DISSERTATION

Dynamic Contact Centers with Impatient Customers and Retrials

Von der Wirtschaftswissenschaftlichen Fakultät der
Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

Doktorin der Wirtschaftswisssenschaften
- Doctor rerum politicarum -

genehmigte Dissertation

von
Dipl.-Math. Kirsten Henken
geboren am 31. Juli 1977 in Bremerhaven

2007

Erstgutachter: Professor Dr. Stefan Helber

Zweitgutachter : Professor Dr. Michael H. Breitner

Tag der Disputation: 19.Januar 2007

Preface

The matters of this PhD thesis in Business Administration are the results of my research in the area of contact center management at the Technical University of Clausthal and the University of Hanover from 2002 to 2006. The focus of this work is the analysis and staff requirement planning of dynamic inbound contact centers with impatient customers who may retry after a while. Contact centers are dynamic in two ways. On the one hand the arrivals, service completions, and abandonments are random events and on the other hand the number of arrivals, service durations and patience of customers depend strongly on the time of the day. Therefore, the strong approximation is used to incorporate both aspects of dynamics into the analysis and the staff requirement planning.

I am indepted to a lot of people for their support during the time of working at this thesis. Firstly, I would like to thank my advisor, Professor Dr. Stefan Helber, for motivating my research und many fruitful discussions. He fortified me to search in very mathematics related directions and fetched me gently back to the problems of business administration. Secondly, I have to thank Professor Avishai Mandelbaum from the University of Haifa for his support and advise. He taught me a lot of the mathematical background I needed for this thesis and turned my attention to the strong approximations. I am grateful to Professor Dr. Michael H. Breitner for referring this thesis.

Furthermore, I thank my sister Ines and my colleague Dr. Raik Stolletz for their personal support and their comments on the draft versions of this thesis. Especially, the discussions with Raik gave me a reality check for the problems and shortcomings related to the models. I am grateful to all my colleagues at both Universities and Anne Menis who helped to improve my English.

Finally, I am deeply indepted to my parents and my husband Lars. They steadily encouraged me during all the time. Without their help and patience I never had got so far.

Midlum, August 2007 Kirsten Henken

Abstract

The focus of this thesis is on the management of dynamic inbound contact centers with impatient customers and retrials. We consider contact centers with both homogeneous and heterogenous customers and agents. The term dynamic refers to the processes in a contact center in two ways. Firstly, these processes are random and secondly they depend on the time of day and the day of the week. The processes are the successive arrivals of customers to the contact center, the consecutive services by agents, and the successive abandonments and retrials of customers. In order to consider both aspects in the queueing model we use the so-called strong approximation which consists of a fluid approach and a diffusion refinement. By means of the fluid approximation we are able to derive an initial value problem for the number of customers in the system and the so-called orbits, which are virtual queues of recalling customers. The diffusion refinement is used to deduce differential equations for the variances and covariances of the queueing processes. From the solution of the initial value problem we analyse the performance measures of the different contact center models.

Thereby an essential point of this thesis is the influence of the retrial behaviour of impatient customers on the performance and shift scheduling decision. For this purpose, time-dependent and aggregated technical and economic performance measures are derived from the fluid approximation. Afterwards the influence of the service, abandonment and retrial parameters of the contact center model on the performance is shown. Furthermore, the results of the approximated measures are compared to the results of a simulation. It is shown how the fluid approximation can easily be extended to much more complex models with multiple customers classes and agent groups. The variances and covariances derived from the diffusion refinement are used to present the stochastic effects which influence the processes. Furthermore, the impact of the parameters on the variances and covariances is shown.

Based on the fluid approximation an integrative staff requirement planning and shift scheduling approach for contact centers with impatient customers and retrials is developed. This optimisation approach can easily be extended to even more complex models. In order to solve the optimisation problem an initial procedure and a heuristic is used. It is shown that the initial algorithm already leads to remarkably good results with respect to the profit and the technical performance measures which are slightly improved by the heuristic algorithm.

Key words: Contact Center Management, Retrial Queues, Strong Approximations

Zusammenfassung

Das Thema dieser Dissertation sind dynamische Contact Center mit ungeduldigen Kunden, von denen ein Teil während des Wartens auflegt und nach einer gewissen Zeit erneut anruft. Dabei betrachten wir sowohl homogene als auch heterogene Kunden und Agenten. Contact Center sind in zweierlei Hinsicht dynamisch. Einerseits sind die Zeitspannen zwischen den Ankünfte der Kunden, die Bediendauern und die Geduld der Kunden aus der Sicht des Contact Center Managements zufällige Größen, andererseits hängen insbesondere die Ankunftraten sehr stark von der Tageszeit ab.

Die Contact Center werden mit Hilfe von Warteschlangenmodellen modelliert. Um dabei sowohl die stochastischen als auch die zeitabhängigen Einflüsse berücksichtigen zu können, die durch den traditionellen stationären Warteschlangenansatz nur schwer abgebildet werden können, wird in dieser Arbeit die sogenannte Starke Approximation verwendet. Die Starke Approximation setzt sich aus einer Fluid Approximation und einer Diffusionserweiterung zusammen. Die Fluid Approximation führt auf ein Anfangswertproblem für die Anzahl der Kunden im System und in den sogenannten Orbits, den imaginären Warteschlangen der Wahlwiederholer. Aus der Diffusionserweiterung werden Differentialgleichungen für die Varianzen und Covarianzen der Prozesse hergeleitet. Mit Hilfe der Lösung des Anfangswertproblems berechnen wir die Leistung der Contact Center.

Dabei ist ein Schwerpunkt dieser Arbeit, den Einfluss des Wahlwiederholungsverhaltens der ungeduldigen Kunden auf die Leistungsfähigkeit und Wirtschaftlichkeit sowie die Personaleinsatzplanung darzustellen. Daher werden aus der Fluid Approximation zeitabhängige und aggregierte technische und ökonomische Leistungskenngrößen für die Analyse und Bewertung des Contact Centers hergeleitet. Anschließend wird mit Hilfe der Fluid Approximation der Einfluss der verschiedenen Parameter auf diese Leistungskenngrößen dargestellt und diese Ergebnisse mit jenen einer Simulation verglichen. Es zeigt sich außerdem, dass sich die Fluid Modelle sehr einfach auf sehr komplexe Modelle mit multiplen Kunden- und Agentengruppen erweitern lassen. Mit den aus der Diffusionserweiterung hergeleiteten Varianzen und Covarianzen für die Prozesse der Anzahl Kunden im System und den sogenannten Orbits werden die stochastischen Einflüsse dargestellt, die auf die Prozesse einwirken. Ferner werden die Wirkungen der verschiedenen Parameter auf die Varianzen und Covarianzen gezeigt.

Aufbauend auf der Fluid Approximation wird eine integrierte Personalbedarfs- und Schichteinsatzplanung für die verschiedenen Modelle mit heterogenen Kunden und Agenten und Wahlwiederholern entwickelt. Es zeigt sich, dass auch der Optimierungsansatz sehr leicht auf weitaus komplexere Modelle zu erweitern ist. Für das Optimierungsproblem wird ein zweigeteilter Optimierungsansatz vorgestellt. Dieser setzt sich aus einem Startalgorithmus und einer Verbesserungsheuristik zusammen. Dabei führt bereits der Startalgorithmus zu sehr guten Ergebnissen hinsichtlich des Gewinns und der technischen Leistungskenngrößen, die durch die Heuristik nur noch leicht verbessert werden können.

Schlagworte: Contact Center Management, Warteschlangentheorie, Fluid-Approximation

Contents

1 Introduction

The topic of this thesis are so-called contact centers. Contact centers deliver tele-services to distant customers via internet communication, e-mail, fax, phone, or other channels. Therefore, contact centers can be regarded as the successors of call centers which offer their service solely via phone. The palette of service of contact centers ranges from acquiring new customers to providing general information. In general, contact centers belong to the tertiary service sector. During recent years the number of contact and call centers has been growing steadily, stressing the importance of this service industry.

Two operation methods of contact centers are distinguished according to the direction of the contact. If the contact is initiated by a customer, the contact is said to work in the inbound mode. If an agent, the person working in a contact center, contacts the customers via any channel we speak of an outbound contact center. If both methods are combined, the contact center is called a blended or hybrid contact center.

Inbound contact centers are driven by the random contact arrivals of customers and their service durations. These random events depend highly on the time of day and the day of the week. Therefore, the arrival and the service rate will vary over the day. If the number of agents cannot be perfectly adjusted to the varying demand of customers, this will lead to waiting by both customers and agents.

In general, customers prefer to be served immediately or accept at most very short waiting times. Therefore the waiting time of customers is a very important performance measure of contact centers. Other technical performance measures are, e.g., the utilisation of the agents, the percentage of customers served or percentage of abandoning customers. These measures can be improved if more seats, the workstations of agents, are equipped and more agents are staffed by the management, but additional facilities and employees cause costs. The main part of all costs of a contact center are brought about by wages. Hence, the management has to balance the technical and the economical aspects of performance.

By means of forecasts the expected time-dependent arrival and service rates can be determined, which are needed for staffing and scheduling. Staffing and scheduling is the process of determining the number of agents needed to achieve predefined performance requirements and arrange the shifts associated with the agents. The staffing and scheduling problem covers the main part of the short-time operational planning in a contact center environment. A performance analysis relates the different parameters, e.g., the time-dependent rates or the schedule of agents, of a contact center to the technical and economical performance measures.

Based on the performance analysis of dynamic contact center models with retrials, we develop a staffing and scheduling approach. The term dynamic summarises both the time-dependency of the processes, e.g. the arrivals, and the randomness of the events, e.g. a service completion, in a contact center. As the arrivals of customer requests are highly time-dependent and random, this thesis utilises fluid and diffusion approximations of the contact center models. Thereby an essential point of this thesis is the influence of the retrial behaviour of customers on the performance and scheduling decision.

In the next chapter we introduce the characteristics of contact centers by means of a first simple model. Beyond that, we explain the various aspects of dynamics occurring in contact centers and their causes. In order to measure the performance of contact centers, technical performance measures are explained as well as the basic decision problem appearing in contact centers.

Different queueing-theory approaches have been developed to deal with these different kinds of dynamics. We introduce three different approaches in the third chapter, however the last approach is more a refinement of the second one than an independent approach. The stationary approach is widely discussed in literature, but its ability to deal with time-dependencies is very limited. Therefore, we explain the fluid approach and diffusion refinement on the basis of a well known call center model, the so-called Erlang-A model.

The fourth chapter is dedicated to the analysis of contact center models with impatient customers and retrials by means of the fluid and diffusion approximation. There we distinguish contact center models with statistically identical customers and agents from those with different kinds of agents and customers. The number of customers in the contact center and waiting for a retrial as well as other technical and economical performance measures are determined. The results of the approximations are compared to simulation results, and the influence of the different parameters of the contact centers on the performance measures is presented.

Based on the fluid approximation of the dynamic contact center models, a generic staffing and shift scheduling problem is formulated in the sixth chapter. We investigate the objective function and adapt a heuristic optimisation method to the requirements of the optimisation problem. The influence of the parameters of the heuristic algorithm on the solution is shown. In the second part of the chapter we extend the optimisation approach to a more realistic contact center with different kinds of agents and customers.

Finally, we summarise the results of this thesis and give some directions for further research. In the appendix the differential equations of the contact center models in the fourth chapter are derived in detail.

2

Functions and Structure of Contact Centers

2.1 Characteristics of a Contact Center

In this section we give an overview over the relevance, the functionality, and the components of contact centers. Contact centers are service units which deliver tele-services to distant customers via internet communication, e-mail, fax, phone, or other channels[1]. Therefore, contact centers can be regarded as the successors of call centers which offer their service solely via phone[2]. Simultaneity of production and consumption characterises the service of call centers, which is known as the *uno actu principle*[3]. However, in a contact center the principle is weakened, since e-mails and faxes can be stored and processed later, so that the communication works asynchronously.

Contact centers are used in both the private and public sectors to communicate with customers. Communication with customers can be regarded as a special form of public relations management[4] which includes after-sales support, accessory advertising, complaint handling, and other forms of activities aiming to strengthen customer's loyalty.

Emergency hotlines or public utilities are examples of contact centers in the public sector. Contact centers in the private sector are found in banks, insurance companies, mail order businesses, and many other industries. In Germany, 58% of all call centers belong to insurance and banking services[5].

In this thesis we focus on the private sector. Therefore, all kinds of requests are associated with customers, whatever the form of the contact might be. In this sector customer perception of the company is often driven by the experience of the customers gained in communication with the agents work-

[1] See, Sisselman and Whitt (2004) p. 2, Koole and Mandelbaum (2002) Section 1.1, and Koole (2002), Chapter 6

[2] See Hawkins et al. (2001).

[3] See Gross and Badura (1977); Maas and Graf (2005).

[4] See, e.g., Winer (2001).

[5] See, Stockmann (2005).

ing in contact centers[6]. Agents or (telephone) service representatives talk to customers on the phone or communicate via the internet and answer e-mails or faxes.

In February 2005 about 330,000 employees were working in about 5,500 contact centers in Germany[7]. By 2010 the number of employees in contact centers is assumed to grow further, where 80% of all new jobs are assumed to be part-time work.

Agents can be characterised according to their skills and contracts of employment. The set of skills summarises the number of tasks the agent is trained for and her or his[8] level of experience[9]. The contract of employment regulates the availability, the duration of shifts, and the frequency of rest periods.

In Germany 50% of the call and contact centers are so-called inhouse call centers, i.e., these contact centers belong to firms which use the contact center for customer relationship management purposes[10]. If agents need special knowledge of the products or services of the company, this form is considered to be preferable.

Besides those inhouse contact centers, about 18% are operated by pure contact center service providers. In these contact centers the processes are often standardised, such that the agents need less special knowledge. An example of standardised service is the mail order service of big trading companies. In these contact centers the information technology system has to guide the agents through a predefined menu.

In call centers two forms of operating methods are distinguished according to the initiator of the call. If customers call the call center, it is said to operate in inbound mode, otherwise the operating method is outbound. If both forms of communication are combined, the call center is called a blended or hybrid call center[11].

The durations and the times of the events are random. If an agent has just received a call he will neither know how long the service will take nor when he will receive the next call. The same holds true for e-mails and other contacts[12].

Besides the agents, the second most important component of a contact center is the so-called computer telephony integration (CTI). CTI interconnects

[6] See, e.g. Maas and Graf (2004), Borchardt et al. (2005), and Evenson et al. (1999).

[7] See, Stockmann (2005). For some facts about the call center industry in the USA, see Wharton Business School (2002), Mandelbaum et al. (2000), and Koole and Mandelbaum (2002).

[8] From now on we use both the female and the male form for a customer or agent equivalently.

[9] See, e.g. Stolletz (2003) pp. 29–32.

[10] See, e.g. Wharton Business School (2000), Lüde and Nerlich (2002), and Witness Systems (2002a).

[11] See, Gans et al. (2003) and references therein.

[12] See, Koole (2005) and references therein.

classical telecommunication and electronic data processing by an information technology system. The CTI system is the major instrument of customer experience management (CEM)[13]. On the one hand it handles all sorts of incoming and outgoing communication including phone calls, faxes, and e-mails. On the other hand it reports and supervises the processes in the contact center via a database system. In order to conduct the communication processes, the major constituents of the CTI system are the number of trunks, the automatic call distribution unit (ACD), the interactive voice response system (IVR), the voice mail server (VMS), the predictive dialling unit, and the user interfaces for the agents. The number of trunks limits the number of customers who are able to enter the system by phone at the same time. In the case of e-mail communication, the capacity of the contact center is limited by the amount of memory, by the rate of transmission, and the bandwidth[14] in the case of internet communication.

In a call center the ACD routes incoming calls to available agents according to the required skill (Skills-based routing (SBR)). Furthermore, the ACD is used to record various data, e.g., the duration of the service process or the number of arrivals frequently based on half-hour intervals.

In order to preselect the requests of different calling customers, often a technology known as interactive voice response[15] is used. This technology can also be used to fulfil standard requests, e.g., bank account inquiries via speech recognition[16].

Predictive dialling is a tool for outbound communication[17]. A telephone number of a customer is dialled just before an agent is going to be available to talk to the customer. This mechanism is used to shorten dialling and query times of the agents and thereby increase their productivity.

2.2 A Dynamic Contact Center Model

In Figure 2.1 a basic model of a contact center with a single type of contact and impatient callers is presented. This model is called the Erlang-A model, where the **A** stands for abandonment[18]. Besides the Erlang-A model, the Erlang-B[19] and Erlang-C[20] model are distinguished.

[13] See, Wharton Business School (2000), Schmitt (2003), and Witness Systems (2002b).

[14] See, e.g., Kelly and Williams (2004) and Kang et al. (2004).

[15] See, e.g., International Engineering Consortium (2005) for a definition.

[16] See, e.g., Helber and Stolletz (2003) Chapter 5.

[17] See, e.g., Chamberlain (2001) for a definition.

[18] See Garnett et al. (2002).

[19] The **B** stands for balking, i.e. impatient customers renege on instance, if they are not served immediately.

[20] The **C** stands for the number of homogeneous agents serving a single class of patient customers. This basic model was named after A.K. Erlang.

