International Journal of Production Economics 197 (2018) 174–185

Contents lists available at ScienceDirect

International Journal of Production Economics

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

Improving efficiency in service systems by performing and storing "preliminary services"

Gabi Hanukov[a], Tal Avinadav[a,*], Tatyana Chernonog[a], Uriel Spiegel[a,b,d], Uri Yechiali[c]

[a] *Department of Management, Bar-Ilan University, Ramat Gan, 5290002, Israel*
[b] *Department of Economics, University of Pennsylvania, Philadelphia, USA*
[c] *Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv, 6997801, Israel*
[d] *Zefat College, Zefat, Israel*

A R T I C L E I N F O

Keywords:
Queueing systems
Server's idle time
Preliminary services
Performance measures
Cost analysis

A B S T R A C T

We propose a novel approach to improve efficiency in service systems. The idea is to utilize the server's idle time to perform and store "preliminary services" for customers who will arrive in the future. Such a model is relevant to settings in which service consists of multiple consecutive tasks, some of which are generic and needed by all customers (and thus can be performed even in their absence), while other require the customer's presence. To show the model's benefits, we formulate a two-dimensional single-server queueing-inventory system for which we derive closed-form expressions for the system's steady-state probabilities, as well as for its performance measures. Assuming linear costs for customers waiting in line and for stored preliminary services, a cost analysis determines the optimal maximal number of stored preliminary services in the system. Numerical examples illustrated with graphs demonstrate the advantages of our approach, in terms of cost savings, as compared with the classical M/M/1 model.

1. Introduction

Operations managers frequently face the difficult challenge of reducing service systems' "idle" time in order to improve those systems' efficiency. Two sources of idleness characterize such systems: either customers wait in line to be served, or servers stay idle while waiting for customers to arrive. Because of the stochastic nature of queues, neither of the two sources of idleness can be entirely eliminated. It is estimated that the annual monetary loss due to idleness of employees in organizations reaches billions of dollars per year (Malachowski and Simonini, 2006), which further emphasizes the importance of improving the efficiency of service systems.

The literature discusses two common approaches that might be used to mitigate the two sources of idleness. (i) Increasing the number of servers in order to reduce the waiting times of customers. The drawback of this approach is that it leads to an increase in the servers' idle time and thus reduces each server's utilization. (ii) Increasing servers' utilization. Many studies propose achieving this goal by adopting a so-called vacation model, in which, instead of being allowed to remain idle, servers perform ancillary duties ("vacations") that are not directly related to their main task (see, e.g., Levy and Yechiali, 1975, 1976; Doshi, 1986; Kella and Yechiali, 1988; Takagi, 1991; Rosenberg and Yechiali, 1993; Boxma et al., 2002; Yechiali, 2004; Jain and Jain, 2010; Wei et al., 2013b; Yang and Wu, 2015; Mytalas and Zazanis, 2015; and Guha et al., 2016). However, the need to wait for a server to complete such tasks may increase customers' waiting times. Thus, each of these two approaches (more servers or server vacations) improves one source of idleness at the expense of the other.

Idleness of servers has been analyzed in the literature from additional perspectives. For example, Armony (2005), Armony and Ward (2010, 2013), and Mandelbaum et al. (2012) investigated fair routing of customers to idle servers in large-scale systems with heterogeneous customers. Cachon and Zhang (2007) investigated allocation of jobs to strategic servers (state-dependent as well as state-independent policies) under a capacity choice game played between the servers. They showed that there are cases in which it is beneficial to allocate a job to

[*] Corresponding author.
E-mail addresses: german.khanukov@live.biu.ac.il (G. Hanukov), tal.avinadav@biu.ac.il (T. Avinadav), tatyana.chernonog@biu.ac.il (T. Chernonog), uriel.spiegel@biu.ac.il (U. Spiegel), uriy@post.tau.ac.il (U. Yechiali).

https://doi.org/10.1016/j.ijpe.2018.01.004
Received 7 August 2016; Received in revised form 5 January 2018; Accepted 6 January 2018
Available online 10 January 2018
0925-5273/© 2018 Elsevier B.V. All rights reserved.

a busy fast server rather than to an idle slow server. Clearly, these allocation approaches can serve to mitigate idleness of servers and customers in multi-server systems, but they are not applicable to single-server systems.

In this paper we propose a novel approach to improve the performance of service systems by utilizing servers' idle time in cases where the service can be decomposed into two stages. The first stage, denoted “preliminary service” (PS henceforth), can be performed in the absence of customers, and its outcome can be preserved until an actual service is requested. The second stage, denoted “complementary service” (CS henceforth), requires the presence of the customer to be completed. In such settings, in contrast to the case of a vacation model, in which servers are diverted to ancillary duties during their idle time, an idle server can be utilized to accumulate PSs and store them until customers arrive and require service. This approach leads to a reduction of customers' mean sojourn time, since a certain fraction of the customers receive only a CS upon arrival (as part of their service was prepared prior to their arrival), and do not have to wait for the full service (FS henceforth).

