

**ANALYSIS OF A QUEUEING MODEL FOR A CALL CENTER
WITH IMPATIENT CUSTOMERS AND AFTER-CALL WORK**

Hideaki Takagi¹ §, Yutaro Taguchi²

¹Faculty of Engineering
Information and Systems
University of Tsukuba

Tsukuba Science City, Ibaraki, 305-8573, JAPAN

²Graduate School of Systems and Information Engineering
University of Tsukuba

Tsukuba Science City, Ibaraki, 305-8573, JAPAN

Abstract: Two essential features needed for queueing models of call centers are “impatient customers” and “after-call work” (ACW) of operators. Their effects on the performance are comparable to those of service times. We propose and analyze a queueing model of a call center with impatient customers and ACW. We consider a two-dimensional birth-and-death process for the state combining the number of calls waiting or being served in the system and the number of operators working on ACW. The steady-state distribution is obtained numerically, from which we compute a variety of performance measures, including the blocking probability, the probability of wait, the mean waiting time, the ratio of getting service and abandonment, and the fraction of operators who are serving customers, working on ACW, or being idle at an arbitrary time. Furthermore, we show a method of finding the mean waiting time for those customers who get served and for those who abandon while waiting. Our analysis is validated by numerical examples for a call center model of realistic size, namely 40 operators and 30 incoming telephone lines.

Received: October 12, 2013

© 2014 Academic Publications, Ltd.
url: www.acadpubl.eu

§Correspondence author

AMS Subject Classification: 60K25, 68M20, 90B22

Key Words: queue, call center, abandonment, after-call work, performance evaluation

1. Introduction

A telephone call center is one of typical service places where a combination of service equipments and human operators provides service to customers. If managers of such facilities try to increase the utilization of equipments and answering operators without proper dimensioning, it would result in long waiting times leading to the dissatisfaction of not only customers but also operators employed at the center. Queueing models have been widely used to model the performance of call centers with impatient customers, which means that customers in the waiting line may leave before getting service. The multiserver queue $M/M/m$ with impatient customers is called *Erlang-A* model, “A” for “abandonment”, in contrast with the well-known *Erlang-B* model, $M/M/m/m$, and the *Erlang-C* model, $M/M/m$ with only patient customers. See survey articles, for example, by Fischer et al. [2], Gans et al. [3], Koole and Mandelbaum [11], Mandelbaum and Zeltyn [12] as well as a recent article on capacity management issues by Hishinuma and Takagi [6].

Through the measurements at real call centers, we have observed that after operators finish conversation with customers, they must spend similar amount of time to complete additional work such as entering profiles of customers into the customer management database. This extra work of operators is called the *after-call work* (ACW). According to Cleveland and Harne [1, Section 8],

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 operator is unavailable to receive another inbound call while in this mode.

The ACW is also called “post call activity” [5, 7, 8], “wrap-up times” [2], “after-hung-up times” [4], and “postservice activity” [9, 10]. According to Harris and Phillips [5],

The post call activity is a phase in which the operator may fill out docket, 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).

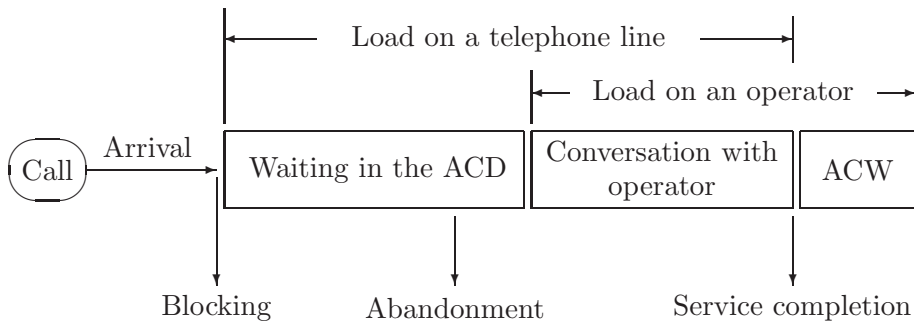


Figure 1: Processing of an inbound call in a call center [6]

We note that all previous work cited above treats ACW approximately. For example, Harris et al. [4] present a model that incorporates retrieval of blocked customers, abandonment and ACW together, but they assume a single exponential distribution for the combination of service time and ACW. More recent studies [9, 13] exactly deal with queues with ACW but without impatient customers. However, we are not aware of any analysis of a queueing model for a call center with impatient customers and ACW correctly together.

In this paper, as both features are essential to the performance evaluation of call centers, we propose and analyze such a queueing model that is formalized as a two-dimensional birth-and-death process. The steady-state joint distribution for the number of calls present in the system and the number of operators working on ACW is obtained numerically, from which we compute a variety of performance measures. Furthermore, we show a method of finding the conditional mean waiting times for those calls that are served and for those abandoned while waiting. We show numerical examples for a call center model of realistic size, namely 40 operators and 30 incoming telephone lines.

2. Queueing Model

Figure 1 displays a sequence of processing an inbound call in a call center which we assume for modeling in this paper. If a customer calls the center while all the telephone lines are in use, he receives a busy signal and his call is blocked (we ignore the retrieval of blocked calls in this paper). However, if a line is free, his call is accepted at an Automatic Call Distributor (ACD), which distributes the call to an available operator. If all the operators are busy, the caller receives

a recorded voice message to ask waiting and then some music possibly. He may hang up while waiting, which results in abandonment of the call, before he is connected to an operator. He leaves the system after completing his conversation with the operator, at which point the telephone line is released. However, the operator may have additional work to perform associated with this call (ACW) before another caller can be taken.

The basic framework of our model is an $M/M/K/J$ queue in the Kendall notation, where K , the number of servers, represents the total number of operators working in the center and J , the maximum number of customers accommodated in the system, stands for the number of incoming telephone lines. In addition, as often with a queueing model of a call center, customers in the waiting room may depart before getting service (abandonment). A new feature of our model is that each server (operator) must spend some amount of time for post-processing work after finishing the service with a customer. This corresponds to the after-call work (ACW). The customer leaves the system as soon as his service with an operator has finished. However, during the ACW the server cannot give service to another customer. Note that this feature makes our model different from the one assuming that each customer has effectively a service time consisting of two exponentially distributed phases. 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. Our queueing model is drawn in Figure 2, where calls and operators are represented by customers and servers, respectively.

2.1. Model and Parameters

We assume that the maximum number of customers accommodated in the total system is limited to J and the total number of servers is assumed to be K . Customers (calls) arrive in a Poisson process with rate λ . The service time, i.e. the time each customer is processed by a server, is assumed to be exponentially distributed with mean $1/\mu$. After completion of service, the customer leaves the system, while the server does not become free but starts the ACW of which the duration is assumed to be exponentially distributed with mean $1/\sigma$. After the completion of ACW, the server accepts a new customer, if any, from the waiting room. If there are no customers waiting, the server becomes idle. When a customer (call) arrives, if there are already J customers in the system, he is blocked and lost immediately and forever (we do not consider retrials of customer call). Otherwise, he is accepted. However, if all the servers (operators) are busy by either serving (processing a call) or working for ACW, the customer