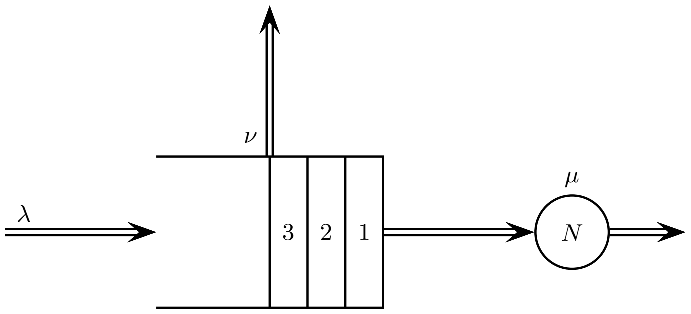

Fig. 2.1. A basic call center model with impatient customers

Contact centers are highly stochastic systems, i.e., from a managerial point of view events in a contact center happen accidentally[21]. An event is, e.g., the arrival of an e-mail or the completion of a service.

Therefore, contact centers are modelled by means of stochastic processes and queueing theory. Mathematically speaking, a stochastic process is a sequence of random variables defined on the same probability space, e.g., the number of arrivals in a half-hour interval can be represented by a random variable. If we consider a sequence of times between subsequent arrivals, we get a realisation or sample path[22] of a stochastic process describing the arrivals.

By means of many realisations, statements about the general behaviour of the processes can be made. An extensive study on empirical data in a call center of a bank was done by Brown et al. (2002). In addition to being observed, realisations in a real system can be generated in a virtual system via simulation[23]. However, simulations are often very time-consuming and expensive.

In addition to empirical analysis, the theory of stochastic processes and queueing is a powerful tool to analyse contact center models. It enables the investigation of the influence of changes in various parameters, e.g., the number of agents or the duration of service, on the performance of the contact center very quickly and accurately.

The three basic processes appearing in a contact center are the arrival process, the service process and the abandonment process. The service and abandonment process can also be summarised to a departure process.

Arrivals to a call center occur at random times. Usually, the arrival process is assumed to be a Poisson process with mean rate λ, which means the times between successive arrivals are exponentially distributed. This assumption

[21] See also Jongbloed and Koole (2001).

[22] See Whitt (2002a) Section 1.1 or Karlin and Taylor (1975).

[23] See Oh (1999), Mehrotra and Fama (2003), Cezik and L'Ecuyer (2005) and references therein.

is justified as customers arrive independently of each other and stressed by the results of Brown et al. (2002). They find that the arrival process is well modelled by a non-homogeneous Poisson process[24] with a rate function which depends on the date, the time of day, the type and priority of the call, and a variety of other facts. The exponential distribution has the so-called *lack-of-memory* property. This means that the probability of observing the next arrival within the next minute is always the same whether a call has just arrived or some time has already passed by. This describes the actual arrival of calls quite well as customers arrive independently of each other.

Beside the stochastic influence, the time-dependency of the arrival processes and its rate is very important. The arrival rate varies drastically within a day and over a week. In Figure 2.2 the number of call arrivals per half-hour interval in a call center of the Telegate AG in October, 1998, is presented. On Mondays the arrival rate is typically higher than on the other days. At the weekend the arrival rate is much lower than on working days. However, the shape of the arrival rate curve for each day looks similar. In the morning the arrival rate rises quickly to a first maximum before lunch. During the lunch break the arrivals decrease and increase afterwards. In the late afternoon a second maximum is reached. In the evening the number of arrivals decreases almost as fast as it rises in the morning[25].

By means of trigonometric functions the typical arrival rate pattern of a single day can be approximated. The sinusoidal form of the time-dependent arrival rate in Figure 2.3 is a common assumption in modelling arrival rate functions.[26]

If agents are available, arriving customers are served immediately. The duration of an individual service is also a random variable. The duration depends on the kind of request, the customer, and the agent. In the Erlang-A model the service time of each customer is assumed to be exponentially distributed with mean service time μ^{-1}.

This assumption stands in contrast to the results of Brown et al. (2002) who find that the service times are lognormally distributed. Brown et al. (2002) estimate the service time distribution under the assumption of statistically identical agents and requests. The phrase statistically identical means that agents and requests have on average the same preferences, skill levels and behaviour with respect to the service. Their analysis of the empirical data stresses early proposals of Rahko (1991) and Bolotin (1994) that service times are lognormally distributed[27].

The lognormal distribution differs from the exponential distribution commonly assumed in particular with respect to the variability of the process.

[24] See also Lariviere and Van Mieghem (2004) for a argument for the usefulness and Koole and Talim (2000).

[25] See also Haas Margolius (1999).

[26] See, e.g., Jennings et al. (1996); Feldman et al. (2005).

[27] See also Mandelbaum, Sakov, and Zeltyn (2000).

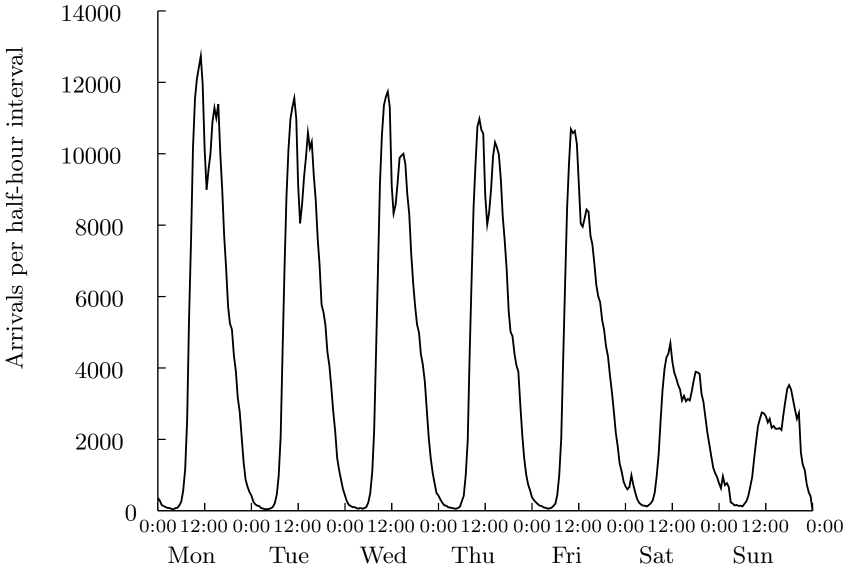

Fig. 2.2. Number of arriving calls to a call center of the Telegate AG from Monday, October 19th, 1998 to Sunday, October 25th, 1998 (See, Helber and Stolletz (2003), p.5).

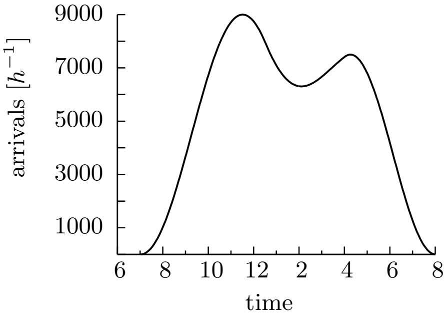

Fig. 2.3. A sinusoidal arrival rate function of a single day

However, a carefully chosen Erlang-A model with exponentially distributed service times often approximates the performance measures accurately[28]. Additionally, Brown et al. (2002) find that the service times and arrivals are positively correlated, i.e., the service times increase if more customers arrive.

If no agent is available upon an arrival, the customer has to wait. Each customer has a finite waiting time limit which describes the maximum time he is willing to wait. For each individual customer his waiting time limit is

[28] See, e.g. Whitt (2005a).

fixed, but it is unknown and random from the point of view of the contact center.

Contrary to the service and arrival rate functions and their distributions, the observation of the distribution of the waiting time limits is a serious problem[29]. The reason for the difficulties is that the waiting time limits of served customers cannot be measured directly. The waiting times of served customers are lower bounds for their waiting time limits. For customers who have abandoned, the waiting time limit can be recorded and evaluated. We assume that the waiting time limits are exponentially distributed for each customer with mean waiting time limit ν^{-1}, i.e., each customer abandons after a random amount of time with rate ν. If we assume that the customers are patient, i.e. no customer abandons ($\nu = 0$), this model turns into the Erlang-C model.

Similar to the arrival rate, the other parameters of the model depend on the time of day as well. However, their time-dependency is usually not as strong as those of the arrival rates[30], e.g., the agents might slow down during the day if they become tired. Then the average service time will increase. Other agents might become faster if they get used to their tasks. Furthermore, contrarily to the arrival rate, the time-dependency of service and abandonment rate is negligibly small. Therefore, we suppose that the mean service time and the mean time to abandonment are constant.

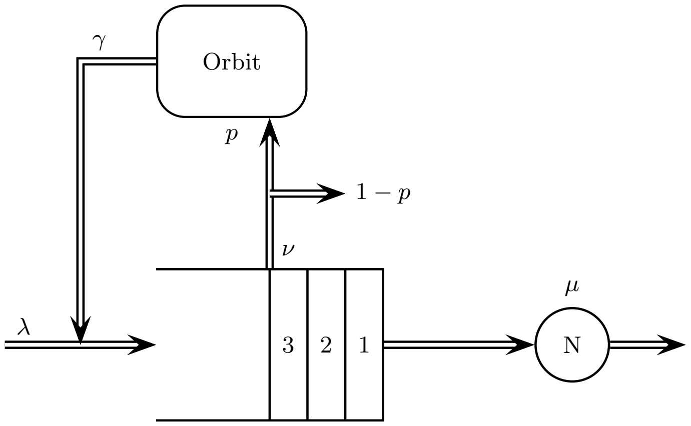

Fig. 2.4. A basic call center model with retrials of impatient customers

Besides these basic processes in Figure 2.1, the retrial process is considered in Figure 2.4. The retrial process is directly linked to the patience of customers. A customer who has abandoned might recall after a while. A fraction p of all

[29] See Whitt (1999b), Brown et al. (2002), and Mandelbaum et al. (2000) for a detailed discussion of the difficulties of reporting and analysis.

[30] See Brown et al. (2002).

customers having reneged is assumed to recall after a random time in the so-called *orbit.* This process is a connection of the arrival and the abandonment process. The abandonment process becomes the arrival process to the orbit. However, Aguir et al. (2004) find that retrials are almost independent of the mean waiting time limit, if it is exponentially distributed or deterministic. Similar to the service and abandonment process, the time-dependency of the retrial process is small.

The departure process of the orbit is an additional arrival process to the contact center. Because of the assumptions about the arrival, service and abandonment processes, the arrival process to the orbit is a Poisson process. If the number of customers in the system constantly exceeds the number of agents on duty, $\lambda - \mu N$ is the rate by which unattended customers leave the system because of the *Poisson arrival see time average* principle[31]. If a fraction of p of these customers is willing to retry, this leads to an orbit arrival rate of $p(\lambda - \mu N)$.

The orbit is modelled as an infinite server queue, because each customer can be regarded as his own server, as each customer can define his own personal sojourn time in orbit. If we assume that the durations of stay are also exponentially distributed with mean sojourn time γ^{-1} and the number of customers in the system stays above the number of agents on duty, the queueing model becomes a Jackson network[32]. Equivalently, the return of successfully served customers can be modelled by means of Jackson networks.

Other customers might not be willing to wait at all as shown in Figure 2.5, i.e., they leave the system as soon as they recognise that they are not served immediately. Those customers are said to balk. The fraction of balking customers is denoted by β. Some of these customer p_β may also recall and the residual fraction $1 - p_\beta$ of customers is assumed to be lost. Contrarily to the abandonment, balking reduces the arrival rate to the contact center, if all agents on duty are busy.

If we consider two types of requests, e.g. calls and e-mails as in Figure 2.5, or two different types of customers, priority rules for the attendance of customers must be established. If both types of customers are waiting and an agent ends a service, the question arises which request should be served next. In the case of e-mails and calls it might be reasonable to serve a call first as the e-mails are more patient. If delaying the e-mails is expensive, it might even be reasonable to serve the e-mails first. Another priority rule might put the e-mail out of the service process whenever a call arrives, i.e., the calls have *preemptive priority.* If the service of e-mails is finished before a new call

[31] See, Wolff (1982) and Artalejo (1995).

[32] See Gelenbe and Pujolle (1999, Chapter 2) and Asmussen (2003, pp. 117-122) A Jackson network is a network of queueing systems, in which the customers circulate, such that each customer enters another system with a certain, fixed probability after leaving the previous system.

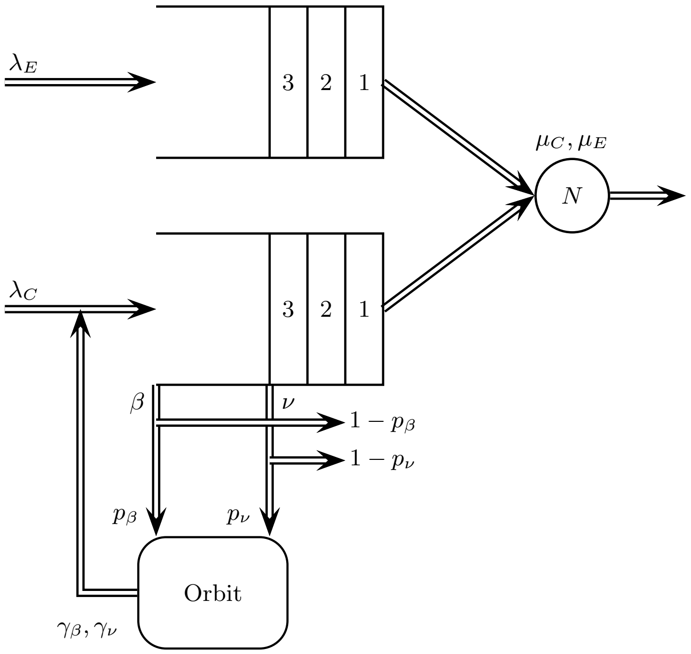

Fig. 2.5. A contact center with two types of requests

is attended to, the priority is *non-preemptive*. The service discipline of each queue is assumed to be *first come first served* (FCFS).

If different agent groups are considered, routing rules for routing the requests to the free agents have to be established[33]. In general, we assume that requests are first served by specialists for this type of contact and second by generalists.

Unfortunately, beside the extensive papers by Mandelbaum et al. (2000) and Brown et al. (2002), the number of studies on time-dependent rates is very small. All these results are based on a fixed set of empirical data. In a real-world contact center the determined distribution, mean values, and other parameters are influenced by many facts and may change permanently. Therefore, the empirically determined distributions and parameters should be checked regularly. However, some of the effects of random or changing parameters[34] can be considered in a analysis of a contact center.

[33] Stolletz (2003) gives an overview of different routing rules. Problems related to routing are also considered, e.g., byKoole and Pot (2006), Armony and Maglaras (2004), Gans and Zhou (2003), and Sisselman and Whitt (2004)

[34] Stochastic arrival rates are considered, e.g., in Harrison and Zeevi (2004, 2005); Bassamboo et al. (2005), and Whitt (2005b).

2.3 Technical Performance Measures

In order to measure the service quality of a contact center, performance measures are needed. We distinguish between technical and economical performance measures. Technical performance measures can take different perspectives on the service process depending on the individual group involved. Economical performance measures focus on the viewpoint of the management.

On the one hand such performance measures can be calculated based on empirical data collected by the reporting unit of the contact center by means of statistical methods. On the other hand performance measures can be derived from contact center models as introduced in the previous section and the assumption associated to these models. These performance measures correspond to each other because the theoretic model is a mapping of the real-world contact center. However, the statistical analysis can only be made if enough data has already been collected, against which the performance measures of the theoretic model can be calculated before any customer has called. Because of this advantage, the theoretic model can be used to determine staffing levels for different demand scenarios. These performance measures are functions of the parameters and assumptions of the stochastic processes.

A performance measure very popular in practice is the so-called X/Y *service level*[35]. It describes the percentage X of customers who wait at most Y seconds. If Y is zero, this performance measure gives the percentage of customers served immediately. This empirical performance measure corresponds to the probability of a single customer being served within Y seconds. This probability can be determined from the distribution of the waiting time W which is derived from the distribution of the service times and waiting time limits. Formally, the percentage of customers X whose service starts within at most Y seconds and the probability of being served within Y seconds are related as follows:

$$P(W \leq Y) \cdot 100 = X. \tag{2.1}$$

Therefore, the probability of delay is related to the percentage of customers being served immediately and given by

$$P(\text{delay}) = 1 - P(W = 0). \tag{2.2}$$

If a customer has to wait longer than Y seconds, the actual duration of the waiting time of this special customer does not influence the so-called service level. Only waiting times within the limit of Y are considered in this performance measure[36].

A performance measure which takes into account the tail of the waiting time distribution as well, i. e. long waiting times of customers, is the expected

[35] See, Koole (2003).

[36] For a discussion of the advantages and disadvantages of the service level see, e.g., Jackson (2002), Helber and Stolletz (2003), and Koole (2003).

waiting time $\mathbf{E}[W]$. In the case of the mathematical model the mean waiting time is derived from the waiting time distribution, while the empirical mean waiting time is the average waiting time of all customers. In order to distinguish the waiting times of served customers and those who have abandoned, these performance measures can be conditioned on the case of being served and abandoning as presented in Table 2.1. In the mathematical model the conditioned waiting time distribution has to be determined. In a real-world contact center the waiting times of customers who have abandoned and those who are served have to be collected separately.

Performance measure	Description
$P(W \leq Y \mid \text{served})$	Probability of waiting less than Y seconds given the customer is served
$P(W \leq Y \mid \text{abandon})$	Probability of waiting less than Y seconds given the customer abandons
$P(\text{delay} \mid \text{served})$	Probability of delay given the customer is served
$\mathbf{E}[W \mid \text{served}]$	Expected waiting time of served customers, also called average speed of answer (ASA)
$\mathbf{E}[W \mid \text{abandon}]$	Expected waiting time of abandoning customers.

