Operations Research Perspectives 2 (2015) 81–96

Contents lists available at ScienceDirect

Operations Research Perspectives

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

Fluid approximation analysis of a call center model with time-varying arrivals and after-call work

Yosuke Kawai [a], Hideaki Takagi [b,*]

[a] Graduate School of Systems, Information and Engineering, University of Tsukuba, Tsukuba Science City, Ibaraki 305-8573, Japan
[b] Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba Science City, Ibaraki 305-8573, Japan

ARTICLE INFO

Article history:
Available online 4 April 2015

Keywords:
Fluid approximation
Multiserver queue
Call center
Time-varying arrival process
After-call work
Abandonment

ABSTRACT

Important features to be included in queueing-theoretic models of the call center operation are multiple servers, impatient customers, time-varying arrival process, and operator's after-call work (ACW). We propose a fluid approximation technique for the queueing model with these features by extending the analysis of a similar model without ACW recently developed by Liu and Whitt (2012). Our model assumes that the service for each quantum of fluid consists of a sequence of two stages, the first stage for the conversation with a customer and the second stage for the ACW. When the duration of each stage has exponential, hyperexponential or hypo-exponential distribution, we derive the time-dependent behavior of the content of fluid in each stage of service as well as that in the waiting room. Numerical examples are shown to illustrate the system performance for the cases in which the input rate and/or the number of servers vary in sinusoidal fashion as well as in adaptive ways and in stationary cases.

1. Introduction

Queueing models have been widely used to model the performance of call centers with impatient customers [1–4], which means that customers in the waiting line may leave before getting service. The multiserver queue M/M/s with impatient customers is called *Erlang-A* model, "A" for "abandonment", in contrast with the well-known *Erlang-B* model (M/M/s/s) and the *Erlang-C* model (M/M/s with only patient customers).

Through the measurements at real call centers, however, we observe that operators usually spend sizable amount of time to complete additional work after finishing conversation with customers. For example, they enter customer profiles and summary of conversation into the customer management database after conversation. Such extra work of operators is called the *after-call work* (ACW). Cleveland and Harne [5, Section 8] describe:

> The ACW is the work that is necessitated by and immediately follows an inbound transaction. Often includes entering data, filling out forms and making outbound calls necessary to complete the transaction. The agent is unavailable to receive another inbound call while in this mode.

The ACW is also called "post call activity" [6–8], "wrap-up times" [1], "after-hung-up times" [9], and "postservice activity" [10,11]. Harris and Phillips [6] mention:

> The post call activity is a phase in which the operator may fill out dockets, make supplementary phone calls or perform other clerical activities before pressing a key to indicate that he/she is able to accept another call from the queue (if such a call is present).

Takagi and Taguchi [12] study a two-dimensional birth-and-death process for the M/M/K/J queue with ACW, where K, the number of servers, represents the total number of operators working in the call center and J, the maximum number of customers accommodated in the system, stands for the number of incoming telephone lines. Unlike usual queueing models, we do not necessarily assume that $J \geq K$, because servers may be working on ACW while some customers are present in the waiting room. Phung-Duc and Kawanishi [13] present a matrix-geometric analysis for a queueing model with retrial arrivals of blocked and abandoned customers. All models in these pieces of work assume the steady state of the system.

Another realistic feature of call center operation is that the call input process is time-varying. However, the exact stochastic analysis of a queueing model with multiple servers with generally distributed service times and/or time-varying arrival process is not easy. The *fluid approximation* technique has been exploited to

[*] Corresponding author. Tel.: +81 29 853 5414; fax: +81 29 853 7291.
E-mail address: takagi@sk.tsukuba.ac.jp (H. Takagi).

http://dx.doi.org/10.1016/j.orp.2015.03.003

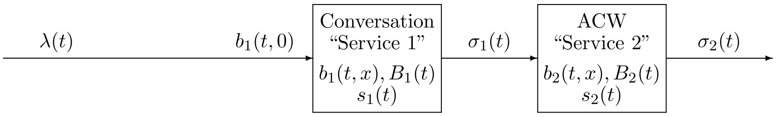

Fig. 1. System model and state variables in the underloaded state.

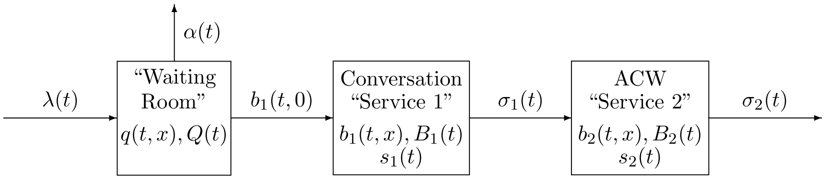

Fig. 2. System model and state variables in the overloaded state.

deal with such models traditionally [14]. More recently, the fluid approximation is applied to stationary multiserver queues with impatient customers [15,16] as well as those with the time-varying input rate and number of servers [17–19].