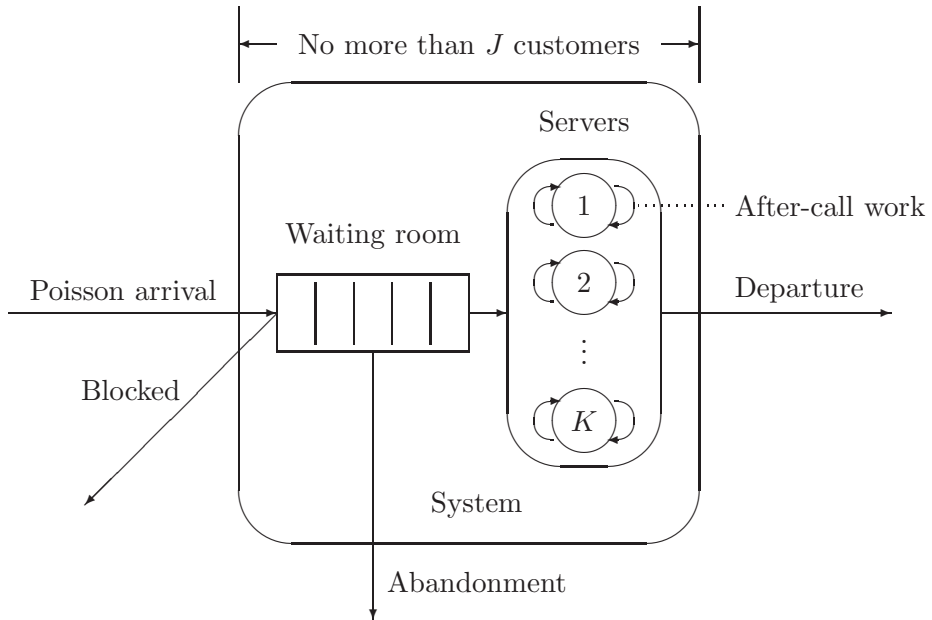
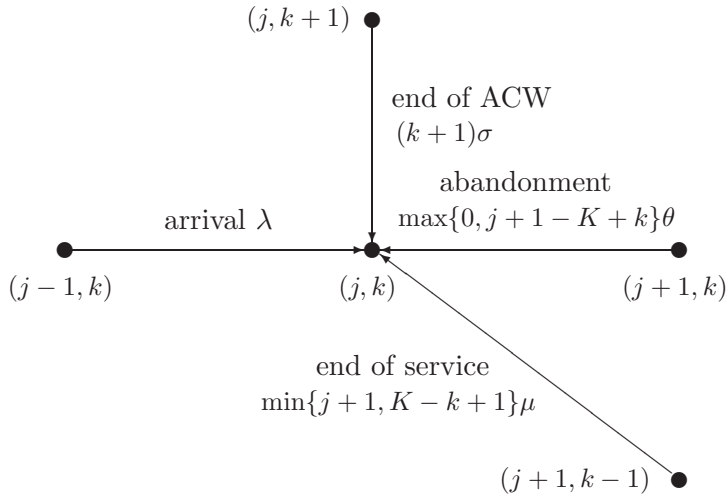
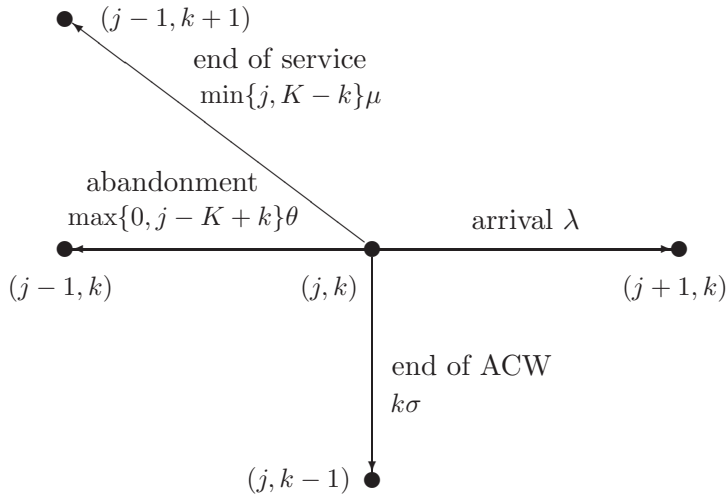


Figure 2: M/M/K/J queueing model with impatient customers and after-call work

must wait in the waiting room. Waiting customers are impatient such that they may leave the system before getting service at rate θ , i.e. the patience time of each waiting customer is exponentially distributed with mean $1/\theta$. If all customers are patient, i.e. they never leave the system once accepted until service completion, our model reduces to the one studied by Harris et al. [5, 7, 8].



(a) Transitions to state (j, k)



(b) Transitions from state (j, k)

Figure 3: State transitions for the system state (j, k)

2.2. Transition of System State

Let $N(t)$ be the number of customers present in the system, and let $A(t)$ be the number of servers working for ACW at time t . Then the two-dimensional process $\{(N(t), A(t)), t \geq 0\}$ is a continuous-time Markov process with a finite state space $0 \leq N(t) \leq J$ and $0 \leq A(t) \leq K$.

We consider the steady-state distribution

$$P_{jk} := \lim_{t \rightarrow \infty} P\{N(t) = j, A(t) = k\} \quad 0 \leq j \leq J, 0 \leq k \leq K.$$

Let us denote by (j, k) the state $\{N(t) = j, A(t) = k\}$. In this state, the number of operators serving customers is $\min\{j, K - k\}$, and the number of waiting customers is $j - \min\{j, K - k\} = \max\{0, j - K + k\}$. The state transition rates to and from state (j, k) are shown in Figure 3. Four kinds of events to happen in the system state are the arrival of a new customer, the abandonment of a customer in the waiting room, the end of service, and the end of ACW. The rates at which these events happen along with the states to and from the system state (j, k) are tabulated in Table 1.

Table 1: Transition rates to and from the system state (j, k)

(a) Transition to state (j, k)		
From state	Event	Rate
$(j - 1, k)$	arrival	λ
$(j + 1, k)$	abandonment	$\max\{0, j + 1 - K + k\}\theta$
$(j + 1, k - 1)$	end of service	$\min\{j + 1, K - k + 1\}\mu$
$(j, k + 1)$	end of ACW	$(k + 1)\sigma$
(b) Transition from state (j, k)		
To state	Event	Rate
$(j + 1, k)$	arrival	λ
$(j - 1, k)$	abandonment	$\max\{0, j - K + k\}\theta$
$(j - 1, k + 1)$	end of service	$\min\{j, K - k\}\mu$
$(j, k - 1)$	end of ACW	$k\sigma$

According to Figure 3 and Table 1, we have the following set of balance equations for $\{P_{jk}; 0 \leq j \leq J, 0 \leq k \leq K\}$:

$$\lambda P_{00} = \sigma P_{01},$$

$$\begin{aligned}
& [\lambda + \min\{j, K\}\mu + \max\{0, j - K\}\theta]P_{j0} \\
& = \lambda P_{j-1,0} + \sigma P_{j1} + \max\{0, j + 1 - K\}\theta P_{j+1,0} \\
& \quad (1 \leq j \leq J - 1),
\end{aligned}$$

$$(\lambda + k\sigma)P_{0k} = \mu P_{1,k-1} + (k + 1)\sigma P_{0,k+1} \quad (1 \leq k \leq K - 1),$$

$$(\lambda + K\sigma)P_{0K} = \mu P_{1,K-1} + \theta P_{1,K},$$

$$\begin{aligned}
(\lambda + K\sigma + j\theta)P_{jK} & = \lambda P_{j-1,K}(t) + \mu P_{j+1,K-1} + (j + 1)\theta P_{j+1,K} \\
& \quad (1 \leq j \leq J - 1),
\end{aligned}$$

$$\begin{aligned}
& [\lambda + \min\{j, K - k\}\mu + k\sigma + \max\{0, j - K + k\}\theta]P_{jk} \\
& = \lambda P_{j-1,k} + \min\{j + 1, K - k + 1\}\mu P_{j+1,k-1} + (k + 1)\sigma P_{j,k+1} \\
& + \max\{0, j + 1 - K + k\}\theta P_{j+1,k} \\
& \quad (1 \leq j \leq J - 1, 1 \leq k \leq K - 1),
\end{aligned}$$

$$[\min\{J, K\}\mu + \max\{0, J - K\}\theta]P_{J0} = \lambda P_{J-1,0} + \sigma P_{J1},$$

$$\begin{aligned}
& [\min\{J, K - k\}\mu + k\sigma + \max\{0, J - K + k\}\theta]P_{Jk} \\
& = \lambda P_{J-1,k} + (k + 1)\sigma P_{J,k+1} \quad (1 \leq k \leq K - 1),
\end{aligned}$$

and

$$(K\sigma + J\theta)P_{JK} = \lambda P_{J-1,K}.$$

Note that there are $(J + 1)(K + 1)$ homogeneous equations, one of which is redundant. The normalization condition is given by

$$\sum_{j=0}^J \sum_{k=0}^K P_{jk} = 1.$$

We assume that this set of linear equations is solved numerically. The set of these linear equations can be written in a matrix form as shown in [14].

3. Performance Measures and Numerical Examples

In this section, we consider several performance measures of our queueing system for a call center with ACW that can be calculated in terms of the state distribution probabilities obtained in the previous section as follows:

- Probability that an arriving customer is blocked (blocking probability): P_b ,
- Probability that an arriving customer is accepted but waits (probability of wait): P_W ,
- Mean number of customers present in the waiting room at an arbitrary time: $E[L]$,
- Mean waiting time for both customers who are served and who abandon: $E[W]$,
- Probability that an waiting customer abandons before getting service (probability of abandonment): $P\{\text{Ab}\}$,
- Fraction of time that each operator is either serving a customer or working for ACW (server utilization): U , and
- Probabilities that an arriving customer is blocked, abandons, and gets served.

We note that Harris et al. [5, 7, 8] show the calculation of some performance measures for a queueing model with ACW and without impatient customers.

The reader is referred to Figure 4 for the domain of summation for calculating these performance measures in the two-dimensional state space shown separately for the cases $K < J$ and $J < K$. The domain named “wait” indicates the set of states in which an arriving customer waits because all servers are busy either serving a customer or working for ACW. The domain named “no wait” indicates the set of states in which an arriving customer can enter service immediately. They are defined in Table 2.