Table 2.1. Conditional performance measures

Other interconnected empirical and theoretical performance measures related to waiting are the expected queue length $\mathbf{E}[L]$ and the expected number of customers in the system $\mathbf{E}[Q]$.

From the manager's point of view it is interesting how many out of all customers are served at all. This empirical percentage of customers corresponds to the probability of an individual customer being served $P(\text{served})$.
The individual probability of abandoning $P(\text{abandon})$ equals the long-term percentage of customers who abandon. Besides these two measures, a certain percentage of customers might balk. If the number of trunks is limited, some customers might be blocked, i.e., the customer is not able to enter the system because all telephone lines are occupied. These probabilities add up to one, i.e.,

$$1 = P(\text{blocked}) + P(\text{balk}) + P(\text{abandon}) + P(\text{served}). \qquad (2.3)$$

Finally, from the point of view of both the agents and the management the mean utilisation of the agents is important. Too high as well as too low utilisation might lead to a bad performance in economic terms. The goal is to slightly balance the utilisation of agents, because high utilisation may cause stress whereas low utilisation is expensive and can result in boredom[37]. As

[37] Compare, Wharton Business School (2004)

the number of arriving requests and the duration of service are random, the utilisation is a random variable U in the theoretic model. The utilisation is the number of busy agents divided by the number of agents on duty, which can be observed at each time in real-world systems. In order to take periods of high utilisation as well as periods of low utilisation into account, the average utilisation of the contact center can be calculated by dividing the expected number of busy agents by the total number of agents on duty. This empirical and technical measure for the workload of agents is related to the mean utilisation $\mathbf{E}[U]$ calculated by means of the distribution of the utilisation.

All these performance measures just introduced are so-called stationary measures because they average the performance of a given long time period. But real-world contact centers are highly time-dependent systems, i.e., the operating conditions change over time. Hence, these performance measures should be considered only for small time intervals to get good estimates of the dynamic performance of the contact center. However, these time intervals are interdependent in the majority of cases, e.g., a service started in one interval is continued in the subsequent interval. Therefore, each performance measure has its time-dependent form which refers to the time t of observation. The time of a virtual arrival was chosen for calculating the performance measures because the performance experienced by an arriving customer corresponds to the performance observed from an outsider[38]. The notation for time-dependent performance measures is presented in Table 2.2.

Performance measure	Description
$\mathbf{E}[W(t)]$	Expected waiting time of a customer entering the system at time t
$P(\text{delay}, t)$	Probability of delay for a customer entering the system at time t
$P(\text{served}, t)$	Probability of being served for a customer entering the system at time t
$\mathbf{E}[U(t)]$	Expected utilisation of agents observed by an arrival at time t.

Table 2.2. Time-dependent performance measures

If one wants to estimate the performance of a certain period consisting of several subsequent time intervals, performance measures should be aggregated. Periods of high load and periods of low load are considered differently with respect to their importance for the contact center[39]. Therefore we must weight the performance measure of each moment in time. If performance mea-

[38] See, Wolff (1982) and Mandelbaum et al. (1999b).

[39] See also Helber and Stolletz (2004), pp. 30-32.

sures with respect to the point of view of customers are considered, the arrival or the departure rates serve as weights. If the focus lies on the point of view of agents, the performance measures are weighted by the number of agents on duty.

In the case of constant arrival rates λ_i within each time interval $i = 1, \dots, \mathcal{J}$, the expected waiting time in the interval i is $\mathbf{E}[W_i]$. Then the aggregated mean waiting time is defined as

$$\boldsymbol{E}_{agg}[W] = \frac{\sum_{i=1}^{\mathcal{J}} \lambda_i \mathbf{E}[W_i]}{\sum_{i=1}^{\mathcal{J}} \lambda_i}. \tag{2.4}$$

If the arrival rates $\lambda(t)$, $t \in [0, T]$ are continuous functions, the aggregated waiting time can be calculated by

$$\boldsymbol{E}_{agg}[W] = \frac{\int_0^T \lambda(t) \mathbf{E}[W(t)]\ dt}{\int_0^T \lambda(t) dt}. \tag{2.5}$$

Other aggregated technical performance measures are described in Table 2.3. These performance measures are derived equivalently to Equations (2.4) and (2.5), respectively.

Performance measure	Description
$P_{agg}(W \leq Y)$	Aggregated probability that waiting time exceeds Y seconds
P_{agg}(delay)	Aggregated probability of delay
P_{agg}(served)	Aggregated probability of being served
$\boldsymbol{E}_{agg}[U]$	Aggregated mean utilisation.

Table 2.3. Aggregated performance measures

Thereby, in the case of the aggregated mean utilisation, the arrival rates are substituted by the number of agents. If the number of agents N_i is constant within the time interval i, $i = 1, \dots, \mathcal{J}$, the expected utilisation of agents in this interval is $\mathbf{E}[U_i]$. Alternatively, if the number of agents $N(t)$ depends on time, the expected utilisation at time t is $\mathbf{E}[U(t)]$. Equivalently to Equations (2.4) and (2.5), the aggregated mean utilisation is given by

$$\boldsymbol{E}_{agg}[U] = \frac{\sum_{i=1}^{\mathcal{J}} N_i \mathbf{E}[U_i]}{\sum_{i=1}^{\mathcal{J}} N_i} \tag{2.6}$$

and by

$$\boldsymbol{E}_{agg}[U] = \frac{\int_0^T N(t) \mathbf{E}[U(t)]\ dt}{\int_0^T N(t) dt}. \tag{2.7}$$

Obviously, many empirical performance measures of a real-world contact center can be calculated by means of mathematical models as introduced in the previous section. These models allow us to study the influence of the various parameters of a contact center on the performance and give rise to suggestions for improvement[40].

2.4 Operational Decision Problems in Contact Center Management

In a contact center various decisions have to be made[41]. These decisions are mostly related to information technology, the number of agents to be hired and scheduled, the number of seats and trunks, the training of agents, shift types and timing of rests.

The decision on the information technology and telephone equipment including the number of trunks and seats can be regarded as mid-term planning. The determination of the number of agents to hire, the design of their contract of employment, and the training belong to mid-term planning as well.

If these decisions have been made, shifts and breaks for rest and training have to be scheduled and agents have to be assigned to shifts. These tasks belong to the operational or short-term personnel planning[42] in a contact center. Figure 2.6 below presents a schematic view of the operational planning process.

Forecasting is the foundation of the decision process. Based on forecasts for the demand and empirical analysis of the arrival and service processes, the staffing requirements have to be determined. For this purpose, in a contact center various empirical data about arrivals, service times, abandonment, and waiting times are collected and reported by means of the automatic call distribution (ACD) unit and the voice mail server (VMS).

The recorded data represent realisations of unknown random processes and can be used to determine the distributions of the processes and their parameters[43]. Unfortunately, the stored statistical data do not represent a single process or effect in isolation but are a mixture of different events. Arrivals to the contact center may be primary attempts or retrials [44]. Furthermore, the patience or waiting time limit of served customers cannot be recorded exactly, because their waiting time is ended by service. Other processes may be composed of two processes. For example, the service time includes the call handling time and the after call work, which are reported separately. During

[40] See, e.g., Stolletz (2003) and Gans et al. (2003).

[41] See, e.g., Koole (2005) Section 3.2.

[42] See, also, Stolletz (2003) Section 2.3 or Koole (2004).

[43] See the description in Section 2.2.

[44] Aguir et al. (2004) developed a method to distinguish primary contacts from retrials.

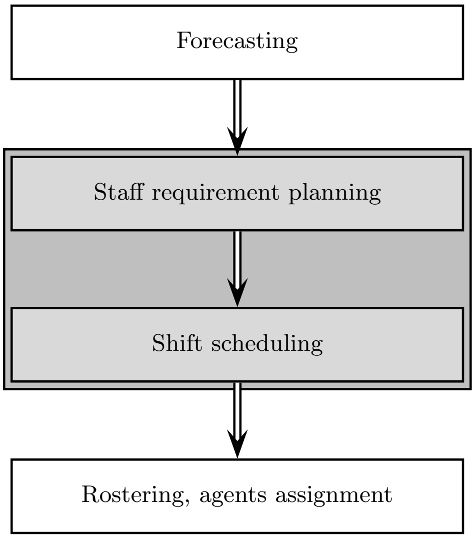

Fig. 2.6. Phases of operational decision problems in a contact center

the after call work agents take notes or complete forms. Consequently, reporting on the different processes and influences is difficult. Often additional data is needed to separate composed informations[45].

Empirical analysis of this data aims to determine the distributions of the different processes as well as mean values, variances, correlations and other parameters. Especially the mean values are used for the theoretical analysis of a contact center model. Furthermore, the calculated mean values play a central role in forecasting future demand and determining the future staff requirements[46].

Usually, the number of arrivals in separate half-hour intervals is reported and forecasted[47]. If the arrival rates, the mean service times for all types of customers and agents, and the other parameters of a contact center have been estimated, the staffing requirements can be determined. The determination of personnel requirements can be regarded as the second phase in the operational planning process depicted schematically in Figure 2.6. In this phase the minimum number of agents N_i needed in time interval $i, i = 1, \ldots, \mathcal{J}$ for all intervals of the planning period subject to performance constraints has to be determined. Commonly, the so-called X/Y service level (2.1) on Page 14 is used as a performance constraint. But also all the other performance measures presented in Section 2.3 may serve as constraints[48].

[45] See, e.g., Brown et al. (2002) and Koole (2005), Chapter 5.
[46] See, e.g., Antipov and Meade (2002).
[47] See, e.g., Helber and Stolletz (2004).
[48] See, also, Koole (2003).

If performance measure constraints are formulated and must be fulfilled in each time interval separately, we call these constraints *hard constraints*. If performance constraints are given by means of aggregated performance measures, such that intervals with high performance compensate intervals with low performance, the constraints are called *soft* constraints[49]. Furthermore, the number of agents may be limited and has to be non-negative and integer valued.

Based on the solution of such decision problems, rules of thumb have been developed which lead to quite accurate results[50]. The most famous is the so-called *square-root staffing rule*, which relates the number of agents in a time interval to the offered load and the probability of delay[51]. However, this rule relies on quite restrictive assumptions about the contact center model: A single group of agents and homogeneous customers are assumed who do not recall.

In the traditional hierarchical planning process the determination of staffing requirements is followed by the scheduling of shifts. A *shift* is a sequence of time intervals during which an agent is present in the contact center. A shift contains both periods of service and rest breaks.

The scheduling of such shifts is the major task of the third phase presented in Figure 2.6. The aim is to maximise a profit function or to minimise the costs such that the staffing requirements in each period are fulfilled for a given set of shifts. The solution of this decision problem leads to a schedule of shifts, so that in each time interval at least as many agents are on duty as determined in the previous phase of the planning process.

Finally, in the last operational planning phase, individual agents have to be assigned to the shifts of a schedule. This has to be done over consecutive days such that the conditions given by contracts of employments, personal preferences of agents and labour law are met. This decision problem is also called rostering.

In current literature[52] mostly the third and fourth phase are solved together. In the pure shift assignment problem, the days off and the location of shifts are fixed, but in the integrated decision problem the schedules of shifts, days off and allocation of agents are determined simultaneously.

Contrarily, in this thesis the staff requirement planning and the shift scheduling are integrated in one optimisation approach[53], i.e., the second and third phase in the planning process are combined as depicted in gray colour

[49] See Koole and Van der Sluis (2003) and Koole (2002, pp. 85-87).

[50] See, Jennings et al. (1996), Borst et al. (2002), Feldman et al. (2005) and references therein.

[51] See Garnett et al. (2002).

[52] See Ernst et al. (2004) for an overview over scheduling and staffing literature and Koole and Pot (2006) for an overview over staffing approaches related to contact centers.

[53] See also Ingolfsson et al. (2003), Cezik and L'Ecuyer (2005), and Bhulai et al. (2006).

in Figure 2.6. We assume that agents are available to work according to the resulting schedule. The determined schedule becomes an input of the fourth phase and cannot be changed during rostering.

In the integrative approach interdependencies of consecutive time intervals are better addressed than in the isolated approaches. Furthermore, the constraints concerning shift agreements are considered while determining the staff requirements.

Solving the staff requirement planning problem first and shift scheduling second might lead to suboptimal solutions. In the staffing requirement planning problem fewer interdependencies of consecutive time intervals are considered, because the number of agents needed in each time interval is independent of the other time intervals.

2.5 Literature related to Contact Center Management

The topic of call and contact center management is the object of many scientific research disciplines and professional journals[54]. Even various firms publish so-called white papers[55] on their homepages. Each scientific research discipline as well as all the other publications have their own point of view. Mandelbaum (2004) gives an overview of literature with abstracts covering Mathematics, Statistics, Operations Research, Industrial Engineering, Information Technology, Human Resource Mangement, Psychology and Sociology.

In applied social studies and human resource research the focus is on the behaviour and perceptions of agents and customers[56]. Another discipline is marketing. In this discipline call centers are considered as an instrument of customer relationship management[57]. The publications about contact center management related to mathematics and operations research range from statistical analysis[58] and forecasting[59] to rostering[60]. Several aspects of mathematical and operational analysis are discussed in the publications by Helber and Stolletz (2003) and Koole (2005). Koole (2002), Koole and Mandelbaum (2002), and Whitt (2002b) explain stochastic models and show how to use these models in contact center management. This research is summarised and extended in the tutorial by Gans et al. (2003).

[54] For example, Call Center News, Call Center Profi or Call Center Magazine.

[55] See, e.g., Fukunaga et al. (2002) and Witness Systems (2002a).

[56] See, e.g., Evenson et al. (1999), Lüde and Nerlich (2002), Witness Systems (2004), and references therein.

[57] See, e.g., Schmitt (2003), Borchardt et al. (2005) and references therein.

[58] See, e.g., Mandelbaum et al. (2000), Brown et al. (2002), Zohar et al. (2002), and Brown (2003).

[59] See Daley and Servi (1997), Antipov and Meade (2002) and Helber and Stolletz (2004).

[60] See, e.g., Koole and Pot (2006) for an overview.

The number of publications related to contact center management is enourmous. Therefore, the aspect of management should be specified carefully.

3

Queueing-Theoretic Approaches for Contact Center Analysis

3.1 The Stationary Erlang-A Model

3.1.1 Motivation

The traditional stationary queueing-theoretic approach relies on the assumption that the rates of the different processes, e.g. the arrival process, do not change over quite a long time period. An important condition is that the arrival rate must in the long run always be smaller than the maximum departure rate, i.e. the number of agents multiplied by their service rate in an Erlang-C model[1]. In a Erlang-A system with impatient customers the departure rate is the sum of the service rate multiplied by the number of agents and the abandonment rate multiplied by the number of waiting customers. Consequently, the arrival rate will always be smaller than the maximum departure rate in theory.

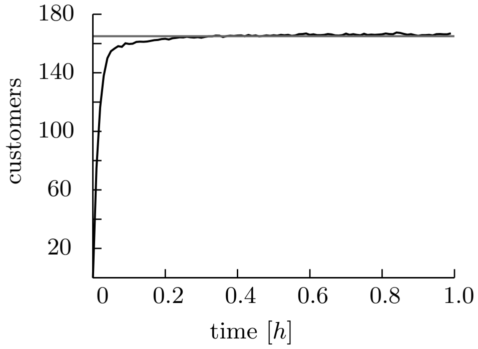

Fig. 3.1. Stabilisation of the average number of customers in the system as the simulation progresses in an Erlang-A queueing model

[1] See, e.g., Kleinrock (1975) or Karlin and Taylor (1975).

Under these conditions the queueing process will stabilise, i.e., the average number of customers in the system will converge to a constant after some time as presented in Figure 3.1.

The state of the underlying contact center queueing model can be represented by the number of orders or requests in the system. In the special case of a call center these requests are associated with the customers. In order to determine various performance measures of the contact center, the probability distribution of the number of requests in the system under stable conditions has to be calculated. Stable conditions mean that the parameters do not change for a sufficiently long time and the arrival rate is on the long run smaller than the departure rate.

3.1.2 Modelling and Justification

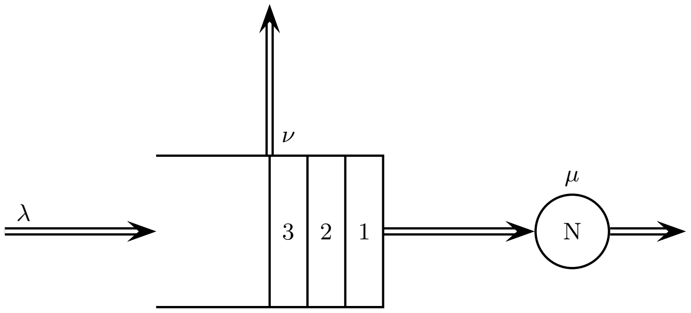

Fig. 3.2. A contact center model with impatient customers

We consider the so-called Erlang-A call center model presented in Figure 3.2. This model has already been described in Section 2.2. It has been widely discussed in the literature[2]. This contact center model is an extension of the so-called Erlang-C model commonly used in the call center industry. If we assume patient customers who do not abandon, the Erlang-A model reduces to the Erlang-C model, i.e., the arrow belonging to the abandonment rate ν in Figure 3.2 is deleted. However, the Erlang-A model has been shown to be a better model to analyse a call center[3].

