İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ
İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ DERGİSİ

IZMIR KATIP CELEBI UNIVERSITY
FACULTY OF ECONOMICS AND ADMINISTRATIVE SCIENCES JOURNAL

OPERATIONS MANAGEMENT FOR DOUBLE-ENDED QUEUES

ÇİFT-TARAFLI KUYRUK (BEKLEME) SİSTEMLERİNDE OPERASYON YÖNETİMİ

Işılay Talay
Sorumlu Yazar, Dr. Öğr. Ü., Antalya Bilim Üniversitesi, , ORCID: 0000-0002-8956-9505, isilay.talay@antalya.edu.tr

Kabul Tarihi / Accepted: 27.11.2019

İletişim / Correspondence: Işılay Talay

Benzerlik Oranı/ Plagiriasim : % 2

ABSTRACT

In this paper we have presented the double-ended queueing model with its generic form, abstract modeling extensions, different types of controls to be assumed by the management of the system, as well as optimization via various methodologies. Since its inception in 1950s, double-ended queuing model has had widespread use and recently it is observed that literature is focusing on social welfare perspectives on queues under strategically acting customers and enhanced information exchange due to advanced technology. It is also reviewed that for the complex and analytically tractable versions of double-ended queues it is possible to obtain approximate or near-optimal results via methodologies such as simulation and fluid and diffusion approximations.

Keywords: double-ended queue, service capacity management, modeling and applications

ÖZET

Bu makalede, çift taraflı kuyruk (bekleme) modelinin genel formunu, soyut modelleme yaklaşımları, sistemin yönetim tarafından üstlenilen farklı kontrol türleri ve çeşitli metotlarla optimizasyonu ile beraber sunulmuştur. 1950'lerde oluşturulmasından bu yana, çift taraflı kuyruk (bekleme) modeli yaygın bir şekilde kullanılmaya başlanmış ve son zamanlarda literatürün stratejik olarak hareket eden müşterilerin ve ileri teknoloji nedeniyle gelişmiş bilgi alışverişinin olduğu kuyruklar üzerindeki sosyal refah perspektiflerine odaklandığı görülmüştür. Ayrıca, çift taraflı kuyrukların (bekleme sırası) karmaşık ve analitik olarak incelenemez versiyonları için, simülasyon ve akışkan ve difüzyon yaklaşımları gibi metodolojilerle yaklaşık olarak optimum ya da en uyguna yakın sonuçların elde edilmesinin mümkün olduğu da belirtilmiştir.

Anahtar Kelimeler: çift-taraflı kuyruk sistemleri, hizmet kapasitesi yönetimi, modelleme ve uygulama

1. INTRODUCTION

Queueing models constitute great importance for all of us since we come across waiting lines or queues most of the time. While in general the experience of waiting provides the core definition of a queue, not all waiting lines or queues have the same characteristics, and these differences require separate analyses to be conducted for different queueing systems. For instance, the queue for a passenger waiting for a taxi and the queue for a customer waiting for a teller at a bank branch will not possess the same system elements and management principles. In this paper we will focus on the double-ended (or double-sided) queues that were first inspired by the setting of taxi-passenger matching at an airport. In the taxi-passenger example, the passengers getting off a plane and will be needing a taxi would go to the stop, and as soon as possible they will be matched with the first taxi waiting in line to get a passenger. If there are no taxis present, then the customers will wait in a queue and the first arriving taxi would be matched with the first passenger waiting in the line. There will never be both taxis and passengers waiting simultaneously. This typical example has been studied since the 1950s (Dobbie, 1961: 756; Kashyap, 1965: 559; Gaur and Kashyap, 1973: 74;) and many extensions had been done to the model since there are numerous other applications, and as times have changed so did the technology, enabling to have queues to be controlled for different objectives and service purposes.

