Barth, Wolfgang; Manitz, Michael; Stolletz, Raik

Working Paper

Analysis of two-level support systems with time-dependent overflow: a banking application

Diskussionsbeitrag, No. 399

Provided in Cooperation with:
School of Economics and Management, University of Hannover

Suggested Citation: Barth, Wolfgang; Manitz, Michael; Stolletz, Raik (2008) : Analysis of two-level support systems with time-dependent overflow: a banking application, Diskussionsbeitrag, No. 399, Leibniz Universität Hannover, Wirtschaftswissenschaftliche Fakultät, Hannover

This Version is available at:
https://hdl.handle.net/10419/27208

Analysis of Two-Level Support Systems with Time-Dependent Overflow — A Banking Application

Wolfgang Barth

Hochschule der Sparkassen-Finanzgruppe
University of Applied Sciences
Professor of Banking, Marketing and Sales
Simrockstr. 4, 53113 Bonn, GERMANY
wolfgang.barth@dsgv.de
Phone: +49(0)228 204-931, Fax: +49(0)228 204-903

Michael Manitz

University of Cologne
Department for Supply Chain Management and Production
Albertus-Magnus-Platz, 50923 Köln, GERMANY
manitz@wiso.uni-koeln.de
Phone: +49(0)221 470-2730, Fax: +49(0)221 470-5140

Raik Stolletz

Leibniz University Hannover
Department for Production Management
Königsworther Platz 1, 30167 Hannover, GERMANY
raik.stolletz@prod.uni-hannover.de
Phone: +49(0)511 762-5649, Fax: +49(0)511 762-4863

May 20, 2008

Analysis of Two-Level Support Systems with Time-Dependent Overflow — A Banking Application

Abstract

In this paper, we analyze the performance of call centers of financial service providers with two levels of support and a time-dependent overflow mechanism. Waiting calls from the front-office queue flow over to the back office, if a waiting-time limit is reached and at least one back-office agent is available. The analysis of such a system with time-dependent overflow is reduced to the analysis of a continuous-time Markov chain with state-dependent overflow probabilities. To approximate the system with time-dependent overflow, some waiting-based performance measures are modified. Numerical results demonstrate the reliability of this Markovian performance approximation for different parameter settings. A sensitivity analysis shows the impact of the waiting-time limit and the dependence of the performance measures on the arrival rate.

Key words Financial Service Operations – Performance Evaluation – Queueing Models of Call Centers – Time-Dependent Overflow

1 Introduction

Customers of financial service institutes use different communication channels for getting service, for example e-mail, telephone, or a visit at the local agency. Recent studies for the German financial sector quote that in 2010 about 60 % of the customers will be multi-channel clients (Association of German Banks (2004)). In the German banking sector, 72 % of the companies are working only in local districts (Deutsche Bundesbank (2008)),

and, therefore, run medium- and small-sized call centers. The design of such customer-oriented service systems has to provide sufficient capacity to achieve a certain service level. About 60–80 % of the total costs for carrying out service operations in a call center result from the staff (Aksin, Armony, and Mehrotra (2007)). To set staffing levels for different time intervals, fast approximation methods are necessary for the performance analysis.

We consider a call center of a financial service provider with two levels of support. All calls require a first-level support at the front office, but the front-office agents are trained to manage only a subset of the service requests of the customers. A fraction b of all customers requires an additional second-level support in the so-called back office. Back offices usually provide special services, and the back-office agents have a higher qualification than the first-level support agents. The mean call time in a back office is often longer than in the front office. In addition, back-office agents provide services not only via phone but also via e-mail, fax, or they do desk work. It is important to note that this work will be preempted if a phone call arrives in the back office.

The special feature of the call center analyzed in this paper is that a time-dependent overflow from the front-office queue to back-office agents may occur. In this system, a waiting customer in the front-office queue will be routed to an available back-office agent if the waiting time exceeds a given limit of t time units. Time-dependent overflow mechanisms are a common feature in many Automatic Call Distribution (ACD) systems, see for example Lucent Technologies (1999). The managerial objective of such a time-

dependent overflow is to avoid large waiting times in the front-office queue while back-office agents are available at the same time.

This paper studies medium- and small-sized call centers of financial service providers. Due to economies of scale, the stochastics in the system due to random inter-arrival- and processing times has a larger impact on the performance measures in medium- and small-sized systems than in large-scale call centers. Therefore, the performance analysis via (quasi-)deterministic fluid models is not appropriate; see Jimenéz and Koole (2004) and Stolletz (2008). In our approach, the performance of the system is analyzed with a stochastic model in steady state. This analysis does not capture time-varying arrival rates and time-varying number of agents. Nevertheless, the performance measures of non-stationary systems can be approximated using stationary models, for example via the Stationary Independent Period by Period (SIPP) approach or the Stationary Backlog Carryover (SBC) approach; see Green and Kolesar (1991), Green, Kolesar, and Whitt (2007), and Stolletz (2008).

There is a large body of literature concerning the operations management of call centers. Surveys on the recent literature can be found in Koole and Mandelbaum (2002), Gans, Koole, and Mandelbaum (2003), Stolletz (2003), and Aksin et al. (2007). Pinedo, Seshadri, and Shanthikumar (2000) give an overview on call-center operations in financial services. Queueing models of call centers with parallel server groups and skill-based routing are analyzed for different system designs, see for example Stolletz and Helber

(2004), Wallace and Whitt (2005), and the references therein. The system analyzed in this paper has serially organized agent groups. Different overflow policies for such serial server groups are described in the literature. Models with an overflow if all agents of a dedicated group are busy are analyzed in Chevalier and Tabordon (2003), Franx, Koole, and Pot (2006), Gans and Zhou (2007), and Sendfeld (2007). Finite waiting-room systems with an overflow of blocked customers are analyzed in Guerin and Lien (1990). All these models assume a state-dependent overflow instead of a time-dependent overflow. The time-dependent overflow in a single-stage system can be interpreted as an impatient call leaving the queue after some (possibly random) waiting time. Brandt and Brandt (2002) present a Markovian approximation for $M/M/c+GI$ queues with generally distributed patience times. With this approach, the special case of a deterministic waiting-time tolerance is equal to the case of an overflow after a fixed waiting-time limit (as assumed in this paper), but without conditions on the availability of agents in other server groups. Down and Lewis (2007) describe a call-center model with exponentially distributed waiting times before an overflow to another agent group occurs. This can be seen as an exponentially distributed abandonment instead of the deterministic abandonment we are analyzing in this paper. To the best of our knowledge a system with an overflow that depends on the waiting time as well as on the availability of agents in the back office has not been analyzed in the literature yet.

The remainder of the paper is organized as follows: Section 2 describes the assump-

tions of the analyzed model with time-dependent overflow. A Markovian performance approximation is developed in Section 3. The time-dependent overflow with constant time limit t is approximated via an overflow with the queue-length dependent probability that the waiting time of an arriving call will exceed the time limit t. We formulate a Markov model for this approximation of the real system and derive different performance measures. To capture the feature of the time-dependent overflow, some waiting-based performance measures are modified. The results for this approximation are compared to simulation results in the numerical study in Section 4. We conclude with suggestions for further research in Section 5.

2 Model Description

We analyze the following model of a two-level service system with time-dependent overflow, as depicted in Figure 1. The front office consists of c_F agents, and c_B agents are working in the back office. The system has two queues, one for the front-office calls, and one for calls that are routed to the back office. The number of customers in the system (in queue and in service) is restricted by K_F and K_B, respectively.

