{(m+1)!m!}$, see Koshy 2008), where m is a non-negative integer. Then,

Theorem 2. The entries of matrix R are:

$$r_{i,1} = \begin{cases} \lambda/(2\mu) & i = 1 \\[2ex] \dfrac{r_{2,2}^2 \beta}{2\mu(1-r_{2,2})} & i = 2 \\[2ex] \dfrac{\left(\begin{array}{l} (\lambda+\mu+\beta) \displaystyle\sum_{k=0}^{0.5(i-1)} r_{i,2+2k} \left(r_{2+2k,2}\beta + 2r_{2+2k,1}\mu \right) \\ + \displaystyle\sum_{k=0}^{0.5(i-3)} \frac{r_{3+2k,2} C_{0.5(i-3)-k} (\beta\lambda)^{0.5(i-1)-k}}{(\lambda+\mu+\beta)^{i-3-2k}} \\ + 2\beta^{-1} \displaystyle\sum_{k=0}^{0.5(i-5)} \frac{r_{3+2k,1} C_{0.5(i-3)-k} \mu(\beta\lambda)^{0.5(i-1)-k}}{(\lambda+\mu+\beta)^{i-3-2k}} \end{array} \right)}{2\mu(\mu+\beta)} & i = 3,5,\ldots,2n-1 \\[2ex] \dfrac{\left(\begin{array}{l} (\lambda+2\beta) r_{i,2} (2\mu r_{2,1} + \beta r_{2,2}) \\ + 0.5 \displaystyle\sum_{k=0}^{0.5(i-4)} \frac{r_{4+2k,2} C_{0.5(i-4)-k} (2\beta\lambda)^{0.5(i-2)-k}}{(\lambda+2\beta)^{i-4-2k}} \\ \beta^{-1} \displaystyle\sum_{k=0}^{0.5(i-6)} \frac{r_{4+2k,1} C_{0.5(i-4)-k} \mu(2\beta\lambda)^{0.5(i-2)-k}}{(\lambda+2\beta)^{i-4-2k}} \end{array} \right)}{4\mu\beta} & i = 4,6,\ldots,2n \end{cases}$$

$$r_{i,j} = \begin{cases} 0 & i = 1,2,\ldots,j-1 \text{ and } i = j+1, j+3, \ldots, 2n \\[2ex] \dfrac{C_{0.5(i-j)} \beta^{0.5(i-j)} \lambda^{0.5(i-j+2)}}{(\lambda+\mu+\beta)^{i-j+1}} & i = j, j+2, \ldots, 2n-1 \end{cases}$$

for $j = 3,5,\ldots,2n-1$.

$$r_{i,j} = \begin{cases} 0 & i = 1,2,\ldots,j-2 \\[2ex] \left(\begin{array}{l} \dfrac{C_{0.5(i-j+3)} \mu \beta^{0.5(i-j+1)} \lambda^{0.5(i-j+5)}}{(\lambda+\mu+\beta)^{i-j+3} (\lambda+2\beta)} \\ + \displaystyle\sum_{k=0}^{0.5(i-j-1)} \frac{r_{i,j+2+2k} C_k (2\beta\lambda)^{k+1}}{(\lambda+2\beta)^{2k+2}} \\ + 2 \displaystyle\sum_{k=0}^{0.5(i-j-1)} \frac{r_{j+1+2k,j+2} C_{0.5(i-j-1)-k} (\beta\lambda)^{0.5(i-j+1)-k}}{(\lambda+2\beta)(\lambda+\mu+\beta)^{i-j-2k}} \end{array} \right) & i = j-1, j+1, \ldots, 2n-1 \\[2ex] \dfrac{2^{0.5(i-j)} C_{0.5(i-j)} \beta^{0.5(i-j)} \lambda^{0.5(i-j+2)}}{(\lambda+2\beta)^{i-j+1}} & i = j, j+2, \ldots, 2n \end{cases}$$

for $j = 4,6,\ldots,2n$.