The necessity for studying double-ended queues come from their capability to capture operations properties of different systems and settings. The double-ended queue application settings range from organ transplants where patients waiting for organs would be matched with donors or organs already obtained, to communication protocols. The management of operations for these versatile settings prioritize distinct objectives, and also as technology advances these systems will be run with different flexibilities and opportunities. Therefore, it is essential to continue the research on double-ended queues which also represents itself with continuously improving literature (Bhardwaj et al., 2014: 259; Wang et al., 2017: 264-265).

Another reason for the continuing research on double-ended queuing systems is the benefits of advances in technology enabling different types of analysis to be executed on the same model. Most queueing systems' analytical derivations become very difficult and intractable as more assumptions are flexed; however, due to the advances in computer technology approximate solutions or simulation-based experiments could be conducted where the traditional analytical derivations become impossible (Kim et al., 2010: 216).

A double-ended queue represents a situation where there is a demand process arriving to the queue as well as a supply process arriving to the queue to match the demand process. In the traditional model (Kashyap, 1965: 559) both demand and supply process has arrivals in single units and demand will need only a single unit to be served. At the time of arrival for both sides, if there is a queue of attendants waiting for the arrival of the unit under consideration, then, the demand and supply is matched instantaneously and leave the system together. It is also assumed that as long as a demand unit and a supply unit come across each other, the match will happen without failure. The arrival processes were initially assumed to be exponential as well. A depiction of this system could also be found in Figure 1 below.

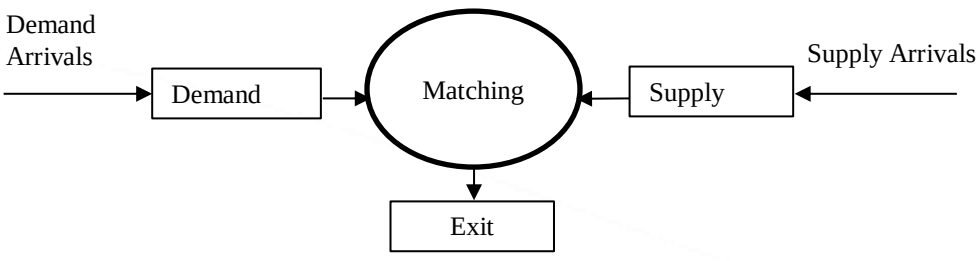

Figure 1. Depiction of a generalized double-ended queue system

This paper focuses on the developments of operations management for this system and reviews and evaluates various extensions to the generic double-ended queuing model. In the next section, we will cover the modeling aspect of double-ended queue analysis where double-ended queue systems with different characteristics, different control mechanisms, and analyzed through various methodologies are presented. In the third section we will describe the application settings for the double-ended queue model in more detail, followed by some representative studies on optimization of double-ended queue systems with diverse objectives prioritized. We will end with the conclusion part.

2. MODELING OF DOUBLE-ENDED QUEUES

In the generic double-ended queue model both demand and supply arrival processes are assumed to be Markovian, i. e. with random exponential interarrival times. Thus, the time period between consecutive arrivals is assumed to be exponentially distributed. Moreover, the queue space is assumed to be unlimited, hence both demand and supply queues had unlimited waiting room space. No priorities were assumed, so the service discipline was assumed first-come-first-serve. The steady-state probabilities for the traditional model could be found in (Dobbie, 1961: 756-757). However, there have been many extensions to this model.

We will present some representative modeling extensions to the double-ended queuing systems based on three perspectives: the system under consideration, controls applied to the system by management, and solution methodologies.

Extensions to the basic double-ended queue model by relaxing the system design assumptions target having less structured demand and supply arrivals, giving the customers the opportunity to balk or renege from the system, having more than one customer class with different priority levels, and accepting the possibility of catastrophic instances which restart the whole system from state zero after the repair time passes.

In most queueing systems, admission principles and issues would be one of the major topics that would determine how the system will perform and how the users will benefit from it. In addition to that, a reality we all face every day is how challenging waiting could be. Not only having an unpleasant feeling, but also due to emergency needs waiting may not be possible at all for some customers as well. Hence, there have been studies with the assumption that customers are impatient (Diamant and Baron, 2019: 220-221; Degirmenci, 2010: 43-46), and they immediately or after some time leave the system if they are not served when they arrive.