All customers arrive at the front office according to a Poisson process with rate λ. If an agent of the front-office team is available, the incoming call will be routed immediately to this agent. If all front-office agents are busy and the capacity $K_F - c_F$ of the front-office queue is not exhausted, the call joins the front-office queue. Otherwise, the call is blocked

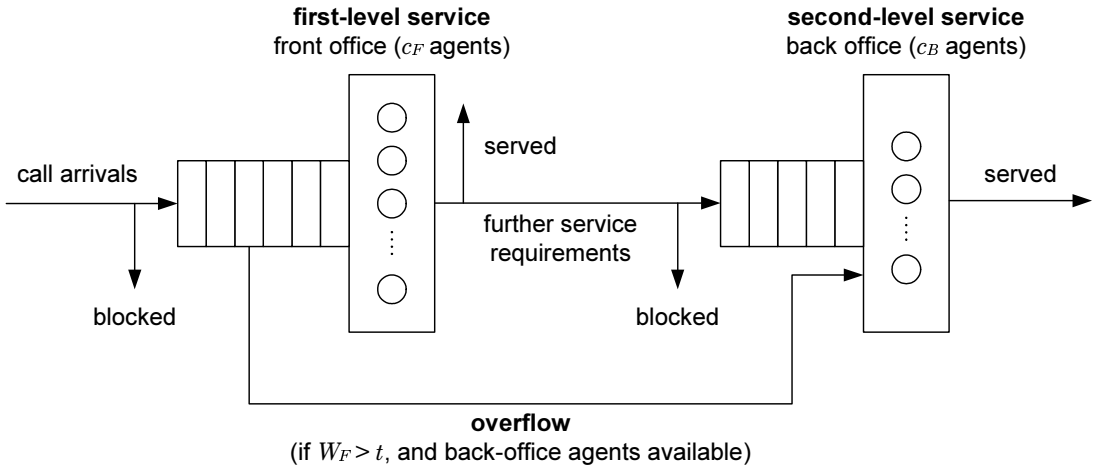

Figure 1: The two-level call center with time-dependent overflow

and leaves the system without getting service. After finishing service in the first level, a fraction b of all customers need additional service in the back office. If a second level agent is available, the call will be routed to him immediately. Otherwise, the call will be queued in the back-office queue upon the back-office capacity K_B is reached. For both queues, we assume a first-come first-served (FCFS) discipline.

If the waiting time W_F of a call in the front-office queue exceeds t time units and at least one back-office agent is available, this call flows over to the back office. If all back-office agents are busy, the call still waits in the front-office queue and can be served by agents of both groups. The back-office agents serve calls from the back-office queue with non-preemptive priority. That means, if a back-office agent becomes available, he takes the longest waiting call from the back-office queue next. If there is no call waiting in the back-office queue, this agent takes the longest waiting call from the front-office queue,

if its waiting time W_F has exceeded t.

The service times in the front office are assumed to be exponentially distributed with rate μ_F. In the considered banking application, the service time in the front office is independent of the requirement of back-office services. In the back office, the service-time distribution usually depends on where the call came from. The service time of a second-level call is assumed to be exponentially distributed with rate μ_{B2}. The rate at which first-level calls from an overflow can be served by a back-office agent is μ_{B1}.

3 Markovian Performance Approximation

3.1 Outline of the Approach

The call center with an overflow as described in Section 2 cannot be analyzed as a continuous-time Markov chain (CTMC) because of the deterministic waiting-time limit. Therefore, we replace the overflow rule of the original system with an overflow that occurs immediately upon arrival with a certain probability. For this system, a CTMC analysis is possible and we derive different performance measures. To approximate the performance measures for the original system with a time-dependent overflow, some of these performance measures have to be modified.

3.2 Modelling of the Overflow Mechanism

The overflow after a deterministic waiting-time limit t is replaced by an overflow immediately upon arrival with the state-dependent probability $p_n^{(t)}$. The probability $p_n^{(t)}$ that

an arriving customer waits longer than t time units until getting service depends on the number n of customers waiting in front of that customer. If all c_F first-level agents are occupied and at least one back-office agent is available, then an arriving call will be routed into the back office immediately upon arrival with this probability $p_n^{(t)}$.

The random waiting time in the front-office queue corresponds to the conditional waiting time of an arrival in a queueing system without overflow and n customers in queue. Upon such an arrival, the new calling customer must wait for $n+1$ service completions before reaching an agent. Hence, the waiting time T_n^{W} of an arriving customer that sees n customers in queue is the sum of $n+1$ exponential service completion times, i. e. an Erlang-k distributed random variable with $k = n+1$ and a service rate $c_F \cdot \mu_F$. For a particular random variable X (with $x > 0$), the Erlang-k distribution with rate λ has the probability density function

$$f_X(x) = \frac{\lambda^k \cdot x^{k-1}}{(k-1)!} \cdot e^{-\lambda x} \quad . \tag{1}$$

Therefore, the probability $p_n^{(t)}$ to wait longer than t time units if n customers are waiting in front of the arriving call is

$$\begin{aligned} p_n^{(t)} = P\left(T_n^{\mathrm{W}} > t\right) = 1 - P\left(T_n^{\mathrm{W}} \leq t\right) &= 1 - \int_0^t \frac{(c_F \cdot \mu_F)^{n+1} \cdot x^n}{n!} \cdot e^{-(c_F \cdot \mu_F)x} \,\mathrm{d}x \\ &= 1 - \frac{1}{n!} \int_0^{c_F \cdot \mu_F \cdot t} x^n \, e^{-x} \,\mathrm{d}x \quad . \\ & \qquad\qquad (n = 0, 1, \ldots, K_F - c_F - 1) \end{aligned} \tag{2}$$

Using a series representation for the incomplete gamma function, one can find (see Press

et al. (1992))

$$\begin{aligned} p_n^{(t)} &= 1 - \frac{1}{n!} \cdot e^{-c_F \cdot \mu_F \cdot t} \left(c_F \cdot \mu_F \cdot t\right)^{n+1} \sum_{i=0}^{\infty} \frac{n!}{(n+1+i)!} \left(c_F \cdot \mu_F \cdot t\right)^i \\ &= 1 - e^{-c_F \cdot \mu_F \cdot t} \sum_{i=0}^{\infty} \frac{\left(c_F \cdot \mu_F \cdot t\right)^{n+1+i}}{(n+1+i)!} = 1 - e^{-c_F \cdot \mu_F \cdot t} \sum_{k=n+1}^{\infty} \frac{\left(c_F \cdot \mu_F \cdot t\right)^k}{k!} \\ &= 1 - e^{-c_F \cdot \mu_F \cdot t} \left(e^{c_F \cdot \mu_F \cdot t} - \sum_{k=0}^{n} \frac{\left(c_F \cdot \mu_F \cdot t\right)^k}{k!} \right) = e^{-c_F \cdot \mu_F \cdot t} \sum_{k=0}^{n} \frac{\left(c_F \cdot \mu_F \cdot t\right)^k}{k!} \\ & \qquad\qquad (n = 0, 1, \ldots, K_F - c_F - 1) \quad (3) \end{aligned}$$

which is obviously a cumulative distribution function of a Poisson-distributed random variable with mean $c_F \cdot \mu_F \cdot t$. The effective service rate $c_F \cdot \mu_F$ represents the mean number of customers that can be served per unit of time. Then, $c_F \cdot \mu_F \cdot t$ is the mean number of customers that can be served during t time units. Hence, the probability $p_n^{(t)}$ that an arriving customer will wait more than t units of time (until n customers in front of him are served) corresponds to the probability that no more than n customers can be served during that time.

To ensure numerical tractability, the overflow probability $p_n^{(t)}$ is developed recursively via

$$p_n^{(t)} = e^{-c_F \cdot \mu_F \cdot t} \cdot S_n \qquad (n = 0, 1, \ldots, K_F - c_F - 1) \quad (4)$$

with

$$\begin{aligned} S_n &:= \sum_{k=0}^{n} \frac{\left(c_F \cdot \mu_F \cdot t\right)^k}{k!} = \sum_{k=0}^{n-1} \frac{\left(c_F \cdot \mu_F \cdot t\right)^k}{k!} + \frac{\left(c_F \cdot \mu_F \cdot t\right)^n}{n!} = S_{n-1} + \zeta_n \\ & \qquad\qquad (n = 1, \ldots, K_F - c_F - 1) \quad (5) \end{aligned}$$

where the last summand of S_n is

$$\zeta_n := \frac{(c_F \cdot \mu_F \cdot t)^n}{n!} = \frac{(c_F \cdot \mu_F \cdot t)^{(n-1)}}{(n-1)!} \cdot \frac{c_F \cdot \mu_F \cdot t}{n} = \zeta_{n-1} \cdot \frac{c_F \cdot \mu_F \cdot t}{n}$$
$$(n = 1, \ldots, K_F - c_F - 1) \quad (6)$$

with $S_0 \equiv 1$. Then, it follows:

$$p_n^{(t)} = \begin{cases} e^{-c_F \cdot \mu_F \cdot t} & (n = 0) \\ e^{-c_F \cdot \mu_F \cdot t} \cdot \left(S_{n-1} + \zeta_{n-1} \cdot \frac{c_F \cdot \mu_F \cdot t}{n} \right) & (n = 1, \ldots, K_F - c_F - 1) \end{cases} \quad (7)$$

The state-dependent overflow probabilities according to the Equations (2)–(7) can be integrated into the CTMC model to approximate the transition rates out of those states at which an overflow is possible.

