Omega 41 (2013) 280–286

Contents lists available at SciVerse ScienceDirect

Omega

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

The impact of a waiting-time threshold in overflow systems with impatient customers

Raik Stolletz [a,*], Michael Manitz [b]

[a] University of Mannheim, Business School, Chair of Production Management, Schloss, 68131 Mannheim, Germany
[b] University of Duisburg/Essen, Chair of Production and Supply Chain Management, Mercator School of Management, Lotharstr. 65, 47057 Duisburg, Germany

ARTICLE INFO

Article history:
Received 26 April 2011
Accepted 1 May 2012
Processed by B. Lev
Available online 15 May 2012

Keywords:
Service operations
Queueing
Call center
Overflow
Impatient customers
Markov Chain

ABSTRACT

This paper analyzes the performance of call centers with impatient customers, two levels of support, and an overflow mechanism. Waiting calls from the front-office queue – if not reneging – are sent to the back office if at least one back-office agent is available and if a certain threshold t on the waiting time is reached. We approximate such systems via a continuous-time Markov chain that allows for overflow immediately upon arrival. Two different approaches for the derivation of the respective probability of an overflow are developed. Numerical results compare the reliability of these Markovian performance approximations for different parameter settings. The impact of the threshold t on different performance measures is shown dependent on the impatience of customers.

1. Introduction

Service systems often offer services in multiple stages. A front office serves a majority of customers with basic requests, while back offices provide special services. The challenging task in planning such systems is to take into account stochastic variations in the arrival process, stochastic service times, and the random patience of waiting customers. To reduce the risk of having customers waiting an excessive amount of time or to lose them due to impatience, overflow to another agent group may be organized. This is a common feature in many automatic call distribution (ACD) systems, see, for example, Lucent Technologies [14]. To improve the service offered to the calling customers, an overflow from the front-office queue to the back office may occur, depending on the availability of back-office agents.

The existing literature discusses three conceptual types of overflows in call centers:

1. In models with *state-dependent overflows*, the overflow depends on the number of customers in the system, but not on the waiting time. Models that implement an overflow mechanism if all agents of a dedicated group are busy are analyzed in Chevalier and Tabordon [5], Franx et al. [8], Gans and Zhou [9], and Sendfeld [16]. Finite waiting-room systems with an overflow of blocked customers are analyzed in Guerin and Lien [10].
2. Another class of approximation methods uses a *random threshold on the waiting time*. Such time-dependent overflows in a single-stage system can be interpreted as impatient calls leaving the queue after a random waiting time. Brandt and Brandt [4] present a Markovian approximation for $M/M/c+GI$ queues with generally distributed patience times, but without conditions on the availability of agents in other server groups. Down and Lewis [7] describe a call-center model with exponentially distributed waiting times before overflow to another agent group is triggered. In this case, overflow can be seen as exponentially distributed abandonment instead of the deterministic overflow we analyze in this paper. In fact, overflows after exponentially distributed waiting times reflect individual patience thresholds that cannot be implemented in automatic call distribution (ACD) systems.
3. Systems with an overflow that depends on a *fixed threshold on the waiting time* are analyzed in Bekker et al. [2] and Koole et al. [12,13] for single-stage systems with server groups working in parallel. Barth et al. [1] analyze a serial system with multiple servers in which overflow depends on a fixed threshold on the waiting time *and* on the availability of agents in the second stage.

Overflow mechanisms with a fixed threshold t on the waiting time are often used in practice, while t being a managerial decision variable. In addition, the impatience of customers influences the

* Corresponding author. Tel.: +49 621 181 1577.
E-mail addresses: stolletz@bwl.uni-mannheim.de (R. Stolletz), michael.manitz@uni-due.de (M. Manitz).

0305-0483/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.omega.2012.05.001

performance measures of call centers and has to be integrated into agent staffing; see for example Dietz [6] for a single-stage system. For staffing decisions, fast performance-evaluation methods are needed. Because simulation experiments require too much computation time, analytical approximation approaches are necessary. To the best of our knowledge, there does not exist any approximation method that considers both the fixed threshold t on the waiting time and the impatience of customers. To analyze the impact of impatient customers onto the performance of these overflow systems, we integrate the impatience into the model studied in Barth et al. [1].

The main contribution of this paper is as follows: Besides considering the impatience, we develop a new approximation for the overflow probabilities in a continuous-time Markov chain: the queue-length based approach. This approximation scheme is computationally simpler and gives very comparable results to the alternative waiting-time based overflow approach. Furthermore, the numerical studies show the impact of setting the waiting-time threshold t onto several performance measures dependent on the impatience of customers.

The remainder of the paper is organized as follows. Section 2 describes the assumptions of the analyzed system. The Markovian performance approximations are developed in Section 3. In Section 4, the numerical results for both approximation approaches are presented. The reliability of both approaches is shown in comparison to a simulation study.