As the interarrival times, the service times and the waiting time limits are assumed to be exponentially distributed, this model can be described by a *birth-death process*[4]. In a birth-death process the state of the system can only change from a state with n requests in the system to a neighbouring state either with $n-1$ or $n+1$ customers in the system. Therefore, no state

[2] See, e.g., Gans et al. (2003), Garnett et al. (2002), Koole and Mandelbaum (2002), Stolletz (2003), and references therein.

[3] See, Garnett et al. (2002), Borst et al. (2002), and Whitt (2005a).

[4] See, e.g., Asmussen (2003, pp. 71-80), or Gross and Harris (1998, pp. 45-47).

can be skipped. A transition from state n to state $n+1$ is called a birth and the intensity is denoted by β_n depending on the number of customers in the system. A transition from state n to state $n-1$ is called a death with intensity δ_n.

The steady-state probability that n customers are in the system is denoted by π_n and given by[5]

$$\pi_n = \begin{cases} \left(1 + \sum_{i=1}^{\infty} \prod_{k=0}^{i} \frac{\beta_k}{\delta_{k+1}}\right)^{-1}, & n = 0 \\ \pi_0 \prod_{i=0}^{n} \frac{\beta_i}{\delta_{i+1}}, & n \geq 1. \end{cases} \tag{3.1}$$

In the case of the Erlang-A model presented in Figure 3.2 the birth rate $\beta_n = \lambda$ is independent of the number of customers in the system, because the waiting room is supposed to be unlimited and balking of customers is excluded. The death rates of this model are given by

$$\delta_n = \begin{cases} n\mu, & n \leq N \\ N\mu + (n-N)\nu, & n > N \end{cases} \tag{3.2}$$

Hence, the stationary distribution of the number of requests in the Erlang-A model considered is given by

$$\pi_n = \begin{cases} \left(1 + \sum_{i=1}^{N} \frac{\lambda^i}{i!\mu^i} + \sum_{i=N+1}^{\infty} \frac{\lambda^i}{N!\mu^N \cdot \prod_{k=0}^{i-N}(N\mu + k\nu)}\right)^{-1}, & n = 0 \\ \frac{\lambda^n}{n!\mu^n}\pi_0, & 0 < n \leq N \\ \frac{\lambda^n}{N!\mu^N \cdot \prod_{k=0}^{i-N}(N\mu + k\nu)}\pi_0, & \text{otherwise.} \end{cases} \tag{3.3}$$

By means of these probabilities many performance measures for contact center analysis can be determined as shown in the following section.

3.1.3 Performance Measures

The probability distribution (3.3) is used to calculate the technical performance measures discussed in Section 2.3. First of all, the mean number of customers in the system[6] is derived by weighting the number of customers by its steady-state probability, i.e.,

[5] See Asmussen (2003, Corollary 2.5) on Page 74

[6] See, e.g., Kleinrock (1975).

$$\mathbf{E}[Q_S] = \sum_{n=1}^{\infty} n\pi_n \tag{3.4}$$

If we subtract the mean number of customers in service, we get the mean number of waiting customers, which is given by

$$\mathbf{E}[Q_W] = \sum_{n=N+1}^{\infty} (n-N)\pi_n \tag{3.5}$$

A single waiting customer abandons with rate ν. Therefore the abandonment rate of all waiting customers is derived by multiplying the mean number of waiting customers $\mathbf{E}[Q_W]$ by the individual abandonment rate. Therefore, in a time interval of length Δt the mean number of abandoning customers is $\nu\mathbf{E}[Q_W]\,\Delta t$. During the same time interval on average $\lambda\Delta t$ customers arrive. Consequently, the percentage of abandonments is given by the mean number of customers abandoning divided by the mean number of customers who have entered the system[7], i.e.,

$$P(\text{abandon}) = \frac{\nu\Delta t}{\lambda\Delta t} \cdot \mathbf{E}[Q_W] = \frac{\nu}{\lambda}\mathbf{E}[Q_W]\,. \tag{3.6}$$

Each customer who has entered the system must leave somehow; the probability of being served is simply calculated by subtracting the probability of abandoning from one, which is the probability that all customers are served.

$$P(\text{served}) = 1 - P(\text{abandon}) = 1 - \frac{\nu}{\lambda}\mathbf{E}[Q_W]\,. \tag{3.7}$$

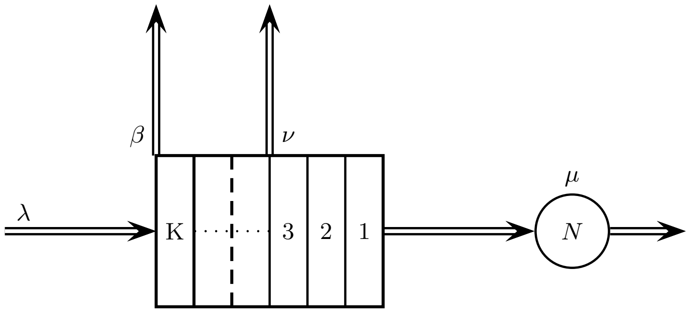

Fig. 3.3. The Erlang-A model with finite waiting room and balking

An extension of the Erlang-A model presented in Figure 3.3 is a model with a finite waiting room of size K and a certain fraction b of balking customers, who tolerate no waiting at all[8]. Under these conditions also the probability of blocking and balking can be calculated. The probability of being

[7] See, e.g., Mandelbaum and Shimkim (2000).

[8] See, e.g., Whitt (1999a) and Stolletz (2003).

blocked $P(\text{blocked})$ is given by the steady-state probability π_K that exactly K customers are in the system, because no additional customer can enter the system. The probability of balking is given by the product of the fraction of customers who balk, if they have to wait, and the probability that an arriving customer will have to wait. This probability is the sum of the stationary probabilities of the states with at least N customers in the system and at most $K-1$. Therefore we get

$$P(\text{balk}) = b \sum_{n=N}^{K-1} \pi_n. \tag{3.8}$$

For this model the probability of being served is

$$P(\text{served}) = 1 - P(\text{blocked}) - P(\text{balk}) - P(\text{abandon}). \tag{3.9}$$

Another important performance measure is the mean waiting time of customers. In the case of an unlimited waiting room, this measure is calculated according to Little's law[9] by dividing the number of waiting customers by the arrival rate

$$\mathbf{E}[W] = \frac{\mathbf{E}[Q_W]}{\lambda}. \tag{3.10}$$

The mean waiting time is the time both abandoning and served customers have been waiting in the system. In order to distinguish the waiting time of served customers from the waiting time of abandoning customers the conditional mean waiting times should be calculated. The derivation of these conditional mean waiting times is presented by Stolletz (2003, pp. 77-79).

Finally, the mean utilisation of agents should be considered, which is a measure of the burden of work the agents undertake. The mean utilisation is calculated by weighting the percentage of occupied agents in each state by the respective steady-state probabilities, i.e.,

$$\mathbf{E}[U] = \sum_{i=0}^{N-1} \frac{i}{N} \pi_i + \sum_{i=N}^{\infty} \pi_i. \tag{3.11}$$

For the analysis and comparison of approaches it often suffices to restrict oneself to the mean number of customers in system, the mean waiting time, the probability of being served, and the utilisation of agents, although further performance measures could be calculated. This restriction is reasonable because these performance measures can present all major dimensions of performance of the call center.

However, for the sake of completeness we present the stationary waiting time distribution, which is related to the empirical so-called X/Y service level[10] widely used in practice. By means of the stationary waiting time distribution, the probability that an arriving customer has to wait less than t

[9] See, e.g., Kleinrock (1975).

[10] See Section 2.3.

seconds before being served can be calculated. The derivation of this waiting time distribution in the case of an Erlang-A model with finite waiting room and balking customers can be found in Stolletz (2003) and is given by

$$\begin{aligned} &P(W \leq t \mid \text{served}) \qquad (3.12) \\ &= \frac{1}{P(\text{served})} \left(\sum_{n=0}^{N-1} \pi_n + \sum_{n=N}^{\infty} \frac{\prod_{i=0}^{n-N} (N\mu + i\nu)}{(n-N)!} \right. \\ &\qquad \left. \int_0^t \left(\frac{1 - e^{-\nu\tau}}{\nu} \right)^{(n-N)} e^{-\tau(N\mu+\nu)} d\tau \right) . \end{aligned}$$

In this section we have presented several performance measures with varying usefulness for the analysis. In more complex contact centers the calculation of performance measures becomes more and more difficult.

3.1.4 Applicability and Limitation for Contact Center Analysis

The stationary approach for call centers is quite popular in the literature[11] as it allows us to calculate a variety of performance measures. Another advantage is the fact that this approach takes randomness of the arrival and departure processes into account. However, only few probability distributions for the service duration, mean time to abandonment and interarrival times can be mathematically analysed. Therefore, restrictive assumptions on these probability distributions have to be made.

As far as the arrival rates are concerned, the assumption of exponentially distributed interarrival times seems to be quite reasonable[12]. However, statistical analysis has shown that the service times are more likely to be lognormally distributed[13]. For the distribution of the abandonment times we are not able to make any statements because of several difficulties in determining the waiting time limits of served customers as mentioned in Section 2.2[14].

The assumption of constant rates is problematic, as in real-world call center environments the arrival rates in particular change drastically[15] over the day. The other rates often vary very little.

A modification of the stationary approach which takes some time-dependencies into account is the so-called SIPP approach[16]. The SIPP approach

[11] See Gans et al. (2003) and references therein.

[12] See Lariviere and Van Mieghem (2004) and Section 2.2.

[13] See, e.g., Brown et al. (2002), Mandelbaum et al. (2000) and Section 2.2.

[14] See also Brown et al. (2002).

[15] See Figure 2.2 and the argumentation in Section 2.2 as well as Helber and Stolletz (2003) and references therein for a discussion of the problem associated with time-dependencies.

[16] A description of the **s**tationary **i**ndependent **p**eriod by **p**eriod approach can be found, e.g., in Green et al. (2001).

approximates the call center as a stationary system in each period, wherein the different periods are supposed to be independent of each other. In such a short interval the arrival rates are supposed to be nearly constant.

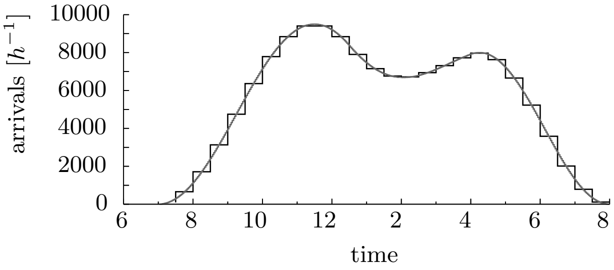

Fig. 3.4. The arrival rate functions for the simulation and the SIPP approach

A comparison of this approach to simulation results of a contact center model with time-dependent arrival rates is given in Figure 3.5. The simulation was done by a computer program in C++ designed by Feldman[17]. The gray sinusoidal arrival rate depicted in Figure 3.4 was generated by the equation

$$\lambda(t) = \begin{cases} \frac{1}{2}m_1 \cdot \left(1 - \cos\left(2\pi \frac{t-t_0}{t_2-t_0}\right)\right) & \text{for } t_0 \le t < t_1 \\ \frac{1}{2}m_1 \cdot \left(1 - \cos\left(2\pi \frac{t-t_0}{t_2-t_0}\right)\right) & \\ \quad +\frac{1}{2}m_2 \cdot \left(1 - \cos\left(2\pi \frac{t-t_1}{t_3-t_1}\right)\right) & \text{for } t_1 \le t < t_2 \\ \frac{1}{2}m_2 \cdot \left(1 - \cos\left(2\pi \frac{t-t_1}{t_3-t_1}\right)\right) & \text{for } t_2 \le t < t_3 \end{cases} \tag{3.13}$$

with the following parameters

$$\begin{array}{lll} m_1 = 9500 & t_0 = 7\,(\text{for 7 am}) & t_2 = 16\,(\text{for 4 pm}) \\ m_2 = 8000 & t_1 = 12.5\,(\text{for 12:30 pm}) & t_3 = 20\,(\text{for 8 pm}). \end{array} \tag{3.14}$$

In order to apply the SIPP approach, the mean arrival rate for half-hour intervals is calculated. Both arrival rate functions are presented in Figure 3.4. The mean service time is assumed to be $\mu^{-1} = 1$ minute. The mean waiting time limit of customers is $\nu^{-1} = 30$ seconds.

Besides the problems related to time-dependencies, if the structure of the contact center becomes more complicated[18], the generator matrix of the Markov chain grows drastically and its structure becomes more complex. Hence, more linear equations have to be generated and solved. If we consider different server groups, customer classes or retrial, the linear equations system is nearly unsolvable. Additionally, numerical instabilities[19] occur with

[17] Feldman (2004)

[18] See, e.g., Figure 2.4 on Page 11.

[19] See, e.g., Stewart (1994), Seneta (1967, 1968, 1980), Tweedie (1973), and Van der Cruyssen (1979).

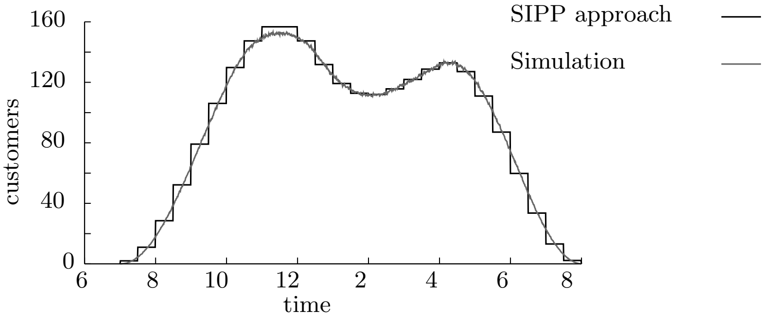

Fig. 3.5. Comparison of the SIPP-approach to simulation results for a contact center model with time-dependent arrival rates

the standard algorithms used for solving systems of linear equation. Therefore different methods have been developed[20].

To overcome some of these shortcomings we use the fluid approach and the diffusion refinement described in the next section, which can deal with time-dependencies and more complex structures.

3.2 A Non-Stationary Fluid Approach for the Erlang-A Model

3.2.1 Motivation

In order to model and analyse contact center models with time-varying parameters as discussed in Section 2.2, other approaches than the stationary approach are needed. A major argument[21] against the stationary approach is that in the case of time-varying parameters, no steady state is reached and in reality the parameters change quickly. Another advantage of the fluid approach over the stationary approach in Section 3.1 is that the fluid approach allows for interdependencies[22] of periods in modelling.

The idea of the fluid approximation is to substitute the random, time-dependent, and discrete processes by deterministic, time-dependent, and continuous processes[23] which represent the mean of the original discrete random processes. Then one is able to describe the change in the system by some physical principles which are used to model liquids moved by pumps. The underlying notion is to think of a continuous amount of customers which runs around the system instead of discrete, individual customers. Then the amount of customers is like some fluid, which flows into the system with some

[20] See, e.g., Seneta (1980), Neuts (1981), Hanschke (1992, 1999), Schmidt (1997), and references therein.

[21] See Mandelbaum et al. (1998).

[22] See, e.g., Mandelbaum and Massey (1995) or Jiménez and Koole (2004).

[23] Jiménez and Koole (2004).

arrival rate and is either processed by the pumps representing the servers or the pumps representing the departure because of impatience. A description of such a system will give rise to a system of differential equations which can be solved efficiently by standard numerical methods.

Based on Mandelbaum et al. (1998), we show how the system of differential equations arises from a modelling of a stochastic contact center system by applying the strong law of large numbers. The differential equations describe the mean number of customers in the contact center with time-varying parameters.

3.2.2 Modelling and Justification

In order to illustrate the fluid approach, we consider the special contact center model depicted in Figure 3.6 which has been analysed within this framework by Mandelbaum et al. (1998). This model is a time-dependent form of the Erlang-A model considered in the previous Section 3.1 in Figure 3.2 on Page 24.

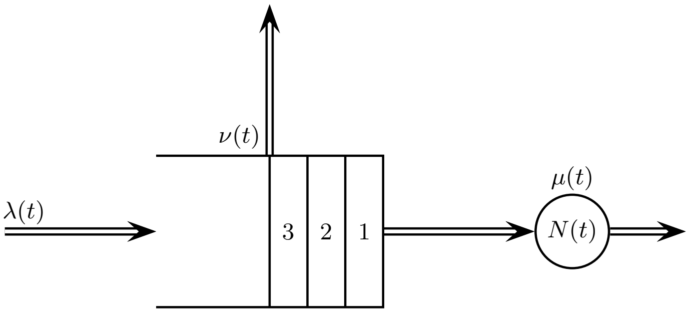

Fig. 3.6. A contact center model with impatient customers and time-dependent rates

Statistically identical customers arrive according to a non-homogeneous Poisson process with time-dependent rate $\lambda(t)$ at time t. A single group of $N(t)$ homogeneous agents attends the customers with time-dependent rate $\mu(t)$. If a customer arrives, she is either served immediately or she has to wait until an agent becomes available or her waiting time limit is reached. On average each customer abandons with rate $\nu(t)$.

In order to derive the fluid approximation, the random number of customers in the system waiting or being served at time t is denoted by $Q(t)$. The stochastic process of the number of discrete customers in the system $Q(t)$ is a composed process. This process consists of three stochastic processes, so-called *counting processes*[24], which are assumed to be stochastically independent. The counting processes are non-homogeneous Poisson processes with

[24] See Karlin and Taylor (1975).

time-dependent rates and describe the arrivals to and departure of customers from the system.