Another extension to the Markovian double-sided queue model would be on the structure of arrival processes for supply and demand as well as patience times of the arrival units on each queue. For example, arrivals being renewal processes with exponential patience times (Liu et al., 2014: 5), arrivals being generally distributed independent and identical interarrival times in random batches (Jain, 2000: 194) have been studied. Moreover, the possibility of having more than one customer class is also recognized previously in the literature (Liu, 2019: 52).

Besides arrival processes, there are other characteristics of the system design the literature has considered. These characteristics are about the nature of the matching process, for instance, the matching may take a random amount of time before the matched units leave the system together or there could also be possibility of matching being unsuccessful, coupled with batch arrivals and a potential demand of more than one unit for service (Kim et al., 2010: 210), these types of systems could be very intractable to study via traditional queueing equations. Up to now the extensions discussed assumed the system will be running in an undisturbed environment. However, this is definitely not always the case; thus, there have been studies assuming catastrophes with exponentially distributed repair times, where the system restarts from zero every time it occurs (Di Crescenzo et al., 2012: 939) or

catastrophes occurring based on a non-homogenous Poisson process with time-varying repairs (Di Crescenzo et al., 2018: 3).

The controls applied to the double-ended queueing system by the management have also been recently considered to be versatile due to the advances in technology. While it could be natural to think the management of the system could easily control admission, priority, and other issues about operations, without traceability these controls could not be applied even if they are intended to. Therefore, as information technologies advanced it became possible to assert more controls over the system. For example, dynamic control of the queue on one side may change the performance measures of the system drastically. If one of the queues become congested due to limited waiting space, the management may choose to delay these arrivals until the system becomes less congested. Hence, there could be a gateway policy with management's decision residing on the maximum allowed congestion level to admit arrivals into the system. For this type of system, the results of interest could be equilibrium and socially optimal strategies with different information levels (Wang and Liu, 2019: 2).

In terms of methodology, the traditional approaches were focusing on steady-state analysis where exact derivation of stationary performance measures such as average queue length, probability of system being empty, and similar were proved (Kashyap, 1967: 169; Jain, 2000: 197; Diamant and Baron, 2019: 222-223). However, the extensions discussed above pointed to many systems being intractable if all or some assumptions were relaxed. Therefore, alternative methods other than steady-state analysis were explored in the literature. One of such methods is the asymptotic analysis where the system presents special characteristics when the dimensions or the scale of the system is assumed to approach infinity. With such perspective, the stochastic processes that govern the system have limiting values that would provide good approximations as the system becomes loaded with very high congestion. These solution mechanisms are named fluid and diffusion approximations under the heavy-traffic regime and research via these methods is being actively conducted (Di Crescenzo et al., 2012: 939; Liu et al., 2014: 5; Liu, 2019: 52). For the systems with limited waiting space on one arrival process regarding the supply, social welfare optimization through equilibrium joining strategy analysis has also been studied (Wang et al., 2017: 264-265).

From a modeling perspective, many extensions to the traditional double-ended queueing system has been done, and these abstract models are very powerful tools to represent many different real life systems. The strength of the abstract expression through queueing theory of waiting systems lies in the ability of these models to express dynamics of versatile systems. In the next section, we will review some of the most popular and representative applications of double-ended queueing mathematical models. The settings will cover many distinct service units, users, and other waiting line system elements.

