Kawai, Yosuke; Takagi, Hideaki

Article

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

Operations Research Perspectives

Provided in Cooperation with:
Elsevier

Suggested Citation: Kawai, Yosuke; Takagi, Hideaki (2015) : Fluid approximation analysis of a call center model with time-varying arrivals and after-call work, Operations Research Perspectives, ISSN 2214-7160, Elsevier, Amsterdam, Vol. 2, pp. 81-96, https://doi.org/10.1016/j.orp.2015.03.003

This Version is available at:
https://hdl.handle.net/10419/178252

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

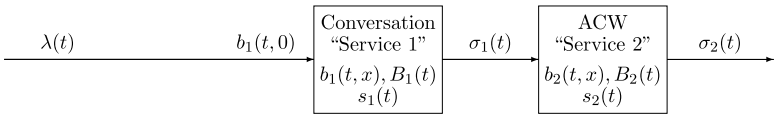


Fig. 1. System model and state variables in the underloaded state.

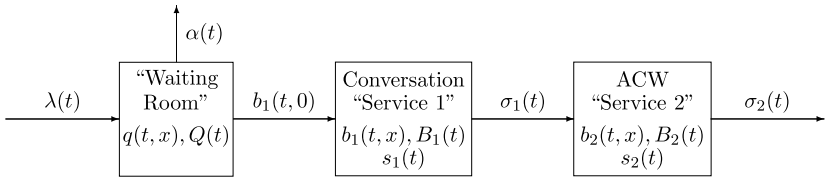


Fig. 2. System model and state variables in the overloaded state.

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 $M_t/GI/s_t + GI$ 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 $M_t/GI/s_t + GI$ 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

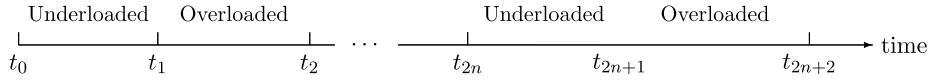


Fig. 3. Alternation of underloaded and overloaded intervals.

equal to x $(i = 1, 2)$. Furthermore, let $q(t, x)$ and $b_i(t, x)$ be their corresponding density functions as

$$Q(t,x) = \int_0^x q(t,u)du, \qquad B_i(t,x) = \int_0^x b_i(t,u)du$$
$$x \geq 0,\ i = 1, 2. \tag{1}$$

Then the fluid content in the waiting room and that in service i at time t are respectively given by

$$Q(t) := Q(t,\infty), \qquad B_i(t) := B_i(t,\infty) \quad i = 1, 2. \tag{2}$$

Let us define $w(t)$ as the largest waiting time for the fluid in the waiting room:

$$w(t) := \inf\{y \geq 0 : q(t,x) = 0 \text{ for all } x > y\} \quad t \geq 0.$$

If we assume that fluid quanta enter service in the order of arrival, i.e., according to the first-come first-served (FCFS) discipline, $w(t)$ is the waiting time of a fluid quantum that enters service at time t. Since

$$q(t,x) = 0 \quad x > w(t),$$

$w(t)$ is called the *boundary waiting time* by Liu and Whitt [19].

The above-defined density functions for the fluid content satisfy the following *fundamental evolution equations* [19, Assumption 6]:

$$q(t+y, x+y) = q(t,x)\frac{\overline{F}(x+y)}{\overline{F}(x)} \quad 0 \leq x < w(t) - y, \tag{3}$$

$$b_i(t+y, x+y) = b_i(t,x)\frac{\overline{G}_i(x+y)}{\overline{G}_i(x)} \quad x \geq 0,\ i = 1, 2. \tag{4}$$

Eq. (3) simply says that the fluid quanta present in the waiting room at time t that have not abandoned for x time units and do not do so for y more time units remain there at time $t + y$. Similarly, Eq. (4) says that the fluid quanta in service i at time t that have not completed service for x time units and do not do so for y more time units remain in service i at time $t + y$ $(i = 1, 2)$.

We define the instantaneous ending rates, i.e., the hazard-rate functions, for the abandonment time and the service times by

$$h_F(x) := \frac{f(x)}{\overline{F}(x)}, \qquad h_{G_i}(x) := \frac{g_i(x)}{\overline{G}_i(x)} \quad x \geq 0,\ i = 1, 2.$$

Then the rate $\alpha(t)$ of fluid that abandons at time t and the output rate $\sigma_i(t)$ of fluid that service i is completed at time t are respectively given by

$$\alpha(t) = \int_0^\infty q(t,x)h_F(x)dx = \int_0^{w(t)} q(t,x)h_F(x)dx$$
$$0 \leq w(t) \leq t, \tag{5}$$

$$\sigma_i(t) = \int_0^\infty b_i(t,x)h_{G_i}(x)dx \quad t \geq 0,\ i = 1, 2. \tag{6}$$

2.3. Alternation of underloaded and overloaded intervals

We assume that the system state, started at time t_0, alternates between the underloaded and overloaded intervals. Let us denote by $\{t_{2n};\ n = 0, 1, 2, \ldots\}$ a sequence of epochs at which underloaded intervals are started, and denote by $\{t_{2n+1};\ n = 0, 1, 2, \ldots\}$ a sequence of epochs at which overloaded intervals are started such that

$[t_{2n}, t_{2n+1}]$: underloaded interval;
$[t_{2n+1}, t_{2n+2}]$: overloaded interval $\quad n = 0, 1, 2, \ldots$.

See Fig. 3 for the alternation of underloaded and overloaded intervals.

The system is said to be in the *underloaded* state if a fluid quantum that arrives enters service 1 immediately because there are more servers than the total fluid content in the system. Therefore, if the system is underloaded at time $t \in [t_{2n}, t_{2n+1}]$, we have

$$Q(t) = 0, \qquad s(t) > B_1(t) + B_2(t) \quad t \in [t_{2n}, t_{2n+1}]. \tag{7}$$

The underloaded interval $[t_{2n}, t_{2n+1}]$ ends when the total fluid content in services 1 and 2 becomes equal to the total number of servers for the first time after t_{2n}. Thus the termination epoch t_{2n+1} of the underloaded interval is determined by the condition

$$t_{2n+1} = \inf\{t \geq t_{2n} : s(t) = B_1(t) + B_2(t)\} \quad n = 0, 1, 2, \ldots. \tag{8}$$

The system is said to be in the *overloaded* state if a fluid quantum that arrives cannot enter service 1 because there are no servers available. Therefore, if the system is overloaded at time $t \in [t_{2n+1}, t_{2n+2}]$, we have

$$Q(t) > 0, \qquad s(t) = B_1(t) + B_2(t) \quad t \in [t_{2n+1}, t_{2n+2}]. \tag{9}$$

The overloaded interval $[t_{2n+1}, t_{2n+2}]$ ends when the fluid content in the waiting room vanishes. Thus the termination epoch t_{2n+2} of the overloaded interval is determined by the condition

$$t_{2n+2} = \inf\{t \geq t_{2n+1} : Q(t) = 0\} \quad n = 0, 1, 2, \ldots. \tag{10}$$

2.4. Fluid in service in the underloaded state

Let us first study the fluid content in service i $(i = 1, 2)$ at time t when the system is in the underloaded interval. For the simplicity of notation, we assume in this subsection that the underloaded interval of our concern is started at time 0 without loss of generality. For service 1, from Eq. (4) with $i = 1$, we can derive the following *transport* partial differential equation for $b_1(t, x)$ (Liu and Whitt [19, online version, Appendix B]):

$$\frac{\partial b_1(t,x)}{\partial t} + \frac{\partial b_1(t,x)}{\partial x} = -h_{G_1}(x)b_1(t,x) \tag{11}$$

with initial condition $b_1(0, x)$ and the boundary condition

$$b_1(t,0) = \lambda(t) \quad t \geq 0, \tag{12}$$

which is the rate of fluid going into service 1. Then the solution is given by (Liu and Whitt [19, Proposition 2]):

$$b_1(t,x) = \overline{G}_1(x)\lambda(t-x)1_{\{x \leq t\}} + \frac{\overline{G}_1(x)}{\overline{G}_1(x-t)}b_1(0, x-t)1_{\{x>t\}}, \tag{13}$$

where $1_{\mathcal{E}}$ is the *indicator function* for event $\mathcal{E}$ defined by

$$1_{\mathcal{E}} := \begin{cases} 1 & \text{if } \mathcal{E} \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

See Fig. 4 for the two cases $x \leq t$ and $x > t$ of contribution to $b_1(t, x)$. If $x \leq t$, the new arrival of fluid quanta with rate $\lambda(t-x)$ at time $t-x$ that stay longer than x time units contributes to $b_1(t, x)$. If

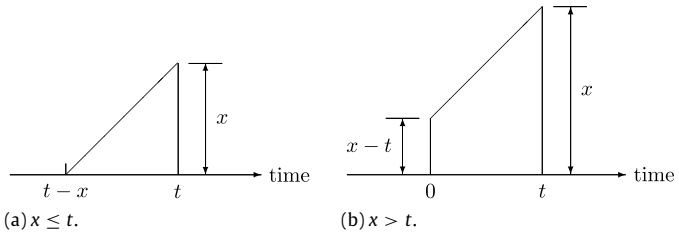


Fig. 4. Two cases for calculating $b_1(t, x)$ at time t in the underloaded state.

$x > t$, the fluid content $b_1(0, x-t)$ at time 0 that stays longer than x time units contributes to $b_1(t, x)$. From Eq. (13) we can derive

$$B_1(t) = \int_0^t \overline{G}_1(x)\lambda(t-x)dx + \int_0^\infty \frac{\overline{G}_1(x+t)}{\overline{G}_1(x)} b_1(0, x)dx. \tag{14}$$

From Eqs. (6) and (13), the output rate from service 1 is given by

$$\sigma_1(t) = \int_0^t g_1(x)\lambda(t-x)dx + \int_0^\infty \frac{g_1(x+t)}{\overline{G}_1(x)} b_1(0, x)dx.$$

For service 2, by similar manipulation of Eq. (4) with $i = 2$, we get the partial differential equation for $b_2(t, x)$:

$$\frac{\partial b_2(t,x)}{\partial t} + \frac{\partial b_2(t,x)}{\partial x} = -h_{G_2}(x) b_2(t, x) \tag{15}$$

with initial condition $b_2(0, x)$ and the boundary condition

$$b_2(t, 0) = \sigma_1(t) \quad t \geq 0. \tag{16}$$

Then we can derive

$$b_2(t,x) = \overline{G}_2(x)\sigma_1(t-x)1_{\{x\leq t\}} + \frac{\overline{G}_2(x)}{\overline{G}_2(x-t)} b_2(0, x-t)1_{\{x>t\}},$$

$$B_2(t) = \int_0^t \overline{G}_2(x)\sigma_1(t-x)dx + \int_0^\infty \frac{\overline{G}_2(x+t)}{\overline{G}_2(x)} b_2(0, x)dx,$$

$$\sigma_2(t) = \int_0^t g_2(x)\sigma_1(t-x)dx + \int_0^\infty \frac{g_2(x+t)}{\overline{G}_2(x)} b_2(0, x)dx. \tag{17}$$

During the underloaded interval, there are $s(t) - B_1(t) - B_2(t) > 0$ idle servers.