A representative example of an application of our model is a fast food restaurant in which food, e.g., hamburger patties, can be prepared before demand occurs, and only upon the arrival of a customer is a hamburger patty heated up, inserted into a bun and served to the customer. Another example is a bicycle shop, which can assemble parts of a bicycle before a purchase occurs, 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 (not necessarily identical) nameplates from wood, clay or glass before an order is placed, and complete a nameplate for a specific customer upon request (e.g., writing the name, adding decorations, etc.).

Hypothetically, a server can produce PSs during its entire idle time to minimize the customers' sojourn time. However, we show, in our model, that when cost considerations are taken into account—such as holding costs of PSs in inventory and costs of customers' presence in the system—there may be a certain number of stored PSs beyond which it is more beneficial to keep the server idle rather than to occupy it with producing additional PSs. Herein, we analyze this innovative queueing-inventory system. Examples of other types of queueing-inventory systems, in which each customer requires a unit from inventory when being served, appear in Zhao and Lin (2011), and in Adacher and Cassandras (2014). We use the classical M/M/1 queue as a baseline for comparison, which is common practice in the literature (e.g., Andritsos and Tang, 2013; Wei et al., 2013a; Güler et al., 2014).

We can summarize the main contributions of this paper as follows:

- A novel single-server queueing-inventory system is formulated as a two-dimensional stochastic process, and a method to derive closed-form expressions for the system's steady-state probabilities and for its performance measures is provided.
- It is shown that under certain conditions related to the duration of the service stages and the cost structure, the performance of a system that produces and stores PSs is superior to that of a similar system but without PSs. Nevertheless, Theorem 1 states that the stability conditions of the two systems (with or without PSs) are the same.
- A condition is established in Proposition 1 under which a server that utilizes some of its idle time to produce PSs actually remains idle for a larger fraction of time compared with a server in a similar system that does not store PSs.
- Assuming linear costs for each waiting customer and each stored PS, results of a cost analysis are provided, demonstrating how the optimal maximal number of stored PSs is affected by the model parameters.

2. Notations and assumptions

The following notations and assumptions are used throughout the paper:

Notations	
FS	Full service rendered continuously
PS	Preliminary service
CS	Complementary service
λ	customers' mean arrival rate
μ	server's mean rate of performing FSs
α	server's mean rate of producing PSs
β	server's mean rate of performing CSs
n	a decision variable denoting the maximal number of stored PSs
L	number of customers in the system in the long run (a random variable)
S	number of PSs in the system in the long run (a random variable)
$p_{i,j}$	steady-state probability of finding the system in state $\{L = i, S = j\}$
R	rate matrix of the matrix geometric analysis
$L(n)$	mean number of customers in the system as a function of n
$L_q(n)$	mean number of customers in queue as a function of n
$W(n)$	mean sojourn time of a customer in the system as a function of n
$W_q(n)$	mean waiting time of a customer in queue as a function of n
$S(n)$	mean number of PSs in the system as a function of n
$S_q(n)$	mean number of PSs in inventory as a function of n
$T(n)$	mean time a PS resides in the system as a function of n
$T_q(n)$	mean time a PS resides in inventory as a function of n
$\alpha_{eff}(n)$	effective production rate of PSs as a function of n
c	cost per unit of time per customer in the system
h	holding cost per unit of time per inventoried PS
$Z(n)$	total expected cost per unit of time as a function of n
η	percentage reduction in total expected cost in comparison to the classical M/M/1 model
ξ	percentage reduction in idle time of the server in comparison to the classical M/M/1 model

Assumptions

1. We consider a single-server system with a Poisson arrival rate λ and exponentially-distributed full-service time with mean $1/\mu$.
2. The service can be split into two consecutive stages. The first stage, PS, can be performed in the absence of customers, and its outcome can be preserved until an actual service is requested. The second stage, CS, requires the actual presence of the customer to be completed.
3. When the system is empty, the server produces PSs at a Poisson rate α. The PSs are stored until the arrival of customers, and can be considered as work-in-process inventory whose aim is to reduce the sojourn time of customers in the system.
4. When the number of stored PSs reaches the value of n, the server stops producing PSs and becomes idle.
5. If a customer arrives at the front of the queue and a PS is available, the server immediately starts rendering a CS for that customer; otherwise, the customer receives an FS.
6. The CS time is assumed to be exponentially distributed with mean $1/\beta(< 1/\mu)$.
7. The decomposition of service into two separate stages (potentially with an intermission between them) does not affect service quality, which implies that customers have no preference between receiving CS or FS. This assumption suits cases in which the storage time of PSs is relatively short in comparison to the shelf-life duration of a PS.

We now justify our assumptions regarding the production rates. We emphasize that although the PSs are standard units, they are not produced in an automatic process (which implies a deterministic preparation time). Specifically, the variability of the PS production durations emerges from three sources: the server, the production process and raw materials. Variability associated with the human server may stem from external