3. APPLICATIONS OF DOUBLE-ENDED QUEUEING MODELS

The taxi-passenger matching problem at an airport has been the original problem setting that has given rise to double-ended queueing models. At the airports, many passengers would be waiting for a taxi to come. Especially at peak hours it is common to see a long queue of passengers waiting unless this system is well operated. Referring to Figure 1, the double-ended queue for the taxi-passenger system would have passengers as the demand process and the taxis as the supply process. Assuming the system starts at zero state, there will be either a queue of passengers waiting for a taxi or a queue of taxis, as soon as there will be a match with an arrival to the opposite side, the arriving passenger (taxi) will be matched with the waiting taxi (passenger) and will leave the system. The trade-offs for this type of system would be between the passengers' loss of time and the opportunity cost of the taxis that end up waiting at the queue. Moreover, the taxis waiting represent a transportation capacity for the city, and the managers could opt to put a buffer or limited waiting area for a certain number of taxis to be allowed to wait at the same time. For this kind of decision, the

social welfare criteria should be considered via accounting for the strategic behavior of passengers that could affect the arrival rate for the demand side (Shi and Lian, 2016: 1025).

A current application of double-ended queueing models is the organ transplant matching. Modern medical science does not have a storage device to keep the organs well enough to wait for a long time. The other option is to find a donor, and sometimes it is also not possible to find a match by the patient himself/herself. However, it is possible to form a queue of patient and donor couples in a region. If there is another patient and donor couple at another region which corresponds to the queue at the opposite side, such that the donors of the couples match with the patients of the couple from the opposite side, then the two couples are matched and they leave the queues (Degirmenci, 2010: 41). In case of an invention that would allow the organs to be stored, the patients will not need to bring a donor to the queue and s/he may directly be placed in the queue herself/himself. In this case the other queue will be formed by the organs stored (Elalouf et al., 2018: 182). In the organ transplant case, a phenomenon to be included in the model would be the probability of the patients dying and hence leaving the queue before receiving service. For the patient-donor couples, the patients at either side could die, for the patient queue vs. stored organ case the stored organ may perish or expire as well. Therefore, the double-ended queue models should have the arrivals renege after a random time passes without service.

The double-ended queueing models have also been used to manage and design shared data structure and communication protocols. These implementations refer to the double-ended queueing model as dequeues and design algorithms to run the system with less congestion (Herlihy et al., 2003: 2). In general, it is also possible two different processes may need to communicate messages between each other and sometimes the messages on hold would be from one process sometimes the other, with messages arriving based on the tasks executed by the processes. The double-ended queue models would be helpful in these cases as well.

Other applications of double-ended queue include baggage claim inventory management at the airport, where the baggage and passenger arrivals are modeled as having uniform rate, with the possibility that a passenger may have a batch of baggage, rather than only one unit (Browne et al., 1970: 65). Diffusion approximations as described in the previous section have also been derived for a double-ended queue model with the setting being a hospital with unlimited bulk arrivals of patients but limited waiting space where a certain number of servers are performing limited number of services (Pandey and Gangeshwer, 2018: 306).

Besides the common and typical application settings, many different systems could also be modeled as a double-ended queue, for example, an automatic storage and retrieval system could be described with a double-ended queue model. For such a representation, the two arrivals coming to the system will be the packages that needs to be stored and the packages that are requested to be retrieved. Storage of the packages will be limited at each instant. Indeed, this type of system has become substantially important especially for online retailers due to their centralized warehouse operations, and any e-retailer company that manages their storage operations with minimum cost could obtain a competitive advantage with the use of double-ended queue models (Dolhun, 1997: 3).

Standard manufacturing and assembly systems could also be modeled with double-ended queues. For instance, for a product or a component comprising of two parts and produced via assembly of these two parts, its production would present a double-ended queueing model, where production of the two assembled parts would form the two arrival processes and their assembly would correspond to the matching of units. In this example, the management has control over the arrival rates since they are formed by the production rates, and the matching time could be nonzero since it would represent the assembly procedure (Som et al., 1994: 472).

As can be seen from above, double-ended queue models have many different application areas, while these examples are not exhaustive, we included them here to provide a representative review of its potential uses. In the next section we will consider the optimization approaches towards double-ended queues, which constitutes the core of the operations management of these systems.