Customers arrive with rate $\lambda(t)$. Therefore the arrival process is given by[25]

$$A_1\left(\int_0^t \lambda(s)\,ds\right), \tag{3.15}$$

with $A_1(\cdot)$ being a standard Poisson process[26] with rate 1. The service

$$A_2\left(\int_0^t \mu(s)\min\{Q(s), N(s)\}\;ds\right) \tag{3.16}$$

and the abandonment process

$$A_3\left(\int_0^t \nu(s)\{Q(s) - N(s)\}^+\,ds\right) \tag{3.17}$$

are also Poisson processes which describe the departure of customers. By $\{X\}^+$ the maximum of zero and X is denoted.

A departure of a customer decreases the number of customers in the system. Therefore, the standard Poisson processes $A_2(\cdot)$ and $A_3(\cdot)$ both have a negative sign in the Equation (3.18) for the composed stochastic process $Q(t)$ which describes the random number of customers in the system at time t, i.e.,

$$\begin{aligned} Q(t) = Q(0) + A_1\left(\int_0^t \lambda(s)\,ds\right) - A_2\left(\int_0^t \mu(s)\min\{Q(s), N(s)\}\;ds\right) \\ - A_3\left(\int_0^t \nu(s)\{Q(s) - N(s)\}^+\,ds\right) \end{aligned} \tag{3.18}$$

All solutions of this stochastic equation define a unique set of the sample paths[27] or realisations of this process.

In order to derive an approximating fluid process for the stochastic process of the number of customers in the system presented in Equation (3.18), we follow the idea of Halfin and Whitt (1981)[28]. For this approach the arrival rate and the number of servers is scaled by the same factor n. For Poisson

[25] See Massey (2002), Mandelbaum et al. (1998), Mandelbaum et al. (1999a,b), and references therein.

[26] See Whitt (2002a)

[27] See Whitt (2002a) and Haas Margolius (1999).

[28] See, also Whitt (2003) and references therein.

processes this proceeding is equivalent to multiplying the arrival rate as well as the number of agents by n. Contrary to the classic fluid approach developed by Newell in 1971[29], the service rate remains unchanged but time-dependent. Therefore, the traffic intensity

$$\rho = \frac{\lambda(t)}{N(t)\mu(t)} \tag{3.19}$$

remains constant.

By means of this so-called *Halfin-Whitt scaling*, we get insights into the behaviour of the stochastic system becoming large. If the arrival rate and the number of servers increase, the number of customers in the system will increase by the same factor. Therefore, this approach is sometimes also called *heavy-traffic* but we prefer to call it *strong approximation*[30] according to Mandelbaum et al. (1998). In conventional heavy-traffic[31] the traffic intensity approaches one in contrast to the Halfin-Whitt scaling.

In order to illustrate the Halfin-Whitt scaling, Figure 3.7 presents some simulation results for the scaled processes of the number of customers in the system $Q^n(t)$ divided by the scaling parameter n. The simulation was done by the same computer program in C++ designed by Feldman (2004) mentioned in the previous section. Figure 3.7 shows a single realisation of the scaled processes for $n = 1, 10, 100$, and 1000. For the simulation the arrival rate function $\lambda(t)$ had the sinusoidal form of Figure 2.3 on Page 10 and Equation (3.13) on Page 29 with parameters

$$\begin{array}{lll} m_1^{(1)} = 800 & t_0^{(1)} = 7 \text{ am} & t_2^{(1)} = 4 \text{ pm} \\ m_2^{(1)} = 750 & t_1^{(1)} = 12{:}30 \text{ pm} & t_3^{(1)} = 8 \text{ pm.} \end{array} \tag{3.20}$$

Furthermore, the mean service time μ^{-1} was assumed to be one minute and the average waiting time limit ν^{-1} thirty seconds.

In Figure 3.7 the scaled processes stabilise around the mean, if the scaling parameter grows[32]. A comparison of the simulation results of the scaled process for $n = 1000$ and the fluid limit is given in Figure 3.8. The convergence depicted in Figure 3.7 can be explained by the *functional strong law of large numbers*[33].

[29] See Newell (1982) and Jiménez and Koole (2004) for a comparison of the two approaches.

[30] Mandelbaum, Massey, and Reiman (1998) introduce the term *strong approximation* for the fluid approach and the diffusion refinement discussed in Section 3.3. They develop a whole theory of deriving fluid and diffusion approximations from stochastic model.

[31] See, Harrison and Zeevi (2004) for a comparison of the system behaviour in the Halfin-Whitt regime and conventional heavy-traffic.

[32] See also Whitt (2002a) for further results due to scaling.

[33] See Asmussen (2003) and Whitt (2002a).

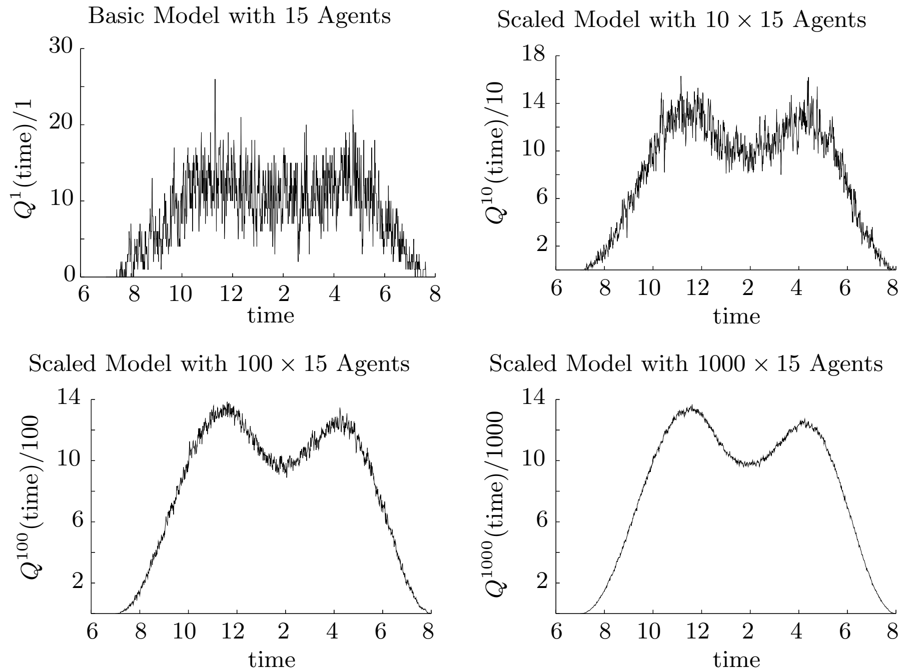

Fig. 3.7. Convergence of the scaled process $Q^n(t)/n$ of the number of customers in a contact center model to the mean

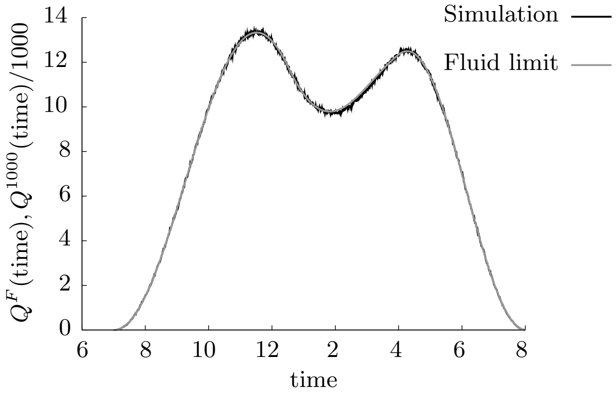

Fig. 3.8. Comparison of the limiting fluid process $Q^F(t)$ for the amount of customers in the system to the scaled process $Q^{1000}(t)/1000$

Formally, the random number of customers in a scaled system with scaling parameter n is equivalent to the sum of n independent Poisson processes[34] defined on the same probability space with the same mean. If $Q^n(0)$ is the number of customers in the scaled system at the beginning of the considered time period, e.g., a day or a week, the number of customers in the scaled sys-

[34] See Whitt (2003) and Mandelbaum et al. (1998).

tem at some time $t > 0$ is given by the number of customers at the beginning $Q(0)$ and the number of customers who have entered the system reduced by the number of departures. In order to apply the functional strong law of large numbers we divide the scaled process $Q^n(t)$ by the scaling parameter n. We get[35]

$$\begin{aligned}\frac{Q^n(t)}{n} = \frac{Q^n(0)}{n} &+ A_1\left(\int_0^t \frac{1}{n}(n\lambda(s))\,ds\right) \\ &- A_2\left(\int_0^t n\mu(s)\min\left\{\frac{1}{n}Q^n(s), N(s)\right\}\,ds\right) \\ &- A_3\left(\int_0^t n\nu(s)\left\{\frac{1}{n}Q^n(s) - N(s)\right\}^+ ds\right).\end{aligned} \tag{3.21}$$

If we pass n to infinity, we can apply the functional strong law of large numbers which states that the term $Q^n(t)/n$ in Equation (3.21) almost surely converges to a mean process $\mathbf{E}[Q(t)]$. The result is a functional equation (3.22) for the mean process which is a deterministic process. In other words, the scaled number of customers in the system is a sum of n independent and identically distributed random variables with the same mean. Therefore, the sum divided by the number of summands n converges to the mean number of customers in the system if n increases. Formally we get

$$\begin{aligned}\mathbf{E}[Q(t)] = \mathbf{E}[Q(0)] + \int_0^t \lambda(s)\,ds - \int_0^t \mu(s)\min\{\mathbf{E}[Q(s)], N(s)\}\,ds \\ - \int_0^t \nu(s)\{\mathbf{E}[Q(s)] - N(s)\}^+\,ds\end{aligned} \tag{3.22}$$

The process of the number of customers in the system no longer depends on either the growing number of agents or the increasing number of customer arrivals according to the scaling parameter. The mean process describes very well the average number of customers determined by statistical analysis of simulation results, which can be confirmed by the comparison of the approximation to simulation results in Figure 3.8.

Denoting the deterministic mean process $\mathbf{E}[Q(t)]$ by $Q^F(t)$ and differentiating (3.22) with respect to t gives rise to the differential equation (3.23), which describes the change in a deterministic fluid system within an infinitesimal time interval.

[35] See Mandelbaum et al. (1998).

$$\frac{d}{dt}Q^F(t) = \lambda(t) - \mu(t)\min\{Q^F(t), N(t)\} - \nu(t)\{Q^F(t) - N(t)\}^+. \tag{3.23}$$

In order to model small changes in the system, the fluid approach assumes that the number of customers in the system becomes a continuous quantity like the mean. Nevertheless, we continue to speak about numbers of customers in general. Then the change in the amount of customers in the system is given by the amount of customers entering the system per time unit reduced by the amount of customers leaving. This gives rise to a more simple derivation of the fluid approximation by describing the change in the amount of customers by means of the rates.

Customers move into the system with a time-dependent arrival rate $\lambda(t)$ and depart after either being served or reaching their waiting time limit. The amount of customers leaving the system per time unit after being served is the service rate $\mu(t)$ multiplied by the minimum of the level of customers and the number of servers. If the amount of customers exceeds the number of servers, the leaving rate of waiting customers is the abandonment rate $\nu(t)$ multiplied by the level of waiting customers in the system.

Contrary to the approach by Halfin and Whitt (1981) described earlier and used in this thesis, the classic method of deriving fluid limits for such systems was first developed by Newell in 1971[36]. Newell averages the arrival process over many realisations. For Poisson arrival and departure processes, this is equivalent to scaling the arrival, the service rate, and the abandonment rate by the same parameter n. A disadvantage of the classical approach is that the scaling leads to identical limits for systems with a single server and with many servers[37]. Furthermore, Altman et al. (2001) show that the fluid limit derived by this classical method is a lower bound of the original system.

Contrarily, Garnett et al. (2002) show that the method of Halfin and Whitt can be applied to multi-server queues with abandonment and lead to different limits for single- and multi-server queues. Mandelbaum et al. (1998) illustrate that the fluid limit derived by the method of Halfin and Whitt differs from the limit derived by the classical method and has a higher value. However, the fluid limit is still a lower bound[38].

To show how accurate the fluid approach approximates the mean number of customers in the system in Figure 3.9, the fluid results are compared to the results of the simulation program of Feldman.

In this example the arrival function is given by Equation (3.13) with the same parameters as for the stationary approach in Equation (3.14) on Page 29:

$$\begin{array}{lll} m_1 = 9500 & t_0 = 7 \text{ am} & t_2 = 4 \text{ pm} \\ m_2 = 8000 & t_1 = 12:30 \text{ pm} & t_3 = 8 \text{ pm}. \end{array} \tag{3.24}$$

[36] See Newell (1982).

[37] This is shown by Mandelbaum et al. (1998).

[38] See, Jiménez and Koole (2004).

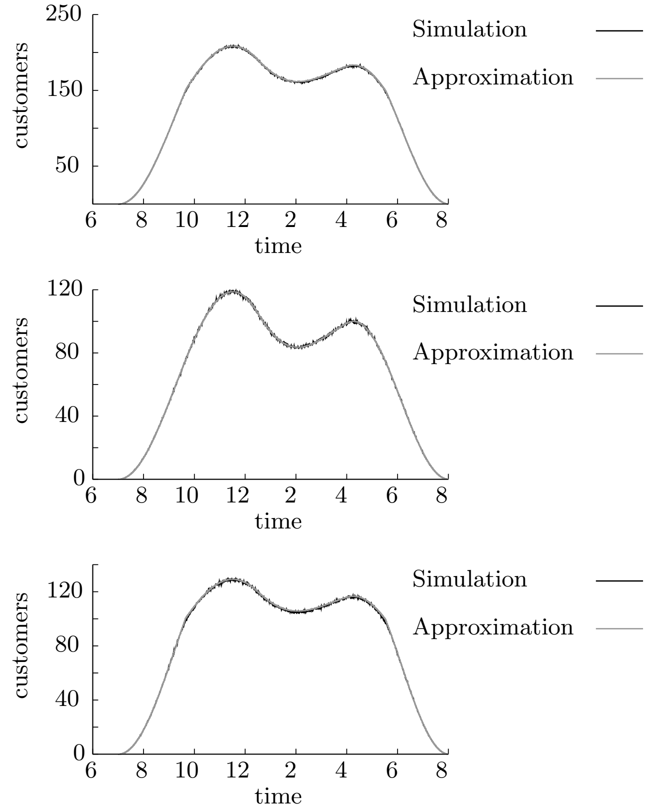

Fig. 3.9. Example for the accuracy of the fluid approximation for contact centers with time-varying arrival rate and impatient customers

The arrival rate function is shown in Figure 3.4 on Page 29. The other parameters of the contact center model analysed in Figure 3.9 are given in Table 3.1.

Picture	Service rate $\mu(t)$	Abandonment rate $\nu(t)$	Number of agents $N(t)$
1	$40\,\mathrm{h}^{-1}$	$60\,\mathrm{h}^{-1}$	150
2	$80\,\mathrm{h}^{-1}$	$60\,\mathrm{h}^{-1}$	150
3	$60\,\mathrm{h}^{-1}$	$120\,\mathrm{h}^{-1}$	100

Table 3.1. Parameters of the contact center model analysed in Figure 3.9.

If the mean service time μ^{-1} is long, as shown in the first picture of Figure 3.9, more customers are on average in the system than in the other cases. The approximation of the mean process calculated by means of 500 simulation runs is very accurate. A comparison of the second and third picture shows that a high abandonment reduces the number of customers further, although fewer agents are on duty. The curve of the mean number of customers in the system follows the shape of the arrival rate function as depicted in Figure 2.3 on Page 10.

In this section we have shown how fluid approximations can be derived by means of the strong law of large numbers. Furthermore, we have observed that the fluid approach approximates the mean number of customers in the system accurately.

3.2.3 Performance Measures

To evaluate the technical and economical performance of the presented contact center, performance measures as described in Section 2.3 are needed. Unfortunately, not all the performance measures illustrated in Section 2.3 can be calculated by means of the fluid approach. However, most important technical measures from the point of view of both customers and agents can be determined as well as a performance measure for the point of view of managers.

3.2.3.1 Technical Performance Measures

First of all, the average number of customers in the system $Q^F(t)$ and the number of customers waiting $\max\{0, Q^F(t) - N(t)\}$ are directly derived by numerically solving the initial value problem given by Equation (3.23) and some initial condition for the number of customers,

$$Q^F(t_0) = Q_0. \tag{3.25}$$

Therefore, these performances are not presented again.

As mentioned before, most customers are impatient and prefer short waiting times. That is why the waiting time of a customer arriving at time t is an important technical performance measure. It is given by dividing the mean number of customers waiting $\max\{0, Q^F(t) - N(t)\}$ by the departure rate at time t.