2.5. Fluid in service and in the waiting room in the overloaded state

We next consider the fluid content present in the system at time t when it is in the overloaded interval. Again we assume in this subsection that the overloaded interval of our concern is started at time 0 without loss of generality. From Eq. (4), the fluid content in service i ($i = 1, 2$) is governed by

$$b_1(t,x) = \overline{G}_1(x) b_1(t-x, 0)1_{\{x\leq t\}} + \frac{\overline{G}_1(x)}{\overline{G}_1(x-t)} b_1(0, x-t)1_{\{x>t\}}, \tag{18}$$

$$b_2(t,x) = \overline{G}_2(x) b_2(t-x, 0)1_{\{x\leq t\}} + \frac{\overline{G}_2(x)}{\overline{G}_2(x-t)} b_2(0, x-t)1_{\{x>t\}}. \tag{19}$$

In these equations, while $b_1(0, x)$ and $b_2(0, x)$ are given as initial conditions, we must know the boundary conditions $b_1(t, 0)$ and $b_2(t, 0)$. This can be done in principle by extending the method of Liu and Whitt [19, Section 6.2] for the solution to the fixed-point equation for our $M_t/GI/s_t + GI$ model as follows. Since the input rate to service 2 equals the output rate of service 1, we have

$$b_2(t, 0) = \sigma_1(t) = \int_0^t g_1(x) b_1(t-x, 0)dx + \int_0^\infty \frac{g_1(x+t)}{\overline{G}_1(x)} b_1(0, x)dx. \tag{20}$$

In the overloaded state, the fluid entrance rate into service 1 is the sum of the rate at which the number of servers is increased and the output rate of service 2. Thus we have

$$b_1(t, 0) = s'(t) + \sigma_2(t) = s'(t) + \int_0^t g_2(x) b_2(t-x, 0)dx + \int_0^\infty \frac{g_2(x+t)}{\overline{G}_2(x)} b_2(0, x)dx. \tag{21}$$

Eqs. (20) and (21) provide the set of simultaneous integral equations for the functions $b_1(t, 0)$ and $b_2(t, 0)$ given $s'(t)$ and the initial conditions $b_1(0, x)$ and $b_2(0, x)$.

During the overloaded interval, all servers are busy, each being engaged either in service 1 or in service 2, so that the fluid content in service i equals the number of servers for service i:

$$B_i(t) = s_i(t) \quad i = 1, 2; \qquad s(t) = s_1(t) + s_2(t). \tag{22}$$

There are no idle servers in the overloaded interval.

Let us derive the equations governing the number $s_i(t)$ of servers in service i ($i = 1, 2$). The fluid quantum whose service 1 is finished at time t enters service 2 immediately with rate $\sigma_1(t)$. At this moment, the number of servers in service 1 decreases with rate $\sigma_1(t)$ and the number of servers in service 2 increases with the same rate. The fluid quantum whose service 2 is finished at time t leaves the system with rate $\sigma_2(t)$. The server then turns to service 1 immediately. At this moment, the number of servers in service 2 decreases with rate $\sigma_2(t)$ and the number of servers in service 1 increases with the same rate. Since the time variation in the total number of servers is assumed to occur in service 1, we have the following set of ordinary differential equations for $\{s_1(t), s_2(t)\}$:

$$\frac{ds_1(t)}{dt} = -\sigma_1(t) + \sigma_2(t) + s'(t),$$

$$\frac{ds_2(t)}{dt} = \sigma_1(t) - \sigma_2(t). \tag{23}$$

Recall our assumption that the number of servers for service i never goes below the level of fluid content in service i ($i = 1, 2$). These conditions are expressed as

$$s_1(t) + s'(t)\Delta t + \sigma_2(t)\Delta t - \sigma_1(t)\Delta t \geq B_1(t) - \sigma_1(t)\Delta t$$

$$\text{and} \quad s_2(t) + \sigma_1(t)\Delta t - \sigma_2(t)\Delta t \geq B_2(t) - \sigma_2(t)\Delta t,$$

where the left-hand side of each equation shows the number of servers while the right-hand side shows the fluid content at time $t + \Delta t$. The second inequality is always satisfied as $\sigma_1(t) \geq 0$. From the first inequality, a sufficient condition for the *feasibility* is given by [19, online version, Appendix G.2]

$$b_1(t, 0) = s'(t) + \sigma_2(t) \geq 0. \tag{24}$$

Therefore, we impose the condition in Eq. (24) for the set of equations in Eq. (23).

We next consider the fluid content in the waiting room during the overloaded interval. From Eq. (3), we can derive the partial differential equation for $q(t, x)$:

$$\frac{\partial q(t,x)}{\partial t} + \frac{\partial q(t,x)}{\partial x} = -h_F(x) q(t, x) \quad 0 \leq x \leq w(t) \tag{25}$$

with initial condition $q(0, x)$ and the boundary condition

$$q(t, 0) = \lambda(t) \quad t \geq 0. \tag{26}$$

Then the density function of the fluid content in the waiting room is given by

$$q(t, x) = \overline{F}(x)\lambda(t - x)1_{\{x \leq \min\{w(t), t\}\}} + \frac{\overline{F}(x)}{\overline{F}(x - t)} q(0, x - t)1_{\{t < x \leq w(t)\}}. \tag{27}$$

This expression has the following meaning; see Fig. 4 for a similar situation in the underloaded state. For $x \leq t$, $q(t, x)dx = \overline{F}(x)\lambda(t - x)dx$ is the quantity of fluid quanta that arrive during $[t-x-dx, t-x]$ and do not abandon for x time units. For $x > t$, $q(t, x)dx = [\overline{F}(x)/\overline{F}(x - t)]q(0, x - t)dx$ is the fluid content in the waiting room at time t which comes from the fluid content $q(0, x-t)dx$ at time 0 that does not abandon for x time units. From Eq. (27) we can derive

$$Q(t) = \int_0^{\min\{w(t), t\}} \overline{F}(x)\lambda(t - x)dx + 1_{\{w(t) > t\}} \int_0^{w(t)-t} \frac{\overline{F}(x + t)}{\overline{F}(x)} q(0, x)dx.$$

2.6. Waiting time of a fluid quantum in the overloaded state

Let us now derive the differential equation for the boundary waiting time $w(t)$ by following Liu and Whitt [19, online version, Appendix D.1]. To do so, for $0 \leq w(t) \leq t$, we note

$$Q(t) = \int_0^{w(t)} \lambda(t - x)\overline{F}(x)dx = \int_{t-w(t)}^{t} \lambda(x)\overline{F}(t - x)dx. \tag{28}$$

Differentiating the rightmost side of this equation with respect to t, we get

$$\begin{aligned} \frac{dQ(t)}{dt} &= \lambda(t)\overline{F}(0) - \lambda[t - w(t)]\overline{F}[w(t)][1 - w'(t)] \\ &\quad - \int_{t-w(t)}^{t} \lambda(x)f(t - x)dx \\ &= \lambda(t) - q[t, w(t)][1 - w'(t)] - \alpha(t), \end{aligned} \tag{29}$$

where, from Eq. (5) we have used

$$\begin{aligned} \alpha(t) &= \int_0^{w(t)} q(t, x)h_F(x)dx = \int_0^{w(t)} \lambda(t - x)f(x)dx \\ &= \int_{t-w(t)}^{t} \lambda(x)f(t - x)dx. \end{aligned} \tag{30}$$

On the other hand, the fluid content in the waiting room varies as

$$\frac{dQ(t)}{dt} = \lambda(t) - \alpha(t) - b_1(t, 0). \tag{31}$$

Comparing Eqs. (29) and (31), we obtain the relation

$$b_1(t, 0) = q[t, w(t)][1 - w'(t)],$$

which leads to the first-order nonlinear differential equation for $w(t)$:

$$\frac{dw(t)}{dt} = 1 - \frac{b_1(t, 0)}{\lambda[t - w(t)]\overline{F}[w(t)]} \quad t \geq 0 \tag{32}$$

with initial condition $w(0) = 0$, where $b_1(t, 0)$ is given as the solution to the set of Eqs. (20) and (21).

In addition, Liu and Whitt [19, Section 7.3] consider the *potential waiting time* $v(t)$. This is defined as the virtual waiting time of an arriving fluid quantum at time t which elects never to abandon. Since the waiting time of the fluid quantum that is entering service 1 at time t is $w(t)$, then this quantum must have entered the waiting room $w(t)$ time units ago. This implies that the potential waiting time at $t - w(t)$ is $w(t)$. It follows that

$$v[t - w(t)] = w(t) \quad \text{or} \quad v(t) = w[t + v(t)]. \tag{33}$$

Then the differential equation for $v(t)$ is given by

$$\frac{dv(t)}{dt} = \frac{\lambda(t)\overline{F}[v(t)]}{b_1[t + v(t), 0]} - 1 \quad t \geq 0 \tag{34}$$

with initial condition $v(0) = 0$.

2.7. Stationary model

We show the results for the stationary fluid model of the current model, which is again an extension of the analysis by Whitt [16, Section 3]. In this case we assume that

$$\lambda(t) \equiv \lambda, \qquad s_i(t) \equiv s_i \quad t \geq 0, \ i = 1, 2$$

and that the distribution functions $F(x)$, $G_1(x)$, and $G_2(x)$ are given.

We first consider the case in which both services 1 and 2 are underloaded. In this case, for service 1 we have the density of fluid content given by

$$b_1(x) = \lambda\overline{G}_1(x) \quad x \geq 0,$$

in particular, $b_1(0) = \lambda$. Thus we get the total fluid content in service 1:

$$B_1 = \int_0^{\infty} b_1(x)dx = \lambda \int_0^{\infty} \overline{G}_1(x)dx = \lambda E[G_1].$$

Then the rate of fluid at which service 1 is completed equals the input rate:

$$\sigma_1 = \int_0^{\infty} b_1(x)h_{G_1}(x)dx = \lambda \int_0^{\infty} g_1(x)dx = \lambda.$$

Similarly, for service 2 we have the density of fluid content given by

$$b_2(x) = \sigma_1\overline{G}_2(x) \quad x \geq 0,$$

in particular, $b_2(0) = \sigma_1 = \lambda$. Thus we get the total fluid content in service 2:

$$B_2 = \int_0^{\infty} b_2(x)dx = \sigma_1 \int_0^{\infty} \overline{G}_2(x)dx = \sigma_1 E[G_2] = \lambda E[G_2].$$

Then the rate of fluid at which service 2 is completed also equals the input rate:

$$\sigma_2 = \int_0^{\infty} b_2(x)h_{G_2}(x)dx = \sigma_1 \int_0^{\infty} g_2(x)dx = \sigma_1 = \lambda.$$

This case occurs if $B_i < s_i$ $(i = 1, 2)$, or

$$\lambda < \min\left\{\frac{s_1}{E[G_1]}, \frac{s_2}{E[G_2]}\right\}.$$

We next consider the overloaded case for service 1. Due to the abandonment of fluid in the waiting room, there is certainly a stationary state in the overloaded case for service 1. There is no stationary state in the overloaded case for service 2.