4. DOUBLE-ENDED QUEUE OPTIMIZATION APPLIED THROUGH VARIOUS OBJECTIVES

With the modeling, methodology, and application settings being versatile, different objectives are pursued for optimization of double-ended queues. For the optimization studies on double-ended queues that pursue balancing of the arrivals to the two opposite queues, the controls to be used by the management could either be reducing the arrivals that are causing excess waiting at the respective queue or increasing the arrivals at the other queue. For these types of optimization problems on double-ended queues, the objective function would usually be minimizing costs that are imposed on the system due to waiting. For instance, there could be costs incurred per arrival unit per time unit spent waiting in the queue, and the constraints could be limited queue space and the level of controls imposed on the system (Mendoza, 2009: 93).

Another perspective on balancing supply and demand in a double-ended queueing setting would be to control production rate of a manufacturing system where arrivals on the demand side would form the backorders and the arrivals to the opposite side would form the inventory. For non-stationary demand, impatient customers who may cancel orders if backordered and perishable products that would exit the system due to becoming obsolete after a random exponential amount of time, and the inflexibility of the system resulting in additional costs if production rate is fluctuated spontaneously, fluid and diffusion approximations provide bounds on the performance measures and asymptotically optimal solutions (Lee et al., 2019: 5).

The balancing of supply and demand considerations could involve more than the costs incurred via waiting. For many real-life systems, throughput, which, for these settings correspond to the number of matchings executed, would constitute a reward mechanism for the system. Hence, considering throughput in the optimization would result in turning the objective into maximizing the profit consisting of rewards minus the waiting costs (Koca et al., 2014: 47). There could be differing assumptions about limited waiting space, arrival processes, impatience of the arrival units, and similar.

Going back to the passenger-taxi example with the advances of information technologies it is now possible to inform the passengers of the queue length, and therefore giving them an option to balk if the queue length exceeds a threshold. Such strategic behavior coupled with the technology makes it possible to derive equilibrium behavior that would optimize social welfare, which would point to the design of the system in a way to reach the equilibrium. Currently this perspective is explored with the traditional passenger-taxi problem having different assumptions for arrival processes, limited waiting space, impatience of the customers and so on.

In this section we have covered representative optimization approaches towards double-ended queues since optimization constitutes the core of operations management, it is seen that despite being a well-known problem, as times change, various new conditions and perspectives are added into the setting and studies are continuously conducted.

5. CONCLUSION

In this paper double-ended queues were the research topic, and our approach was to first introduce the basic model, then discuss the relaxed assumptions on the model as literature has progressed, this was followed by the introduction of the many application settings the models was used to represent, and finally we have reviewed some of the main optimization approaches to the systems demonstrated by double-ended queues. Operations

management of double-ended queues would require modeling of the system setting through an abstract queueing model as we have discussed here, such that the model would fit to the real life situation at hand. After the model is constructed, optimization through different methodologies should be conducted to design and derive operating principles for the system. This paper provides a demonstration of model in general, its use, and as well as the recent advances from this perspective.

The current research trend on double-ended queues is leaning towards topics such as deriving equilibria for social welfare with strategic customers, developing models for futuristic technology use to advance the double-ended queue systems, and developing approximate or near-optimal solutions for analytically intractable problems via methods such as simulation or fluid and diffusion approximations.

REFERENCES

Bhardwaj, R., Singh, T. P., & Kumar, V. (2014). A generalized double ended stochastic queue system with excess customer demand in real world situations. *Arya Bhatta Journal of Mathematics and Informatics*, *6*(2), 247-260.

Browne, J. J., Kelly, J. J., & Le Bourgeois, P. (1970). Maximum inventories in baggage claim: a double ended queuing system. *Transportation Science*, *4*(1), 64-78.

Degirmenci, I. T. (2010). Asymptotic analysis and performance-based design of large scale service and inventory systems (Doctoral dissertation, Department of Business Administration, Duke University).

Di Crescenzo, A., Giorno, V., Kumar, B. K., & Nobile, A. G. (2012). A double-ended queue with catastrophes and repairs, and a jump-diffusion approximation. Methodology and Computing in Applied Probability, 14(4), 937-954.