$$r_{i,2} = \begin{cases} 0 & i = 1 \\[2ex] \dfrac{\lambda+\mu+\beta-\sqrt{(\lambda+\mu+\beta)^2 - 4\lambda\mu}}{2\mu} & i = 2 \\[2ex] \dfrac{\left(\begin{array}{l} \dfrac{C_{0.5(i-1)} \beta^{0.5(i-1)} \lambda^{0.5(i+1)}}{(\lambda+\mu+\beta)^{i-2}} + \beta^{-1} \displaystyle\sum_{k=0}^{0.5(i-5)} \left(\frac{r_{3+2k,2} C_{0.5(i-3)-k} \mu(\beta\lambda)^{0.5(i-1)-k}}{(\lambda+\mu+\beta)^{i-3-2k}} \right) \\ + 2 \displaystyle\sum_{k=0}^{0.5(i-3)} \left(\frac{r_{3+2k,4} C_{0.5(i-3)-k} (\beta\lambda)^{0.5(i-1)-k}}{(\lambda+\mu+\beta)^{i-3-2k}} \right) \\ + \displaystyle\sum_{k=0}^{0.5(i-3)} \frac{r_{i,4+2k} (\lambda+\mu+\beta)(C_k (2\beta\lambda)^{k+1} + \mu r_{4+2k,2} (\lambda+2\beta)^{2k+1})}{(\lambda+2\beta)^{2k+1}} \end{array} \right)}{(\lambda+\mu+\beta)((\lambda+\beta+\mu) - \mu r_{2,2}) - \lambda\mu} & i = 3,5,\ldots,2n-1 \\[2ex] \dfrac{\left(\begin{array}{l} \dfrac{2^{0.5(i-2)} C_{0.5(i-2)} \beta^{0.5(i-2)} \lambda^{0.5i}}{(\lambda+2\beta)^{i-3}} \\ + (2\beta)^{-1} \displaystyle\sum_{k=0}^{0.5(i-6)} \frac{r_{4+2k,2} C_{0.5(i-4)-k} \mu(2\beta\lambda)^{0.5(i-2)-k}}{(\lambda+2\beta)^{i-4-2k}} \end{array} \right)}{(\lambda+2\beta)((\lambda+\beta+\mu) - \mu r_{2,2}) - \lambda\mu} & i = 4,6,\ldots,2n \end{cases}$$

Proof. The solution has been obtained using induction, and can be verified by substitution in (5).

As for calculating $\vec{p}_i$, $i = 0,1,\ldots,n+1$, we use a subset of equations from $\vec{p}Q = 0$ (where $\vec{p} \equiv (\vec{p}_0, \vec{p}_1, \vec{p}_2, \ldots)$) combined with the normalization equation $\vec{p}_0 \cdot \vec{e} + \vec{p}_1 \cdot \vec{e} + \sum_{i=2}^{\infty} \vec{p}_i \cdot \vec{e} = 1$.

Consequently, we solve the following linear set:

$$\vec{p}_0 B_{1,1} + \vec{p}_1 B_{2,1} = \vec{0}$$

$$\vec{p}_0 B_{1,2} + \vec{p}_1 B_{2,2} + \vec{p}_2 B_{3,1} = \vec{0}$$

$$\vec{p}_{i-1} B_{i,3} + \vec{p}_i B_{i+1,2} + \vec{p}_{i+1} A_2 = \vec{0} \qquad i = 2,3,\ldots,n$$

$$\vec{p}_n B_{n+1,3} + \vec{p}_{n+1} \left(A_1 + RA_2 \right) = \vec{0}$$

$$\vec{p}_0 \cdot \vec{e} + \vec{p}_1 \cdot \vec{e} + \sum_{i=2}^{n} \vec{p}_i \cdot \vec{e} + \vec{p}_{n+1} \left[I - R \right]^{-1} \cdot \vec{e} = 1.$$

4. PERFORMANCE MEASURES

Based on the steady state probabilities, we show how to calculate relevant systems performance measures. For a given capacity of n PSs, let $L(n)$ and $L_q(n)$ be the mean number of customers in the system and in queue, respectively. Similarly, let $W(n)$ and $W_q(n)$ be the mean sojourn time of a customer in the system and in queue, respectively; $S(n)$ and $S_q(n)$ be the mean number of PSs in the system and in inventory, respectively; and $T(n)$ and $T_q(n)$ be the mean time a PS resides in the system and in inventory, respectively.