In this case, the density of the fluid content in the waiting room is given by

$$q(x) = \lambda\overline{F}(x)1_{\{x \leq w\}} \quad x \geq 0,$$

in particular, $q(0) = \lambda$. We determine w later by Eq. (37). Then the total fluid content in the waiting room is given by

$$Q = \int_0^\infty q(x)dx = \lambda \int_0^w \overline{F}(x)dx.$$

The rate of fluid that abandons in the waiting room is given by

$$\alpha = \int_0^\infty q(x)h_F(x)dx = \lambda \int_0^w f(x)dx = \lambda F(w). \tag{35}$$

Considering the stationary case in Eq. (31), we get the balance relation

$$\lambda = \alpha + b_1(0). \tag{36}$$

We can find $b_1(0)$ as follows. Since $b_1(x) = b_1(0)\overline{G}_1(x)$ for $x \geq 0$, we get

$$B_1 = b_1(0) \int_0^\infty \overline{G}_1(x)dx = b_1(0)E[G_1].$$

It follows from the overloaded condition $B_1 = s_1$ that

$$b_1(0) = \frac{s_1}{E[G_1]}.$$

Substituting this result into Eqs. (35) and (36), we obtain the equation to solve for w:

$$\overline{F}(w) = \frac{s_1}{\lambda E[G_1]}. \tag{37}$$

The rate of fluid that service 1 is completed is given by

$$\sigma_1 = \int_0^\infty b_1(x)h_{G_1}(x)dx = b_1(0) \int_0^\infty g_1(x)dx = b_1(0) = \frac{s_1}{E[G_1]}.$$

Then the rate of abandonment is given by

$$\alpha = \lambda F(w) = \lambda \left[1 - \overline{F}(w)\right] = \lambda - \frac{s_1}{E[G_1]}.$$

Note that service 2 must be underloaded because there is no waiting room for service 2. In this case, we again have $b_2(x) = \sigma_1\overline{G}_2(x)$ for $x \geq 0$, which leads to

$$\sigma_2 = \sigma_1 = b_1(0); \qquad B_2 = \sigma_1 E[G_2] = b_1(0)E[G_2] = \frac{s_1 E[G_2]}{E[G_1]}.$$

This case occurs if $\alpha > 0$, or

$$\lambda > \frac{s_1}{E[G_1]}.$$

3. System with exponentially distributed conversation and ACW times

We study a special system with exponentially distributed conversation and ACW times for more explicit analysis. For such a system, we assume

$$G_i(x) = 1 - e^{-\mu_i x} \quad x \geq 0, \ i = 1, 2, \tag{38}$$

which means that the associated hazard-rate functions are constant: $h_{G_i}(x) = \mu_i$ $(i = 1, 2)$. Then the mean conversation time is $1/\mu_1$ and the mean ACW time is $1/\mu_2$. The solution in this case is straightforward as we show below for the underloaded interval $[t_{2n}, t_{2n+1}]$ and the overloaded interval $[t_{2n+1}, t_{2n+2}]$, $n = 0, 1, 2, \ldots$, defined in Section 2.3.

3.1. Solution for the underloaded state

Let us consider the underloaded state at time $t \in [t_{2n}, t_{2n+1}]$. From Eq. (13), we have

$$\begin{aligned} b_1(t,x) &= \begin{cases} \overline{G}_1(x)\lambda(t-x) & 0 \leq x \leq t - t_{2n}, \\ \dfrac{\overline{G}_1(x)}{\overline{G}_1(x - t + t_{2n})} b_1(t_{2n}, x - t + t_{2n}) & x > t - t_{2n} \end{cases} \\ &= \begin{cases} e^{-\mu_1 x}\lambda(t-x) & 0 \leq x \leq t - t_{2n}, \\ e^{-\mu_1(t-t_{2n})} b_1(t_{2n}, x - t + t_{2n}) & x > t - t_{2n}. \end{cases} \end{aligned}$$

In particular, we can confirm Eq. (12). Therefore, or from Eq. (14), we obtain the fluid content $B_1(t)$ in service 1 at time t as

$$\begin{aligned} B_1(t) &= \int_0^{t-t_{2n}} \overline{G}_1(x)\lambda(t-x)dx \\ &\quad + \int_0^\infty \frac{\overline{G}_1(x + t - t_{2n})}{\overline{G}_1(x)} b_1(t_{2n}, x)dx \\ &= \int_0^{t-t_{2n}} \lambda(t-x)e^{-\mu_1 x}dx + e^{-\mu_1(t-t_{2n})}B_1(t_{2n}). \end{aligned} \tag{39}$$

The output rate of service 1 is simply given by

$$\begin{aligned} \sigma_1(t) &= \int_0^\infty b_1(t,x)h_{G_1}(x)dx = \mu_1 \int_0^\infty b_1(t,x)dx \\ &= \mu_1 B_1(t). \end{aligned} \tag{40}$$

Similarly, from Eq. (17), we have

$$\begin{aligned} b_2(t,x) &= \begin{cases} \overline{G}_2(x)\sigma_1(t-x) & 0 \leq x \leq t - t_{2n}, \\ \dfrac{\overline{G}_2(x)}{\overline{G}_2(x - t + t_{2n})} b_2(t_{2n}, x - t + t_{2n}) & x > t - t_{2n} \end{cases} \\ &= \begin{cases} \mu_1 e^{-\mu_2 x}B_1(t-x) & 0 \leq x \leq t - t_{2n}, \\ e^{-\mu_2(t-t_{2n})} b_2(t_{2n}, x - t + t_{2n}) & x > t - t_{2n}. \end{cases} \end{aligned}$$

Thus we obtain the fluid content $B_2(t)$ in service 2 at time t as

$$\begin{aligned} B_2(t) &= \int_0^{t-t_{2n}} \sigma_1(t-x)e^{-\mu_2 x}dx \\ &\quad + e^{-\mu_2(t-t_{2n})} \int_{t-t_{2n}}^\infty b_2(t_{2n}, x - t + t_{2n})dx \\ &= \mu_1 \int_0^{t-t_{2n}} B_1(t-x)e^{-\mu_2 x}dx + e^{-\mu_2(t-t_{2n})}B_2(t_{2n}). \end{aligned} \tag{41}$$

The output rate of service 2 is given by

$$\sigma_2(t) = \mu_2 B_2(t). \tag{42}$$

We note a great merit of exponentially distributed conversation and ACW times that we can obtain $B_i(t)$ in Eqs. (39) and (41) and then $\sigma_i(t)$ in Eqs. (40) and (42) without finding $b_i(t,x)$, $i = 1, 2$.

The initial values $B_i(t_{2n})$, $i = 1, 2$, for the underloaded interval $[t_{2n}, t_{2n+1}]$ are given from the solution for the preceding overloaded interval $[t_{2n-1}, t_{2n}]$ as

$$\begin{aligned} B_1(t_{2n}) &= s(t_{2n}) - B_2(t_{2n}), \\ B_2(t_{2n}) &= e^{-(\mu_1+\mu_2)(t_{2n}-t_{2n-1})}B_2(t_{2n-1}) \\ &\quad + \mu_1 e^{-(\mu_1+\mu_2)t_{2n}} \int_{t_{2n-1}}^{t_{2n}} e^{(\mu_1+\mu_2)u}s(u)du \end{aligned}$$

from Eq. (47) in the sequel. The termination epoch t_{2n+1} of the underloaded interval $[t_{2n}, t_{2n+1}]$ is found by the condition in Eq. (8).

3.2. Solution for the overloaded state

Let us consider the overloaded state at time $t \in [t_{2n+1}, t_{2n+2}]$. The output rate of fluid at which service i is completed at time t is obtained from Eqs. (6) and (22) as

$$\sigma_i(t) = \int_0^\infty b_i(t,x)h_{G_i}(x)dx = \mu_i \int_0^\infty b_i(t,x)dx = \mu_i B_i(t) = \mu_i s_i(t) \quad i = 1, 2. \tag{43}$$

Substituting Eq. (43) into Eq. (23), we have the following set of simultaneous linear differential equations for $\{B_1(t), B_2(t)\}$:

$$\frac{dB_1(t)}{dt} = -\mu_1 B_1(t) + \mu_2 B_2(t) + s'(t), \tag{44}$$

$$\frac{dB_2(t)}{dt} = \mu_1 B_1(t) - \mu_2 B_2(t). \tag{45}$$

Using $s(t) = B_1(t) + B_2(t)$, we get the first-order differential equation for $B_2(t)$:

$$\frac{dB_2(t)}{dt} = \mu_1 s(t) - (\mu_1 + \mu_2)B_2(t) \quad t_{2n+1} \le t \le t_{2n+2}, \tag{46}$$

whose solution is given by

$$B_2(t) = e^{-(\mu_1+\mu_2)(t-t_{2n+1})}B_2(t_{2n+1}) + \mu_1 e^{-(\mu_1+\mu_2)t}\int_{t_{2n+1}}^t e^{(\mu_1+\mu_2)u}s(u)du, \tag{47}$$

with the initial condition $B_2(t_{2n+1})$ from the fluid content $B_2(t)$ at the end of the preceding underloaded interval $[t_{2n}, t_{2n+1}]$. We then get $B_1(t) = s(t) - B_2(t)$.

In the differential equation (32) for $w(t)$, we have noted that $b_1(t, 0)$ should be given as the solution to the fixed-point equations (20) and (21) in general. However, it can be obtained easily in the present case in which both conversation and ACW times are exponentially distributed. Since service 1 is started by the servers who have completed service 2 as well as by the servers who are added to the system, we have the relation

$$b_1(t, 0) = \mu_2 B_2(t) + s'(t), \tag{48}$$

which is assumed to be nonnegative from the feasibility condition in Eq. (24). Substituting Eq. (48) into Eq. (32), we get the differential equation for $w(t)$:

$$\frac{dw(t)}{dt} = 1 - \frac{\mu_2 B_2(t) + s'(t)}{\lambda[t - w(t)]\bar{F}[w(t)]} \quad t_{2n+1} \le t \le t_{2n+2} \tag{49}$$

with initial condition $w(t_{2n+1}) = 0$. Substituting Eq. (48) into Eq. (34), we get the differential equation for $v(t)$:

$$\frac{dv(t)}{dt} = \frac{\lambda(t)\bar{F}[v(t)]}{\mu_2 B_2[t + v(t)] + s'[t + v(t)]} - 1 \quad t_{2n+1} \le t \le t_{2n+2} \tag{50}$$

with initial condition $v(t_{2n+1}) = 0$, which is consistent with $w(t_{2n+1}) = 0$.

Once $w(t)$ is obtained, we get the fluid content $Q(t)$ in the waiting room at time t by Eq. (28). The abandonment rate $\alpha(t)$ is calculated by Eq. (30). The termination epoch t_{2n+2} of the overloaded interval $[t_{2n+1}, t_{2n+2}]$ is determined by the condition in Eq. (10), or equivalently

$$t_{2n+2} = \inf\{t \ge t_{2n+1} : w(t) = 0\}. \tag{51}$$

4. Numerical examples for systems with exponentially distributed abandonment time

The nonlinear differential equations (49) for $w(t)$ and (50) for $v(t)$ are to be solved numerically when the function $\bar{F}(t)$ is given. Therefore, a simple form for $\bar{F}(t)$ does not help much for analytical solution. Nevertheless, we show some numerical examples for systems with the exponentially distributed conversation and ACW times and the exponentially distributed abandonment time. Thus we assume Eq. (38) for the service times along with

$$F(x) = 1 - e^{-\theta x}, \qquad f(x) = \theta e^{-\theta x} \quad x \ge 0,$$

where θ is the constant hazard rate of the abandonment time, and $1/\theta$ is the mean abandonment time. Green et al. [18] mention that it often suffices to work with an exponentially distributed time-to-abandon approximation as it was empirically justified.

In this case, the rate $\alpha(t)$ of fluid that abandons at time t is proportional to the amount $Q(t)$ of the fluid in the waiting room during overloaded intervals. Indeed, from Eqs. (28) and (30), we have

$$\alpha(t) = \int_0^{w(t)} \lambda(t-x)f(x)dx = \theta\int_0^{w(t)} \lambda(t-x)e^{-\theta x}dx = \theta\int_0^{w(t)} \lambda(t-x)\bar{F}(x)dx = \theta Q(t),$$

which is reasonable as θ is the rate at which each quantum of fluid in the waiting room abandons.