For each of these performance measures, we present the analytical expression followed by numerical examples. Numerical examples are plotted against the arrival rate λ in numbers/minute for three different values of the mean time of ACW $1/\sigma = 5, 15$, and 25 minutes (so that we can see the effects of ACW) with the following set of fixed values for other parameters:

- Number of servers (operators), $K = 40$,

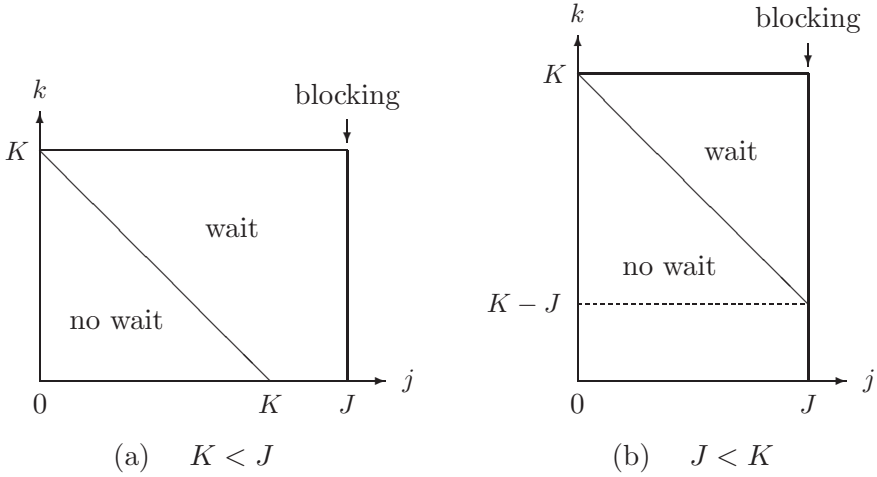


Figure 4: Domain of system state space

Table 2: Domain of summation

$K < J$	
no wait:	$0 \leq k \leq K, 0 \leq j \leq K - k - 1$
wait:	$0 \leq k \leq K, K - k \leq j \leq J$
$J < K$	
no wait:	$0 \leq j \leq J, 0 \leq k \leq K - j - 1$
wait:	$0 \leq j \leq J, K - j \leq k \leq K$

- Total number of customers that can be accepted in the system (capacity of the system, total number of incoming telephone lines), $J = 30$,
- Mean service time, $1/\mu = 30$ minutes, and
- Mean patience time, $1/\theta = 15$ minutes (note that $\mu/\theta = 0.5$).

For calculating each performance measure, we take the summation of relevant numbers to that measure, which is summarized in Table 3, over the appropriate domain.

Table 3: Counting of performance measures ($L+S = N$ and $S+A+I = K$)

	L	S	A	I	N
no wait:	0	j	k	$K - j - k$	j
wait:	$j + k - K$	$K - k$	k	0	j

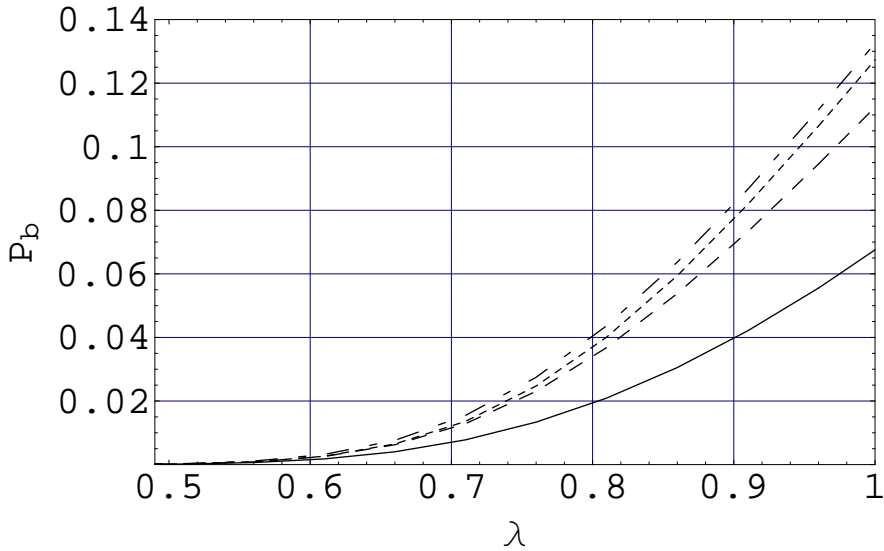


Figure 5: Blocking probability (“no ACW”, $\sigma = \frac{1}{5}, \frac{1}{15}$, and $\frac{1}{25}$ from above)

3.1. Blocking Probability

Those customers who arrive to find that all the waiting room is full, i.e there are J customers in the system are blocked and lost. Owing to the PASTA (Poisson arrivals see time averages) property, the state distribution seen by an arriving customer is the same as the state distribution at an arbitrary time in the steady state given above. Therefore, we obtain the blocking probability as

$$P_b = \sum_{k=0}^K P_{Jk}.$$

The numerical values for the blocking probability are plotted in Figure 5 for cases of “no ACW”, $\sigma = 1/5, 1/15$, and $1/25$. The case “no ACW” corresponds

to $\sigma = \infty$. This figure indicates the trends that the longer the ACW is the lower the blocking probability is. These trends occur when the mean patience time is less than the mean service time ($\mu/\theta < 1$). When customers are impatient, many of them abandon waiting, which reduces the number of customers present in the system. For longer ACW, more customers abandon waiting, which makes the blocking probability lower. In contrast, when the mean patience time is longer than the mean service time ($\mu/\theta > 1$), the longer the ACW is the higher the blocking probability is, as shown in Table 4. If the mean patience time is the same as the mean service time ($\mu/\theta = 1$), the blocking probability does not depend on the length of ACW. The situation in this case is as if all customers in the system, including the ones in the waiting room, are being served. Since the blocking probability depends on the number of customers in the system, not on the number of operators working for ACW, it does not depend on ACW.

Table 4: Blocking probability ($\lambda = 1$ customer/minute)

σ		μ/θ		
		0.5 (impatient)	1	2 (patient)
0.2	(short ACW)	0.132458	0.132460	0.132461
0.067		0.116902	0.132460	0.143091
0.04	(long ACW)	0.070554	0.132460	0.190545

3.2. Probability of Wait

When a customer arrives and gets accepted, if all the servers are busy by either serving a customer or working for ACW, the customer must wait in the waiting room. Therefore, in order to calculate the probability of wait, we must know the probability distribution for the number of customers in the system seen by an arriving customer. Owing to the PASTA property again, the state probability distribution \hat{P}_{jk} seen by an arbitrary customer who is accepted upon arrival is given by

$$\hat{P}_{jk} = \frac{P_{jk}}{1 - P_b} \quad 0 \leq j \leq J - 1, 0 \leq k \leq K.$$

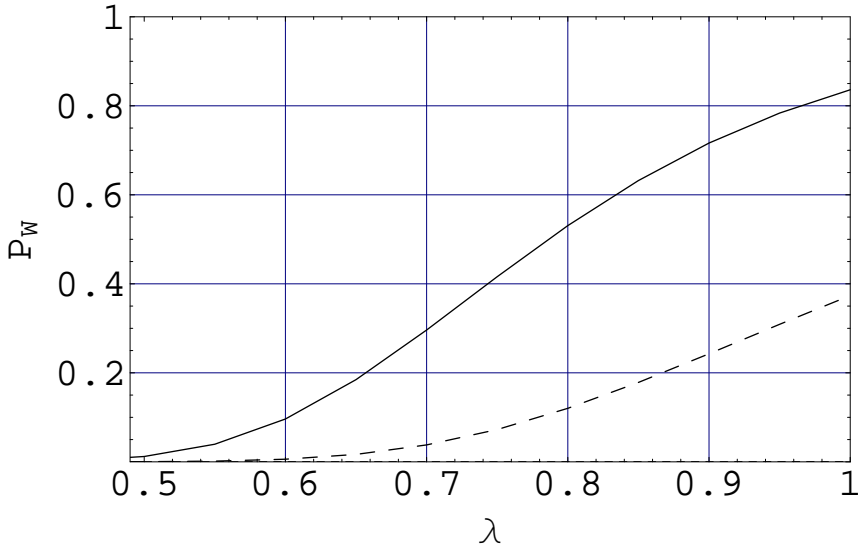


Figure 6: Probability of wait ($\sigma = \frac{1}{25}, \frac{1}{15},$ and $\frac{1}{5}$ from above)

Thus, the probability of wait of an accepted customer is given by

$$P_W = \begin{cases} \sum_{k=0}^K \sum_{j=K-k}^{J-1} \hat{P}_{jk} & J \geq K, \\ \sum_{j=0}^{J-1} \sum_{k=K-j}^K \hat{P}_{jk} & J < K. \end{cases}$$

Numerical values for P_W are plotted in Figure 6 for $\sigma = 1/25, 1/15,$ and $1/5$. We observe that the shorter the ACW is the less the probability of wait for accepted customers is. If the ACW is less than 5 minutes, the probability of wait is almost zero. Therefore, most arriving customers can get service without waiting.