Di Crescenzo, A., Giorno, V., Krishna Kumar, B., & Nobile, A. (2018). A time-non-homogeneous double-ended queue with failures and repairs and its continuous approximation. Mathematics, 6(5), 81, 1-23.

Diamant, A., & Baron, O. (2019). Double-sided matching queues: Priority and impatient customers. *Operations Research Letters*, *47*(3), 219-224.

Dobbie, J. M. (1961). Letter to the Editor—A Doubled-Ended Queuing Problem of Kendall. *Operations Research*, *9*(5), 755-757.

Dolhun, K. L. (1997). *A double-ended single server queueing system*. Unpublished master thesis. Faculty of Graduate Studies, University of Manitoba, Canada.

Elalouf, A., Perlman, Y., & Yechiali, U. (2018). A double-ended queueing model for dynamic allocation of live organs based on a best-fit criterion. *Applied Mathematical Modelling*, *60*, 179-191.

Gaur, K. N., & Kashyap, B. R. K. (1973). The double-ended queue with limited waiting space. *Indian Journal of Pure and Applied Mathematics*, *4*, 73-81.

Herlihy, M., Luchangco, V., & Moir, M. (2003, May). Obstruction-free synchronization: Double-ended queues as an example. In *23rd International Conference on Distributed Computing Systems, 2003. Proceedings.* (pp. 522-529). IEEE.

Jain, M. (2000). GX/GY/1 double ended queue: diffusion approximation. *Journal of Statistics and Management Systems*, *3*(2), 193-203.

Kashyap, B. R. K. (1965). A double-ended queueing system with limited waiting space. In *Proc. Nat. Inst. Sci. India* (Vol. 31, No. 6, pp. 559-570).

Kashyap, B. R. K. (1967). Further results for the double ended queue. *Metrika*, *11*(1), 168-186.

Kim, W. K., Yoon, K. P., Mendoza, G., & Sedaghat, M. (2010). Simulation model for extended double-ended queueing. *Computers & Industrial Engineering*, *59*(2), 209-219.

Koca, E., Sedaghat, M., & Yoon, K. P. (2014). Optimal Supply & Demand Balance In Service Environments. *Journal of Service Science (Online)*, *7*(1), 43-52.

Lee, C., Liu, X., Liu, Y., & Zhang, L. (2019). Optimal Control of a Time-Varying Double-Ended Production Queueing Model. *Available at SSRN 3367263*.

Liu, X. (2019). Diffusion approximations for double-ended queues with reneging in heavy traffic. *Queueing Systems*, *91*(1-2), 49-87.

Liu, X., Gong, Q., & Kulkarni, V. G. (2014). Diffusion models for double-ended queues with renewal arrival processes. Stochastic Systems, 5(1), 1-61.

Mendoza, G., Sedaghat, M., & Yoon, K. P. (2009). Queuing models to balance systems with excess supply. International Business & Economics Research Journal (IBER), 8(1), 91-104.

Pandey, M. K., & Gangeshwer, D. K. (2018). Applications of the Diffusion Approximation to Hospital Sector Using G∞/GM/1 Double Ended Queue Model. *Journal of Computer and Mathematical Sciences*, *9*(4), 302-308.

Shi, Y., & Lian, Z. (2016). Optimization and strategic behavior in a passenger–taxi service system. *European Journal of Operational Research*, *249*(3), 1024-1032.

Som, P., Wilhelm, W. E., & Disney, R. L. (1994). Kitting process in a stochastic assembly system. *Queueing Systems*, *17*(3-4), 471-490.

Wang, Y., & Liu, Z. (2019). Equilibrium and Optimization in a Double-Ended Queueing System with Dynamic Control. *Journal of Advanced Transportation*, *2019*, 1-13.

Wang, F., Wang, J., & Zhang, Z. G. (2017). Strategic behavior and social optimization in a double-ended queue with gated policy. *Computers & Industrial Engineering*, *114*, 264-273.