Using (4) and the equations $\sum_{i=0}^{\infty} R^i = [I-R]^{-1}$ and $\sum_{i=0}^{\infty}(i+1)R^i = [I-R]^{-2}$, we obtain:

$$L(n) = \vec{p}_1\vec{e} + \sum_{i=2}^{n} i\vec{p}_i\vec{e} + \vec{p}_{n+1}\left(n[I-R]^{-1} + [I-R]^{-2}\right)\vec{e} . \quad (6)$$

$$L_q(n) = L(n) - 2(1-\vec{p}_0\vec{e}) + \vec{p}_1\vec{e} . \quad (7)$$

In order to calculate the mean sojourn time of customer in system and in queue, we first have to calculate the effective customer's arrival rate λ_{eff} . Let p_δ be the proportion of time when customers arrive with rate δ , and $p_\lambda = 1 - p_\delta$ the proportion of time when customers arrive with rate λ . Then, the effective customers' arrival rate is obtained as follows

$$\lambda_{eff} = \delta p_\delta + \lambda p_\lambda = (\delta - \lambda) p_\delta + \lambda , \quad (8)$$

where

$$p_\delta = \sum_{i=0}^{1}\sum_{j=1}^{n} p_{i,j,0} + \sum_{j=2}^{n} p_{1,j,1} + \sum_{i=2}^{n}\sum_{j=i}^{n} p_{i,j,1} + \sum_{i=2}^{n-1}\sum_{j=i+1}^{n} p_{i,j,2} . \quad (9)$$

Then, $W(n) = L(n)/\lambda_{eff}$ and $W_q(n) = L_q(n)/\lambda_{eff}$. In order to obtain $S(n)$, we define three column vectors $\vec{v}_{n+1} = \begin{pmatrix}0 & 1 & 2 & 3 & \cdots & n\end{pmatrix}^T$, $\vec{v}_{2n} = \begin{pmatrix}0 & 1 & 2 & 2 & 3 & 3 & \cdots & n & n\end{pmatrix}^T$ and $\vec{v}_{2n+1} = \begin{pmatrix}0 & 1 & 1 & 2 & 2 & 3 & 3 & \cdots & n & n\end{pmatrix}^T$. Thus, by using equation (4) and algebraic manipulations, we get

$$S(n) = \vec{p}_0\vec{v}_{n+1} + \vec{p}_1\vec{v}_{2n+1} + \sum_{i=2}^{n}\vec{p}_i\vec{v}_{2n} + \vec{p}_{n+1}[I-R]^{-1}\vec{v}_{2n} . \quad (10)$$

Similarly, in order to get $S_q(n)$, we define three column vectors $\vec{u}_{n+1} = \begin{pmatrix}0 & 1 & 2 & 3 & \cdots & n\end{pmatrix}^T$, $\vec{u}_{2n} = \begin{pmatrix}0 & 0 & 1 & 0 & 2 & 1 & 3 & 2 & \cdots & n-1 & n-2\end{pmatrix}^T$ and $\vec{u}_{2n+1} = \begin{pmatrix}0 & 1 & 0 & 2 & 1 & 3 & 2 & \cdots & n & n-1\end{pmatrix}^T$, so that

$$S_q(n) = \vec{p}_0\vec{u}_{n+1} + \vec{p}_1\vec{u}_{2n+1} + \sum_{i=2}^{n}\vec{p}_i\vec{u}_{2n} + \vec{p}_{n+1}[I-R]^{-1}\vec{u}_{2n} . \quad (11)$$