3.3. Mean Number of Waiting Customers

If there are j customers present in the system and k servers are working for ACW such that $j + k \geq K$, $j - (K - k)$ customers are waiting in the waiting room. Therefore, the mean number of customers present in the waiting room

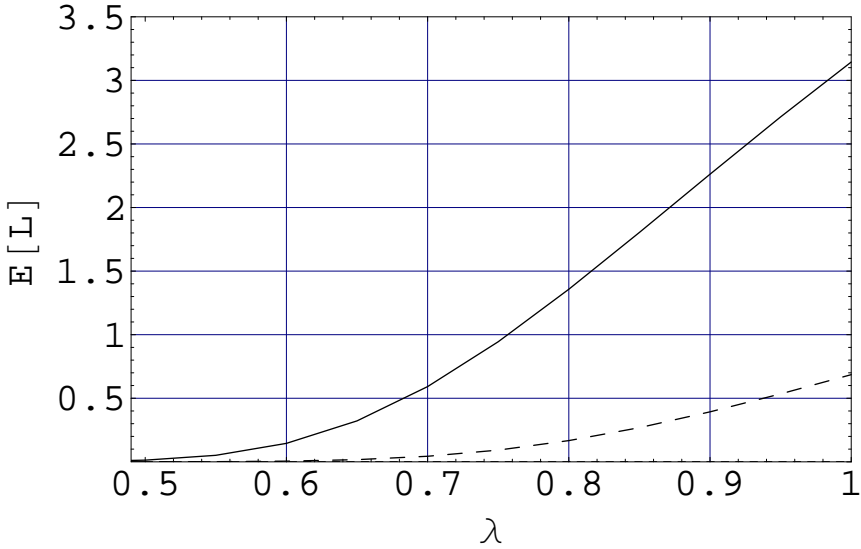


Figure 7: Mean number of customers in the waiting room ($\sigma = \frac{1}{25}, \frac{1}{15},$ and $\frac{1}{5}$ from above)

at an arbitrary time is given by

$$E[L] = \begin{cases} \sum_{k=0}^K \sum_{j=K-k}^J (j+k-K)P_{jk} & J \geq K, \\ \sum_{j=0}^J \sum_{k=K-j}^K (j+k-K)P_{jk} & J < K. \end{cases}$$

Numerical values for $E[L]$ are plotted in Figure 7 for $\sigma = 1/25, 1/15,$ and $1/5$. We observe that the shorter the ACW is the less the mean number of customers in the waiting room is. If the ACW is less than 5 minutes, few customers wait in the waiting room.

3.4. Mean Waiting Time

By applying Little’s law to the set of customers present in the waiting room, we get the relation

$$E[L] = \lambda(1 - P_b)E[W],$$

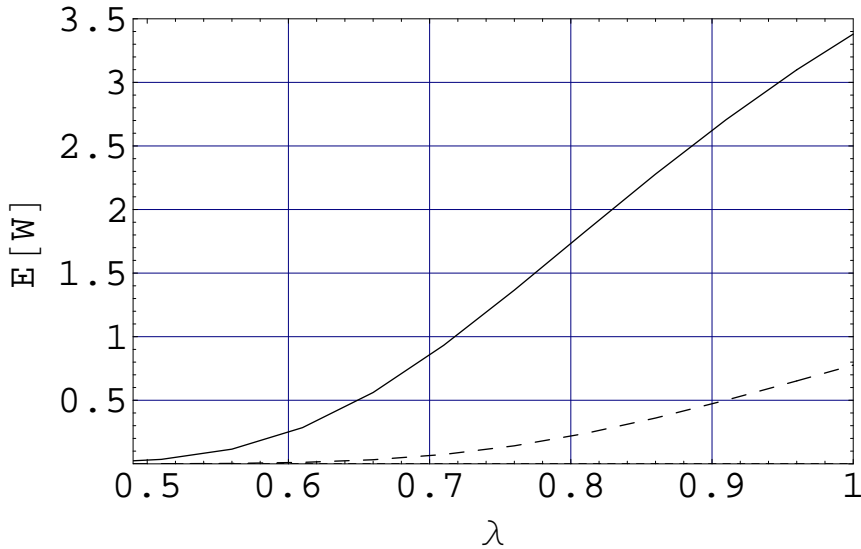


Figure 8: Mean waiting time ($\sigma = \frac{1}{25}, \frac{1}{15}$, and $\frac{1}{5}$ from above)

where $\lambda(1 - P_b)$ is the arrival rate to the waiting room and $E[W]$ is the mean waiting time for both customers who are served and who abandon before getting service. It follows that

$$E[W] = \frac{E[L]}{\lambda(1 - P_b)}.$$

Numerical values for $E[W]$ are plotted in Figure 8 for $\sigma = 1/25, 1/15$, and $1/5$. We observe that the shorter the ACW is the shorter the mean waiting time is. If the ACW is less than 5 minutes, the mean waiting time is almost zero.

3.5. Probability of Abandonment

If we consider a long time τ , the mean number of customers who arrive to the waiting room during τ is $\lambda(1 - P_b)\tau$, while the mean number of customers who abandon during τ is $\theta\tau \times E[L]$. The ratio of the two gives the probability that an arbitrary customer in the waiting room abandons waiting as follows:

$$P\{\text{Ab}\} = \frac{\theta\tau E[L]}{\lambda(1 - P_b)\tau} = \theta E[W].$$

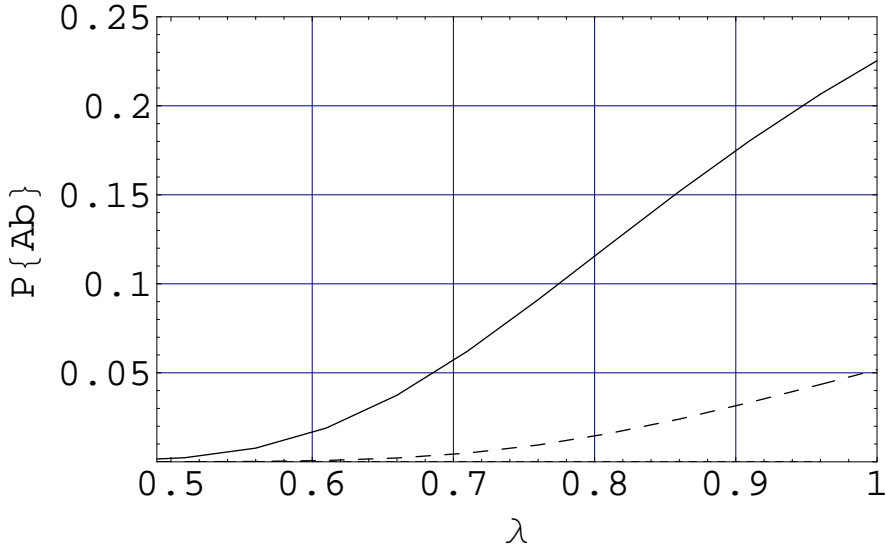


Figure 9: Probability of abandonment ($\sigma = \frac{1}{25}, \frac{1}{15},$ and $\frac{1}{5}$ from above)

We note that the same relation holds for models without ACW as shown by Mandelbaum and Zeltyn [12].

Numerical values for $P\{Ab\}$ are plotted in Figure 9 for $\sigma = 1/25, 1/15,$ and $1/5$. It is not surprising that the shapes of the plots are similar to those for $E[W]$ in Figure 8 due to the above proportional relation.

3.6. Probability of Successful Service

Let S be the number of customers being served at an arbitrary time. We can calculate $E[S]$ as follows. Suppose that there are j customers present in the system and k servers are working for ACW. If $j + k \geq K$, then $K - k$ customers are attended by the servers. (In Section 3.3, we have claimed that $j - (K - k)$ customers are in the waiting room.) If $j + k \leq K$, then the waiting room is

empty and all the j customers are attended by the servers. Therefore, we have

$$E[S] = \begin{cases} \sum_{k=0}^{K-1} \left[\sum_{j=0}^{K-k-1} jP_{jk} + \sum_{j=K-k}^J (K-k)P_{jk} \right] & J \geq K, \\ \sum_{j=0}^J \left[\sum_{k=0}^{K-j-1} jP_{jk} + \sum_{k=K-j}^K (K-k)P_{jk} \right] & J < K. \end{cases}$$

Considering a long time τ as above, the mean number of customers who are served during τ is given by $\mu\tau \times E[S]$. Then we get the probability that an accepted customer is served as

$$P\{\text{Sr}\} = \frac{\mu\tau E[S]}{\lambda(1 - P_b)\tau} = \frac{\mu E[S]}{\lambda(1 - P_b)},$$

which also results from Little's law applied to those customers who get service:

$$\lambda(1 - P_b)P\{\text{Sr}\} \cdot \frac{1}{\mu} = E[S].$$

We show later (Section 3.8) the relation $P\{\text{Sr}\} + P\{\text{Ab}\} = 1$.

3.7. Server Utilization

The utilization of each server (operator), denoted by U , is defined as the fraction of time that he is either serving a customer or working for ACW. Note that $E[S]$ is the mean number of servers who are serving customers at an arbitrary time. If A denotes the number of servers who are working for ACW at an arbitrary time, we have

$$U = \frac{E[S] + E[A]}{K},$$

where we obviously have

$$E[A] = \sum_{k=0}^K \sum_{j=0}^J kP_{jk}.$$

We must note the relation

$$\mu E[S] = \sigma E[A],$$

which means that the rate of finishing service equals the rate of finishing ACW, because each service completion must be followed by exactly one ACW.