2. Description of the queueing system

We analyze a two-level service system with overflow after a waiting-time threshold. A similar model is already analyzed with completely patient customers in Barth et al. [1]. The system (as depicted in Fig. 1) provides two levels of service. A first level of support is offered by a total of c_F agents in the front office, whereas c_B agents are working at the second level, that is, the back office. The capacity of the front office is limited to K_F customers (either waiting or in service). A fraction b of all customers requires additional second-level support in the back office. Customers with further service requirements are routed to the back office with an overall capacity of K_B inbound calls in queue and in service. Systems with an unlimited capacity in the back office could be analyzed by setting K_B large enough. There is no queue for overflow calls in the back office. These calls are sent to the back office only if there are agents are available.

We assume a Poisson call-arrival process in the front office with an arrival rate λ. A call is blocked and not accepted if the capacity K_F of the front office is exhausted. Otherwise, the customer joins the queue until receiving service as soon as a front-office agent is available to him/her. We assume a first-come first-served (FCFS) queueing discipline. If the waiting time W_F of a customer in the front-office queue exceeds the threshold t, the customer will be routed to the back office by an overflow mechanism if back-office agents are available. If not, overflow is not allowed, and the customer must wait even longer than t time units in the front-office queue.

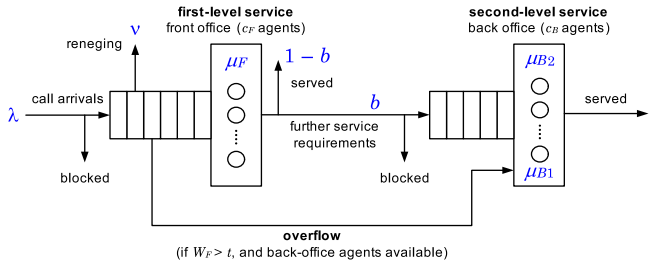

Fig. 1. The two-level call center with time-dependent overflow and impatience.

The customers have a limited patience in the front-office-queue (i.e., time to abandonment), which is assumed to be exponentially distributed with rate ν. Customers with a waiting time that exceeds the individual patience threshold leave the queue without receiving service. Usually, those customers that have been routed to the back office do not renege. Therefore, we consider unlimited patience in the back-office queue.

All service times are exponentially distributed random variables. For the front office, the service rate of an agent is described by μ_F. After finishing service at the first level, a fraction b of all customers need additional service that is offered in the back office. The service rate for such an original second-level call with further service requirements is denoted by μ_{B2}. Since this special service usually is time consuming, the second-level service rates are typically lower than the service rates in the front office. In comparison to that, the rate μ_{B1} at which back-office agents serve overflow calls with first-level service requirements is larger than μ_{B2} but usually lower than that of trained front-office agents. Hence, $\mu_F < \mu_{B1} < \mu_{B2}$ often is observed. Although the quality of our performance-evaluation procedure does not depend on this assumption, the numerical study in Section 4 is designed according to this relation.

We study systems with a time-dependent overflow of first-level calls from the front-office queue to the back office. The back-office agents serve second-level calls with non-preemptive priority. As long as this queue is not empty, overflow cannot occur. Only if this queue is empty, a back-office agent picks up the longest waiting customer from the front-office queue if the customer's waiting time W_F has exceeded the threshold t.

3. CTMC analysis with impatient customers

3.1. Idea of the performance approximation

The basic idea behind the performance analysis of call centers with overflow is to model the queueing system presented in Section 2 as a continuous-time Markov chain (CTMC). To ensure the Markovian property of memorylessness, the overflow rule with a deterministic threshold t on waiting time is replaced by an immediate overflow upon arrival with a certain probability $p_n^{(t)}$. This probability depends on the threshold t and on the number n of customers waiting in front of the arriving one.

Two different approaches for determining these overflow probabilities $p_n^{(t)}$ are discussed in Section 3.2, namely, the waiting-time based overflow (WTBO) and the queue-length based overflow (QLBO). In the first approach, the probability $p_n^{(t)}$ is approximated by the probability that the waiting time of an arrival in a queueing system without overflow and n customers in queue would exceed the limit t. This approach integrates the impatience of customers into the performance-analysis method that is presented by Barth et al. [1]. This extension to a system with impatient customers requires a new derivation of the probability $p_n^{(t)}$ that an arriving customer would wait more than t seconds until getting service (WTBO approach). Apart from that, in the second approach (QLBO), the overflow – for an arriving customer – is possible as long as a certain critical queue length $\overline{n}$ is reached, i.e., $p_n^{(t)}$ is set to 1 for queue lengths $n \geq \overline{n}$, and 0 otherwise. The critical queue length $\overline{n}$ is defined as the expected number of customers that can be served during t time units.

Using these probabilities $p_n^{(t)}$ of an immediate overflow allows for a CTMC-based performance approximation. While the state space of the CTMC is the same as in Barth et al. [1], the transition