In all the numerical examples in this section, we specifically assume that

$$G_1(x) = 1 - e^{-5x/4}, \qquad G_2(x) = 1 - e^{-5x},$$
$$F(x) = 1 - e^{-2x} \quad x \ge 0$$

which means that $\mu_1 = \frac{5}{4}$, $\mu_2 = 5$ (meaning that the total mean service time is $1/\mu_1 + 1/\mu_2 = 1$), and $\theta = 2$ in the above formulation. If the unit of time is 1 h, the mean conversation time is 48 min, the mean ACW time is 12 min, and the mean abandonment time is 30 min.

4.1. Sinusoidal input rate and constant number of servers

In the first example, we consider a system in which the input rate changes as a sinusoidal function of time t while the total number of servers is kept constant:

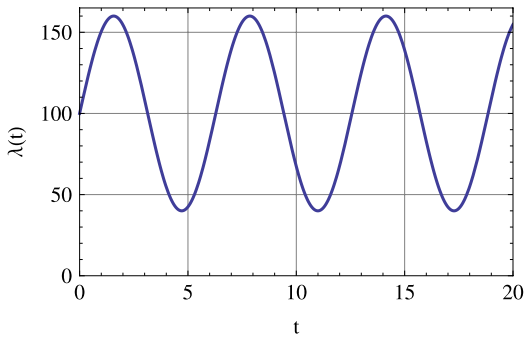
$$\lambda(t) = 100[1 + 0.6\sin(t)], \qquad s(t) = 100 \quad t \ge 0. \tag{52}$$

Green et al. [18] mention that the dynamic behavior of the demand function (for call centers) is reasonably characterized by a sinusoidal function. The input rate function $\lambda(t)$ in Eq. (52) is used as a base example by Liu and Whitt [19, Section 2]. According to them, by making the mean input rate coincide with the fixed number of servers, we ensure that the system will alternate between underloaded and overloaded states.

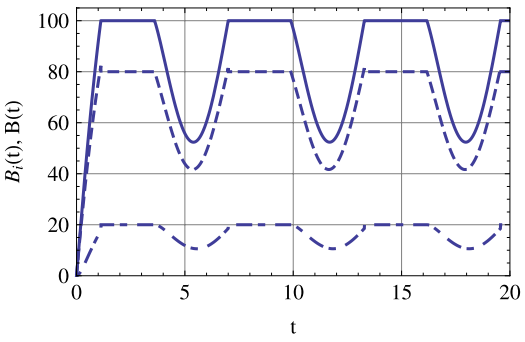
The result of numerical analysis is shown in Table 1 and Fig. 5. Table 1 shows a sequence of epochs at which the state changes from underloaded to overloaded and vice versa. Fig. 5(a) shows the sinusoidal arrival rate $\lambda(t)$ given above. Fig. 5(b) plots the fluid content $B_i(t)$ in service i $(i = 1, 2)$ (dashed line for $i = 1$ and dashed-and-dotted line for $i = 2$) as well as $B(t) = B_1(t) + B_2(t)$ (solid line) at time t. We observe that $B_1(t) = 80$ and that $B_2(t) = 20$ if the system is overloaded at time t. The ratio $B_1(t)/B_2(t) = 4$ coincides with the ratio of mean service times $(1/\mu_1)/(1/\mu_2) = 4$. We also plot the fluid content $Q(t)$ in the waiting room and the total fluid content in the system $B(t) + Q(t)$ with superposed input rate $\lambda(t)$ in Figs. 5(c) and (d), respectively. We observe that $B(t) + Q(t)$ lags in time behind $\lambda(t)$. The boundary and potential waiting times are shown in Fig. 5(e). The fluid entrance rate $b_1(t, 0)$ into service 1 is shown in Fig. 5(f).

Table 1
Epochs of state change in the system with sinusoidal input rate $\lambda(t) = 100[1 + 0.6\sin(t)]$ and constant number of servers $s(t) = 100$.

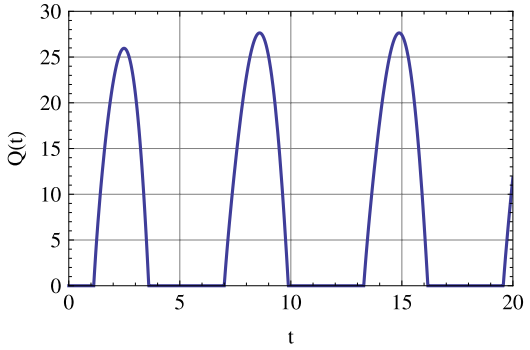
t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
0	1.12283	3.60165	6.99994	9.88785	13.28309	16.17103	19.56628	22.45422



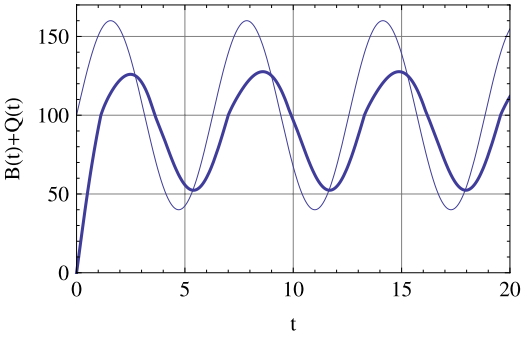
(a) Input rate $\lambda(t)$.



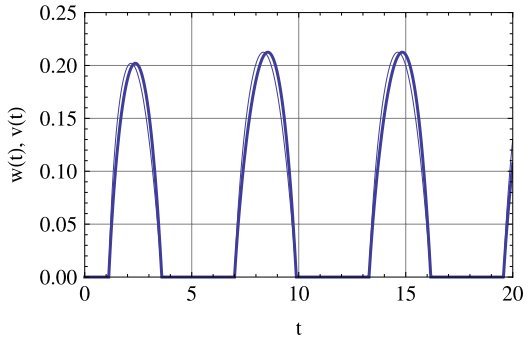
(b) Fluid in service $B_i(t)$ and $B(t)$.



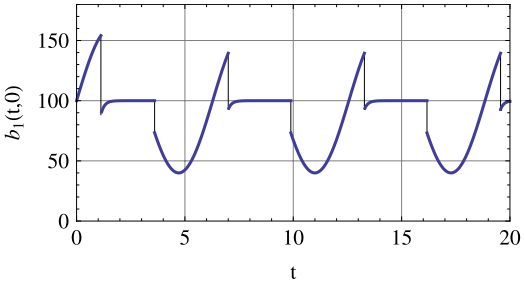
(c) Waiting fluid $Q(t)$.



(d) Total fluid $B(t) + Q(t)$.



(e) Boundary waiting time $w(t)$ (thick) and potential waiting time $v(t)$ (thin).



(f) Fluid entrance rate into service $b_1(t, 0)$.

Fig. 5. Performance of the system with sinusoidal input rate $\lambda(t) = 100[1 + 0.6\sin(t)]$ and constant number of servers $s(t) = 100$.

4.2. Constant input rate and sinusoidal number of servers

In the second example, we consider a system in which the input rate is constant while the total number of servers changes in time:

$$\lambda(t) = 100, \qquad s(t) = 100[1 + 0.6\sin(t)] \quad t \geq 0,$$

where the non-integer number of servers is thought of as approximation.

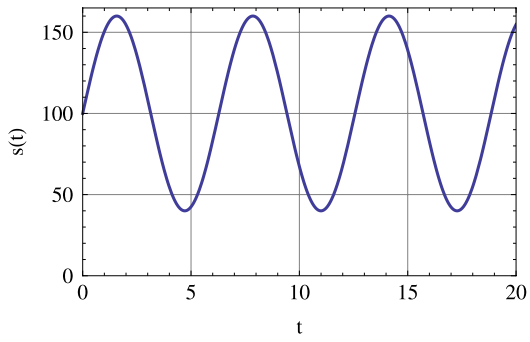
The result of numerical analysis is shown in Table 2 and Fig. 6. Table 2 shows a sequence of epochs at which the state changes from underloaded to overloaded and vice versa. Fig. 6(a) shows the sinusoidal number of servers $s(t)$ given above. Fig. 6(b) plots the fluid content $B_i(t)$ in service i $(i = 1, 2)$ as well as $B(t) = B_1(t) + B_2(t)$ at time t with superposed $s(t)$. We observe that $B(t) = s(t)$ if the system is overloaded at time t. We also plot the fluid content $Q(t)$ in the waiting room and the total fluid content $B(t) + Q(t)$ in Figs. 6(c) and (d), respectively. The boundary and potential waiting times are shown in Fig. 6(e). The fluid entrance rate $b_1(t, 0)$ into service 1 is shown in Fig. 6(f), which is always positive so that the sufficient condition for feasibility in Eq. (24) is satisfied.

4.3. Identical sinusoidal input rate and number of servers

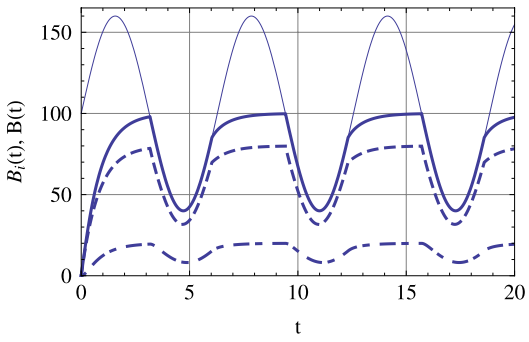
In the third example, we consider a system in which the input rate and the total number of servers change as identical sinusoidal

Table 2
Epochs of state change in the system with constant input rate $\lambda(t) = 100$ and sinusoidal number of servers $s(t) = 100[1 + 0.6\sin(t)]$.

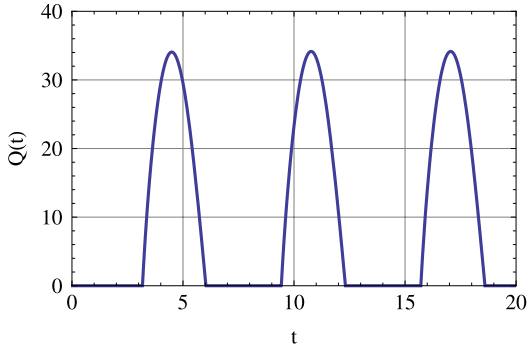
t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7
0	3.17519	6.03265	9.42802	12.31596	15.71120	18.59914	21.99439



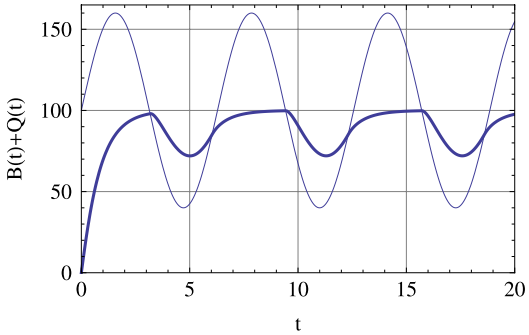
(a) Number of servers $s(t)$.