3.3 The Markov Model with State-Dependent Overflow Probabilities

The CTMC with state-dependent overflow probabilities is described by the three-dimensional state vector (n_F, n_{B1}, n_{B2}). The number of customers waiting in the front-office queue or in service by front-office agents is denoted by $n_F = 0, 1, \ldots, K_F$. If all back-office agents are occupied, no first-level call flow over to the back office. Therefore, the number n_{B1} of first-level customers in service by back-office agents is bounded by c_B. The number n_{B2} describes the second-level calls that are waiting or being served in the back office with $n_{B2} = 0, 1, \ldots, K_B - n_{B1}$.

We divide the state space into three regions. In states of Region I, there is at least

one available front-office agent and at least one available back-office agent. In states of Region II, all c_F front-office agents are busy and at least one back-office agent is available, i. e. an overflow may occur. For states in Region III, it holds that all back-office agents are occupied. The steady-state probabilities are denoted by $\mathrm{P}\{n_F, n_{B1}, n_{B2}\}$. To capture the transitions into boundary states correctly, we use an indicator function $\mathit{1}(x)$ which is 1 if $x > 0$, or 0 otherwise.

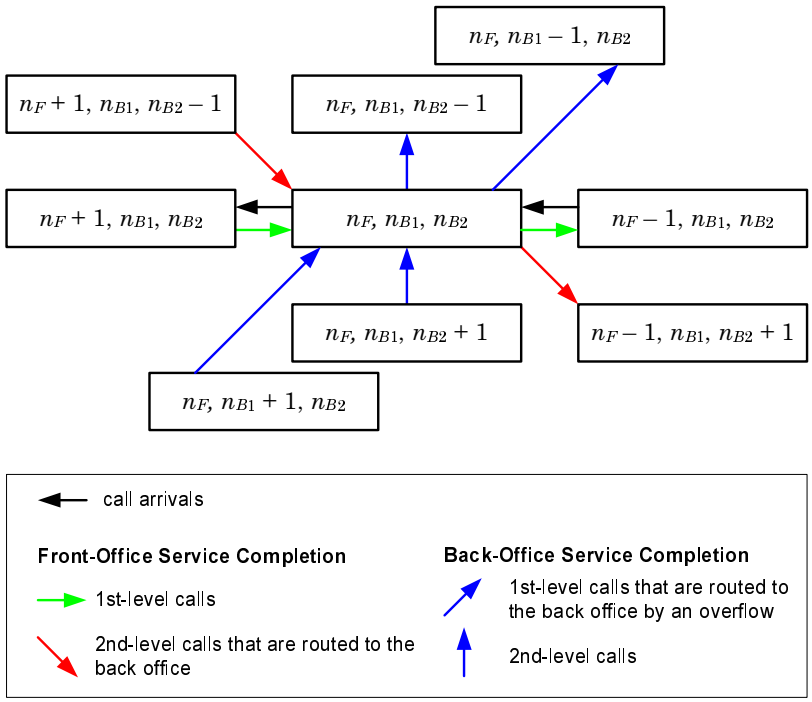

Figure 2: Connected states in case of Region I ($n_F < c_F$, $n_{B1} + n_{B2} < c_B$)

Region I: Available agents in both groups For states (n_F, n_{B1}, n_{B2}) of Region I with $n_F < c_F$ and $n_{B1} + n_{B2} < c_B$ an overflow cannot occur. The first-level customers

in the back-office service — if any — are remains of former overflows. The connected states around the state (n_F, n_{B1}, n_{B2}) are depicted in Figure 2. Starting from this state, the number of calls in the front-office queue is increased by 1 if customers arrive (with rate λ). It is decreased by 1 after a service completion of a customer in the front office. If n_F agents are busy the time until next service completion is exponentially distributed with rate $n_F \cdot \mu_F$. A fraction b of customers completing service in the front office will get additional service in the back office. With rate μ_{B2} such a second-level service is completed. The state (n_F, n_{B1}, n_{B2}) can be reached from state $(n_F + 1, n_{B1}, n_{B2} - 1)$ due to a call coming from the front office with rate $b \cdot (n_F + 1) \cdot \mu_F$. Starting from state $(n_F + 1, n_{B1}, n_{B2})$, customers complete the service by a front-office agent and leave the system with rate $(1-b) \cdot (n_F+1) \cdot \mu_F$. A service completion at the back office leads to a decrease of the number of back-office customers from $n_{B2} + 1$ down to n_{B2} for second-level calls, or from $n_{B1} + 1$ down to n_{B1} for first-level calls, respectively. Finally, the state (n_F, n_{B1}, n_{B2}) can be reached from state $(n_F - 1, n_{B1}, n_{B2})$ by an arrival. Equating the flow into the state (n_F, n_{B1}, n_{B2}) with the flow out of this state, it follows:

$$
\begin{aligned}
& b \cdot (n_F + 1) \cdot \mu_F \cdot \mathrm{P}\left\{n_F + 1, n_{B1}, n_{B2} - 1\right\} \cdot \mathit{1}(n_{B2}) \\
& + (1 - b) \cdot (n_F + 1) \cdot \mu_F \cdot \mathrm{P}\left\{n_F + 1, n_{B1}, n_{B2}\right\} \\
& + (n_{B2} + 1) \cdot \mu_{B2} \cdot \mathrm{P}\left\{n_F, n_{B1}, n_{B2} + 1\right\} \\
& + (n_{B1} + 1) \cdot \mu_{B1} \cdot \mathrm{P}\left\{n_F, n_{B1} + 1, n_{B2}\right\} \\
& + \lambda \cdot \mathrm{P}\left\{n_F - 1, n_{B1}, n_{B2}\right\} \cdot \mathit{1}(n_F) \\
& = (\lambda + n_F \cdot \mu_F + n_{B2} \cdot \mu_{B2} + n_{B1} \cdot \mu_{B1}) \cdot \mathrm{P}\left\{n_F, n_{B1}, n_{B2}\right\} \\
& \qquad\qquad (n_F = 0, \ldots, c_F - 1;\, n_{B1} + n_{B2} < c_B) \quad (8)
\end{aligned}
$$

The indicator variables $\mathit{1}(\cdot)$ allow to apply Equation (8) for internal as well as for boundary states. The boundary state $(0, n_{B1}, n_{B2})$ cannot be reached by an arrival. In this situation, all front-office agents are available. The state $(n_F, n_{B1}, 0)$ cannot be reached via a transfer of front-office calls with additional service requirements to the back office.