The departure rate $d(t)$ is given by the amount of customers served per time unit plus the amount of customers abandoning per time unit. If fewer customers than agents are in the system, i.e., $Q^F(t) \leq N(t)$, the departure rate is the product of the service rate and the number of customers in the system. If the amount of customers $Q^F(t)$ in the system exceeds the number of available servers, the amount of customers served is the product of the service rate $\mu(t)$ and the number of agents $N(t)$. The amount of customers leaving the system per time unit because of their impatience is the product of the abandonment rate $\nu(t)$ and the amount of customers that exceeds the number of agents on duty $(Q^F(t) - N(t))$. Formally, the departure rate is given by

$$d(t) = \begin{cases} \mu(t)Q^F(t), & \text{if } Q^F(t) \leq N(t) \\ \\ \mu(t)N(t) + \nu(t)(Q^F(t) - N(t)), & \text{if } Q^F(t) > N(t). \end{cases} \tag{3.26}$$

As the waiting time of customers is zero, if fewer customers are in the system than agents are on duty, the expected waiting time of a customer entering at time t is given by

$$W^F(t) = \frac{\left\{Q^F(t) - N(t)\right\}^+}{\mu(t)N(t) + \nu(t)(Q^F(t) - N(t))}, \tag{3.27}$$

where $\{X\}^+$ again denotes the maximum of zero and X.

Formally, a similar equation for the virtual waiting time, which is the time a customer arriving at time t has to wait for service under the condition that this special customer does not abandon, is derived in Mandelbaum et al. (1999b, 2002). They assume that at the beginning there are more customers in the system than available servers as otherwise the virtual waiting time would be zero. After the arrival of a chosen customer the arrival rate is supposed to become zero, such that there are no further arrivals to the system. Under these conditions the virtual waiting time is given by the time until the number of customers in the system becomes equal to the number of servers on duty, which is equivalent to the fact that the chosen customer starts his service.

For an analysis of the contact center it is useful to determine the aggregated or mean waiting time of a certain period, e.g. a day. A possibility of aggregation is averaged over the considered period. This is done by integrating the waiting time function over the time horizon and dividing by the length of the period:

$$W_T^F = \frac{1}{T}\int_0^T W^F(t)dt \tag{3.28}$$

However, this does not take into account the varying congestion and number of customers in the system[39]. Consequently, the waiting time should be weighted by a factor, which describes the congestion in the system, e.g., the departure or arrival rate. Both describe how the congestion in the system increases and decreases. For a customer entering the system the departure rate refers to the events in front of him, against which the arrival rate characterises the congestion behind the waiting customer. At each moment in time an arriving customer will be more interested in the number of customers in front. Therefore, we aggregate the waiting time by weighting by the departure rate. This gives rise to

$$\begin{aligned} W_{agg}^F(T) &= \frac{\int_0^T d(t)W^F(t)\,dt}{\int_0^T d(t)\,dt} \\ &= \frac{\int\limits_0^T d(t)\frac{\left\{Q^F(t) - N(t)\right\}^+}{d(t)}dt}{\int_0^T \mu(t)N(t) + \nu(t)(Q^F(t) - N(t))\,dt} \end{aligned}$$

[39] See Helber and Stolletz (2004), pp. 30-32, for a discussion, or Koole (2005) and Section 2.3.

$$= \frac{\int_0^T \big\{Q^F(t) - N(t)\big\}^+ dt}{\int_0^T \mu(t)N(t) + \nu(t)(Q^F(t) - N(t))\, dt}. \tag{3.29}$$

A technical performance measure closely related to economic performance measures is the probability $P(\text{served}, t)$ that a customer is served at time t and the probability $P(\text{abandon}, t)$ that a customer abandons at time t. In a fluid model the rate of customers being served at time t is given by the service rate of the agents multiplied by the minimum of the amount of customers in the system and the number of agents on duty, i.e., $\mu \min\{Q^F(t), N(t)\}$. The probability that a customer is served is given by dividing the number of customers served by the number of customers leaving the system $d(t)$. Similarly, the probability of abandoning in (3.31) is derived. This gives rise to

$$P^F(\text{served}, t) = \frac{\mu(t) \min\big\{Q^F(t), N(t)\big\}}{\mu(t)\min\{Q^F(t), N(t)\} + \nu(t)\{Q^F(t) - N(t)\}^+} \tag{3.30}$$

$$P^F(\text{abandon}, t) = \frac{\nu(t)\big\{Q^F(t) - N(t)\big\}^+}{\mu(t)\min\{Q^F(t), N(t)\} + \nu(t)\{Q^F(t) - N(t)\}^+}. \tag{3.31}$$

Both probabilities add to one, because a customer can either leave after being served or abandon. If no customer is waiting in the system, the probability of being served in (3.30) is one and the probability of abandoning is zero. Furthermore, if the service rate and abandonment rate are constant, these probabilities are both almost independent of the abandonment rate[40], although this result might not be obvious at once especially for the second probability.

If the aggregated probability of being served $P^F_{agg}(\text{served}, T)$ in the time interval starting at time $t = 0$ and ending at time $t = T$ is derived by accumulating the number of served customers $\mu(t) \min\big\{Q^F(t), N(t)\big\}$ at each time t and dividing by the accumulated departure rate, then we get the probability of being served for all customers who have left the system in the considered time interval. Formally, we get for the aggregated probability of being served

$$P^F_{agg}(\text{served}, T) = \frac{\int_0^T \mu(t)\min\{Q(t), N(t)\}\ dt}{\int_0^T d(t)\, dt}. \tag{3.32}$$

Equivalently the aggregated probabilities of abandoning can be derived:

$$P^F_{agg}(\text{abandon}, T) = \frac{\int_0^T \nu(t)\{Q(t) - N(t)\}^+ dt}{\int_0^T d(t)\, dt}. \tag{3.33}$$

In addition to the aggregated probability of all customers who leave the system in the time interval being served, we can determine the aggregated probability $P^F_\lambda(\text{served}, T)$ of all customers who have entered the system in the time interval $[0, T]$ being served. This probability is derived by dividing

[40] See Aguir et al. (2004).

the number of served customers until time T by the accumulated number of arrivals $\int_0^T \lambda(t)\,dt$, i.e.,

$$P_\lambda^F(\text{served}, T) = \frac{\int\limits_0^T \mu(t) \min\{Q(t), N(t)\}\; dt}{\int_0^T \lambda(t)\,dt} \tag{3.34}$$

This probability differs from the first aggregated probability, if customers are still served or waiting at the end of the considered time interval, i.e., the system is not empty. Table 3.2 compares the different probabilities of being served for different numbers of agents and different service rates. The abandonment rate ν is $120\,\text{h}^{-1}$ and the arrival rate function is as in the previous section. In order to rule out equal probabilities for both cases because the call center is empty at the beginning and the end of the considered time period, we aggregated the probabilities from 7 am to 1:30 pm. Obviously, the aggregated probabilities do not differ much. Therefore, the analyses can be restricted to one case.

μ	$N(t)$	$P_{agg}^F(\text{served}, T)$	$P_\lambda^F(\text{served}, T)$
10	50	0.0788	0.0786
10	100	0.1527	0.1521
10	150	0.2235	0.2224
30	50	0.2253	0.2247
30	100	0.4240	0.4226
30	150	0.6036	0.6010
60	50	0.4247	0.4239
60	100	0.7659	0.8764
60	150	0.9884	0.9856

Table 3.2. Comparison of the different aggregated probabilities of being served for different service rates and numbers of servers

Finally, the perspective of the agents is chosen to measure the technical performance of the system. As agents suffer on the one hand from stress if the system is overloaded and on the other hand from boredom if the system is underloaded, the utilisation of agents should be carefully balanced[41]. The utilisation is given by the number of busy agents divided by the number of agents on duty. The number of busy agents is the minimum number of agents on duty and the number of customers in the system, which gives rise to

$$U^F(t) = \frac{\min\{Q^F(t), N(t)\}}{N(t)}. \tag{3.35}$$

In the course of the day agents will experience periods of high load and periods of low load. As long as neither of these two outweigh the other to a high extent, the system can be evaluated as balanced. To measure the balance it is again useful to aggregate the utilisation for a day or another time interval $[0, T]$.

[41] Compare, also Wharton Business School (2004).

$$U^F_{agg}(T) = \frac{\int\limits_0^T N(t)\dfrac{\min\{Q^F(t), N(t)\}}{N(t)}\, dt}{\int_0^t N(t)\, dt}$$
$$= \frac{\int_0^T \min\{Q^F(t), N(t)\}\ dt}{\int_0^T N(t)\, dt}. \qquad (3.36)$$

Obviously, the utilisation of the agents in the fluid model is independent of the abandonment rate as the utilisation will be one, whenever abandonment occurs. The question arising from these observations is whether these independencies are consistent with simulation results for this model.

3.2.3.2 Economical Performance Measures

Technical performance measures are strongly connected with economical performance measures which estimate the profit or the cost. Each served customer might lead to revenue. Furthermore, the contact center will have to pay for salaries of the agents and for usage of telephone lines. The cost for agents of the contact center can be calculated by summing up the product of hourly wage w and the number of agents working over the considered time period, e.g., a day. The costs for telephone trunks are given by the hourly payments per occupied line ℓ multiplied by the aggregated number of customers present. The sum of both components defines the costs of the contact center in the time period $[0, T]$, i.e.,

$$\mathrm{cost}(T) = \int_0^T \big(wN(t) + \ell Q^F(t)\big)\ dt. \qquad (3.37)$$

If a revenue of r monetary units[42] is gained from each customer served, the profit of the contact center for a period of length T is given by the revenue minus the costs, i.e.,

$$\mathrm{profit}(T) = \int_0^T \big(r\mu(t)\min\{Q^F(t), N(t)\} - \ell Q^F(t) - wN(t)\big)\ dt. \qquad (3.38)$$

The profit and the costs depend strongly on the number of agents $N(t)$ staffed at each moment in time. Furthermore, the profit function is closely connected to the aggregated probability of being served. The aggregated probability of being served is given by Equation (3.32). In Equation (3.38) the cumulative revenue

$$r\int_0^T \mu(t)\min\{Q^F(t), N(t)\}\ dt \qquad (3.39)$$

is gained from served customers. The aggregated probability of being served (3.32) is the number of customers served

[42] Later on we will use the € as the monetary unit.

$$\int_0^T \mu(t)\min\{Q^F(t), N(t)\} \tag{3.40}$$

divided by the cumulative number of departures. Therefore, we can substitute the number of customers served in the profit function (3.38) by the aggregated probability of being served multiplied by the cumulative number of departures. We get

$$\text{profit}(T) = rP^F_{agg}(\text{served}, T)\int_0^T d(t)\,dt - \int_0^T \big(\ell Q^F(t) + wN(t)\big)\,dt. \tag{3.41}$$

Hence, the profit and the aggregated probability of being served correspond to each other.

3.2.4 Numerical Solution and Results

The advantage of this model lies in the easiness of handling and interpreting the equations. Furthermore, the model is easily extended to complicated models without loss of accuracy. If the rate functions are integrable, the differential equation (3.23) can be solved analytically, otherwise they can efficiently be solved by means of numerical methods.

However, most more complicated models suffer from interdependencies of the equations such that numerical methods are needed. We used Euler and fourth order Runge-Kutta methods[43] to solve the differential equations. Fortunately, the systems of differential equations considered in this thesis are all ordinary first order non-linear differential equations and the numerical solutions to the initial value problems

$$\begin{aligned}\frac{d}{dt}Q^F(t) &= \lambda(t) - \mu(t)\min\{Q^F(t), N(t)\} - \nu(t)\{Q^F(t) - N(t)\}^+ \\ Q^F(t_0) &= Q_0\end{aligned} \tag{3.42}$$

are very accurate.

In general we assume as an initial condition (3.25) an empty contact center at the beginning of the working day. This is equivalent to the assumption that all e-mails, calls, faxes and other contacts either have been processed the day before or will never return and are lost. However, our approach also allows us to treat initial conditions differing from zero. By means of the solution of the initial value problem, we are able to calculate the technical performance measures just introduced.

For the purpose of numerical illustrations we created time-varying arrival rates as shown in Figure 3.10 using the function in Equation 3.13 on Page 29. The function roughly reproduces the arrival rate function found in real-world call centers with a peak before and after lunch time[44].

[43] see Abramowitz and Stegun (1974) pp. 896-897.

[44] Compare also Figure 2.2 on Page 10.

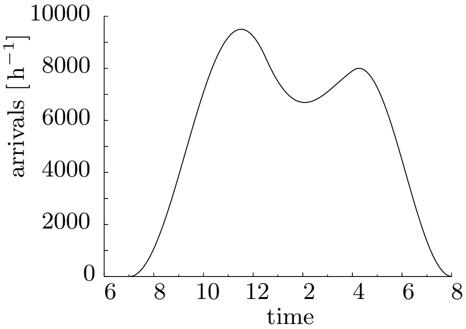

Fig. 3.10. A typical, sinusoidal arrival rate function

In the following examples the parameters of the arrival rate function were chosen to be

$$m_1 = 9500 \qquad t_0 = 7 \text{ am} \qquad t_2 = 4 \text{ pm}$$
$$m_2 = 8000 \qquad t_1 = 12{:}30 \text{ pm} \qquad t_3 = 8 \text{ pm.}$$

By varying the other parameters, different workloads of the system can be modelled without changing the arrival rate function.

First of all, the influence of the different parameters on the number of customers in the system is investigated in Figure 3.11. The default parameters of the following examples are given in Table 3.3.

Service rate $\mu(t)$	Abandonment rate $\nu(t)$	Number of agents $N(t)$
variable	$120\,\mathrm{h}^{-1}$	150
$60\,\mathrm{h}^{-1}$	$120\,\mathrm{h}^{-1}$	variable
$60\,\mathrm{h}^{-1}$	variable	100

Table 3.3. Default parameters of the examples analysed in this section

Differing from Table 3.3 in the first picture, the mean time to abandon ν^{-1} was assumed to be one minute, i.e., $\nu = 60\,\mathrm{h}^{-1}$, in order to make the influence of the service rate more clearly. This is quite a long waiting time limit. Later on an average patience of half a minute is assumed. In the second picture the average service time μ^{-1} is one minute and in the last picture the number of agents $N(t)$ is fixed at 100 agents, which means that the system is overloaded during lunch and in the late afternoon.

As the number of customers in the system has already been compared in Figure 3.9 on Page 37 to simulation results, these results are not reported once more. Obviously, all parameters have a major influence on the number of customers in the system. If the service rate μ increases, fewer customers have to wait. The same happens if the customers are more patient shown in the last picture. The number of servers $N(t)$ is important, depicted in the picture in the middle, if some customers cannot be served immediately. If the arrival rate $\lambda(t)$ is smaller than the number of available servers multiplied by the service

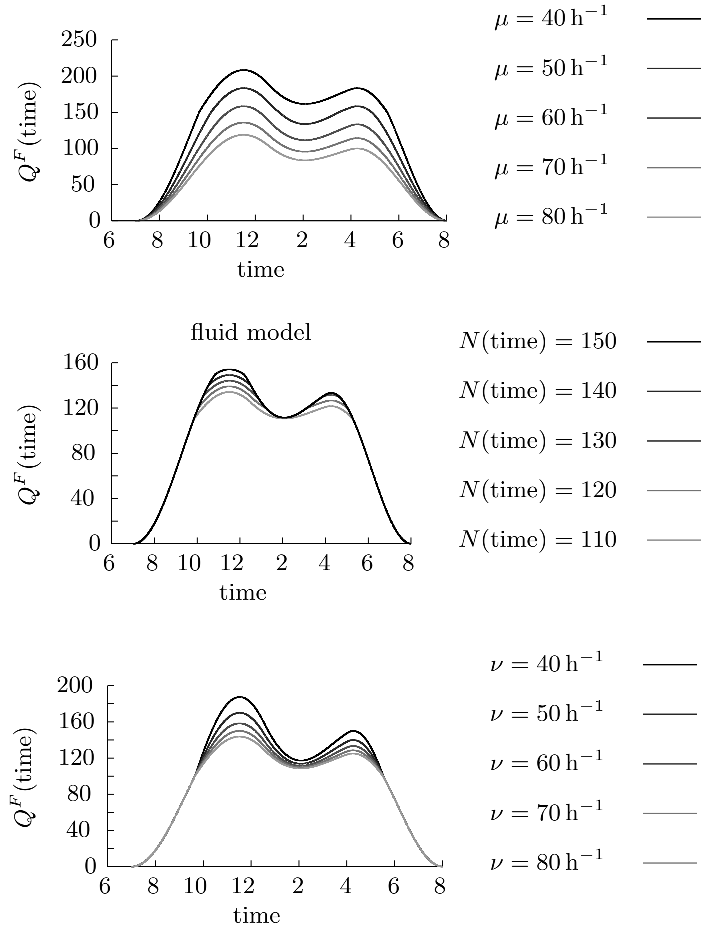

Fig. 3.11. Influence of parameters on the number of customers Q^F(time) in the system calculated by the fluid approach

rate, i.e., $\lambda(t) < \mu N(t)$, the congestion in the system is independent of the number of servers. If the customers are very impatient, such that the mean waiting time limit is smaller than the mean service time as in this example, the number of customers in the system increases if more agents are on duty. In this case a served customer stays longer in the system than an abandoning customer. If the mean service time is smaller than the average patience, i.e., $\mu > \nu$, an increasing number of agents leads to less congestion.

3.2.4.1 The Waiting Times Calculated by the Approximation and the Simulation

In Figure 3.12 the influence of the service rate on the waiting time of all customers given by Equation (3.27) on Page 39 is analysed and the fluid results are compared to simulation results. For this purpose, we used the simulation tool developed by Feldman (2004) mentioned before. We used the results of 500 independent repetitions to estimate the average number of customers in the system at time t and the other time-dependent performance measures. The parameters of the simulation are given in Table 3.3 on Page 44.