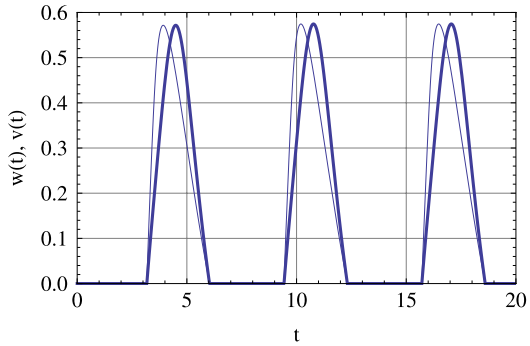
(b) Fluid in service $B_i(t)$ and $B(t)$.



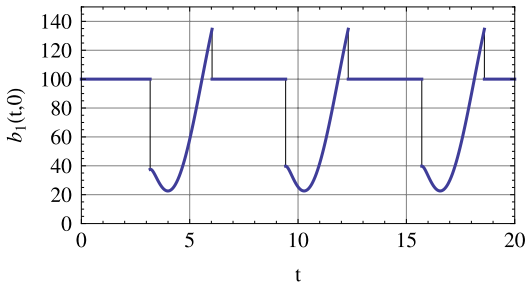
(c) Waiting fluid $Q(t)$.



(d) Total fluid $B(t) + Q(t)$.



(e) Boundary waiting time $w(t)$ (thick) and potential waiting time $v(t)$ (thin).



(f) Fluid entrance rate into service $b_1(t, 0)$.

Fig. 6. Performance of the system with constant input rate $\lambda(t) = 100$ and sinusoidal number of servers $s(t) = 100[1 + 0.6\sin(t)]$.

functions in time:

$$s(t) = \left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right)\lambda(t) = 100[1 + 0.6\sin(t)] \quad t \geq 0.$$

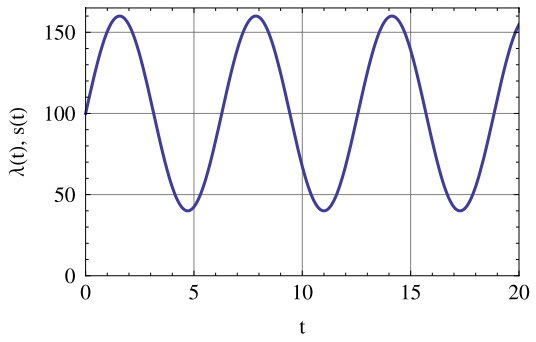
One might expect good performance because the exactly necessary and sufficient number of servers are provided with the input rate at each instant. However, this is not the case as the result of numerical analysis is shown in Table 3 and Fig. 7. Table 3 shows a sequence of epochs at which the state changes from underloaded to overloaded and vice versa. Fig. 7(a) shows $\lambda(t) = s(t)$ given above. Fig. 7(b) plots the fluid content $B_i(t)$ in service i ($i = 1, 2$) as well as $B(t) = B_1(t) + B_2(t)$ at time t. Fig. 7(c) displays the fluid content $Q(t)$ in the waiting room, which is rather significant in the overloaded state. The reason is that there is a time delay because each arrival of the fluid quantum remains in the system for the duration of its service time. Therefore, the fluid content in the system lags in time behind the input rate.

4.4. Partially adaptive number of servers

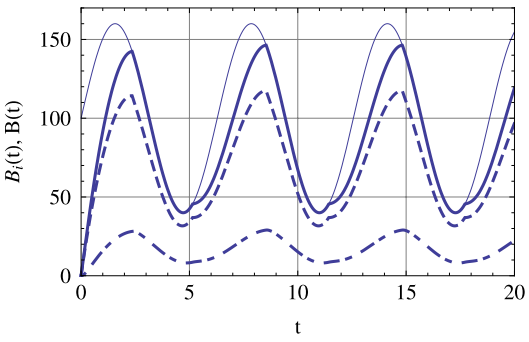
A method of incorporating the time lag in the fluid content behind the input was suggested by Eick et al. [24,25] in terms of the infinite-server model in order to determine the proper number of servers adaptively depending on the past input rate. In the $M_t/GI/\infty$ system with a time-varying Poisson arrival process with

Table 3
Epochs of state change in the system with identical sinusoidal arrival process and number of servers $\lambda(t) = s(t) = 100[1 + 0.6\sin(t)]$.

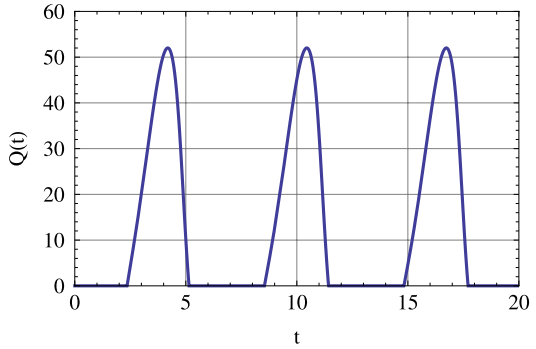
t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
0	2.35592	5.14545	8.54114	11.42908	14.82432	17.71226	21.10751	23.98545



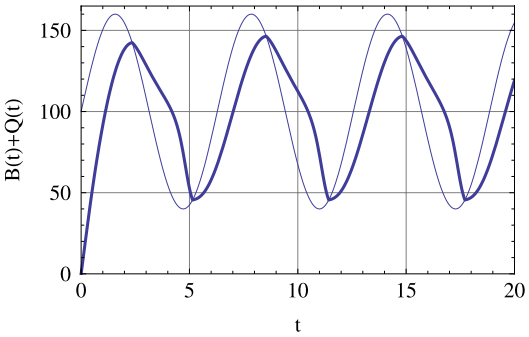
(a) Identical $\lambda(t)$ and $s(t)$.



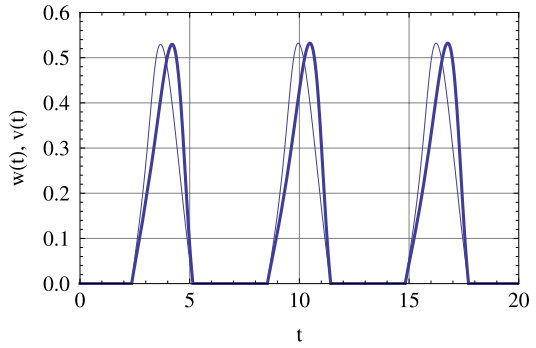
(b) Fluid in service $B_i(t)$ and $B(t)$.



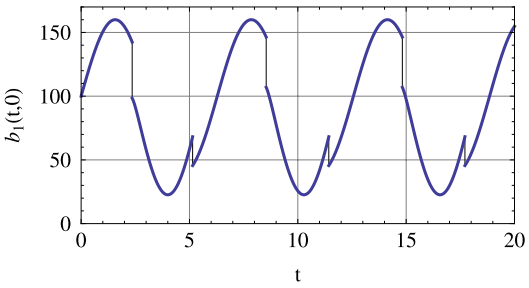
(c) Waiting fluid $Q(t)$.



(d) Total fluid $B(t) + Q(t)$.



(e) Boundary waiting time $w(t)$ (thick) and potential waiting time $v(t)$ (thin).



(f) Fluid entrance rate into service $b_1(t, 0)$.

Fig. 7. Performance of the system with identical sinusoidal input rate and number of servers $\lambda(t) = s(t) = 100[1 + 0.6\sin(t)]$.

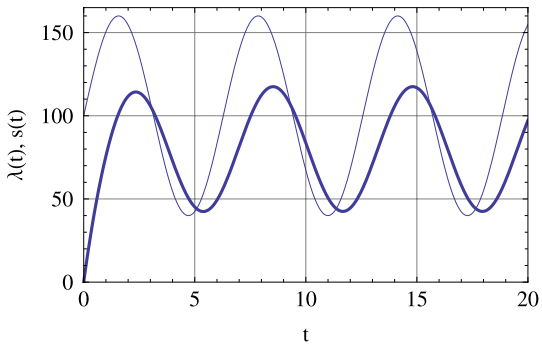
rate $\lambda(t)$, the number of customers present in the system (= the number of busy servers) at time t has a Poisson distribution with mean

$$\int_0^t \overline{G}(x)\lambda(t-x)dx,$$

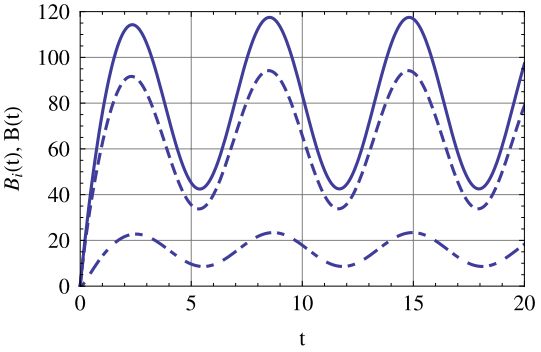
where $\overline{G}(x)$ is the CDF of the service time. The integrand accounts for those arrivals during $[t - x, t - x + dx]$ that have service times longer than x time units collected over the interval $[0, t]$ of x. This form also appears in the transient version of Little's law for non-stationary queueing systems by Bertsimas and Mourtzinou [26].

In our fourth example, we examine the performance of the system in which the number of servers is determined adaptively to cope with the load of service 1 only. As before, let us assume the sinusoidal input rate $\lambda(t)$ given in Eq. (52). We then provide the following number of servers:

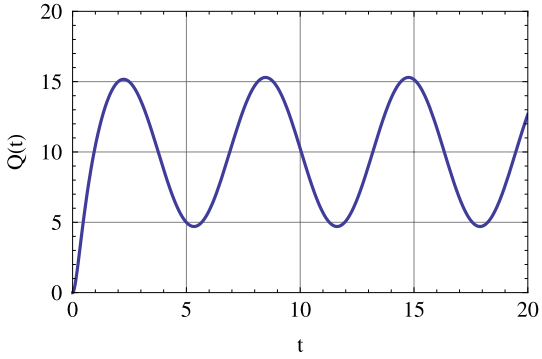
$$\begin{aligned} s(t) &= \int_0^t \overline{G}_1(x)\lambda(t-x)dx = \int_0^t \lambda(t-x)e^{-\mu_1 x}dx \\ &= 80 - \frac{80}{41}\left(29e^{-\frac{5}{4}t} + 12\cos t - 15\sin t\right) \\ &= \frac{\lambda(t)}{\mu_1} - \frac{16}{41}\left(145e^{-\frac{5}{4}t} + 60\cos t + 48\sin t\right). \end{aligned}$$



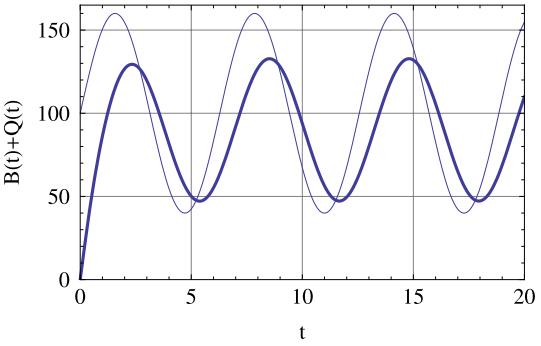
(a) $\lambda(t)$ and partially adaptive $s(t)$.



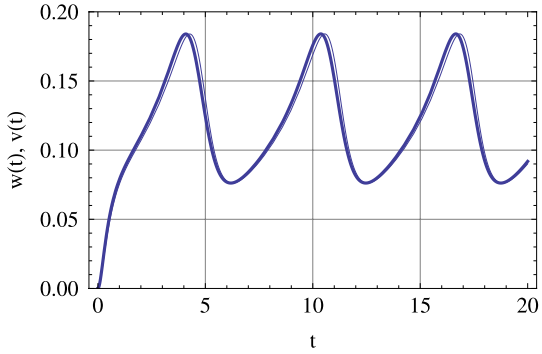
(b) Fluid in service $B_i(t)$ and $B(t)$.