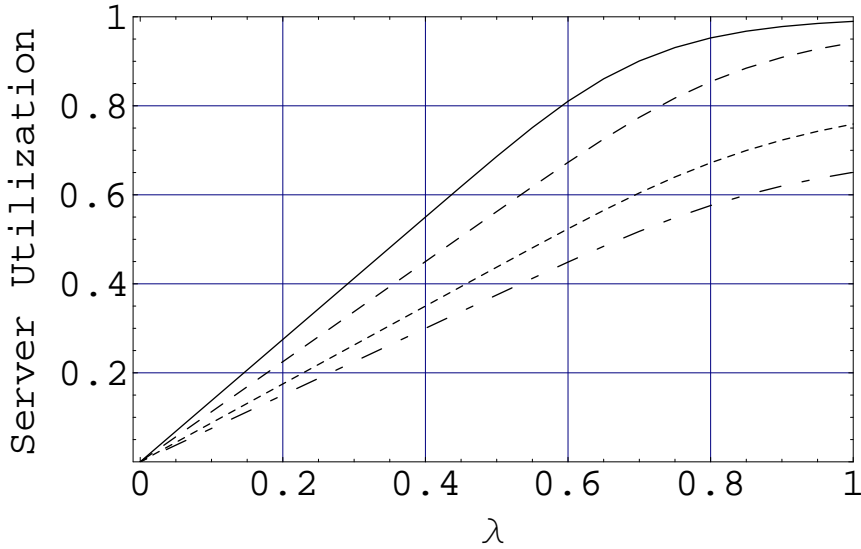


Figure 10: Server utilization ($\sigma = \frac{1}{25}, \frac{1}{15},$ and $\frac{1}{5}$ and “no ACW” from above)

Numerical values of the server utilization U are plotted in Figure 10 for $\sigma = 1/25, 1/15,$ and $1/5$ and “no ACW”. The longer the ACW is the higher the server utilization is, meaning that more operators are working for ACW.

The mean number of idle (neither serving customers nor working for ACW) servers at an arbitrary time is given by

$$E[I] = \begin{cases} \sum_{k=0}^{K-1} \sum_{j=0}^{K-k-1} (K-j-k)P_{ik} & J \geq K, \\ \sum_{j=0}^J \sum_{k=0}^{K-j-1} (K-j-k)P_{ik} & J < K, \end{cases}$$

which leads to the relation

$$E[S] + E[A] + E[I] = K.$$

3.8. Classification of Customers

We can classify all customers into the following three categories with respect to their treatment by the system:

- Customers who are blocked upon arrival.

The mean number of blocked customers per unit time is given by λP_b .

- Customers who abandon in the waiting room.

The mean number of abandoned customers per unit time is given by $\theta E[L]$.

- Customers who are served.

The mean number of served customers per unit time is given by $\mu E[S]$.

These should sum to the total arrival rate:

$$\lambda = \lambda P_b + \theta E[L] + \mu E[S],$$

which is equivalent to the relation

$$P\{\text{Ab}\} + P\{\text{Sr}\} = 1.$$

Finally, the mean number of customers present in the system at an arbitrary time is given by

$$E[N] = \sum_{k=0}^K \sum_{j=0}^J j P_{jk} = E[L] + E[S].$$

In Figure 11, we plot the ratio of three categories of customers for $\sigma = 1/25$. The bottom curve shows the fraction of customers who are served, which is given by $\mu E[S]/\lambda$. For $\lambda < 0.6$, most customers are served. The difference between the middle curve and the bottom curve shows the fraction of customers who abandon waiting, which is given by $\theta E[L]/\lambda$. For $\lambda > 0.6$, the fraction of such customers grows as λ increases. The difference between the top curve and the middle curve shows the fraction of customers who are blocked upon arrival, which is given by P_b . For $\lambda > 0.7$, the fraction of blocked customers grows as λ increases, which agrees with the observation in Figure 5.

4. Waiting Times of Served and Abandoned Customers

The mean waiting time computed by the formula in Section 3.4 covers the waiting time for both customers who get served and who abandon waiting. However, it is more interesting to study the waiting time of those customers who are served and the waiting time of those who abandon separately. In particular, the latter is significant as it is related to the dissatisfaction of customers [12].

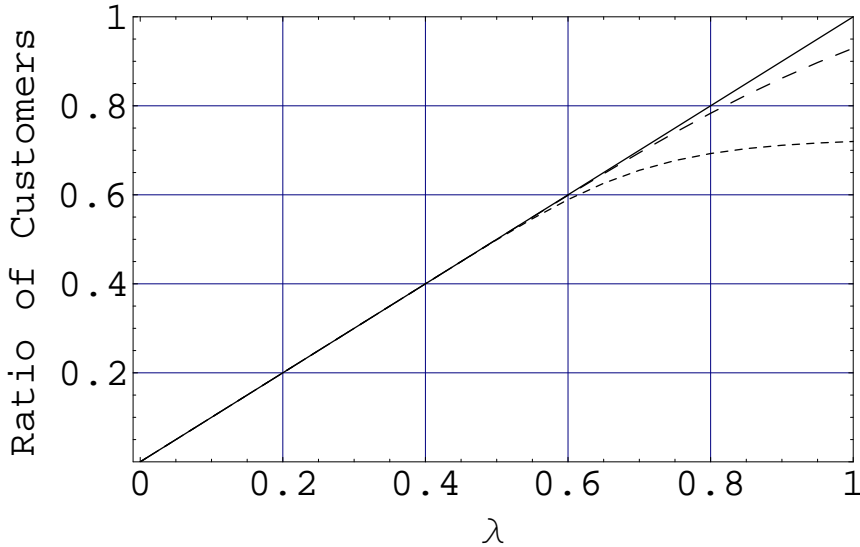


Figure 11: Accumulated ratio of served, abandoned and blocked customers (from below)

In the following, we present a method of finding the mean waiting time for those customers who get served and for those who abandon while being in line in the waiting room. To do so, we keep track of the state of a customer after he arrives and accepted until he leaves the waiting room by either entering the service or by abandoning. The state of such a customer consists of a combination of the number of customers either being served or waiting before him and the number of servers working for ACW. Then we can calculate the distribution of the time until this customer comes to the front of the waiting room or he abandons waiting. The idea of tracking the customer state comes from [15] for the queueing model with impatient customers without ACW.

4.1. State Transition of the Waiting Customer

We consider the transition in the state of a tagged customer in the waiting room after he arrives and gets accepted until his service is started or he abandons waiting whichever occurs first. The state of the tagged customer is denoted by (l, m) when there are l customers present in the system before him, including those being served and those present in the waiting room before him, and there

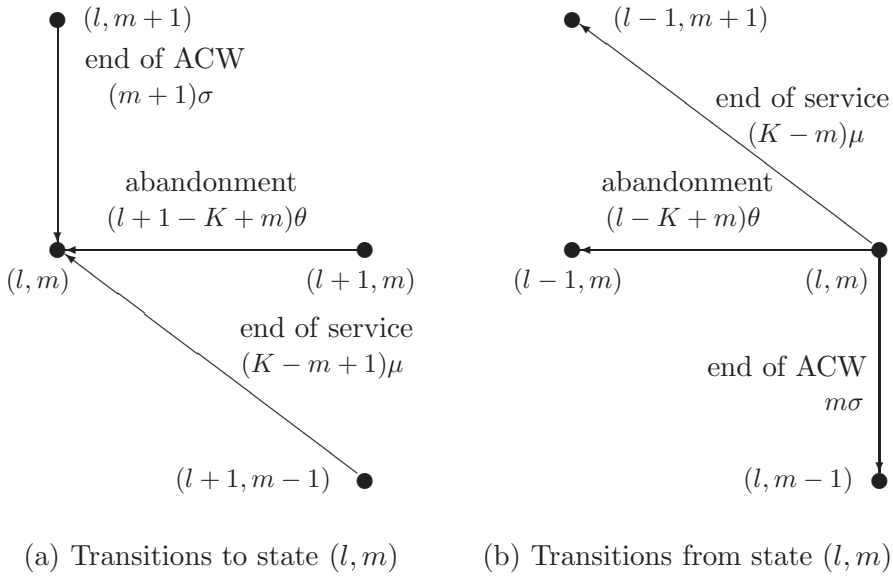


Figure 12: State transitions for the customer state (l, m)

are m servers working for ACW. We assume the first-come first-served (FCFS) discipline.

The state transition rates to and from the customer state (l, m) are shown in Figure 12. Three kinds of events to happen on the state of the tagged customer in the waiting room are the abandonment of a customer ahead of him in the waiting room, the end of service, and the end of ACW. Note that the arrival of a customer does not affect the state of the tagged customer in the waiting room. The rates at which these events to happen along with the states to and from state (l, m) are tabulated in Table 5.

