Computers & Operations Research 76 (2016) 33–42

Contents lists available at ScienceDirect

Computers & Operations Research

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

Predicting the performance of queues–A data analytic approach

Kum Khiong Yang [a,*], Tugba Cayirli [b], Joyce M.W. Low [a]

[a] Lee Kong Chian School of Business, Singapore Management University, Singapore
[b] School of Economics and Administrative Sciences, Ozyegin University, Turkey

ARTICLE INFO

Article history:
Received 9 July 2014
Received in revised form
3 September 2015
Accepted 7 June 2016
Available online 14 June 2016

Keywords:
Data Analytics for Queues
Simulation
Nonlinear regression
Alternating Conditional Expectation

ABSTRACT

Existing models of multi-server queues with system transience and non-standard assumptions are either too complex or restricted in their assumptions to be used broadly in practice. This paper proposes using data analytics, combining computer simulation to generate the data and an advanced non-linear regression technique called the Alternating Conditional Expectation (ACE) to construct a set of easy-to-use equations to predict the performance of queues with a scheduled start and end time. Our results show that the equations can accurately predict the queue performance as a function of the number of servers, mean arrival load, session length and service time variability. To further facilitate its use in practice, the equations are developed into an open-source online tool accessible at http://singlequeuesystemstool.com/. The proposed procedure of data analytics can be used to model other more complex systems.

1. Introduction

Waiting in line is a common phenomenon in daily life. Customers in banks, postal offices and supermarkets wait in line for services. Cars queue up for gas re-fills and jobs wait for idle machines to start production. Long waiting time leads to unhappy customers and congestion in production facilities. Waiting line or queue management is hence a critical part of both service and manufacturing firms alike. In general, queue management involves a trade-off between the cost of waiting and the cost of additional service capacity [26,38]. The former is often measured by the mean number of customers or mean waiting time in queue while the latter is measured by the mean server utilization or mean overtime per customer served. The probability of a customer being served immediately on arrival is also an important indicator of customer service, and it measures the proportion of time that at least one server is idle. Management needs to work on a strategy that reduces waiting times and delights its customers without excessive capacity cost.

Queueing theory is a long-standing discipline that is used to derive formulae to compute performance measures, such as mean queue length and customer waiting time under various queueing configurations that can be represented by the Kendall's notation [19,20]. In its simplest form, the *M*/*M*/*C* model represents a queueing system with *C* servers and an infinite queue capacity such no customer is rejected. The mean customer arrival rate is characterized as stationary and does not vary with time. In addition, both customer inter-arrival times and service times are assumed to follow exponential distributions, and customers wait in a single queue and are served in the order of first come-first served. The sophistication of such formulae increases if the inter-arrival and service times follow other distributions, such as a General (*G*) or Erlang (*E*) distribution, in *M*/*G*/*C*, *G*/*M*/*C*, *M*/*E*/*C*, *E*/*M*/*C*, and *E*/*E*/*C* queueing models.

Nevertheless, many real systems operate in environments where the core assumptions of the above queueing models are violated. Many systems experience nonstationary customer arrival rates that vary both within and across days. For examples, a post office may encounter higher arrival rates during lunch hours and on Mondays. Similarly, an Accident and Emergency Department may receive more patients on Sundays when other clinics are closed. Service systems may also offer services with different service time variability. A specialist clinic, for example, may encounter higher variable service times than a family clinic. Last but not least, many service systems do not operate continuously, but encounter opening and closing transience every day. Under such circumstances, Yang et al. [66] show that the formulae derived for the standard *M*/*M*/*C* queueing models fail to provide satisfactory estimates of the system performance.

Whilst our literature review identifies a number of methods that can model the effects of system transience and non-standard queueing assumptions, these procedures are highly complex. The need for an easy-to-use procedure thus remains. This research

[*] Correspondence to: Lee Kong Chian School of Business, Singapore Management University, 50 Stamford Road, 17889 Singapore.
E-mail addresses: kkyang@smu.edu.sg (K.K. Yang), tugba.cayirli@ozyegin.edu.tr (T. Cayirli), joycelow@smu.edu.sg (J.M.W. Low).

http://dx.doi.org/10.1016/j.cor.2016.06.005
0305-0548/© 2016 Elsevier Ltd. All rights reserved.

proposes combining computer simulation and a non-linear regression technique called Alternating Conditional Expectation (ACE) as a procedure for analyzing complex queueing systems. As an illustration, we apply our procedure on a multi-server queueing system that operates with non-exponential service times and with opening and closing transience. Computer simulation is used to generate the data while ACE is used on the data to construct a set of easy-to-use analytical equations to predict the system performance. Once the number of servers, arrival load, session length, and service time variability for such systems are specified, the ACE equations can predict the key performance measures, such as the mean queue length, probability of no waiting on arrival and mean overtime per customer served. Our results show that our proposed procedure is highly accurate in predicting the performance of the queueing systems that are tested. To facilitate the use of the ACE equations, an online open source tool http://singlequeuesystemstool.com/ is developed for practitioners who can use it to estimate the performance of queuing systems that violate the assumptions of exponential service times and continuous operations. The proposed procedure of combining computer simulation and ACE can be used for modeling other more complex systems.