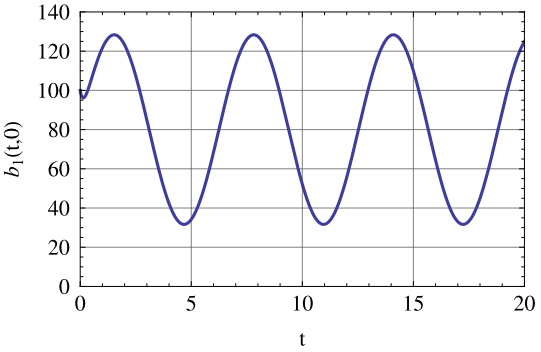
(c) Waiting fluid $Q(t)$.



(d) Total fluid $B(t) + Q(t)$.



(e) Boundary waiting time $w(t)$ (thick) and potential waiting time $v(t)$ (thin).



(f) Fluid entrance rate into service $b_1(t, 0)$.

Fig. 8. Performance of the system with sinusoidal input rate $\lambda(t) = 100[1 + 0.6\sin(t)]$ and partially adaptive number of servers.

In this case, the system is always overloaded. Thus we obtain

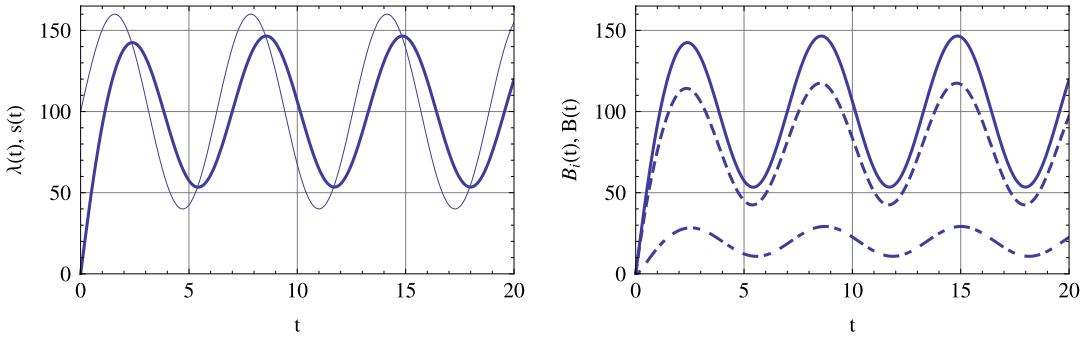
$$\begin{aligned} B_2(t) &= s_2(t) = \mu_1 e^{-(\mu_1+\mu_2)t} \int_0^t e^{(\mu_1+\mu_2)u} s(u)du \\ &= 16 + \frac{4}{26281}\left(23821e^{-\frac{25}{4}t} - 92945e^{-\frac{5}{4}t}\right. \\ &\quad \left. - 36000\cos t + 32700\sin t\right), \\ B_1(t) &= s_1(t) = s(t) - s_2(t) \\ &= 64 - \frac{4}{26281}\left(23821e^{-\frac{25}{4}t} + 278835e^{-\frac{5}{4}t}\right. \\ &\quad \left. + 117840\cos t - 159600\sin t\right), \\ b_1(t,0) &= \mu_2 s_2(t) + s'(t) \\ &= 80 + \frac{20}{641}\left(581e^{-\frac{25}{4}t} + 60\cos t + 1548\sin t\right). \end{aligned}$$

The result of this analysis is shown in Fig. 8. Fig. 8(a) shows that $s(t)$ lags behind $\lambda(t)$ as intended. Fig. 8(b) plots the fluid content $B_i(t)$ in service i $(i = 1, 2)$ as well as $B(t) = B_1(t) + B_2(t)$, where $B_i(t) = s_i(t)$ $(i = 1, 2)$ and $B(t) = s(t)$, at time t because the system is overloaded. Fig. 8(c) plots the fluid content $Q(t)$ in the waiting room which never vanishes but it does not grow to infinity. Therefore the system is stable, which occurs because more customers abandon as more customers wait. The peak values of $Q(t)$ in Fig. 8(c) and those of waiting times $w(t)$ and $v(t)$ in Fig. 8(e) are much less than the corresponding values in Figs. 7(c) and (e), respectively, for the system with non-adaptive number of servers.

4.5. Perfectly adaptive number of servers

In the fifth example, we examine the performance of the system in which the number of servers is determined adaptively to cope with the load of both services 1 and 2. The pdf of service i is given by

$$g_i(x) = \mu_i e^{-\mu_i x} \quad x \geq 0,\ i = 1, 2.$$



(a) $\lambda(t)$ and perfectly adaptive $s(t)$. (b) Fluid in service $B_i(t)$ and $B(t)$.

Fig. 9. Performance of the system with sinusoidal input rate $\lambda(t) = 100[1 + 0.6\sin(t)]$ and the perfectly adaptive number of servers.

If these services are assumed to be independent of each other and $\mu_1 \neq \mu_2$, their sum has the density function given by the convolution

$$g(x) = \int_0^x g_1(t)g_2(x-t)dt = \frac{\mu_1\mu_2}{\mu_1-\mu_2}\left(e^{-\mu_2 x} - e^{-\mu_1 x}\right) \quad x \geq 0$$

and the corresponding CDF given by

$$\overline{G}(x) = \int_x^\infty g(t)dt = \frac{\mu_1 e^{-\mu_2 x} - \mu_2 e^{-\mu_1 x}}{\mu_1-\mu_2} \quad x \geq 0.$$

For the sinusoidal input rate in Eq. (52), we provide the following number of servers:

$$\begin{aligned} s(t) &= \int_0^t \overline{G}(x)\lambda(t-x)dx \\ &= 100 + \frac{10}{1599}\left(943e^{-5t} - 12064e^{-\frac{5}{4}t}\right. \\ &\quad \left. - 4869\cos t + 5625\sin t\right) \\ &= \left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right)\lambda(t) - \frac{10}{1599}\left(-943e^{-5t}\right. \\ &\quad \left. + 12064e^{-\frac{5}{4}t} + 4869\cos t + 3969\sin t\right). \end{aligned}$$

This is exactly the number of servers which makes it possible to accept all input without queueing. Therefore, the system is always underloaded. Thus we have

$$\begin{aligned} B_1(t) &= \int_0^t \lambda(t-x)e^{-\mu_1 x}dx = \int_0^t \lambda(t-x)e^{-\frac{5}{4}x}dx \\ &= 80 - \frac{80}{41}\left(29e^{-\frac{5}{4}t} + 12\cos t - 15\sin t\right), \\ B_2(t) &= \mu_1\int_0^t B_1(t-x)e^{-\mu_2 x}dx = \frac{5}{4}\int_0^t B_1(t-x)e^{-5x}dx \\ &= 20 + \frac{10}{1599}\left(943e^{-5t} - 3016e^{-\frac{5}{4}t}\right. \\ &\quad \left. - 1125\cos t + 945\sin t\right), \end{aligned}$$

which leads to

$$B_1(t) + B_2(t) = s(t).$$

The result of this analysis is shown in Fig. 9. Fig. 9(a) again shows that $s(t)$ lags behind $\lambda(t)$. Comparing with Fig. 8(a), we observe that many more servers are needed in order to achieve the underloaded system all the time. Fig. 9(b) plots the fluid content $B_i(t)$ in service i $(i = 1, 2)$ as well as $B(t) = B_1(t) + B_2(t)$. There is more fluid in service in the present system than in the partially adaptive system shown in Fig. 8(b).

4.6. Constant input rate and constant number of servers

Finally we consider the stationary model of Section 2.7 with

$$\lambda(t) = \lambda, \qquad s_i(t) = s_i \quad t \geq 0,\ i = 1, 2,$$

where we assume that

$$\mu_1 s_1 \leq \mu_2 s_2.$$

Then the system is underloaded if $\lambda \leq \mu_1 s_1$, and it is overloaded if $\lambda > \mu_1 s_1$.

We then have the following stationary performance:

$$B_1 = \begin{cases} \lambda/\mu_1 & \lambda \leq \mu_1 s_1, \\ s_1 & \lambda > \mu_1 s_1 \end{cases}; \qquad B_2 = \begin{cases} \lambda/\mu_2 & \lambda \leq \mu_1 s_1, \\ (\mu_1 s_1)/\mu_2 & \lambda > \mu_1 s_1 \end{cases}$$

$$\alpha = \theta Q = \begin{cases} 0 & \lambda \leq \mu_1 s_1, \\ \lambda - \mu_1 s_1 & \lambda > \mu_1 s_1 \end{cases};$$

$$b_1(0) = \sigma_1 = \sigma_2 = \begin{cases} \lambda & \lambda \leq \mu_1 s_1, \\ \mu_1 s_1 & \lambda > \mu_1 s_1 \end{cases}$$

$$w = \begin{cases} 0 & \lambda \leq \mu_1 s_1, \\ \dfrac{1}{\theta}\log\dfrac{\lambda}{\mu_1 s_1} & \lambda > \mu_1 s_1. \end{cases}$$

5. System with non-exponentially distributed conversation and ACW times

As an example of applying our method to the $M_t/GI/s_t$ + GI model of a call center with non-exponentially distributed conversation and ACW times, let us consider a system in which the duration of service 1 for conversation has a hyperexponential distribution (denoted by H_2) and the duration of ACW has a hypo-exponential distribution (denoted by E_2). Then this system may be denoted by $M_t/(H_2+E_2)/s_t+GI$, where the meanings of "M_t", "s_t", and "+GI" are given in Section 1.

We assume that the conversation time has a two-phase hyperexponential distribution whose pdf and CDF are given by

$$g_1(x) = pg_{11}(x) + qg_{12}(x); \qquad \overline{G}_1(x) = p\overline{G}_{11}(x) + q\overline{G}_{12}(x),$$
$$x \geq 0,\ p + q = 1,$$

where

$$g_{1j}(x) = \mu_{1j}e^{-\mu_{1j}x}; \qquad \overline{G}_{1j}(x) = e^{-\mu_{1j}x} \quad x \geq 0,\ j = 1, 2$$

and

$$E[G_1] = \frac{p}{\mu_{11}} + \frac{q}{\mu_{12}}; \qquad E[\{G_1\}^2] = \frac{2p}{\mu_{11}^2} + \frac{2q}{\mu_{12}^2};$$

$$C_{G_1}^2 = \frac{\mathrm{Var}[G_1]}{(E[G_1])^2} \geq 1.$$

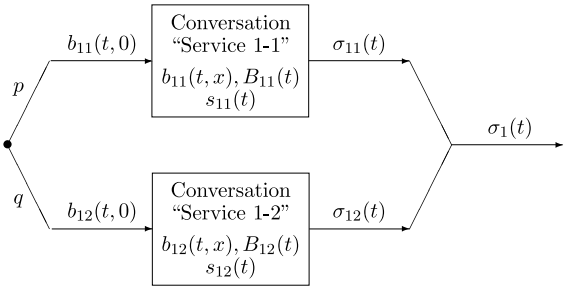


Fig. 10. Hyperexponentially distributed conversation time.