In this paper, we present the fluid approximation for a multiserver queueing model with impatient customers, two stages of service time (representing the conversation and ACW in a call center), and time-varying input rate and number of servers. Our approach is an extension of the method for the $\mathrm{M}_t/\mathrm{GI}/s_t + \mathrm{GI}$ model originally developed by Liu and Whitt [20,19] (they later extended the analysis to networks of fluid queues [21,22]). In this notation, "M_t" means a Poisson arrival process with time-varying arrival rate, the first "GI" an independent, generally distributed service time, "s_t" a time-varying number of servers, and "+GI" a general abandonment-time distribution.

We show that the system state alternates between the *underloaded interval* in which there are idle servers and the *overloaded interval* in which arriving fluid quanta must wait for service because all servers are busy. We study the dynamics of the fluid content in service in both underloaded and overloaded states. We also study the dynamics of the fluid content in the waiting room and the waiting time of a fluid quantum that arrives in the overloaded state. Our analysis is applied to several illustrative cases with time-varying input rate and number of servers. If the number of servers is determined adaptively to cope with only the load of conversation, the system is always overloaded but it remains stable. If the number of servers is determined adaptively in accordance with the load of both conversation and ACW, the system is always underloaded.

To the best of the authors' knowledge, this paper is the first work in which the fluid approximation is applied to the $\mathrm{M}_t/\mathrm{GI}/s_t+$ GI model with two stages of service time as a model of the call center operation with ACW. This paper is partly based on the Master Thesis of the first author [23] submitted to the Graduate School of Systems, Information and Engineering of the University of Tsukuba, Japan.

2. Fluid model of call center operation with after-call work

In this section, we introduce a fluid model of the call center operation with ACW by extending the model and analysis by Liu and Whitt [19,16].

2.1. Definition of the system model

We consider a fluid queueing system with multiple servers where incoming calls in a call center are modeled by quanta of fluid. We assume that the service time a server, representing an operator, spends on each fluid quantum consists of a sequence of two stages, called "service 1" for the conversation with a customer and "service 2" for ACW, each having independent duration. We assume that the same server continues to provide service 2 immediately after service 1 for each fluid quantum. Let there be $s(t)$ servers in the system at time $t \geq 0$. The *staffing function* $s(t)$ is given exogenously or adaptively somehow depending on the input rate of fluid. At any time, each server is either in service or being idle such that $s_i(t)$ servers are engaged in service i $(i = 1, 2)$, where $s(t) \geq s_1(t) + s_2(t)$.

The input of fluid quantum directly enters service 1 if there is a server available; this state is called *underloaded*. Otherwise, the input flows into the "waiting room" for service 1; this state is called *overloaded*. No waiting room is needed for service 2 because the same server takes care of the ACW for the fluid quantum that he has just given service 1. The server who has finished service 2 can start service 1 for another fluid quantum if any in the waiting room, or he becomes idle otherwise. The fluid quanta leave the system either by completing service 2 or by abandonment while being in the waiting room. The fluid quanta never leave the system during services 1 and 2. For a system with time-varying staffing function, we assume that the time variation in the total number of servers is solely turned to the time variation in the number of servers assigned to service 1. We also assume that the number of servers assigned to each service never goes below the level of fluid content in that service at any moment so that no fluid quanta are forced out of the system once they have entered service. This model is schematically depicted along with relevant state variables in Figs. 1 and 2 for the underloaded and overloaded states, respectively. The state variables are introduced in the following subsection.

2.2. Definition of state variables and their relations

We assume that the fluid quanta arrive at service 1 according to a deterministic process with time-varying rate $\lambda(t),\ t \geq 0$. We denote by $F(x)$ and $f(x)$ the distribution function and the probability density function (pdf), respectively, for the abandonment time of each fluid quantum in the waiting room. Also, we denote by $G_i(x)$ and $g_i(x)$ the distribution and density functions, respectively, for the service time of each fluid quantum in service i $(i = 1, 2)$. Thus we have

$$F(x) := \int_0^x f(u)du, \qquad G_i(x) := \int_0^u g_i(u)du \quad x \geq 0,\ i = 1, 2.$$

Furthermore, let $\overline{F}(x)$ and $\overline{G}_i(x)$ be their complimentary distribution functions (CDF's) defined by

$$\overline{F}(x) := 1 - F(x), \qquad \overline{G}_i(x) := 1 - G_i(x) \quad x \geq 0,\ i = 1, 2.$$

These functions are assumed to be given in the model.

At time $t(\geq 0)$, we denote by $Q(t, x)$ the fluid content that has been waiting for the time units less than or equal to x in the waiting room. Similarly, we denote by $B_i(t, x)$ the fluid content in service that has been in service i for the time units less than or