Initially the state of the tagged customer is (j, k) if he arrives when the system state is (j, k) . The domain of the customer state space is shown in Figure 13 separately for the cases $j < K < J$, $J < K$, and $K < j < J$. While the tagged customer is in the waiting room, l decreases monotonically, while m increases or decreases. However, $l + m$ decreases monotonically as the servers keep working. When $l + m = K$ the tagged customer is at the head in the waiting room. Therefore, the following constraint holds:

$$K \leq l + m \leq j + k.$$

Table 5: Transition rates to and from the customer state (l, m)

(a) Transition to state (l, m)		
From state	Event	Rate
$(l + 1, m)$	abandonment	$(l + 1 - K + m)\theta$
$(l + 1, m - 1)$	end of service	$(K - m + 1)\mu$
$(l, m + 1)$	end of ACW	$(m + 1)\sigma$
(b) Transition from state (l, m)		
To state	Event	Rate
$(l - 1, m)$	abandonment	$(l - K + m)\theta$
$(l - 1, m + 1)$	end of service	$(K - m)\mu$
$(l, m - 1)$	end of ACW	$m\sigma$

4.2. Governing Equations for the Waiting Time

Let us denote by $g_{l,m}^{j,k}(t)$ the probability density function for the waiting time of the tagged customer in the waiting room in state (l, m) who arrived when the system state was (j, k) . Then, according to Table 5 and Figure 13, we have the following set of differential equations for $\{g_{l,m}^{j,k}(t)\}$:

$$\frac{dg_{j,k}^{j,k}(t)}{dt} = -[(K - k)\mu + k\sigma + (j - K + k)\theta]g_{j,k}^{j,k}(t),$$

$$\begin{aligned} \frac{dg_{j,m}^{j,k}(t)}{dt} &= (m + 1)\sigma g_{j,m+1}^{j,k}(t) \\ &- [(K - m)\mu + m\sigma + (j - K + m)\theta]g_{j,m}^{j,k}(t) \end{aligned} \quad (\max\{0, K - j\} \leq m \leq k - 1),$$

$$\begin{aligned} \frac{dg_{j+k-m,m}^{j,k}(t)}{dt} &= (K - m + 1)\mu g_{j+k-m+1,m-1}^{j,k}(t) \\ &- [(K - m)\mu + m\sigma + (j - K + k)\theta]g_{j+k-m,m}^{j,k}(t) \end{aligned} \quad (k + 1 \leq m \leq K),$$

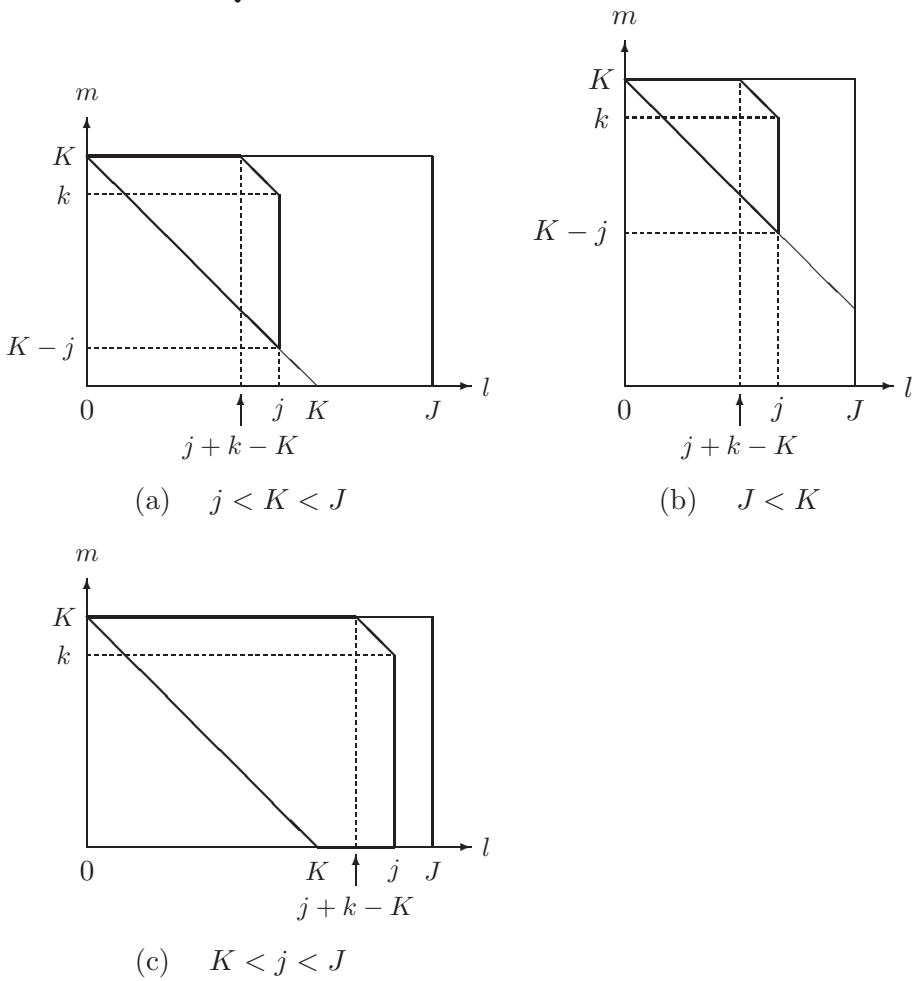


Figure 13: Domain of customer state space

$$\frac{dg_{l,0}^{j,k}(t)}{dt} = \sigma g_{l,1}^{j,k}(t) + (l+1-K)\theta g_{l+1,0}^{j,k}(t) - [K\mu + (l-K)\theta]g_{l,0}^{j,k}(t) \quad (K \leq l \leq j-1),$$

$$\frac{dg_{l,K}^{j,k}(t)}{dt} = \mu g_{l+1,K-1}^{j,k}(t) + (l+1)\theta g_{l+1,K}^{j,k}(t) - (K\sigma + l\theta)g_{l,K}^{j,k}(t) \quad (0 \leq l \leq j-(K-k)-1),$$

and

$$\begin{aligned} \frac{dg_{l,m}^{j,k}(t)}{dt} &= (K - m + 1)\mu g_{l+1,m-1}^{j,k}(t) + (m + 1)\sigma g_{l,m+1}^{j,k}(t) \\ &+ (l + 1 - K + m)\theta g_{l+1,m}^{j,k}(t) \\ &- [(K - m)\mu + m\sigma + (l - K + m)\theta]g_{l,m}^{j,k}(t) \end{aligned}$$

$$(1 \leq l \leq j - 1, \max\{1, K - l + 1\} \leq m \leq \min\{j + k - l - 1, K - 1\}).$$

The initial condition is given by

$$g_{l,m}^{j,k}(0) = \begin{cases} 1 & \text{if } l = j \text{ and } m = k, \\ 0 & \text{otherwise.} \end{cases}$$

We note that $\{g_{l,m}^{j,k}(t)\}$ does not involve λ , because the waiting time of a customer who has arrived does not depend on the arrivals afterwards.

The solution of the first equation is given by

$$g_{j,k}^{j,k}(t) = \exp\{-[(K - k)\mu + k\sigma + (j - K + k)\theta]t\} \quad t \geq 0.$$

The structure of the above set of equations allows us to compute other $g_{l,m}^{j,k}(t)$'s successively as follows. Starting with $g_{j,k}^{j,k}(t)$, we can compute $g_{j+k-m,m}^{j,k}(t)$ for $m = k + 1, k + 2, \dots, K$ recursively. After all $g_{l,m}^{j,k}(t)$'s such that $l + m = n$, we can first obtain $g_{l,m-1}^{j,k}(t)$, and then obtain $g_{l,m}^{j,k}(t)$'s such that $l + m = n - 1$ recursively until we get $g_{l,m}^{j,k}(t)$'s such that $l + m = K$. However, we have been unable to find analytic expressions for general $\{g_{l,m}^{j,k}(t)\}$ explicitly.

We may convert these equations to those for the Laplace transforms:

$$\begin{aligned} g_{l,m}^{*j,k}(s) &:= \int_0^\infty e^{-st} g_{l,m}^{j,k}(t) dt, \\ sg_{l,m}^{*j,k}(s) - g_{l,m}^{j,k}(0) &= \int_0^\infty e^{-st} \frac{dg_{l,m}^{j,k}(t)}{dt} dt. \end{aligned}$$

Then the set of equations for $\{g_{l,m}^{*j,k}(s)\}$ is linear with constant coefficients. Therefore, again starting with

$$g_{j,k}^{*j,k}(s) = \frac{1}{s + (K - k)\mu + k\sigma + (j - K + k)\theta},$$

we can find all $\{g_{l,m}^{*j,k}(s)\}$ recursively. For example, we have

$$g_{j,m}^{*j,k}(s) = \frac{(k!/m!) \sigma^{k-m}}{\prod_{i=m}^k [s + (K-i)\mu + i\sigma + (j-K+i)\theta]}$$

$$\max\{0, K-j\} \leq m \leq k,$$

$$g_{j+k-m,m}^{*j,k}(s) = \frac{[(K-k)!/(K-m)!] \mu^{m-k}}{\prod_{i=k}^m [s + (K-i)\mu + i\sigma + (j-K+k)\theta]}$$

$$k \leq m \leq K.$$

However, it seems difficult to obtain explicit expressions for all $\{g_{l,m}^{*j,k}(s)\}$.

Alternatively, it is easy to find all $\{g_{l,m}^{*j,k}(s)\}$ for each specific value of s , such as $\{g_{l,m}^{*j,k}(\theta)\}$, by recursive numerical computation. It is also clear that the set of first derivatives $\{dg_{l,m}^{*j,k}(s)/ds\}$ can be expressed in terms of $\{g_{l,m}^{*j,k}(s)\}$. Therefore, we can calculate $\{dg_{l,m}^{*j,k}(s)/ds\}$ for any specific value of s as well. The first derivatives at $s = \theta$ are needed for computing the conditional mean waiting times as shown below. Derivatives of any order can be computed similarly in principle.

4.3. Waiting Time of A customer Who is Served

For the tagged customer who arrives when there are j other customers in the system and k servers are working for ACW, the joint probability that he is served and the distribution of his waiting time is given by

$$P\{W > t, \text{Sr} \mid N = j, A = k\} = \sum_{m=1}^K m\sigma \int_t^\infty e^{-\theta x} g_{K-m,m}^{j,k}(x) dx.$$

This expression leads to several performance indicators for the waiting time of the tagged customer.