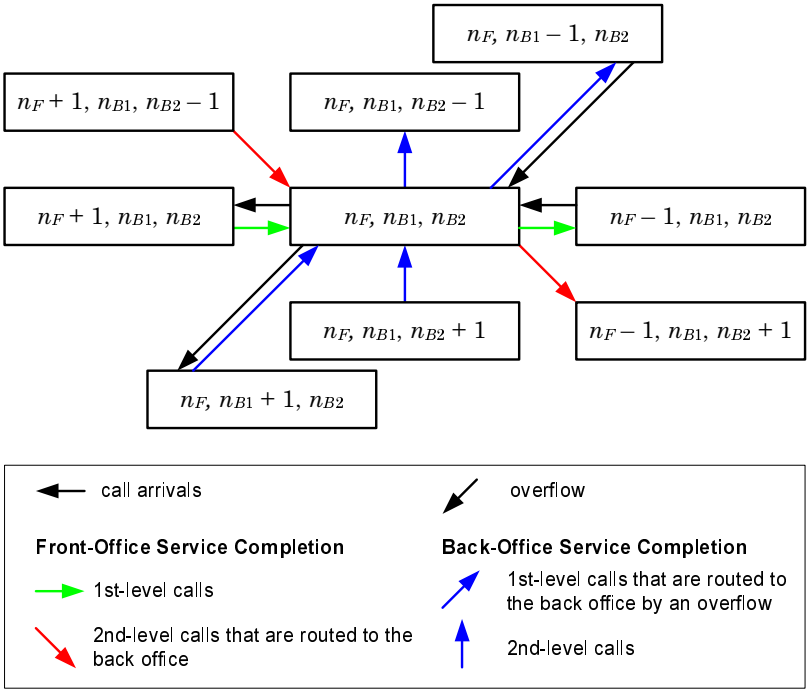

Figure 3: Connected states in case of Region II ($n_F > c_F$, $n_{B1} + n_{B2} < c_B$)

Region II: All front-office agents are busy with available back-office agents

Now, we consider the case when all first-level agents are occupied, but some back-office agents are available, i. e., we have $c_F \leq n_F \leq K_F$ and $n_{B1} + n_{B2} < c_B$. Then, an arriving customer flow over to the back office with probability $p^{(t)}_{n_F - c_F}$, i. e., a call is routed to the back office with rate $p^{(t)}_{n_F - c_F} \cdot \lambda$. With rate $(1 - p^{(t)}_{n_F - c_F}) \cdot \lambda$, a new call will be queued

in the front office. The state (n_F, n_{B1}, n_{B2}) can be reached by such transitions from the states $(n_F, n_{B1}-1, n_{B2})$ and (n_F-1, n_{B1}, n_{B2}), see Figure 3. Using indicator functions, some additional boundary states can be considered. If the system capacity is exhausted (indicated by $1(K_F - n_F)$), a calling customer is blocked and lost, regardless of whether back-office agents are available or not. He will not even be routed to the back-office immediately, i. e., an overflow is not allowed at this moment. States with $n_{B1} = 0$ cannot have been reached by an overflow. Hence, it follows for the states of Region II:

$$\begin{aligned}
&b \cdot c_F \cdot \mu_F \cdot \mathrm{P}\{n_F+1, n_{B1}, n_{B2}-1\} \cdot 1(K_F - n_F) \cdot 1(n_{B2}) \\
&+ (1-b) \cdot c_F \cdot \mu_F \cdot \mathrm{P}\{n_F+1, n_{B1}, n_{B2}\} \cdot 1(K_F - n_F) \\
&+ (n_{B2}+1) \cdot \mu_{B2} \cdot \mathrm{P}\{n_F, n_{B1}, n_{B2}+1\} \\
&+ (n_{B1}+1) \cdot \mu_{B1} \cdot \mathrm{P}\{n_F, n_{B1}+1, n_{B2}\} \\
&+ (1 - p^{(t)}_{n_F-1-c_F}) \cdot \lambda \cdot \mathrm{P}\{n_F-1, n_{B1}, n_{B2}\} \\
&+ p^{(t)}_{n_F-c_F} \cdot \lambda \cdot \mathrm{P}\{n_F, n_{B1}-1, n_{B2}\} \cdot 1(K_F - n_F) \cdot 1(n_{B1}) \\
&= (\lambda \cdot 1(K_F - n_F) + c_F \cdot \mu_F + n_{B2} \cdot \mu_{B2} + n_{B1} \cdot \mu_{B1}) \cdot \mathrm{P}\{n_F, n_{B1}, n_{B2}\} \\
&\qquad\qquad (n_F = c_F, \ldots, K_F;\ n_{B1} + n_{B2} < c_B) \quad (9)
\end{aligned}$$

Notice, in the case of $n_F < c_F$ an overflow cannot take place. Therefore, we set $p^{(t)}_n \equiv 0$ if $n < 0$. For states with $n_F = c_F$, the transition from $(n_F - 1, n_{B1}, n_{B2})$ to (n_F, n_{B1}, n_{B2}) remains possible with the original arrival rate λ. In the state $(c_F - 1, n_{B1}, n_{B2})$ an overflow cannot take place, because, with $c_F - 1$ busy agents, there is still one available.

Region III: All the back-office agents are busy If all the back-office agents are occupied, arriving calls cannot flow over to the back office. Nonetheless, a few overflow calls may still remain in the back office. Hence, the state transitions are essentially the

same as for states of Region I, see Figure 2. A few differences remain: Because further overflow is not possible, no matter how many front-office agents ever are busy, we do not distinguish the cases $n_F < c_F$ and $n_F \geq c_F$ in the front office. Nevertheless, the service rates are proportional to the number of busy agents, i. e. at most c_F or c_B, respectively, or even lower in the back office if n_{B1} overflow calls are served by back-office agents. For $n_{B1} = c_B$, a tranisition to the state $(n_F, n_{B1}+1, n_{B2})$ is impossible. More than c_B front-office calls cannot be routed into the back office by an overflow. Hence, the third type of steady-state equations is:

$$\begin{aligned}
& b \cdot \min\{n_F+1, c_F\} \cdot \mu_F \cdot \mathrm{P}\{n_F+1, n_{B1}, n_{B2}-1\} \cdot \mathit{1}(K_F - n_F) \cdot \mathit{1}(n_{B2}) \\
& + (1-b) \cdot \min\{n_F+1, c_F\} \cdot \mu_F \cdot \mathrm{P}\{n_F+1, n_{B1}, n_{B2}\} \cdot \mathit{1}(K_F - n_F) \\
& + \min\{n_{B2}+1, c_B - n_{B1}\} \cdot \mu_{B2} \cdot \mathrm{P}\{n_F, n_{B1}, n_{B2}+1\} \cdot \mathit{1}(K_B - n_{B1} - n_{B2}) \\
& + (n_{B1}+1) \cdot \mu_{B1} \cdot \mathrm{P}\{n_F, n_{B1}+1, n_{B2}\} \cdot \mathit{1}(K_B - n_{B1} - n_{B2}) \cdot \mathit{1}(c_B - n_{B1}) \\
& + \lambda \cdot \mathrm{P}\{n_F - 1, n_{B1}, n_{B2}\} \cdot \mathit{1}(n_F) \\
& + p^{(t)}_{n_F - c_F} \cdot \lambda \cdot \mathrm{P}\{n_F, n_{B1}-1, n_{B2}\} \cdot (1 - \mathit{1}(n_{B1} + n_{B2} - c_B)) \cdot \mathit{1}(n_{B1}) \cdot \mathit{1}(K_F - n_F) \\
& = \big(\lambda \cdot \mathit{1}(K_F - n_F) + \min\{n_F, c_F\} \cdot \mu_F \\
& \qquad + \min\{n_{B2}, c_B - n_{B1}\} \cdot \mu_{B2} + n_{B1} \cdot \mu_{B1}\big) \cdot \mathrm{P}\{n_F, n_{B1}, n_{B2}\} \\
& \qquad\qquad (n_F = 0, \ldots, K_F;\ c_B \leq n_{B1} + n_{B2} \leq K_B) \quad (10)
\end{aligned}$$

A special case in Equation (10) is $n_{B1} + n_{B2} = c_B$. If $n_F \geq c_F$, these states can be reached by an overflow. But, because the back-office capacity is fully occupied from such a state on, it cannot lead to further overflow. Again, we use that $p^{(t)}_n \equiv 0$ if $n < 0$.

Together with

$$\sum_{n_F=0}^{K_F} \sum_{n_{B1}=0}^{c_B} \sum_{n_{B2}=0}^{K_B - n_{B1}} \mathrm{P}\{n_F, n_{B1}, n_{B2}\} = 1 \qquad (11)$$

the Equations (8)–(10) form a system that has to be solved to determine the stationary state probabilities.