The fit of the approximation to the simulation is remarkably good. Furthermore, the time-dependent waiting time depicted in Figure 3.12 is strongly influenced by the service rate of the agents. In the fluid model, waiting occurs whenever the system is overloaded, i.e., more customers arrive than can be served by the agents. In other words the arrival rate is greater than the product of the service rate and the number of servers. Consequently, a varying arrival rate would have similar effects on the waiting times.

In the picture referring to $\mu = 70\,\mathrm{h}^{-1}$ in the fluid model, no-one has to wait because

$$\mu N(t) = 10500\,\mathrm{h}^{-1} \tag{3.43}$$

is always greater than the maximum arrival rate

$$\lambda_{\max} = \max\{\lambda(t) | t \in [7\,\mathrm{am}, 8\,\mathrm{pm}]\} = 9500\,\mathrm{h}^{-1}. \tag{3.44}$$

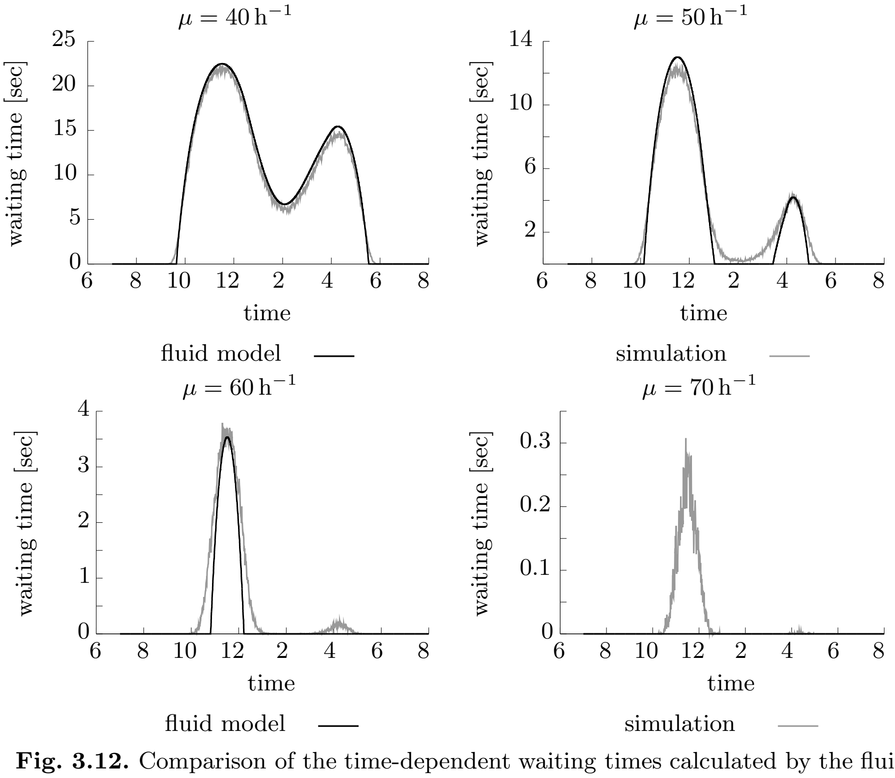

Fig. 3.12. Comparison of the time-dependent waiting times calculated by the fluid approach and the simulation tool for different service rates μ

The waiting times calculated by the simulation model are due to random effects which do occur in reality but these waiting times are negligibly small. In the graphs referring to $\mu = 50\,\mathrm{h}^{-1}$ and $\mu = 60\,\mathrm{h}^{-1}$ the effects of the random processes are even stronger. Visibly, the fluid approximation is a lower bound of the waiting time in the case of critical loading as pointed out by Altman et al. (2001) and Jiménez and Koole (2004). Critical loading means that the arrival rate and the number of agents multiplied by the service rate are almost identical, such that all agents are very busy.

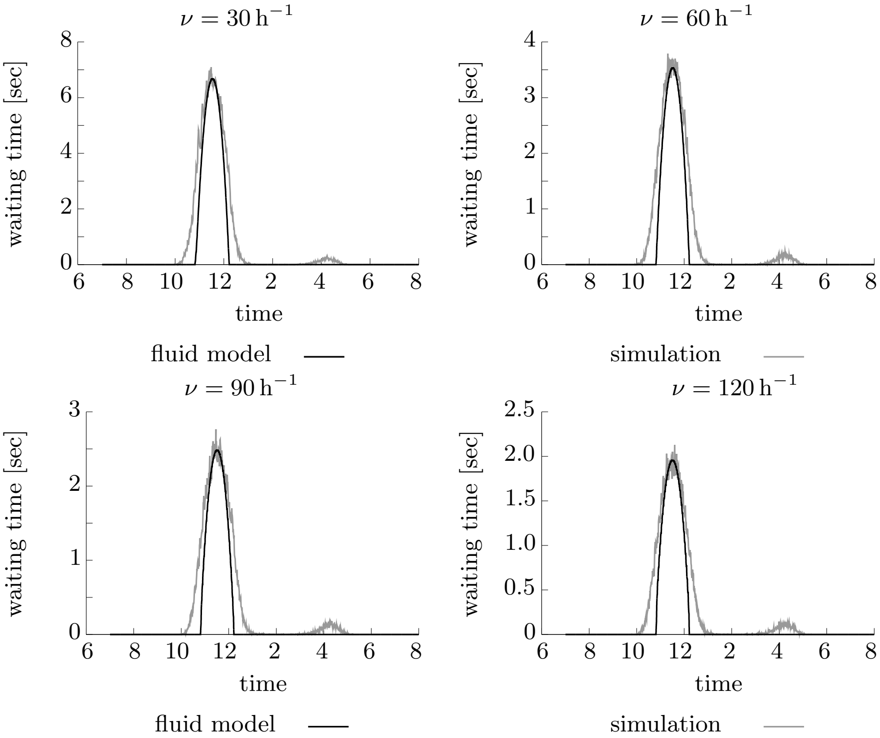

Fig. 3.13. Comparison of the waiting times calculated by the fluid approach and the simulation tool for different abandonment rates ν

Next the effects of varying abandonment rates are investigated in Figure 3.13. Comparing the simulation results and the fluid approach, the approximation seems to be accurate in the given example. Figure 3.13 shows that the more impatient customers are, the shorter waiting times are. This is reasonable as for customers who abandon the waiting time limit is reached earlier, i.e., the time spent in the queue waiting is shorter. The served customers do not influence the waiting time any longer, because their waiting time limit is greater than the time they actually have waited. The deviation of the simulation model at about 4 pm can be explained by the neglected randomness in the fluid model.

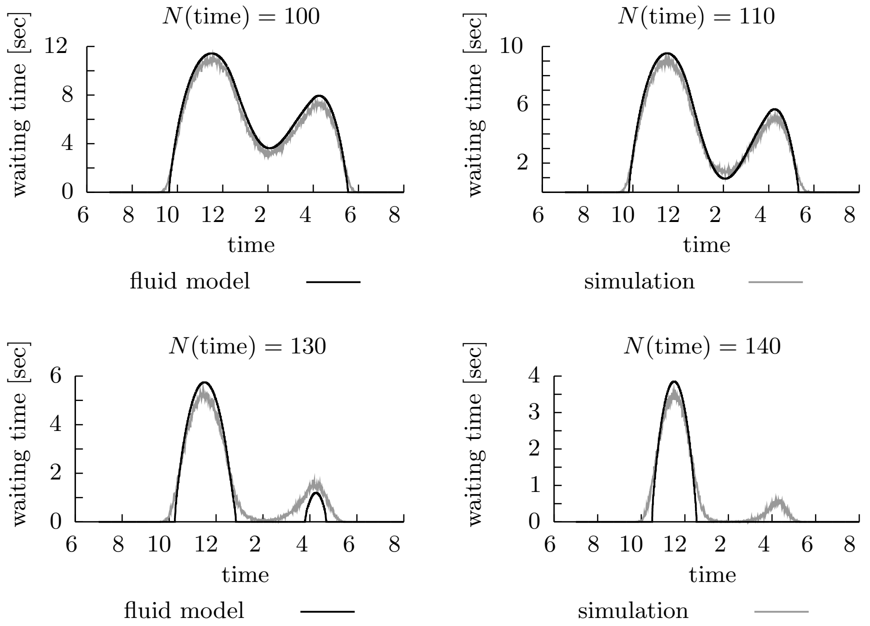

Fig. 3.14. Comparison of the waiting times calculated by the fluid approach and the simulation tool for different numbers of agents N(time)

Finally, we investigate the influence of the number of working agents on the waiting times in Figure 3.14. Again a comparison of the simulation and fluid results stresses the accuracy of the fluid approach. If fewer agents are staffed, the waiting time of the customers increases. However, a waiting time of approximately 12 seconds in the case of 100 agents on duty is still very small. During the time interval [1 pm, 5 pm] for values of 130 and 140 agents on duty, the call center is critically loaded and the waiting time of the simulation model is underestimated by the fluid approach. Contrarily, during time intervals of very high load the fluid results overestimate the simulation result slightly. In general the waiting time of customers decreases, if the number of agents increases.

In addition to the time-dependent waiting times, we study the impact of system parameters on the aggregated waiting times calculated by means of the fluid approach and Equation (3.29) on Page 40. In Figure 3.15 the influence of an increasing number of agents and a varying service rate on the aggregated waiting times is shown. The number of agents is assumed to be constant for the whole length of the considered period [7 am, 8 pm], i.e. $T = 8\,\text{pm}$. If almost no agent is staffed, the aggregated waiting time approaches the waiting time limit of thirty seconds. In this case almost no customer is ever served. A growing number of agents as well as a growing service rate lead to shorter waiting times.

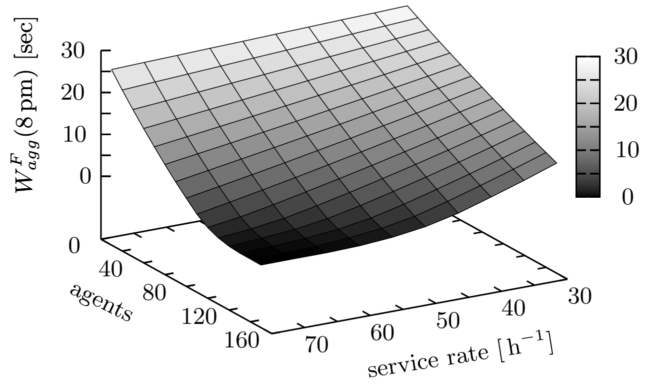

Fig. 3.15. Aggregated waiting time $W^F_{agg}(T)$ for varying service rates μ and numbers of agents $N(t)$

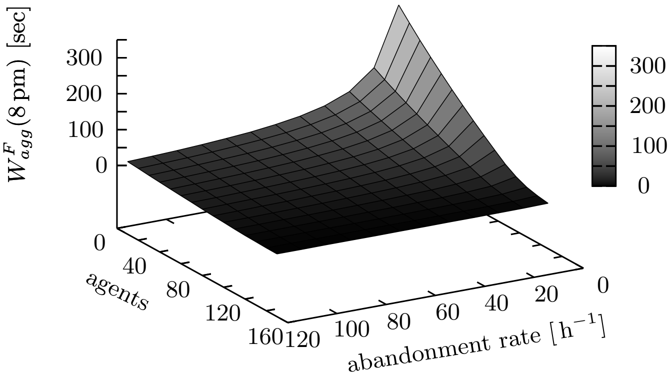

Fig. 3.16. Aggregated waiting time $W^F_{agg}(T)$ for varying abandonment rates ν and numbers of agents $N(t)$

The influence of varying abandonment rates and numbers of agents on the aggregated waiting time is depicted in Figure 3.16. Obviously, abandonment rates can only have an impact on the waiting time if the system is overloaded.

3.2.4.2 The Probability of Being Served Calculated by the Approximation and the Simulation

A second very important performance measure is the probability of all customers who have left the system being served given by Equation (3.30) on Page 40. As the probabilities of being served and abandoning are complementary in the system considered in this section, i.e., the sum of the two probabilities is one, the second probability is not presented. The simulation results are compared to the fluid results and the influences of the different parameters on the probability of being served are explained. The arrival rate is supposed to be sinusoidal as in Figure 3.10 on Page 44 and the other parameters are given in Table 3.3 on Page 44.

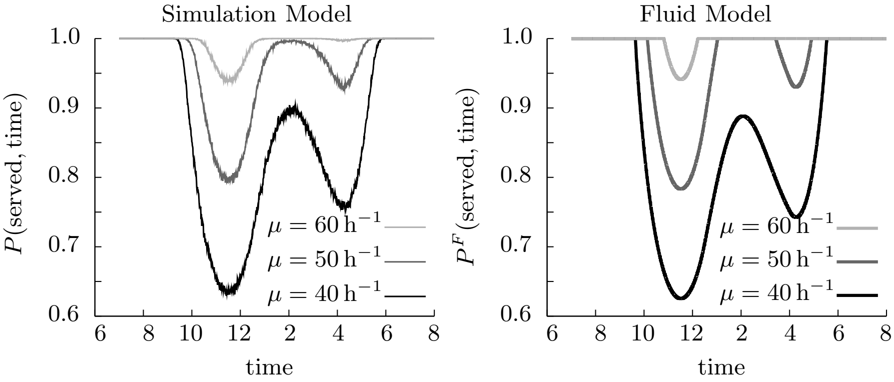

Fig. 3.17. Comparison of the probability of being served calculated by the fluid approach and the simulation tool for different service rates μ

In Figure 3.17 the fluid and simulation results are compared for different service rates. As before, the results of the fluid approach are very good. The time-dependent probability of being served increases, if the service rate increases.

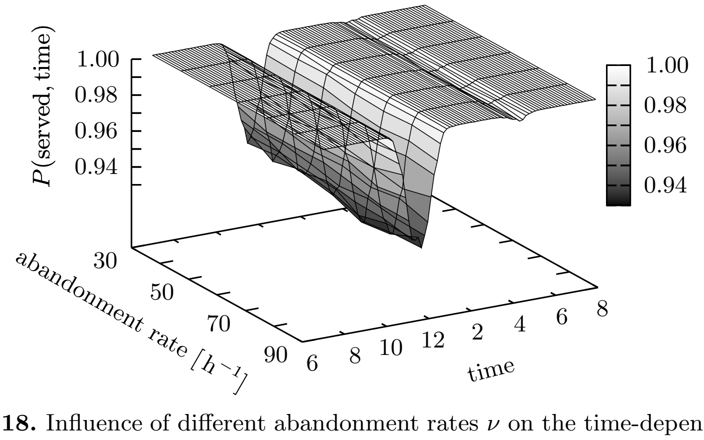

Fig. 3.18. Influence of different abandonment rates ν on the time-dependent probability of being served P(served, time) in the simulation model

Next we analyse the influence of the abandonment rate. As mentioned before[45], Aguir et al. (2004) show analytically that the abandonment rate has no impact on the probability of being served or abandoning in the fluid approach. The question arising is whether this is also true for original discrete systems and its simulation model. Therefore, in Figure 3.18 the simulation results are reported. In this figure the different abandonment rates do not affect

[45] See Page 40.

the time-dependent probability of being served calculated by the simulation tool. The varying probability of being served during the lunch hours is due to stochastic effects but not due to the varying abandonment rate. Consequently, the simulation results underline the accuracy of the fluid approach.

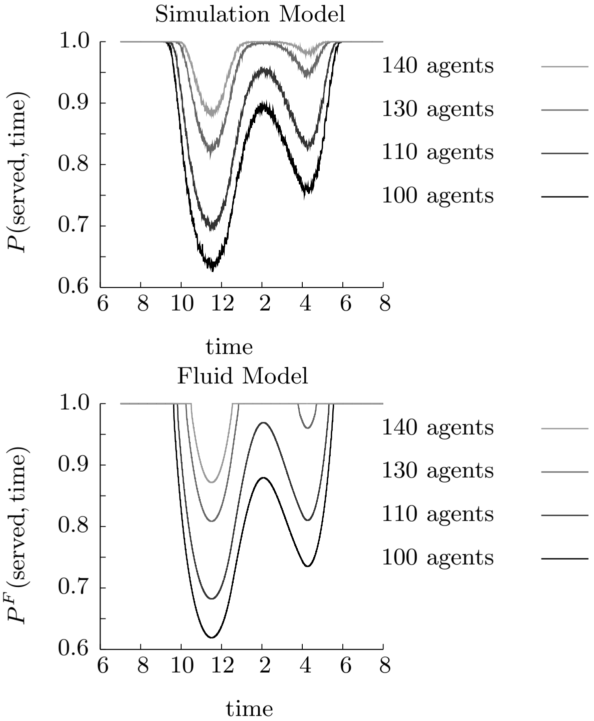

Fig. 3.19. Comparison of the probability of being served with varying numbers of agents N(time)

Finally, the influence of the number of staffed agents is depicted in Figure 3.19. The probability of being served depends on the number of servers as well as on the service rates because the number of servers has a direct influence on the load. If many agents are scheduled, the load of the system decreases such that more customers are served and the probability of service approaches one. Contrarily, if only few agents are scheduled, many customers have to wait and consequently more abandon before they are attended to. Comparing the Figures 3.17, 3.18, and 3.19 depicting the probability of being served, it can be concluded that the approximation works fine except for the phases of critical loading. In our example, these phases are diminishingly small time intervals in most pictures.

In the previous subsection two different aggregated probabilities of being served have been derived in Equation (3.34) and Equation (3.32) on Page 41. In the examples of Table 3.2 on Page 41, it was shown that the aggregated probabilities do not differ much for the system investigated in this section.