First, the probability of getting service is given by

$$P\{W > 0, \text{Sr} \mid N = j, A = k\} = \sigma \sum_{m=1}^K m \int_0^\infty e^{-\theta x} g_{K-m,m}^{j,k}(x) dx$$

$$= \sigma \sum_{m=1}^K m g_{K-m,m}^{*j,k}(\theta).$$

The integrand $e^{-\theta x} g_{K-m,m}^{j,k}(x)$ is the probability that the customer at the head of the waiting room has waited x time units without abandonment. The summand indicates that one of m servers who have been working for ACW finishes ACW in the next dx time units with probability $m\sigma dx$. Then he enters the service by terminating the wait. We have numerically confirmed that the unconditional probability of getting service

$$\begin{aligned} P\{\text{Sr}\} &= P\{W = 0\} + \sum_{j=0}^{J-1} \sum_{k=\max\{0, K-j\}}^K P\{W > 0, \text{Sr} \mid N = j, A = k\} \hat{P}_{j,k} \\ &= 1 - P_W + \sigma \sum_{m=1}^K m \sum_{j=0}^{J-1} \sum_{k=\max\{0, K-j\}}^K g_{K-m,m}^{*j,k}(\theta) \hat{P}_{j,k} \end{aligned}$$

agrees with the value calculated by the formula in Section 3.6.

Second, the joint probability that the tagged is served and his mean waiting time is given by

$$\begin{aligned} E[W; W > 0, \text{Sr} \mid N = j, A = k] &= \int_0^\infty P\{W > t, \text{Sr} \mid N = j, A = k\} dt \\ &= \sigma \sum_{m=1}^K m \int_0^\infty dt \int_t^\infty e^{-\theta x} g_{K-m,m}^{j,k}(x) dx. \end{aligned}$$

This can be evaluated through the Laplace transform:

$$\begin{aligned} W_{j,k}^*(s; W > 0, \text{Sr}) &:= \int_0^\infty e^{-st} P\{W > t, \text{Sr} \mid N = j, A = k\} dt \\ &= \sigma \sum_{m=1}^K m \int_0^\infty e^{-st} dt \int_t^\infty e^{-\theta x} g_{K-m,m}^{j,k}(x) dx \\ &= \sigma \sum_{m=1}^K m \int_0^\infty e^{-\theta x} g_{K-m,m}^{j,k}(x) dx \int_0^x e^{-st} dt \\ &= \frac{\sigma}{s} \sum_{m=1}^K m \int_0^\infty e^{-\theta x} (1 - e^{-sx}) g_{K-m,m}^{j,k}(x) dx \\ &= \frac{\sigma}{s} \sum_{m=1}^K m \left[g_{K-m,m}^{*j,k}(\theta) - g_{K-m,m}^{*j,k}(s + \theta) \right]. \end{aligned}$$

Thus we get

$$\begin{aligned}
 E[W; W > 0, \text{Sr} \mid N = j, A = k] &= W_{j,k}^*(0; W > 0, \text{Sr}) \\
 &= -\sigma \sum_{m=1}^K m \left[\frac{dg_{K-m,m}^{*j,k}(s)}{ds} \right]_{s=\theta}.
 \end{aligned}$$

As explained above, we can compute this quantity numerically.

Then we can obtain the conditional mean waiting time of a customer who waits and gets served by

$$\begin{aligned}
 E[W \mid W > 0, \text{Sr}] &= \frac{E[W; W > 0, \text{Sr}]}{P\{W > 0, \text{Sr}\}} \\
 &= \frac{\sum_{j=0}^{J-1} \sum_{k=\max\{0, K-j\}}^K E[W; W > 0, \text{Sr} \mid N = j, A = k] \hat{P}_{j,k}}{\sum_{j=0}^{J-1} \sum_{k=\max\{0, K-j\}}^K P\{W > 0, \text{Sr} \mid N = j, A = k\} \hat{P}_{j,k}},
 \end{aligned}$$

where $\hat{P}_{j,k}$ is given in Section 3.2. The conditional mean waiting time of a customer who is served is given by

$$E[W \mid \text{Sr}] = \frac{E[W; W = 0] + E[W; W > 0, \text{Sr}]}{P\{W = 0\} + P\{W > 0, \text{Sr}\}} = \frac{E[W; W > 0, \text{Sr}]}{P\{\text{Sr}\}}.$$

4.4. Waiting Time of a Customer Who Abandons

For a customer who arrives when there are j other customers in the system and k servers are working for ACW, the joint probability that he abandons waiting and the distribution of his waiting time is given by

$$P\{W > t; \text{Ab} \mid N = j, A = k\} = \sum_{l=0}^j \sum_{m=\max\{0, K-j\}}^{\min\{j+k-l, k\}} \int_t^\infty \theta e^{-\theta x} g_{l,m}^{j,k}(x) dx,$$

where the summation is taken over all possible customer states shown in Figure 13, because the abandonment can take place at any time. The integrand $\theta e^{-\theta x} g_{K-m,m}^{j,k}(x) dx$ is the probability that the customer has waited x time units and he abandons in the next dx time units.

The probability of abandonment is given by

$$P\{\text{Ab} \mid N = j, A = k\} = P\{W > 0, \text{Ab} \mid N = j, A = k\}$$

$$\begin{aligned}
&= \theta \sum_{l=0}^j \sum_{m=\max\{0, K-j\}}^{\min\{j+k-l, k\}} \int_0^{\infty} e^{-\theta x} g_{l,m}^{j,k}(x) dx \\
&= \theta \sum_{l=0}^j \sum_{m=\max\{0, K-j\}}^{\min\{j+k-l, k\}} g_{l,m}^{*j,k}(\theta),
\end{aligned}$$

which leads to the unconditional probability of abandonment

$$P\{\text{Ab}\} = \sum_{j=0}^{J-1} \sum_{k=\max\{0, K-j\}}^K P\{\text{Ab} \mid N = j, A = k\} \hat{P}_{j,k}.$$

The joint probability that the tagged customer abandons waiting and his mean waiting time until abandonment is given by

$$\begin{aligned}
E[W; \text{Ab} \mid N = j, A = k] &= \int_0^{\infty} P\{W > t, \text{Ab} \mid N = j, A = k\} dt \\
&= \theta \sum_{l=0}^j \sum_{m=\max\{0, K-j\}}^{\min\{j+k-l, k\}} \int_0^{\infty} dt \int_t^{\infty} e^{-\theta x} g_{l,m}^{j,k}(x) dx.
\end{aligned}$$

Again, this can be evaluated through the Laplace transform:

$$\begin{aligned}
W_{j,k}^*(s; \text{Ab}) &= \int_0^{\infty} e^{-st} P\{W > t, \text{Ab} \mid N = j, A = k\} dt \\
&= \theta \sum_{l=0}^j \sum_{m=\max\{0, K-j\}}^{\min\{j+k-l, k\}} \int_0^{\infty} e^{-st} dt \int_t^{\infty} e^{-\theta x} g_{l,m}^{j,k}(x) dx \\
&= \theta \sum_{l=0}^j \sum_{m=\max\{0, K-j\}}^{\min\{j+k-l, k\}} \int_0^{\infty} e^{-\theta x} g_{l,m}^{j,k}(x) dx \int_0^x e^{-st} dt \\
&= \frac{\theta}{s} \sum_{l=0}^j \sum_{m=\max\{0, K-j\}}^{\min\{j+k-l, k\}} \int_0^{\infty} e^{-\theta x} (1 - e^{-sx}) g_{l,m}^{j,k}(x) dx \\
&= \frac{\theta}{s} \sum_{l=0}^j \sum_{m=\max\{0, K-j\}}^{\min\{j+k-l, k\}} \left[g_{l,m}^{*j,k}(\theta) - g_{l,m}^{*j,k}(s + \theta) \right].
\end{aligned}$$

Thus we get

$$E[W; \text{Ab} \mid N = j, A = k] = W_{j,k}^*(0; \text{Ab})$$

$$= -\theta \sum_{l=0}^j \sum_{m=\max\{0, K-j\}}^{\min\{j+k-l, k\}} \left[\frac{dg_{l,m}^{*j,k}(s)}{ds} \right]_{s=\theta}$$

and then

$$E[W | Ab] = \frac{E[W; Ab]}{P\{Ab\}}$$

$$= \frac{\sum_{j=0}^{J-1} \sum_{k=\max\{0, K-j\}}^K E[W; Ab | N = j, A = k] \hat{P}_{j,k}}{\sum_{j=0}^{J-1} \sum_{k=\max\{0, K-j\}}^K P\{Ab | N = j, A = k\} \hat{P}_{j,k}}$$

In Figure 14, we plot the conditional mean waiting times $E[W | Sr]$ and $E[W | Ab]$ together with the unconditional mean waiting time $E[W]$ (already shown in Figure 8) for the case $\mu = \frac{1}{30}$, $\sigma = \frac{1}{25}$, and $\theta = \frac{1}{15}$. For $\lambda < 0.5$, most customers are served as shown in Figure 9. Therefore, we have $E[W | Sr] \approx 0$. Note that $E[W]$ always falls between $E[W | Sr]$ and $E[W | Ab]$, which results from the convex equality

$$E[W | Sr] \cdot P\{Sr\} + E[W | Ab] \cdot P\{Ab\} = E[W],$$

$$P\{Sr\} + P\{Ab\} = 1.$$

For example, at the arrival rate $\lambda = 0.8$, we have

$$E[W | Sr] = 1.6173 \quad ; \quad P\{Sr\} = 0.8845,$$

$$E[W | Ab] = 2.6178 \quad ; \quad P\{Ab\} = 0.1155,$$

and $E[W] = 1.7329$.