3.4 Performance Measures

Using the steady-state probabilities for the CTMC model as described in Section 3.3, some standard performance measures can be derived. In the CTMC model, the time-dependent overflow as in the original system is replaced by a state-dependent overflow immediately upon arrival. Therefore, we modify the measures for the expected waiting time and the expected number of customers in the front-office queue to approximate the system with time-dependent overflow.

To derive the expected utilizations for the front office and the back office, we sum up the average number of customers being served per agent over all possible states, weighted by the respective stationary-state probability, i. e.

$$\rho_F = \sum_{n_F=0}^{K_F} \sum_{n_{B1}=0}^{c_B} \sum_{n_{B2}=0}^{K_B-n_{B1}} \frac{\min\{n_F, c_F\}}{c_F} \cdot \mathrm{P}\left\{n_F, n_{B1}, n_{B2}\right\} \text{ and} \tag{12}$$

$$\rho_B = \sum_{n_F=0}^{K_F} \sum_{n_{B1}=0}^{c_B} \sum_{n_{B2}=0}^{K_B-n_{B1}} \frac{\min\{n_{B1}+n_{B2}, c_B\}}{c_B} \cdot \mathrm{P}\left\{n_F, n_{B1}, n_{B2}\right\} \ . \tag{13}$$

The mean number of customers in the front office (waiting and in service) is given by

$$\mathrm{E}\left\{N_F\right\} = \sum_{n_F=0}^{K_F} \sum_{n_{B1}=0}^{c_B} \sum_{n_{B2}=0}^{K_B-n_{B1}} n_F \cdot \mathrm{P}\left\{n_F, n_{B1}, n_{B2}\right\} \ . \tag{14}$$

According to the CTMC representation, the mean number of first-level calls that are

routed directly into the back office to be served there immediately (the overflow case) is

$$\mathrm{E}\left\{N_{B1}\right\} = \sum_{n_F=0}^{K_F} \sum_{n_{B1}=0}^{c_B} \sum_{n_{B2}=0}^{K_B-n_{B1}} n_{B1} \cdot \mathrm{P}\left\{n_F, n_{B1}, n_{B2}\right\} \ . \tag{15}$$

The mean number of customers in the back office with second-level service requirements (waiting and in service) is determined by

$$\mathrm{E}\left\{N_{B2}\right\} = \sum_{n_F=0}^{K_F} \sum_{n_{B1}=0}^{c_B} \sum_{n_{B2}=0}^{K_B-n_{B1}} n_{B2} \cdot \mathrm{P}\left\{n_F, n_{B1}, n_{B2}\right\} \ . \tag{16}$$

The overall number of customers in the back office is $\mathrm{E}\left\{N_B\right\} = \mathrm{E}\left\{N_{B1}\right\} + \mathrm{E}\left\{N_{B2}\right\}$. For the mean number of waiting front-office customers, it follows from the CTMC model

$$\mathrm{E}\left\{Q_F\right\} = \sum_{n_F=c_F}^{K_F} \sum_{n_{B1}=0}^{c_B} \sum_{n_{B2}=0}^{K_B-n_{B1}} (n_F - c_F) \cdot \mathrm{P}\left\{n_F, n_{B1}, n_{B2}\right\} \ . \tag{17}$$

Their mean waiting time can be computed by Little's law. The effective arrival rate λ_{eff} corresponds to the arrival rate λ of accepted customer calls, i. e. if they are not blocked,

$$\lambda_{\text{eff}} = \lambda \cdot \left(1 - \sum_{n_{B1}=0}^{c_B} \sum_{n_{B2}=0}^{K_B-n_{B1}} \mathrm{P}\left\{K_F, n_{B1}, n_{B2}\right\}\right) \ . \tag{18}$$

The summation of probabilities in Equation (18) gives the blocking probability p_F^{B}, i. e. the probability that the capacity K_F of the front-office is exhausted. Using λ_{eff}, it follows for the mean waiting time of front-office customers

$$\mathrm{E}\left\{W_F\right\} = \frac{\mathrm{E}\left\{Q_F\right\}}{\lambda_{\text{eff}}} \ . \tag{19}$$

The performance measures for the back-office service can be derived in a similar way.

In the CTMC model, an overflow may occur if a customer arrives, all front-office agents are occupied, and at least one back-office agent is available, i. e., if $n_{B1} + n_{B2} < c_B$ holds. Therefore, the probability π that a calling customer is immediately routed to a back-office agent (overflow) is given by

$$\pi = \sum_{n_F=c_F}^{K_F-1} \sum_{n_{B1}=0}^{c_B-1} \sum_{n_{B2}=0}^{c_B-1-n_{B1}} p_{n_F-c_F}^{(t)} \cdot \mathrm{P}\left\{n_F, n_{B1}, n_{B2}\right\} \ . \tag{20}$$

In general, $p_n^{(t)}$ gives the probability that an arriving customer waits more than t time units in the case of n customers waiting in front of him. Hence, the probability that a calling customer will wait at least t time units is

$$\mathrm{P}\left\{W_F > t\right\} = \sum_{n_F=c_F}^{K_F} \sum_{n_{B1}=0}^{c_B} \sum_{n_{B2}=0}^{K_B-n_{B1}} p_{n_F-c_F}^{(t)} \cdot \mathrm{P}\left\{n_F, n_{B1}, n_{B2}\right\} \ , \tag{21}$$

with $p_{K_F-c_F}^{(t)} \equiv 1$. Then, $X = 1 - \mathrm{P}\left\{W_F > t\right\}$ is the probability that a calling customer can be served within t time units. This represents the so-called X/t service level which is an often used performance criterion in call centers; see the discussion in Helber and Stolletz (2004). It gives the probability that X % of all calling customers are served within t time units of waiting time.

To approximate the expected queue length and the expected waiting time for the original call center with time-dependent overflow, we modify some of the performance measures of the CTMC model. Note that from the point of view of the CTMC model, the front-office queue contains only waiting customer calls that will be served by front-office agents. In reality, customers that are routed into the back office by an overflow have been

waiting in the front-office queue for at least t time units. For those customers, we increase the waiting time by an amount t, i. e. the modified waiting time is

$$\widehat{\mathrm{E}\{W_F\}} = \mathrm{E}\{W_F\} + \pi \cdot t \ . \tag{22}$$

Applying Little's law results in the modified queue length of

$$\widehat{\mathrm{E}\{Q_F\}} = \widehat{\mathrm{E}\{W_F\}} \cdot \lambda_{\text{eff}} = \mathrm{E}\{Q_F\} + \pi \cdot t \cdot \lambda_{\text{eff}} \ . \tag{23}$$

The same modification applies for the number of all calls in the system, i. e.

$$\widehat{\mathrm{E}\{N\}} = \mathrm{E}\{N\} + \pi \cdot t \cdot \lambda_{\text{eff}} \tag{24}$$

gives the total number of calls in the system with time-dependent overflow.

4 Numerical Results

4.1 Outline and Experimental Design

In order to study the quality of the Markovian performance approximation we set up a number of numerical experiments. The steady-state equations are solved using CPLEX 10.1. The equations and the formulas for the performance measures are implemented in GAMS. The analytical approximations are tested against simulation results derived with ARENA 10.0. The simulation experiments were run over 110000 minutes with a warm-up phase of 10000 minutes. The statistics were based on 100 replications which guarantees quite moderate half-widths of the confidence intervals for all considered performance measures. The results of the CTMC-based calculations are compared with the corresponding

simulation-based point estimates for different parameter settings. After this, the sensitivity of the approximated performance measures is analyzed dependent on the waiting-time limit t and the arrival rate λ.