Given that inventory level is less than *n*, PSs are produced by the two servers when there are no customers in the system, and by one of the servers when there is only one customer in the system. Therefore, the effective production rate of PSs is

$$\alpha_{eff} = 2\alpha\left(\vec{p}_0\vec{e} - p_{0,n,0}\right) + \alpha\left(\vec{p}_1\vec{e} - p_{1,n,0} - p_{1,n,1}\right) . \quad (12)$$

Consequently, the mean time a PS resides in the system and in inventory is obtained by Little's law: $T(n) = S(n)/\alpha_{eff}$ and $T_q(n) = S_q(n)/\alpha_{eff}$, respectively.

5. ECONOMIC ANALYSIS

In this section, we provide an economic analysis of our queueing-inventory model. Let *p* be the revenue from supplying a service to a customer, so that the expected system's revenue rate is equal to $p\lambda_{eff}$. Let *c* be the sojourn cost per unit of time per customer in the system, *h*—the holding cost per unit of time per inventoried PS (assuming no holding cost for a PS that moves on to the CS phase). Assume that the server can publicize the number of available PSs (i.e. make the stock observable) which stimulates demand due to shorter sojourn time. The publication may be generated in different levels. For example, the server can rent screen times for promotion in different locations near its store, where each screen has a potential to increase the number of potential customers. Another example is transmitting promotion messages to customers' smartphones in a larger distance (radius) to reach more customers. Thus, we denote by $A(\delta)$ the cost rate of making the number of PSs observable (hereafter transparency cost) as a function of the desired arrival rate δ. We assume that the transparency cost is a convex increasing function of the gap in the arrival rate. Specifically, we use $A(\delta) = a(\delta-\lambda)^\tau p_\delta$, where $a > 0$ and $\tau > 1$. The objective is to maximize the total expected profit per time unit by controlling the PS capacity *n* and δ . Thus, the profit function is defined as follows

$$\max_{\substack{n\in\{0,1,2\ldots\} \\ \delta\geq\lambda}} \left\{R(n,\delta) = p\lambda_{eff} - A(\delta) - cL(n) - hS_q(n)\right\} . \quad (13)$$

To illustrate the behaviour of the expected profit as a function of PS capacity, *n*, and arrival rate when PSs are available, δ, we use the following parameter values: $\lambda = 16$, $\mu = 10$, $\alpha = 20$, $\beta = 18$, $p = 0.5$, $a = 0.2$, $\tau = 2$, $c = 1$ and $h = 1$. The results are given in Table 1 and Figure 2.

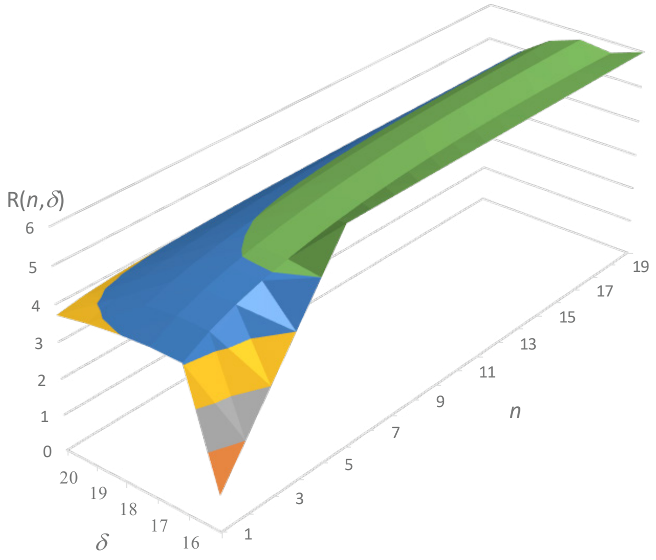

Figure 2. The expected profit as a function of *n* and *δ.*

Table 1. The expected profit for various values of n and δ.