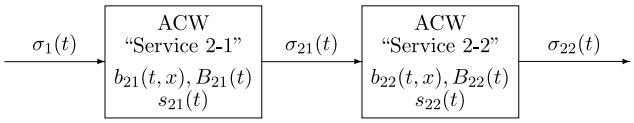


Fig. 11. Hypo-exponentially distributed ACW time.

We assume that the ACW time has a two-phase hypo-exponential distribution whose pdf is given by

$$g_2(x) = \int_0^x g_{21}(t)g_{22}(x-t)dt \quad x \geq 0,$$

where

$$g_{2j}(x) = \mu_{2j}e^{-\mu_{2j}x}; \qquad \overline{G}_{2j}(x) = e^{-\mu_{2j}x} \quad x \geq 0,\ j = 1, 2$$

and

$$E[G_2] = \frac{1}{\mu_{21}} + \frac{1}{\mu_{22}}; \qquad \mathrm{Var}[G_2] = \frac{1}{\mu_{21}^2} + \frac{1}{\mu_{22}^2};$$

$$C_{G_2}^2 = \frac{\mathrm{Var}[G_2]}{(E[G_2])^2} \leq 1.$$

5.1. Solution for the underloaded state

The duration of the conversation time has either pdf $g_{11}(x)$ with probability p (phase 1) or pdf $g_{12}(x)$ with probability q (phase 2) as shown in Fig. 10. Therefore, during the underloaded interval $t \in [t_{2n}, t_{2n+1}]$, the density for the fluid content in each phase satisfies the equation

$$b_{11}(t,x) = \begin{cases} pe^{-\mu_{11}x}\lambda(t-x) & 0 \leq x \leq t - t_{2n}, \\ e^{-\mu_{11}(t-t_{2n})}b_{11}(t_{2n}, x-t+t_{2n}) & x > t - t_{2n} \end{cases}$$

and

$$b_{12}(t,x) = \begin{cases} qe^{-\mu_{12}x}\lambda(t-x) & 0 \leq x \leq t - t_{2n}, \\ e^{-\mu_{12}(t-t_{2n})}b_{12}(t_{2n}, x-t+t_{2n}) & x > t - t_{2n}. \end{cases}$$

In particular, we have

$$b_{11}(t,0) = p\lambda(t); \qquad b_{12}(t,0) = q\lambda(t).$$

It follows that the fluid content in each phase is given by

$$B_{11}(t) = p\int_0^{t-t_{2n}} \lambda(t-x)e^{-\mu_{11}x}dx + e^{-\mu_{11}(t-t_{2n})}B_{11}(t_{2n}),$$

$$B_{12}(t) = q\int_0^{t-t_{2n}} \lambda(t-x)e^{-\mu_{12}x}dx + e^{-\mu_{12}(t-t_{2n})}B_{12}(t_{2n}),$$

where the initial values $B_{11}(t_{2n})$ and $B_{12}(t_{2n})$ are given from the fluid contents $B_{11}(t)$ and $B_{12}(t)$, respectively, at the end of the preceding overloaded interval $[t_{2n-1}, t_{2n}]$.

The amount of fluid content during the conversation time is then given by

$$B_1(t) = B_{11}(t) + B_{12}(t) \quad t \in [t_{2n}, t_{2n+1}].$$

Since the output rate of each phase is given by

$$\sigma_{1j}(t) = \mu_{1j}B_{1j}(t) \quad t \in [t_{2n}, t_{2n+1}],\ j = 1, 2,$$

we get the output rate of the conversation

$$\begin{aligned}\sigma_1(t) &= \sigma_{11}(t) + \sigma_{12}(t) \\ &= \mu_{11}B_{11}(t) + \mu_{12}B_{12}(t) \quad t \in [t_{2n}, t_{2n+1}].\end{aligned}$$

The duration of the ACW time consists of phase 1 with pdf $g_{21}(x)$ serially followed by phase 2 with pdf $g_{22}(x)$ as shown in Fig. 11. Therefore, during the underloaded interval $t \in [t_{2n}, t_{2n+1}]$, the density for the fluid content in phase 1 satisfies the equation

$$b_{21}(t,x) = \begin{cases} e^{-\mu_{21}x}\sigma_1(t-x) & 0 \leq x \leq t - t_{2n}, \\ e^{-\mu_{21}(t-t_{2n})}b_{21}(t_{2n}, x-t+t_{2n}) & x > t - t_{2n}, \end{cases}$$

which leads to the fluid content

$$\begin{aligned}B_{21}(t) &= \mu_{11}\int_0^{t-t_{2n}} e^{-\mu_{21}x}B_{11}(t-x)dx \\ &\quad + \mu_{12}\int_0^{t-t_{2n}} e^{-\mu_{21}x}B_{12}(t-x)dx \\ &\quad + e^{-\mu_{21}(t-t_{2n})}B_{21}(t_{2n}),\end{aligned}$$

where the initial value $B_{21}(t_{2n})$ is given from the fluid content $B_{21}(t)$ at the end of the preceding overloaded interval $[t_{2n-1}, t_{2n}]$. The output rate of phase 1 is given by

$$\sigma_{21}(t) = \mu_{21}B_{21}(t).$$

The density for the fluid content in phase 2 satisfies the equation

$$b_{22}(t,x) = \begin{cases} e^{-\mu_{22}x}\sigma_{21}(t-x) & 0 \leq x \leq t - t_{2n} \\ e^{-\mu_{22}(t-t_{2n})}b_{22}(t_{2n}, x-t+t_{2n}) & x > t - t_{2n}, \end{cases}$$

which leads to the fluid content

$$B_{22}(t) = \mu_{21}\int_0^{t-t_{2n}} e^{-\mu_{22}x}B_{21}(t-x)dx + e^{-\mu_{22}(t-t_{2n})}B_{22}(t_{2n}),$$

where the initial value $B_{22}(t_{2n})$ is given from the fluid content $B_{22}(t)$ at the end of the preceding overloaded interval $[t_{2n-1}, t_{2n}]$. The output rate of phase 2 is given by

$$\sigma_{22}(t) = \mu_{22}B_{22}(t).$$

The amount of fluid content during the ACW time is then given by

$$B_2(t) = B_{21}(t) + B_{22}(t) \quad t \in [t_{2n}, t_{2n+1}].$$

The total amount of fluid content during the underloaded interval is then given by

$$B(t) = B_1(t) + B_2(t) \quad t \in [t_{2n}, t_{2n+1}].$$

The termination epoch t_{2n+1} of the underloaded interval $[t_{2n}, t_{2n+1}]$ is determined by the condition in Eq. (8).

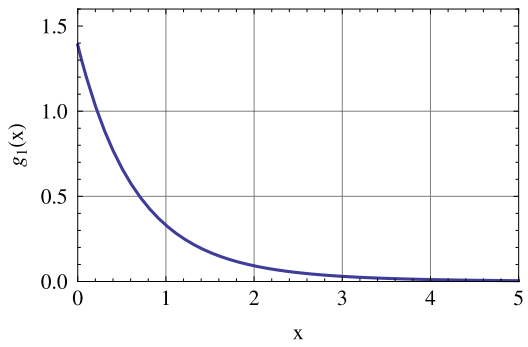
5.2. Solution for the overloaded state

During the overloaded interval $t \in [t_{2n+1}, t_{2n+2}]$, the fluid content in each phase is governed by the following set of simultaneous linear differential equations with constant coefficients:

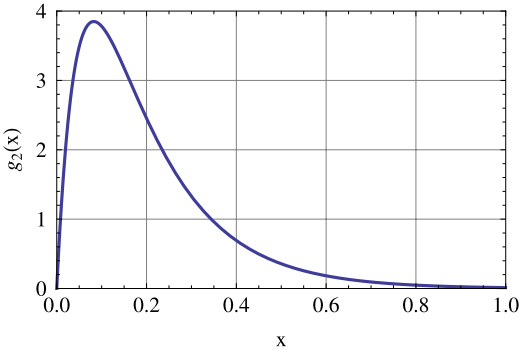
$$\begin{aligned}\frac{dB_{11}(t)}{dt} &= -\mu_{11}B_{11}(t) + p[\mu_{22}B_{22}(t) + s'(t)], \\ \frac{dB_{12}(t)}{dt} &= -\mu_{12}B_{12}(t) + q[\mu_{22}B_{22}(t) + s'(t)], \\ \frac{dB_{21}(t)}{dt} &= -\mu_{21}B_{21}(t) + \mu_{11}B_{11}(t) + \mu_{12}B_{12}(t), \\ \frac{dB_{22}(t)}{dt} &= -\mu_{22}B_{22}(t) + \mu_{21}B_{21}(t)\end{aligned}$$

Table 4
Epochs of state change in the $M_t/(H_2 + E_2)/s/M$ system.

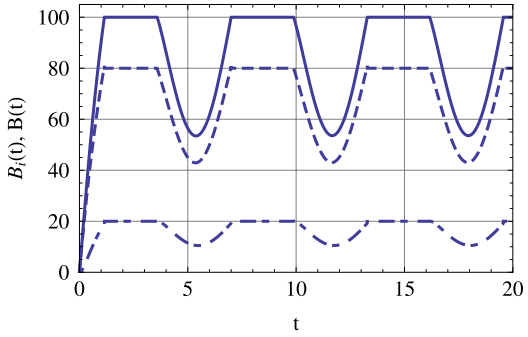
t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
0	1.15041	3.58694	7.00020	9.88379	13.28263	16.16700	19.56582	21.45019



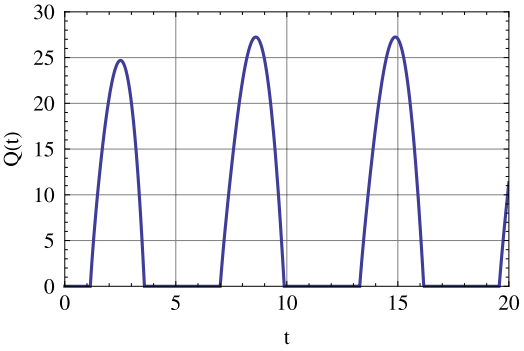
(a) Hyperexponential distribution for the conversation time.



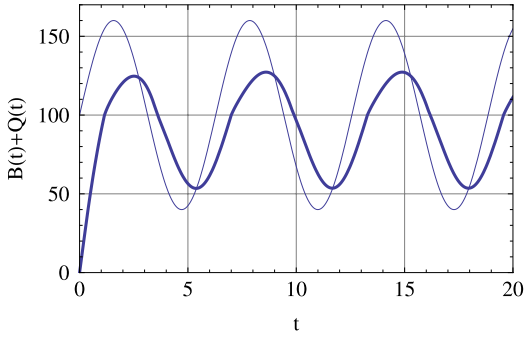
(b) Hypo-exponential distribution for the ACW time.



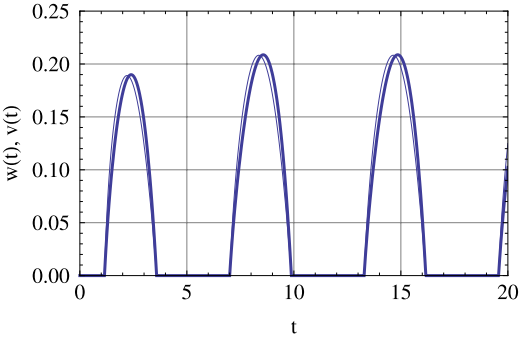
(c) Fluid in service $B_i(t)$ and $B(t)$.