4.2 Comparison with Simulation

First, we consider a small call center with $c_F = 15$ front-office agents and $c_B = 5$ back-office agents. In the medium-sized cases the number of agents is doubled, see Cases 9–16 in Table 1. To analyze these call centers with different loads, the arrival rate λ is adjusted

Case	c_F	c_B	K_F	K_B	λ	b	μ_F	μ_{B1}	μ_{B2}	t
1	15	5	50	20	3.0	0.1	0.25	0.25	0.25	0.25
2	15	5	50	20	4.0	0.1	0.25	0.25	0.25	0.25
3	15	5	50	20	3.0	0.1	0.25	0.25	0.25	2
4	15	5	50	20	4.0	0.1	0.25	0.25	0.25	2
5	15	5	50	20	3.0	0.1	0.25	0.2	0.125	0.25
6	15	5	50	20	4.0	0.1	0.25	0.2	0.125	0.25
7	15	5	50	20	3.0	0.1	0.25	0.2	0.125	2
8	15	5	50	20	4.0	0.1	0.25	0.2	0.125	2
9	30	10	70	30	6.0	0.1	0.25	0.25	0.25	0.25
10	30	10	70	30	8.0	0.1	0.25	0.25	0.25	0.25
11	30	10	70	30	6.0	0.1	0.25	0.25	0.25	2
12	30	10	70	30	8.0	0.1	0.25	0.25	0.25	2
13	30	10	70	30	6.0	0.1	0.25	0.2	0.125	0.25
14	30	10	70	30	8.0	0.1	0.25	0.2	0.125	0.25
15	30	10	70	30	6.0	0.1	0.25	0.2	0.125	2
16	30	10	70	30	8.0	0.1	0.25	0.2	0.125	2

Table 1: Test cases (time unit: minutes)

such that the front office is underloaded ($\lambda < c_F \cdot \mu_F$) or overloaded ($\lambda > c_F \cdot \mu_F$). For the processing rates we assume two scenarios. In the first one, all processing rates are equal to

Case	ρ_F [%]			ρ_B [%]			π [%]		
	CTMC	Sim	Δ	CTMC	Sim	Δ	CTMC	Sim	Δ
1	75.07	75.93	–0.86	37.30	34.97	2.33	6.16	5.09	1.07
2	91.29	91.81	–0.52	73.35	72.04	1.31	14.36	13.90	0.46
3	78.53	79.01	–0.48	27.98	26.64	1.34	1.84	1.22	0.62
4	95.34	96.23	–0.89	62.35	59.88	2.47	10.55	9.69	0.86
5	76.09	76.74	–0.65	60.31	58.40	1.91	4.88	4.11	0.77
6	96.07	96.06	0.01	92.40	92.86	–0.46	8.70	8.80	–0.10
7	78.77	79.18	–0.41	51.88	50.60	1.28	1.54	1.02	0.52
8	97.41	97.73	–0.32	87.87	87.12	0.75	7.36	7.11	0.25

Table 2: Expected utilizations and overflow probabilities (small-sized call centers)

0.25 minutes. In the second one, the back-office agents serve overflowed customers with a smaller rate of $\mu_{B1} = 0.2$, and real back-office customers are served with rate $\mu_{B2} = 0.125$. With $t = 0.25$ and $t = 2$ two different waiting-time limits are analyzed. The combination of all of these parameters results in the 16 cases summarized in Table 1.

In Table 2, the analytical (CTMC) results for the utilizations ρ_F and ρ_B and the overflow probability π for the small-sized system configurations are compared with simulation results (SIM). The absolute deviations Δ are recorded because both measures, the utilization and the overflow probability, are relative measures themselves. In general, in case of very low values, the relative deviation does not give important information. For an underloaded front-office with a long wating-time limit t (Case 7), we correctly estimate a very low overflow probability ($\pi = 1.54\%$). The simulation gives $\pi_{\text{Sim}} = 1.02\%$ for that case. This is the largest relative deviation that we have observed (almost 51 %) whereas the absolute error is only 0.52 %. To summarize the results, all the deviations are very

	$\widehat{\mathrm{E}\{N\}}$			$\mathrm{E}\{Q_B\}$			$\widehat{\mathrm{E}\{Q_F\}}$		
Case	CTMC	Sim	Δ	CTMC	Sim	Δ	CTMC	Sim	Δ
1	13.30	13.37	–0.07	0.03	0.03	0.00	0.19	0.20	–0.01
2	20.67	20.52	0.15	0.18	0.20	–0.02	3.28	2.95	0.32
3	13.80	13.96	–0.16	0.01	0.01	0.00	0.72	0.76	–0.04
4	22.70	23.81	–1.10	0.13	0.14	–0.01	6.00	6.24	–0.24
5	14.96	14.99	–0.04	0.20	0.19	0.01	0.37	0.37	0.00
6	30.87	30.39	0.48	0.84	0.89	–0.05	11.08	10.45	0.63
7	15.25	15.38	–0.12	0.14	0.13	0.01	0.80	0.84	–0.04
8	31.78	32.14	–0.37	0.78	0.81	–0.03	12.58	12.32	0.26

Table 3: Expected number of calls in the system and queue lengths (small-sized call centers)

moderate. Therefore, the overflow representation in the model — although the events are time-shifted — leads to nearly the same utilizations and overflow probabilities as in the real system.

With this overflow representation, the expected number of calls in the back office is modeled adequately. Hence, the analytical results for $\mathrm{E}\{Q_B\}$ taken from the Markov model are quite accurate for the real system, see Table 3. The results for the modified performance measure $\widehat{\mathrm{E}\{Q_F\}}$ are shown in the last three columns in Table 3. The first columns compare the approximated with the simulated number $\widehat{\mathrm{E}\{N\}}$ of all calls in the system. The deviations are within an acceptable tolerance of maximum 5 % with an absolute deviation of 1.1 customers.

Table 4 shows the expected waiting times in the front-office queue and the service level approximation. Due to Little's law, the deviations from simulation results for the average

Case	$\widehat{\mathrm{E}\{W_F\}}$			$\mathrm{P}\{W_F > t\}$ [%]			Service Level [%]		
	CTMC	Sim	Δ	CTMC	Sim	Δ	CTMC	Sim	Δ
1	0.06	0.07	0.00	9.19	10.84	–1.65	90.81	89.16	1.65
2	0.82	0.74	0.08	47.51	48.35	–0.84	52.49	51.65	0.84
3	0.24	0.25	–0.01	2.34	2.51	–0.17	97.66	97.49	0.17
4	1.50	1.56	–0.06	29.86	32.58	–2.72	70.14	67.42	2.72
5	0.12	0.12	0.00	12.73	13.98	–1.25	87.27	86.02	1.25
6	2.81	2.64	0.16	76.10	75.33	0.77	23.90	24.67	–0.77
7	0.27	0.28	–0.01	3.09	3.29	–0.20	96.91	96.71	0.20
8	3.19	3.12	0.07	58.87	59.51	–0.64	41.13	40.49	0.64

Table 4: Waiting time and service levels (small-sized call centers)

waiting time are comparable to the results for the average queue lengths. The second measure in the center of Table 4 gives the probability that a customer may overflow, i. e. that his waiting time exceeds t time units. Using the waiting-time probability $p_n^{(t)}$ of an arriving customer as described in Chapter 3.2, we can quite accurately estimate the waiting-time probability $\mathrm{P}\{W_F > t\}$ for the critical value t. The probability $\mathrm{P}\{W_F > t\}$ is larger than the overflow probability π, i. e., just a fraction of customers waiting longer than t time units in the front-office queue finally flow over to the back office. This can be explained by the overflow condition, that at least one back-office agent has to be available. In our overflow model, the waiting-time limit t from which on an overflow may occur is a crucial management decision variable. Using the complementary formulation of $\mathrm{P}\{W_F > t\}$, the common used service-level can be given, as recorded in the right column of Table 4. Naturally, the absolute deviations $|\Delta|$ are the same.