$n\backslash\delta$	16	17	18	19	20
1	4.0455	4.0526	3.9930	3.8762	3.7099
2	4.4359	4.4397	4.3282	4.1208	3.8330
3	4.7169	4.7113	4.5564	4.2807	3.9071
4	4.9323	4.9172	4.7261	4.3955	3.9553
5	5.1057	5.0833	4.8618	4.4858	3.9913
6	5.2476	5.2203	4.9735	4.5589	4.0189
7	5.3639	5.3344	5.0667	4.6192	4.0403
8	5.4591	5.4299	5.1450	4.6692	4.0567
9	5.5363	5.5098	5.2110	4.7107	4.0689
10	5.5981	5.5764	5.2666	4.7451	4.0778
11	5.6465	5.6316	5.3134	4.7736	4.0838
12	5.6832	5.6767	5.3525	4.7970	4.0875
13	5.7097	5.7132	5.3851	4.8161	4.0890
14	5.7272	5.7419	5.4117	4.8313	4.0889
15	5.7367	5.7637	5.4332	4.8431	4.0872
16	5.7391	5.7794	5.4501	4.8521	4.0843
17	5.7352	5.7895	5.4629	4.8584	4.0802
18	5.7257	5.7947	5.4720	4.8624	4.0752
19	5.7111	**5.7954**	5.4777	4.8643	4.0694
20	5.6919	5.7919	5.4804	4.8644	4.0629

Table 1 shows that the optimal PS capacity and enhanced customers' arrival rate are $n=19$ and $\delta=17$, and that this point is the unique optimal solution of the expected profit maximization problem over the domain $n,\delta \geq 20$.

6. CONCLUSIONS

This paper analyzes an M/M/2-type system, which utilizes the server's idle time to produce preliminary services for incoming customers. In addition, the server can publicize the number of available PSs (i.e. make the stock observable) which stimulates demand due to shorter sojourn time. In order to investigate such a system, a three-dimensional state-space is constructed, where the variables are (i) number of customers, (ii) inventory level of PSs and (iii) number of CSs. For a Markovian process and by applying matrix geometric analysis, closed-form expressions for the steady-state probabilities of the system states are derived, leading to the calculation of various performance measures. Due to the structure of the process and regardless of the size of the rate-matrix R, it is possible in this model to derive closed-form expressions of all the entries of R. The relation of these expressions to Catalan numbers is shown. This result is notable in light of the fact that, in typical applications of the matrix geometric analysis, explicit calculation of the entries of R is rarely possible. Furthermore, it is proved analytically that the stability condition of our model is identical to that of a standard M/M/2 queue. In order to provide economic analysis, an optimization problem is formulated, where the objective is to maximize the server's expected profit by controlling the capacity of preliminary services and the investment in increasing customers arrival rate when preliminary services are available. Numerical examples reveal that the investigated profit function is quasi-concave.

The following two directions for future research are proposed: (i) considering time deterioration of inventoried PSs; (ii) assuming customers arrival rates that depend on the inventory level of PSs.

Acknowledgements

This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 1448/17).

References

Benjaafar, S., Cooper, W. L., Mardan, S. (2011). Production-inventory systems with imperfect advance demand information and updating. *Naval Research Logistics*, 58(2), 88-106.

Boxma, O. J., S. Schlegel, U. Yechiali. (2002). A note on the M/G/1 queue with a waiting server, timer and vacations. *American Mathematical Society Translations*, *Series 2* **207** 25-35.

Cox, D. R. (2017). *The theory of stochastic processes*. Routledge.

Flapper, S. D. P., Gayon, J. P., Vercraene, S. (2012). Control of a production–inventory system with returns under imperfect advance return information. *European Journal of Operational Research*, 218(2), 392-400.