The rest of the paper is organized in the following manner: The next section briefly reviews the related literature on existing methods for predicting the performance of multi-server queues with system transience. These methods are generally complex with assumptions, which limit their use and accuracy in practice. Section 3 introduces Alternating Conditional Expectation (ACE) as an advanced nonlinear regression technique and develops a set of regression equations for predicting the key performance measures of multi-server systems in the presence of opening and closing transience. Section 4 presents the results on the accuracy of the ACE equations, while Section 5 ends with the conclusions, limitations and suggestions for future research.

2. Literature review

Queueing theory dates back to the works by Erlang in the early 1900 s, and its theoretical development has grown substantially since 1950 s with the advances in Operations Research [18]. The earliest models assume a stationary $M/M/C$ queue[1] that operates continuously and are popular for their simplicity and ease of use. They are proposed for a wide range of applications such as telecommunications, manufacturing, and services. Readers can refer to Lakshmi and Iyer [39] for a comprehensive taxonomy of queueing applications in healthcare services and to Gans et al. [9] for applications in call centers.

As the literature on queueing is extremely broad, we have to limit our review to works related to our current study. Therefore, we first review literature on the steady-state $M/G/C$ and $G/G/C$ queues, with stationary (i.e. constant) mean arrival and service rates that do not vary with time. In the second part of our review, we discuss the added difficulties of analyzing transient queues with opening and closing of the system, and nonstationary arrival and/or service process.

2.1. Steady-state queues with stationary arrival and service process

Beyond the stationary $M/M/C$ queues, several papers have highlighted the challenges in analyzing non-Markovian queues with non-exponential inter-arrival and/or service times. The emphasis is on finding approximate solutions given that such generalized models are analytically intractable [65]. The earlier works on the $M/G/C$ queues include Takahashi [53], which provides approximate formulae for the first two moments of waiting times. Hokstad [22] proposes an approximation based on Laplace transform for the steady-state queue length distribution. Several approximations in both light and heavy traffics are introduced and tested by Boxma et al. [3], Tijms et al. [55], and Kimura [33,36]. Recognizing that most of the earlier approximations fail when both the number of servers (C) and the coefficient of variation (CV) are large, especially when traffic intensity is low, Ma and Mark [42] present a computationally efficient approximation for the mean queue length of $M/G/C$ queues. In their simulation study, Mandelbaum and Schwartz [43] show that results on $M/G/C$ queues are highly sensitive to the service time distribution.

The more generalized $G/G/C$ queues with general inter-arrival and service time distributions have also been studied extensively. Approximations are the most common methods, apart from studies that apply exact numerical methods to phase-type distributions [49,52] or hyper-exponential and Erlang distributions [57]. Kimura [34,35,37] provides a number of approximations for the $G/G/C$ queues with infinite and finite queue capacity. It is reported that these approximations become less accurate when the variability of the inter-arrival and service times increases and when the traffic intensity is low or moderate. Whitt [60,61] uses an infinite-server approximation and notes that the quality of approximation varies across different performance measures.

A group of related papers addresses the QED (Quality and Efficiency Driven) regime for a special class of queues with high server-utilization and short waiting times relative to service times. This regime, also known as the Halfin-Whitt regime, is introduced by Halfin and Whitt [17] in the analysis of a $G/M/C$ queue. Several extensions are developed by Puhalskii and Reiman [47], Whitt [63,64], Jelenkovic et al. [28], Mandelbaum and Momcilovic [44], and Reed [48]. This is a well-studied problem in the context of call-centres, where relevant trade-offs exist between servers' utilization (i.e. efficiency) and customers' waiting times (i.e. quality) (see [9]).

2.2. Mean state of transient queues

In Section 2.1, it is concluded that no easy-to-use formulae exist for computing the steady-state performance of both the $M/G/C$ and $G/G/C$ queues. Analysis of transient queues with the presence of opening and closing of the system and/or nonstationary arrival and service process is even more difficult, as discussed in the following sections.

2.2.1. Queues with opening and closing of the system

In systems that do not operate continuously, the steady-state models may not apply. For example, a clinic may accept patients only from 9 am to 5 pm such that each day begins and ends with an empty queue. Several queueing papers offer transient solutions on how various measures, such as queue length and probability of no waiting, can vary over time. Such transient solutions can then be used to compute the time-average or mean state of the system. While there are many studies on the single-server transient queues, such as Abate and Whitt [1], Bertsimas and Nakazato [2], Gong and Hu [10], Lee and Roth [41] and Wang [58], few studies exist on the multi-server transient queues. Kelton and Law [31] and Murray and Kelton [46] use an embedded discrete-time Markov chain to calculate the probability of transient queue-length of $M/Ph/C$ queues with exponential inter-arrival times and phase-type service times. Chaudhry and Zhao [7] provide an analysis of a transient queue with finite queue capacity. Whitt [62] analyzes the steady-state and transient performance of an $M/G/C$ queue with a heavy-tailed service time distribution, and shows that the waiting time distribution is also heavy-tailed. Using

[1] Note that for consistency, we use a common notation "C" for multiple servers, regardless of the original notations used in the discussed papers which may include "n", "m" or "s".