The results for the medium-sized systems are summarized in Tables 5–7. The quality

Case	ρ_F [%]			ρ_B [%]			π [%]		
	CTMC	Sim	Δ	CTMC	Sim	Δ	CTMC	Sim	Δ
9	77.80	78.21	–0.41	29.94	28.80	1.14	2.75	2.22	0.53
10	93.07	93.76	–0.69	68.66	66.81	1.85	12.73	12.10	0.63
11	79.79	79.88	–0.09	24.57	24.34	0.23	0.26	0.15	0.11
12	97.68	98.29	–0.61	56.11	54.41	1.70	8.38	7.78	0.60
13	78.05	78.42	–0.37	54.15	53.02	1.13	2.44	1.98	0.46
14	96.68	96.84	–0.16	92.04	91.76	0.28	8.51	8.40	0.11
15	79.81	79.89	–0.08	48.61	48.35	0.26	0.24	0.14	0.10
16	98.52	98.83	–0.31	85.47	84.13	1.34	6.59	6.21	0.38

Table 5: Expected utilizations and overflow probabilities (medium-sized call centers)

Case	$\widehat{\mathrm{E}\{N\}}$			$\mathrm{E}\{Q_B\}$			$\widehat{\mathrm{E}\{Q_F\}}$		
	CTMC	Sim	Δ	CTMC	Sim	Δ	CTMC	Sim	Δ
9	26.41	26.47	–0.06	0.00	0.00	0.00	0.11	0.12	–0.01
10	37.11	37.32	–0.20	0.11	0.12	–0.01	2.47	2.39	0.08
11	26.93	26.99	–0.06	0.00	0.00	0.00	0.56	0.59	–0.03
12	42.65	45.38	–2.73	0.07	0.08	–0.01	9.01	10.38	–1.36
13	29.02	29.06	–0.04	0.06	0.06	0.00	0.16	0.17	–0.01
14	49.84	49.69	0.15	0.73	0.77	–0.04	11.07	10.69	0.38
15	29.38	29.43	–0.06	0.03	0.03	0.00	0.57	0.60	–0.03
16	52.95	54.49	–1.55	0.60	0.61	–0.01	15.29	15.82	–0.53

Table 6: Expected number of calls in the system and queue lengths (medium-sized call centers)

of the approximations is quite similar to those for the small-sized call center. Again, the largest deviation from the simulation results can be observed for the mean number of customers in the system. Although the analytical estimation in Case 12 differs by an amount of almost 3 customers, this is a relative error of 6.0 % comparing to the number of calls in the system.

Case	$\widehat{\mathrm{E}\{W_F\}}$			$\mathrm{P}\{W_F > t\}$ [%]			Service Level [%]		
	CTMC	Sim	Δ	CTMC	Sim	Δ	CTMC	Sim	Δ
9	0.02	0.02	0.00	3.08	3.82	–0.74	96.92	96.18	0.74
10	0.31	0.30	0.01	31.96	34.42	–2.46	68.04	65.58	2.46
11	0.09	0.10	0.00	0.28	0.26	0.02	99.72	99.74	–0.02
12	1.13	1.30	–0.17	17.19	21.36	–4.17	82.81	78.64	4.17
13	0.03	0.03	0.00	4.06	4.73	–0.67	95.94	95.27	0.67
14	1.40	1.35	0.05	66.81	66.97	–0.16	33.19	33.03	0.16
15	0.10	0.10	0.00	0.32	0.32	0.00	99.68	99.68	0.00
16	1.93	2.00	–0.07	43.49	46.13	–2.64	56.51	53.87	2.64

Table 7: Waiting time and service levels (medium-sized call centers)

4.3 Sensitivity Analysis

Another series of simulation experiments studies the sensitivity of the performance measure approximations for varying waiting-time limits t. For Case 6 of Table 1 the waiting-time limit varies between $t = 0.1$ and $t = 3.0$ minutes. With an increasing waiting-time limit t, the utilization ρ_F of the front-office agents increases, and the utilization ρ_B of the back-office agents decreases, see Figure 4. The reason is that the overflow probability π decreases as well; see Figure 6. Figure 5 shows that this leads to an increasing expected waiting time in the front office.

In another sensitivity analysis, we vary the arrival rate for Case 2 between $\lambda = 2.5$ and $\lambda = 4.5$. With an increasing arrival rate, the probability of an overflow increases, see Figures 7 and 8. Again, all performance measures are well approximated by the Markovian performance approximation. The approximated expected waiting time $\widehat{\mathrm{E}\{W_F\}}$ is slightly overestimated only for very large values of the arrival rate λ and very small values of

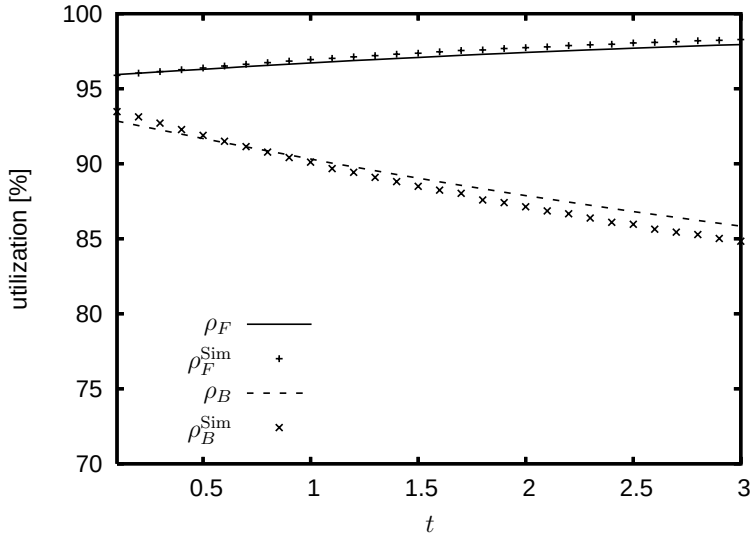

Figure 4: Expected utilization depending on t (Case 6)

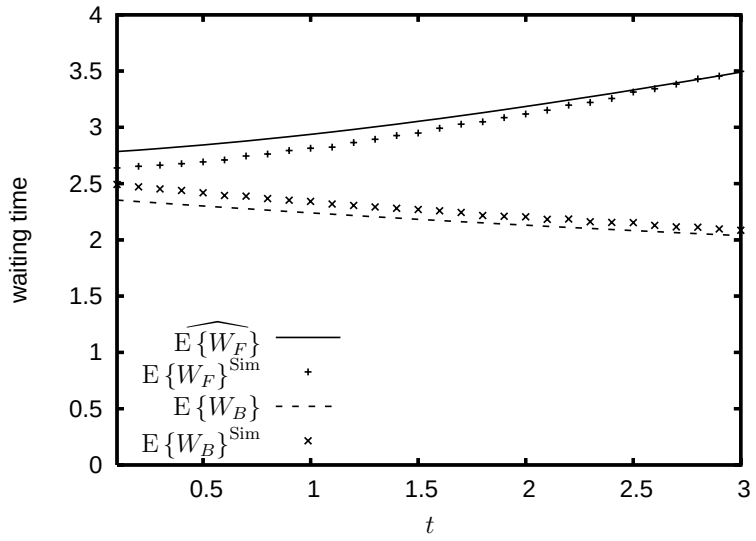

Figure 5: Expected waiting times depending on t (Case 6)

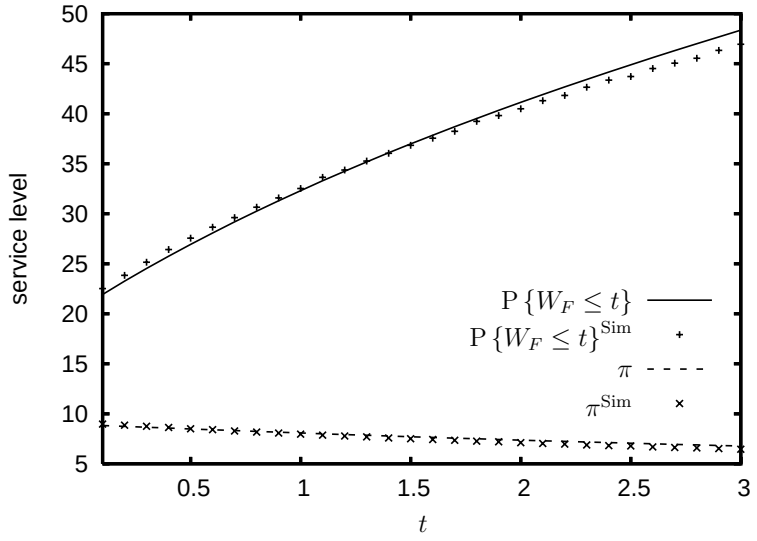

Figure 6: Service Level and overflow probability depending on t (Case 6)