5. Concluding Remarks

In this paper, we have studied a queueing model for a call center with impatient customers and after-call work by formulating the problem as a two-dimensional birth-and-death process. We have derived a variety of performance measures that can be calculated from the state probability of the birth-and-death process. In addition, we have presented a method of finding the mean waiting time for those customers who get served and for those who abandon while waiting

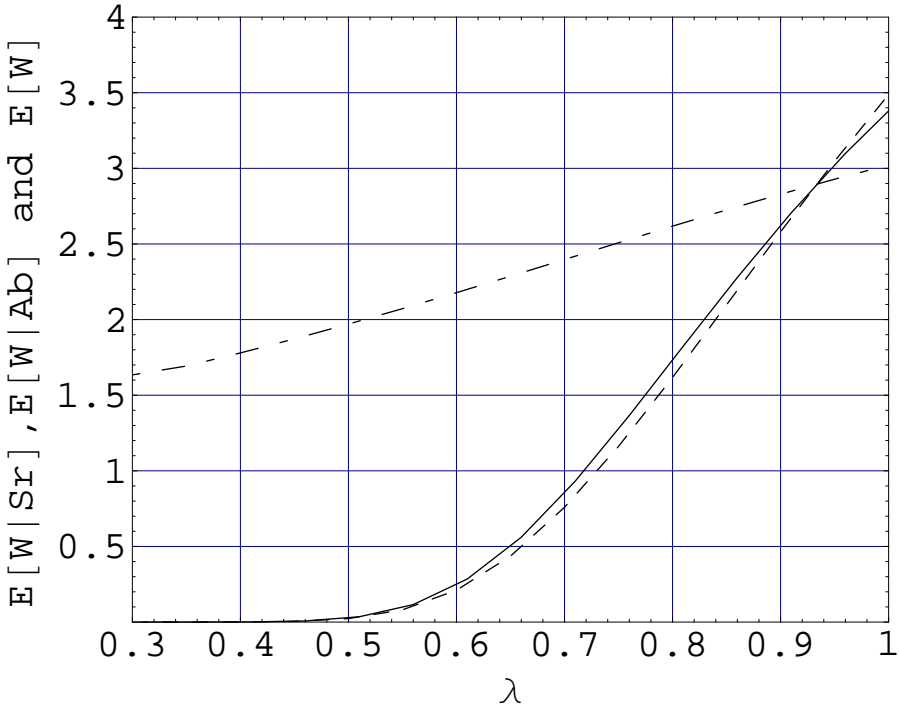


Figure 14: Conditional mean waiting times $E[W | Sr]$ (below) and $E[W | Ab]$ (above) with the unconditional mean waiting time $E[W]$ (middle) ($\mu = \frac{1}{30}, \sigma = \frac{1}{25}$, and $\theta = \frac{1}{15}$)

separately. These analytical results could be used in the capacity planning of a call center as an inverse problem. For example, we can determine the number of operators needed to attain the specified value of the probability of wait.

An important performance indicator of a call center, used in the *service level agreement* between the user and the provider of the service, is the percentage of incoming calls that are answered within a specified threshold such as “ X percent of calls answered in Y seconds” [1, Section 8]. This indicator refers to the joint probability that a customer is served and the distribution of his waiting time, which can be calculated by

$$P\{W > t, Sr\} = \sum_{j=0}^{J-1} \sum_{k=\max\{0, K-j\}}^K P\{W > t, Sr | N = j, A = k\} \hat{P}_{j,k}.$$

An equally interesting measure is the probability of abandonment after a customer has waited t time units as the fraction of poorly-served customers [12]. This is calculated by

$$\begin{aligned}
 P\{\text{Ab} \mid W > t\} &= \frac{P\{W > t, \text{Ab}\}}{P\{W > t\}} \\
 &= \frac{\sum_{j=0}^{J-1} \sum_{k=\max\{0, K-j\}}^K P\{W > t, \text{Ab} \mid N = j, A = k\} \hat{P}_{j,k}}{P\{W > t, \text{Sr}\} + P\{W > t; \text{Ab}\}}.
 \end{aligned}$$

At this moment, we do not have closed-form expressions for these measures, and we can calculate them only numerically.

We have not addressed the effects of retrials by blocked and abandoned calls for two reasons. One reason is that the exact treatment of retrial calls involves very complicated, if not impossible, analysis even for a small number of servers, which makes it infeasible to handle the model of a call center of realistic size with dozens of operators. The other reason is that we have intrinsic difficulty in getting the statistical data about the retrial calls from cooperating call centers. The call center can collect any statistics about accepted calls, but there is no way to determine whether an incoming call is the retrial of a previously blocked call or a new call, to measure the inter-retrial time for a retrial call, or to know the intention of a customer whether he makes a retrial call at all using their equipments inside the center.

We are also aware of the fact that operators spend significant amount of time on things not related to processing calls from customers or ACW such as attending internal meetings, taking a break, and so on while they are in the office. These actions are called the *shrinkage* or *rostered staff factor* (RST) [1, Section 8]. Therefore, we need to develop a model that takes account of the shrinkage for the practical application to workforce management of call centers.

Acknowledgements

The authors are grateful to Professor Emeritas Chiaki Hishinuma of Tokyo University of Technology for his valuable advice on the capacity management of call centers. This work is supported by the Grant-in-Aid for Scientific Research 2008–2010, No.20500061, from the Japan Society for the Promotion of Science (JSPS).

References

- [1] B. Cleveland and D. Harne (editors), *Call Center Operations Management: Handbook and Study Guide*, Version 2.1, ICMI Press, Colorado Springs (2004).
- [2] M. J. Fischer, D. A. Garbin and A. Gharakhanian, Performance modeling of distributed automatic call distribution systems, *Telecommunication Systems*, **9**, No.2 (1998), 133–152, **doi:** 10.1023/A:1019139721840.
- [3] N. Gans, G. Koole and A. Mandelbaum, Telephone call centers: tutorial, review, and research prospects, *Manufacturing & Service Operations Management*, **5**, No.2 (2003), 79–141, **doi:** 10.1287/msom.5.2.79.16071.
- [4] C. M. Harris, K. L. Hoffman and P. B. Saunders, Modeling the IRS telephone taxpayer information system, *Operations Research*, **35**, No.4 (1987), 504–523, **doi:** 10.1287/opre.35.4.504.
- [5] R. J. Harris and M. J. Phillips, 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](http://www-ist.massey.ac.nz/rharris/publicationfiles/pre2000/acdsystems-foatrs.pdf)
- [6] C. Hishinuma and H. Takagi, Capacity management of call centers, *The Journal of the Institute of Electronics, Information and Communication Engineers*, **94**, No.9 (2011), 767–772 (in Japanese).
- [7] W. M. Jolley and R. J. Harris, Analysis of post-call activity in queueing systems, *Proceedings of the 9th International Teletraffic Congress* (Torremolinos, 1979), 1–9.
- [8] W. M. Jolley and R. J. Harris, Analysis and optimal design of queueing systems with post-call activity, *Australian Telecommunication Research*, **15**, No.1 (1981), 12–26.
- [9] K. Kawanishi, On the counting process for a class of Markovian arrival processes with an application to a queueing system, *Queueing Systems*, **49**, No.2 (2005), 93–122, **doi:** 10.1007/s11134-005-6478-7.
- [10] K. Kawanishi, Waiting time distribution of a queueing system with post-service activity. *Computers & Mathematics with Applications*, **51**, No.2 (2006), 209–218, **doi:** 10.1016/j.camwa.2005.11.021.

- [11] G. Koole and A. Mandelbaum, Queueing models of call centers: an introduction, *Annals of Operations Research*, **113**, Nos.1–4 (2002), 41–59, **doi:** 10.1023/A:1020949626017.
- [12] A. Mandelbaum and S. Zeltyn, Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. In D. Spath and K.-P. Fährnich (eds.): *Advances in Services Innovations*, Springer, Berlin, (2007), 17–45.
- [13] T. Phung-Duc and K.Kawanishi, Multiserver retrial queues with after-call work, *Numerical Algebra, Control and Optimization*, **1**, No.4 (2011), 639–656, **doi:** 10.3934/naco.2011.1.639.
- [14] Y. Taguchi, Analysis of the Queueing Model for a Call Center with After-Call Work, Master Thesis, Master Program in Policy and Planning Science, University of Tsukuba, Japan, March 2011 (in Japanese).
- [15] W. Whitt, Engineering solution of a basic call-center model, *Management Science*, **51**, No.2 (2005), 221–235, **doi:** 10.1287/mnsc.1040.0302.