Guha, D., V. Goswami, A. D. Banik. (2016). Algorithmic computation of steady-state probabilities in an almost observable GI/M/c queue with or without vacations under state dependent balking and reneging. *Applied Mathematical Modelling* 40(5) 4199-4219.

Guo, P., & Hassin, R. (2011). Strategic behavior and social optimization in Markovian vacation queues. *Operations research* 59(4) 986-997.

Guo, P., Z.G. Zhang. (2013). Strategic queueing behavior and its impact on system performance in service systems with the congestion-based staffing policy. *Manufacturing & Service Operations Management* 15(1) 118-131.

Gupta, S. K. (1967). Queues with hyper-Poisson input and exponential service time distribution with state dependent arrival and service rates. *Operations Research* 15(5) 847-856.

Hanukov, G., Avinadav, T., Chernonog, T., Spiegel, U., Yechiali, U. (2017). A queueing system with decomposed service and inventoried preliminary services. *Applied Mathematical Modelling*, 47, 276-293.

Hanukov, G., Avinadav, T., Chernonog, T., Spiegel, U., Yechiali, U. (2018a). Improving efficiency in service systems by performing and storing "preliminary services". *International Journal of Production Economics*, 197, 174-185.

Hanukov, G., Avinadav, T., Chernonog, T., Yechiali, U. (2018b). Performance Improvement of a Service System via Stocking Perishable Preliminary Services. *European Journal of Operational Research*.

Hwang, J., L. Gao, W. Jang. (2010). Joint demand and capacity management in a restaurant system. *European Journal of Operational Research* 207(1) 465-472.

Iravani, S. M., Kolfal, B., Van Oyen, M. P. (2011). Capability flexibility: a decision support methodology for parallel service and manufacturing systems with flexible servers. *IIE Transactions*, *43*(5), 363-382.

Koshy, T., (2008). Catalan Numbers with Applications (Oxford University Press, Oxford).

Latouche, G., Ramaswami, V., (1999). Introduction to Matrix Analytic Methods in Stochastic Modeling.

Latouche, G., V. Ramaswami. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*. *ASA-SIAM Series on Statistics and Applied Probability*. SIAM, Philadelphia, PA .

Levy, Y., U. Yechiali. (1975). Utilization of idle time in an M/G/1 queueing system. *Management Science* 22 202-211.

Levy, Y., U. Yechiali. (1976). An M/M/s queue with servers' vacations. *INFOR* 14 153-163.

Harchol-Balter, M. (2013). *Performance modeling and design of computer systems: queueing theory in action*. Cambridge University Press.

Mytalas, G. C., M. A. Zazanis. (2015). An M^X/G/1 queueing system with disasters and repairs under a multiple adapted vacation policy. *Naval Research Logistics* 62 171-189.

Neuts, M. F. (1981). Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.

Rosenberg, E., U. Yechiali. (1993). The M^X/G/1 queue with single and multiple vacations under the LIFO service regime. *Operations Research Letters* 14(3) 171-179.

Urban, T. L. (2005). Inventory models with inventory-level-dependent demand: A comprehensive review and unifying theory. *European Journal of Operational Research*, 162(3) 792-804.

Winston, W. (1978). Optimality of monotonic policies for multiple-server exponential queuing systems with state-dependent arrival rates. *Operations Research* 26(6) 1089-1094.

Yang, D. Y., C. H. Wu. (2015). Cost-minimization analysis of a working vacation queue with N-policy and server breakdowns. *Computers & Industrial Engineering* 82 151-158.

Yechiali, U. (2004). On the M^X/G/1 queue with a waiting server and vacations. *Sankhya* 66 159-174.

Zhao, N., Lian, Z. (2011). A queueing-inventory system with two classes of customers. *International Journal of Production Economics*, 129(1), 225-231.