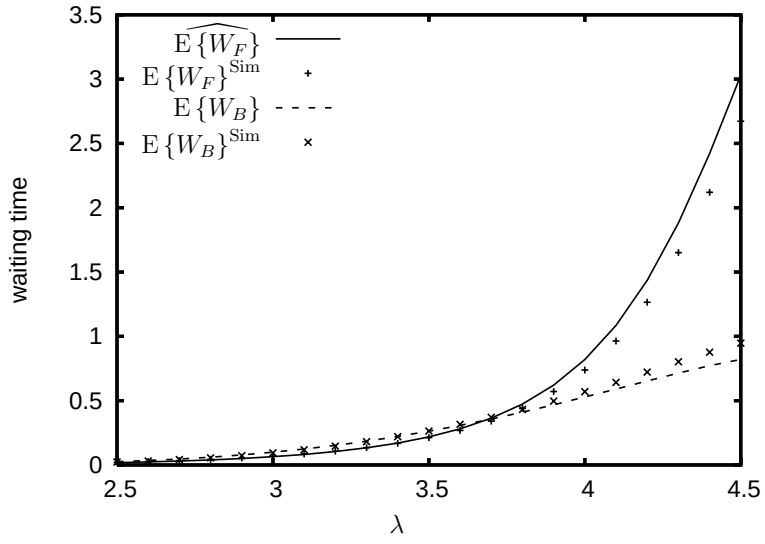

Figure 7: Expected waiting times depending on λ (Case 2)

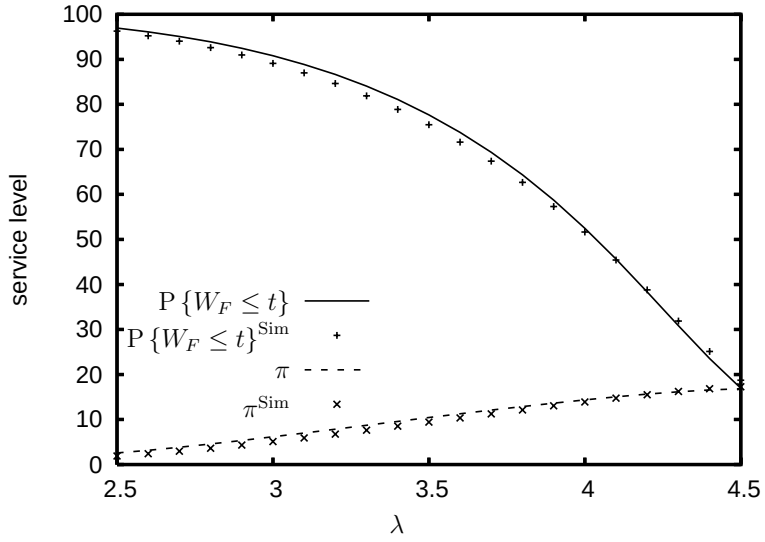

Figure 8: Service Level and overflow probability depending on λ (Case 2)

the waiting-time limit t. This approximation failure may be caused by the fact that an overflow cannot occur if all back-office agents are busy. In the original system, a call with a waiting time larger than t may flow over to available back-office agents later on, while in the approximated system just front-office agents will serve this call. This time-shift in the overflow events leads to a slightly inaccurate representation of the load dynamics in the front office.

5 Conclusion

We have presented a two-level support system which is commonly used in small- and medium-sized call centers. The main feature of this system is the time-dependent overflow of waiting calls from the front-office queue to the back office, under the condition that at

least one back-office agent is available. Such overflow mechanisms are implemented in call centers for financial services and in other industries with multi-level support. The analysis of this system with time-dependent overflow is reduced to the analysis of a CTMC with state-dependent overflow probabilities. Numerical experiments for different parameter settings demonstrate the reliability of this Markovian performance approximation. A sensitivity analysis shows the impact of the waiting-time limit t and the dependence of the performance measures on the arrival rate λ.

Further research should be directed into the extension of this model by impatient customers and by a telephone answering machine to offer call backs during high-load periods. Beyond the application to the considered small- and medium-sized call centers, other approximation methods based on system decomposition are of high interest, especially for large call centers.

References

Aksin, Z., M. Armony, and V. Mehrotra (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management 16*(6), 665–688.

Association of German Banks (2004). Umfragen des Bankenverbandes: Gute Aussichten für Online Banking. *Die Bank Online 10*, Art. 351.

Brandt, A., and M. Brandt (2002). Asymptotic results and a markovian approximation for the $M(n)/M(n)/s + GI$ system. *Queueing Systems 41*, 73–94.

Chevalier, P., and N. Tabordon (2003). Overflow analysis and cross-trained servers. *International Journal of Production Economics 85*, 47–60.

Deutsche Bundesbank (2008, March). Bankendichte sinkt weniger stark. Pressenotiz.

Down, D. G., and M. E. Lewis (2007). A call center model with upgrades. Under review.

Franx, G. J., G. Koole, and A. Pot (2006). Approximating multi-skill blocking systems by HyperExponential decomposition. *Performance Evaluation 63*, 799–824.

Gans, N., G. Koole, and A. Mandelbaum (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management 5*(2), 79–141.

Gans, N., and Y.-P. Zhou (2007). Call-routing schemes for call-center outsourcing. *Manufacturing & Service Operations Management 9*(1), 33–50.

Green, L. V., and P. J. Kolesar (1991). The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science 37*(1), 84–97.

Green, L. V., P. J. Kolesar, and W. Whitt (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management (POMS) 16*(1), 13–39.

Guerin, R., and L. Lien (1990). Overflow analysis of finite waiting room systems. *IEEE Transactions on Communications 38*(9), 1569–1577.

Helber, S., and R. Stolletz (2004). Grundlagen der Personalbedarfsermittlung in Inbound-Call Centern. *Zeitschrift für Betriebswirtschaft, Ergänzungsheft 1/2004*, 67–88.

Jimenéz, T., and G. Koole (2004). Scaling and comparison of fluid limits of queues applied to call centers with time-varying parameters. *OR Spectrum 26*(3), 413–422.

Koole, G., and A. Mandelbaum (2002). Queueing models of call centers: An introduction. *Annals of Operations Research 113*, 41–59.

Lucent Technologies (1999). *CentreVu Release 8 Advocate User Guide*. P.O. Box 4100, Crawfordsville, IN 47933, U.S.A.: Lucent Technologies.

Pinedo, M. L., S. Seshadri, and J. G. Shanthikumar (2000). Call centers in financial services: Strategies, technologies, and operations. In E. L. Melnick, P. R. Nayyar, M. L. Pinedo, and S. Seshadri (Eds.), *Creating Value in Fincancial Services*, pp. 357–388. Boston et al.: Kluwer Academic Publishers.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C — The Art of Scientific Computing* (2 ed.). Cambridge, New York, Melbourne: Cambridge University Press.

Sendfeld, P. (2007). Two queues with weighted one-way overflow. To appear in *Methodology and Computing in Applied Probability*.

Stolletz, R. (2003). *Performance Analysis and Optimization of Inbound Call Centers*, Volume 528 of *Lecture Notes in Economics and Mathematical Systems*. Berlin et al.: Springer.

Stolletz, R. (2008). Approximation of the non-stationary $M(t)/M(t)/c(t)$-queue using stationary queueing models: The stationary backlog-carryover approach. *European Journal of Operational Research 190*(2), 478–493.

Stolletz, R., and S. Helber (2004). Performance analysis of an inbound call center with skills-based routing: A priority queueing system with two classes of impatient customers and heterogeneous agents. *OR Spectrum 26*(3), 331–352.

Wallace, R. B., and W. Whitt (2005). A staffing algorithm for call centers with skill-based routing. *Manufacturing & Service Operations Management* 7(4), 276–294.