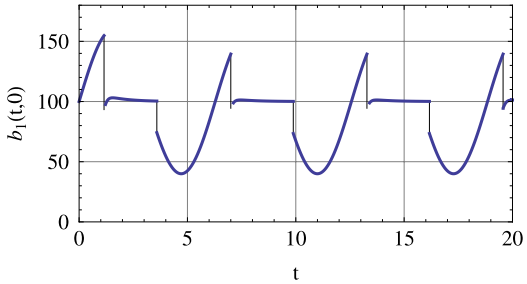
(d) Waiting fluid $Q(t)$.



(e) Total fluid $B(t) + Q(t)$.



(f) Boundary waiting time $w(t)$ (thick) and potential waiting time $v(t)$ (thin).



(g) Fluid entrance rate into service $b_1(t, 0)$.

Fig. 12. Performance of the $M_t/(H_2 + E_2)/s + M$ model.

with the feasibility condition

$$b_1(t, 0) = \mu_{22}B_{22}(t) + s'(t) \geq 0.$$

The initial values $B_{11}(t_{2n+1})$, $B_{12}(t_{2n+1})$, $B_{21}(t_{2n+1})$, and $B_{22}(t_{2n+1})$ are given from their corresponding values at the end of the preceding underloaded interval $[t_{2n}, t_{2n+1}]$. The solution can be easily obtained by the conventional method of Laplace transform for solving a set of simultaneous linear differential equations with constant coefficients.

The differential equation for the boundary waiting time $w(t)$ is given by

$$\frac{dw(t)}{dt} = 1 - \frac{\mu_{22}B_{22}(t) + s'(t)}{\lambda[t - w(t)]\overline{F}[w(t)]} \quad t_{2n+1} \leq t \leq t_{2n+2}$$

with initial condition $w(t_{2n+1}) = 0$. The differential equation for the potential waiting time $v(t)$ is given by

$$\frac{dv(t)}{dt} = \frac{\lambda(t)\overline{F}[v(t)]}{\mu_{22}B_{22}[t + v(t)] + s'[t + v(t)]} - 1 \quad t_{2n+1} \leq t \leq t_{2n+2}$$

with initial condition $v(t_{2n+1}) = 0$.

Once $w(t)$ is obtained, we get the fluid content $Q(t)$ in the waiting room and the abandonment rate $\alpha(t)$ by Eqs. (28) and (30), respectively. The termination epoch t_{2n+2} of the overloaded interval $[t_{2n+1}, t_{2n+2}]$ is determined by the condition in Eq. (51).

5.3. Numerical example

Let us show the numerical results for an $M_t/(H_2 + E_2)/s + M$ system with the sinusoidal input rate function $\lambda(t)$ and constant number of servers $s(t)$ given in Eq. (52). However we assume that the service 1 has a hyperexponential distribution and that service 2 has a hypo-exponential distribution with parameters

$$\mu_{11} = \frac{5}{3}, \quad \mu_{12} = \frac{5}{6}, \quad \mu_{21} = 20,$$
$$\mu_{22} = \frac{20}{3}, \quad p = \frac{2}{3}, \quad q = \frac{1}{3}$$

so that

$$g_1(x) = \frac{10}{9}e^{-\frac{5}{3}x} + \frac{5}{18}e^{-\frac{5}{6}x}, \quad E[G_1] = \frac{4}{5}, \quad C_{G_1}^2 = \frac{5}{4} > 1,$$
$$g_2(x) = 10(e^{-\frac{20}{3}x} - e^{-20x}), \quad E[G_2] = \frac{1}{5}, \quad C_{G_2}^2 = \frac{5}{8} < 1.$$

We also assume the exponentially distributed abandonment time with mean $1/\theta = 0.5$:

$$F(x) = 1 - e^{-2x} \quad x \geq 0.$$

The result of numerical analysis is shown in Table 4 and Fig. 12. Table 4 shows a sequence of epochs at which the state changes from underloaded to overloaded and vice versa. These values are only slightly different from those in Table 1 for the system with exponentially distributed conversation and ACW times with the same means. Figs. 12(a) and (b) show the pdf's of the hyperexponential distribution for the conversation time and the hypo-exponential distribution for the ACW time, respectively. Fig. 12(c) plots the fluid content $B_i(t) = B_{i1}(t) + B_{i2}(t)$ in service $i (i = 1, 2)$ as well as $B(t) = B_1(t) + B_2(t)$ at time t. These curves are not much different from the corresponding curves in Fig. 5(b). Other performance measures are also shown in Figs. 12(c)–(g).

6. Concluding remarks

In this paper, we have studied the fluid approximation to the $M_t/GI/s_t + GI$ model with two stages of service, each being independent and exponentially distributed, for the call center operation. It is true that the same model can also be handled exactly by the $M_t/GI/s_t + GI$ model in [19] with a single stage of service consisting of the sum of two exponentially distributed service times. Then, however, one must solve a fixed-point equation for the function $b(t, 0)$, which is not straightforward as mentioned in Section 6.2 of [19]. By our method of separating the service to exponentially distributed stages, we can do without $b_i(t, 0)$ to obtain the fluid content $B_i(t)$. This merit is already pointed out in Section 2 of [19]. It is well-known that most distributions can be approximated precisely enough by a Coxian distribution for which the Laplace transform of pdf is written as a rational function in the transform parameter. Therefore, we should be able to enjoy the above-mentioned merit by decomposing a non-exponential distribution into the serial–parallel combination of exponential distributions.

In the present case, the two stages of service correspond to the conversation and ACW times in the real operation of call centers. Therefore, we can know the number of operators engaged in the conversation and the number of those in the ACW individually at each moment. This is another merit that cannot be obtained if the service times of the two stages are treated together.

We have shown an example in which the number of servers is determined to cope with the load of only conversation time (service 1). Then the system is *always* overloaded but it seems to remain stable. On the other hand, if the number of servers is determined to cope with the load of both conversation and ACW times (services 1 and 2), the system is *always* underloaded. In this case, we need much more servers than in the former example. Therefore, the staffing rule in the former example may be useful as efficient utilization of the resource (servers).

Acknowledgments

We are grateful to Professor Ward Whitt in the Department of Industrial Engineering and Operations Research of Columbia University, USA, for his valuable comments on the draft of this paper. We also appreciate constructive suggestions by the reviewers of the manuscript who encouraged us to elaborate on the case of non-exponentially distributed conversation and ACW times.

References

[1] Fischer MJ, Garbin DA, Gharakhanian A. Performance modeling of distributed automatic call distribution systems. Telecommun Syst 1998;9(2):133–52. http://dx.doi.org/10.1023/A:1019139721840.
[2] Gans N, Koole G, Mandelbaum A. Telephone call centers: tutorial, review, and research prospects. Manuf Serv Oper Manag 2003;5(2):79–141. http://dx.doi.org/10.1287/msom.5.2.79.16071.
[3] Koole G, Mandelbaum A. Queueing models of call centers: an introduction. Annals of Operations Research 2002;113(1–4):41–59. http://dx.doi.org/10.1023/A:1020949626017.
[4] Mandelbaum A, Zeltyn S. Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. In: Spath D, Fähnrich K-P, editors. Advances in services innovations. Berlin: Springer; 2007. p. 17–45.
[5] Cleveland B, Harne D, editors. Call center operations management. In: Handbook and study guide, Version 2.1. Colorado Springs: ICMI Press; 2004.
[6] Harris R.J., Phillips M.J.. Extensions to a model for post call activity in ACD systems, Preprint, December 1989. http://www-ist.massey.ac.nz/rharris/publicationfiles/pre2000/acdsystems-foatrs.pdf.
[7] Jolley W.M., Harris R.J.. Analysis of post-call activity in queueing systems, In: Proceedings of the 9th international teletraffic congress. (Torremolinos); 1979. p. 1–9.
[8] Jolley WM, Harris RJ. Analysis and optimal design of queueing systems with post-call activity. Austral Telecommun Res 1981;15(1):12–26.
[9] Harris CM, Hoffman KL, Saunders PB. Modeling the IRS telephone taxpayer information system. Oper Res 1987;35(4):504–23. http://dx.doi.org/10.1287/opre.35.4.504.
[10] Kawanishi K. On the counting process for a class of Markovian arrival processes with an application to a queueing system. Queueing Sys 2005;49(2):93–122. http://dx.doi.org/10.1007/s11134-005-6478-7.

[11] Kawanishi K. Waiting time distribution of a queueing system with postservice activity. Comput Math Appl 2006;51(2):209–18. http://dx.doi.org/10.1016/j.camwa.2005.11.021.
[12] Takagi H, Taguchi Y. Analysis of a queueing model for a call center with impatient customers and after-call work. Int J Pure Appl Math 2014;90(2): 205–37. http://dx.doi.org/10.12732/ijpam.v90i2.10.
[13] Phung-Duc T, Kawanishi K. Multiserver retrial queues with after-call work. Numer Algebra Control Optim 2011;1(4):639–56. http://dx.doi.org/10.3934/naco.2011.1.639.
[14] Hall RW. Queueing systems. In: For services and manufacturing. Prentice-Hall; 1991.
[15] Bassamboo A, Randhawa RS. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. Oper Res 2010;58(5): 1398–413. http://dx.doi.org/10.1287/opre.1100.0815.
[16] Whitt W. Fluid models for multiserver queues with abandonments. Oper Res 2006;54(1):37–54. http://dx.doi.org/10.1287/opre.1050.0227.
[17] Feldman Z, Mandelbaum A, Massey WA, Whitt W. Staffing of time-varying queues to achieve time-stable performance. Manag Sci 2008;54(2):324–38. http://dx.doi.org/10.1287/mnsc.1070.0821.
[18] Green LV, Kolesar PJ, Whitt W. Coping with time-varying demand when setting staffing requirements for a service system. Prod Oper Manag 2007;16(1): 13–39. http://dx.doi.org/10.1111/j.1937-5956.2007.tb00164.x.
[19] Liu Y, Whitt W. The $G_t/GI/s_t + GI$ many-server fluid queue. Queueing Syst 2012;71(4):405–44. http://dx.doi.org/10.1007/s11134-012-9291-0. Longer online version with appendix in http://www.columbia.edu/ww2040/allpapers.html.
[20] Liu Y, Whitt W. Fluid models for multiserver queues with abandonments. Oper Res 2006;54(1):37–54. http://dx.doi.org/10.1287/opre.1050.0227.
[21] Liu Y, Whitt W. A network of time-varying many-server fluid queues with customer abandonment. Oper Res 2011;59(4):835–46. http://dx.doi.org/10.1287/opre.1110.0942.
[22] Liu Y, Whitt W. Algorithms for time-varying networks of many-server fluid queues. INFORMS J Comput 2014;26(1):59–73. http://dx.doi.org/10.1287/ijoc.1120.0547.
[23] Kawai Y. Fluid approximation analysis for the queueing model of a call center [Master thesis], Tsukuba, Japan: Master of Engineering, Graduate School of Systems, Information and Engineering, University of Tsukuba; 2013 [in Japanese].
[24] Eick S, Massey WA, Whitt W. The physics of the $M_t/G/\infty$ queue. Oper Res 1993;41(4):731–42. http://dx.doi.org/10.1287/opre.41.4.731.
[25] Eick S, Massey WA, Whitt W. $M_t/G/\infty$ queues with sinusoidal arrival rates. Manag Sci 1993;39(2):241–52. http://dx.doi.org/10.1287/mnsc.39.2.241.
[26] Bertsimas D, Mourtzinou G. Transient laws of non-stationary queueing systems and their applications. Queueing Syst 1997;25(1):115–55. http://dx.doi.org/10.1023/A:1019100301115.