APPENDIX

$$B_{1,1}=\begin{pmatrix} -(\lambda+2\alpha) & 2\alpha & 0 & \cdots & 0 \\ 0 & -(\delta+2\alpha) & 2\alpha & & 0 \\ 0 & 0 & -(\delta+2\alpha) & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 2\alpha \\ 0 & 0 & 0 & \cdots & -\delta \end{pmatrix}_{(n+1)\times(n+1)}$$

$$B_{1,2}=\begin{pmatrix} \lambda & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \delta & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \delta & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \delta & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \delta \end{pmatrix}_{(n+1)\times(2n+1)}$$

$$B_{2,1}=\begin{pmatrix} \mu & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \mu & 0 & 0 & \cdots & 0 & 0 \\ \beta & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \mu & 0 & \cdots & 0 & 0 \\ 0 & \beta & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \mu & & 0 & 0 \\ 0 & 0 & \beta & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \mu \\ 0 & 0 & 0 & 0 & \cdots & \beta & 0 \end{pmatrix}_{(2n+1)\times(n+1)}$$

$$B_{2,3}=\begin{pmatrix} \lambda & 0 & 0 & 0 & \cdots & 0 \\ 0 & \delta & 0 & 0 & \cdots & 0 \\ 0 & \lambda & 0 & 0 & \cdots & 0 \\ 0 & 0 & \delta & 0 & & 0 \\ 0 & 0 & 0 & \delta & & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \delta \end{pmatrix}_{(2n+1)\times 2n}$$

Let *I* be the unit matrix. The matrix $B_{i,2}$ is given by

$$B_{i,2}=-\left(a_1,a_2,\ldots,a_{2n}\right)I_{2n\times 2n}\,,$$

where,

$$a_k=\begin{cases} \lambda+2\mu & k=1 \\ \lambda+\mu+\beta & k=2,3,5,7,\ldots,2i-5 \\ \lambda+2\beta & k=4,6,8,\ldots,2i-4,2i-2 \\ \delta+\mu+\beta & k=2i-3,2i-1,2i+1,\ldots,2n-1 \\ \delta+2\beta & k=2i,2i+2,\ldots,2n \end{cases}.$$

Similarly,

$$B_{i,3}=\left(b_1,b_2,\ldots,b_{2n}\right)I_{2n\times 2n}\,,$$

where

$$b_k=\begin{cases} \lambda & k=1,2,3\ldots,2i-5,2i-4,2i-2 \\ \delta & k=2i-3,2i-1,2i,2i+1,\ldots,2n \end{cases}.$$

A_0, A_1 and A_2 are given by $A_0=\lambda I_{2n\times 2n}$,

$$A_1=-\left(a_1,a_2,\ldots,a_{2n}\right)I_{2n\times 2n}\,,$$

where

$$a_k=\begin{cases} \lambda+2\mu & k=1 \\ \lambda+\mu+\beta & k=2,3,5,7,\ldots,2n-1\,, \\ \lambda+2\beta & k=4,6,8,\ldots,2n \end{cases}$$

and $A_2=[b_{i,j}]_{2n\times 2n}$, where

$$b_{i,j}=\begin{cases} 2\mu & i=j=1 \\ \mu & i=j=2 \text{ and } i=j-1=3,5,7,\ldots,2n-1 \\ \beta & i=j+1=2,3 \text{ and } i=j+2=5,7,9,\ldots,2n-1\,. \\ 2\beta & i=j+2=4,6,8,\ldots,2n \\ 0 & \text{Otherwise} \end{cases}$$

$$B_{2,2}=\begin{pmatrix} -(\lambda+\mu+\alpha) & \alpha & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -(\delta+\mu+\alpha) & 0 & \alpha & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -(\lambda+\beta+\alpha) & 0 & \alpha & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & -(\delta+\mu+\alpha) & 0 & \alpha & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & -(\delta+\beta+\alpha) & 0 & \alpha & & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -(\delta+\mu+\alpha) & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -(\delta+\beta+\alpha) & & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \ddots & 0 & \alpha \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & -(\delta+\mu) & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & & -(\delta+\beta) \end{pmatrix}_{(2n+1)\times(2n+1)},$$