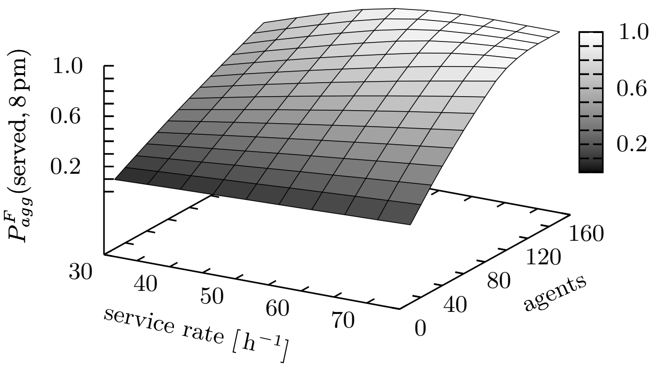

Fig. 3.20. Influence of a varying service rate μ and number of agents $N(t)$ on the aggregated probability of being served $P^F_{agg}(\text{served}, T)$

Therefore, only the probability of all customers who have left the system being served calculated by Equation (3.32) is presented in Figures 3.20.

The aggregated probability of being served is independent of the abandonment rates, because the time-dependent probabilities have been independent. That is why no figure depicting the aggregated probability of being served as a function of the abandonment rate is presented.

However, resulting from the strong influence of the service rate and the number of agents on the time-dependent probabilities, the aggregated probability depends strongly on the number of agents and the service rate as well. Comparing the time-dependent probabilities in Figures 3.17, 3.19 and the aggregated probabilities in Figure 3.20 the averaging effect of aggregation can be observed[46]. The aggregated probability of being served is still high although the probability of being served is very low during some time intervals.

3.2.4.3 The Utilisation of the Agents Calculated by the Approximation and the Simulation

The last performance measure that is calculated is the utilisation of the agents at some time t given in Equation (3.35) on Page 41 or aggregated as in Equation (3.36) on Page 42. The parameters of the underlying model are the same as in the previous examples and shown on Page 44. In Figure 3.21 the utilisation $U^F(t)$ for different service rates is presented. Similar to the probability of being served and the waiting time the impact of the service rates is strong. Growing service rates, i.e., short mean service times, lead to smaller utilisation of the agents.

In order to show that the time-dependent utilisation is independent of the abandonment rate in the simulation, the simulation results are presented

[46] For the discussion about the averaging effect and its role as hard and soft constraints see Page 20.

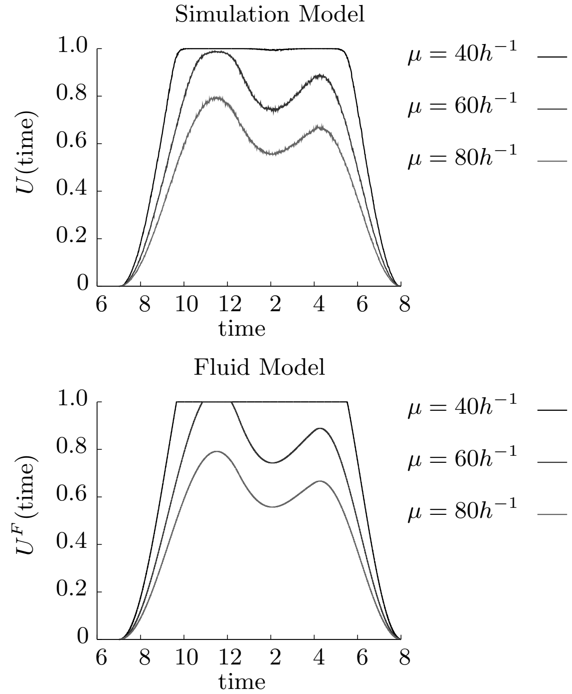

Fig. 3.21. Comparison of the utilisation of the agents calculated by the fluid approach and the simulation tool for different service rates μ

in Figure 3.22. This result is easily explained, because the time-dependent utilisation is one, as long as more customers are in the system than agents on duty. Only in this case do customers abandon. If the number of customers in the system decreases, the utilisation decreases as well. In the second case the utilisation of agents is given by the fraction of the number of busy agents and

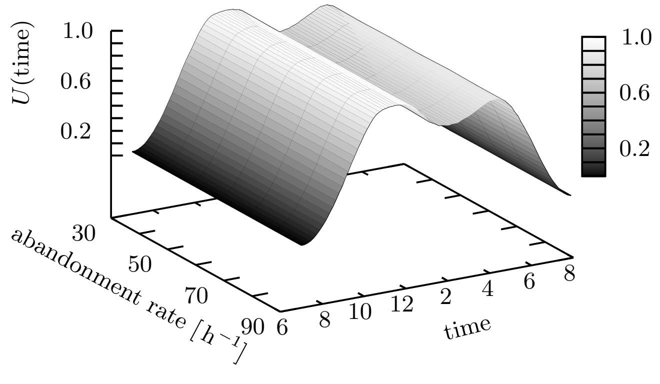

Fig. 3.22. Influence of the abandonment rate ν on the utilisation of agents U(time) in the simulation model

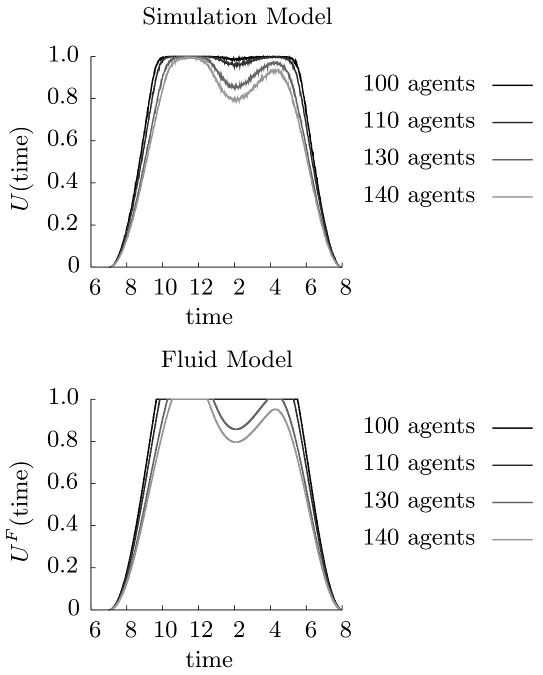

Fig. 3.23. Comparison of the utilisation of the agents calculated by the fluid approach and the simulation tool for different numbers of agents

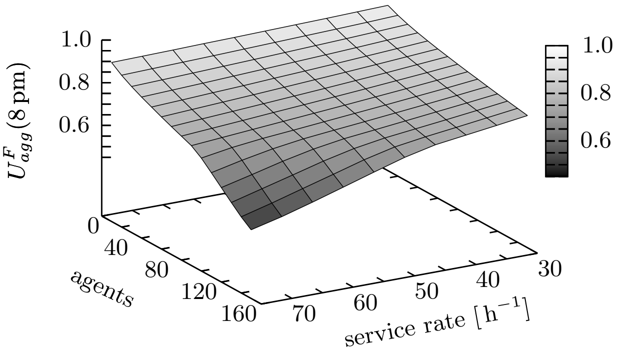

Fig. 3.24. Influence of a varying service rate μ and number of agents $N(t)$ on the aggregated utilisation of agents $U^F_{agg}(T)$

the number of agents staffed, which are both independent of the abandonment rate.

Finally, the influence of the number of agents on the time-dependent utilisation is depicted in Figure 3.23 and on the aggregated utilisation in Figure 3.24. While the time-dependent utilisation varies drastically over time, the aggregated utilisation is more stable. Higher service rates and more agents lead to lower utilisation. As the time-dependent utilisation is independent of the abandonment rate, the aggregated utilisation will be independent as well.

Therefore, no picture showing the aggregated utilisation as a function of the abandonment rate is presented.

These investigations confirm that the service rate and the number of agents working have a high impact on all performance measures. The abandonment rate, however, has a very limited influence on the performance experienced by a customer. Only the waiting times depend on the abandonment rate. The probability of being served and of abandoning as well as the utilisation of agents are not visibly influenced.

3.2.4.4 Influence of the Parameters on the Profit Function

After having studied the technical performance measures, the influence of the service rate μ, the abandonment rate ν and the number of agents $N(t)$ on the profit function of the contact center given in Equation (3.38) on Page 42 are investigated in Figures 3.25 and 3.26. The profit seems to be slightly dependent on the service rate for a low number of staffed servers. If the number of servers increases the influence of the service rate becomes stronger.

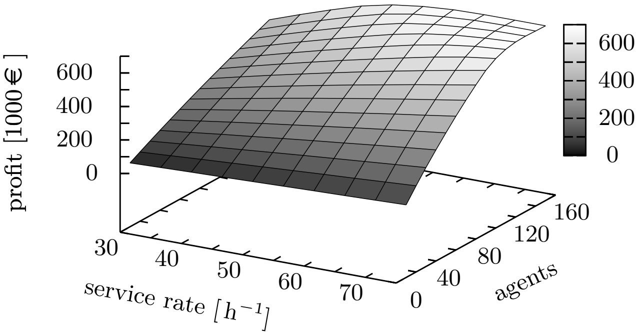

Fig. 3.25. Influence of the service rate μ and the number of agents $N(t)$ on the profit

In Figure 3.26 the abandonment rate seems to have almost no impact on the profit function. This result is in line with the strong links between the aggregated probability of being served and the profit function shown in Equation 3.41 on Page 43. Solely the case of patient customers, i.e. $\nu = 0$, leads to a slightly different profit value.

The numerical results of the fluid approach and the simulation results for both the different technical and economical performance measures just analysed show that the simulation model is well approximated by the solutions of the fluid approach. Furthermore, the abandonment rate has no impact on the probability of being served, on the utilisation of agents and on the profit of the contact center excluding the case of patient customers. Last, the averaging effect of aggregation has been shown by comparing the time-dependent and aggregated technical performance measures.

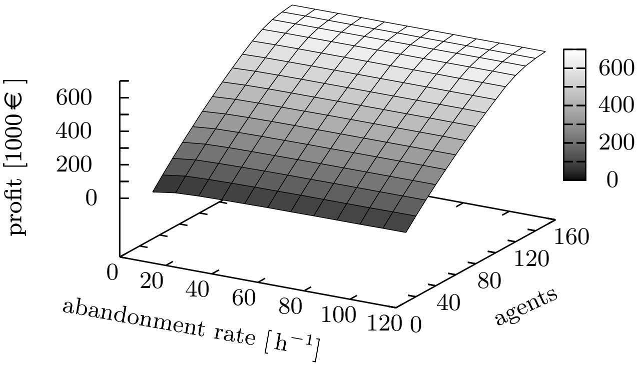

Fig. 3.26. Influence of the abandonment rate ν and the number of agents $N(t)$ on the profit

3.2.5 Applicability and Limitation for Contact Center Analysis

As mentioned before, the major advantage of the fluid modelling approach is that the fluid models are easy to create and interpret. It does not matter how many customer classes or agent groups have to be considered. Adding a new customer class simply means appending another equation to the system of differential equations, and adding a new agent group extends the differential equation by another term which describes the routing and speed of operation of these agents.

Another advantage of this approach with respect to the performance of the contact center is pointed out by Garnett et al. (2002). They show that the probability of delay is strictly smaller than one but greater than zero in the scaling limit[47]. The probability of delay is the probability that an arriving customer will have to wait. Additionally, the probability of abandoning becomes asymptotically negligible, i.e., in the limiting fluid model some customers will have to wait but very few customers abandon and the agents are busy nearly all the time. That is why this limiting regime has become known as *quality and efficiency driven*. Garnett et al. (2002) show that in the *efficiency driven regime* the probability of delay reaches one and the probability of abandoning a limit strictly between zero and one, i.e., in this regime the agents are high utilised and many customers have to wait. In the *quality driven regime* both probabilities are negligible. In this case the utilisation of agents is low and almost no customer has to wait at all[48].

However, the fluid approach eliminates the randomness of the processes[49] which can be regarded as an important reason for congestion in small systems. This fact can be observed if the utilisation of the agents approaches one and the system is almost stable. Then in real-world contact centers there is a

[47] See, Garnett et al. (2002), Borst et al. (2002), and Whitt (2005c)

[48] See also Whitt (2004, 2006b).

[49] See Whitt (2005c).

substantial abandonment whereas in the fluid model everyone is still served and no-one has to wait. This is caused by the fact that in the fluid model the amount of customers is supposed to be continuous such that every piece of work can be done and the fact that the fluid processes are deterministic. The more customers are in the system, the more the discrete nature of customers can be neglected[50].

Consequently, the fluid approach is a accurate approach for large contact centers as long as the individual groups of customers and agents are big enough. As long as the stationary queueing models are simple enough for determining the stationary distribution, more detailed technical performance measures can be calculated by the stationary approach, while in contact centers with a complex structure the fluid approximation appears to be superior. Jiménez and Koole (2004) show that often the combination of a fluid and a stationary model delivers a good approximation for a system with slowly varying rates. They have proven that the fluid limits are lower bounds for the number of customers in the real system and its convex performance measures.

In our models the arrival rate changes quite quickly, which means that we do not have constant arrival rates for a quarter or half an hour even in the moderately loaded system. In such a case the system is not able to become stable and in consequence the stationary approach is less valuable.

3.3 Refinement to a Diffusion Model

3.3.1 Motivation

In this section we consider an extension of the fluid approach derived in the previous section which is able to deal with the stochasticity of the system. Similar to the fluid approach, the diffusion refinement is also based on the heavy traffic scaling by Halfin and Whitt (1981) known as strong approximations. We follow the derivation of the differential equations for the variances and covariance in Mandelbaum et al. (1998).

By means of the scaling, the Poisson processes described in Equations (3.15)-(3.17) on Page 32f. can be approximated by normally distributed diffusion processes with the same mean and variance[51]. The scaling is quite similar to the fluid approach but in spite of applying the functional strong law of large numbers we will apply the extended form of the functional central limit theorem[52].

The notion of deriving heavy traffic approximations for queueing systems based on the central limit theorem was first developed by Kingman in the

[50] Compare, Borst et al. (2004) and Whitt (2003).

[51] See, e.g., Feldman et al. (2005). For the connection between Poisson Processes and normally distributed processes see, e.g., Schoenberg (2002), Ward and Glynn (2003b), and Marchal (2003).

[52] See Mandelbaum, Massey, and Reiman (1998).

early 1960s[53]. The heavy traffic limits were derived by holding the number of servers fixed and letting the service intensity approach one from below.

Another procedure was first used by Iglehart in 1965 and made famous by Halfin and Whitt (1981) who let the number of servers and the arrival rate approach infinity. Both procedures have in common that the discrete process of the number of customers in systems is substituted by a continuous diffusion process[54]. Afterwards this continuous process can be discretised to calculate approximations for steady state distributions.

The underlying idea associated with this approach is the irregular movement of pollen in water. It was first observed by the botanist Robert Brown after whom the process was named *Brownian motion*[55]. The diffusion of the pollen in water is driven by the collision of the water molecules.

We may think of the customers as pollen who fall on water with a little drift towards drains which represent the processing by the agents and the departure due to impatience. These pollen move around in the water until they reach a drain and flow out after having stayed a random time moving back and forth in the basin.

3.3.2 Modelling and Justification

In order to explain this approach we consider the model introduced in Figure 3.6 on Page 31. There we determined the scaled process of the number of customers in a scaled system $Q^n(t)$, which means that the arrival rate and number of servers were multiplied by a scaling parameter n. From Equation (3.21) on Page 35 we derived the functional Equation (3.22) for the limiting mean process. Afterwards we differentiated the process with respect to the time t and got the differential Equation (3.23) describing the fluid process.

In this section we aim to describe the process including the stochasticity and time-varying behaviour by means of the variance of the process according to Mandelbaum et al. (1998). Therefore, *standard Brownian motions* and a stochastic differential equation of the process are needed. A standard Brownian motion is a normally distributed stochastic process with mean zero and variance t at time t[56].

To determine the stochastic differential equations we use the *central limit theorem* and some related results proven by Komlós et al. (1975)[57] and Rao (1973). The central limit theorem states that the sum of n independent and identically distributed random variables converges in distribution to a normally distributed random variable with n-times the mean and n-times the variance of the first random variable as n grows.

[53] See Kingman (1962, 1965).

[54] For a literature review of early developments see Halfin and Whitt (1981).

[55] See Bauer (2001b).

[56] See Whitt (2002a).

[57] See also Kurtz (1978), Grama and Nussbaum (1997), and Brown (2002).

Linked to the functional central limit theorem is a result by Komlós et al. (1975), who show that an approximation of the sum of n Poisson processes by a Brownian motion with n-times the mean and n-times the variance has the remainder term of order $\log n$, where log denotes the natural logarithm.

Rao (1973) proves an extension of the functional central limit theorem, which is needed because the counting process $Q(t)$ given in (3.18) on Page 32 is composed not only of different Poisson processes but also of functions of these processes. This theorem states that if Y_n denotes the sum of independent and identically distributed random variables with finite mean and variance σ^2 and f is a differentiable function in a neighbourhood of $\mathbf{E}[Y_1]$, then $f\left(\frac{Y_n}{n}\right)$ converges to $f(\mathbf{E}[Y_1])$ almost surely according to the strong law of large numbers for n approaching infinity. Furthermore, the central limit theorem can be applied on $f\left(\frac{Y_n}{n}\right)$ and the fraction converges in distribution to a standard normally distributed random variable X multiplied by the derivative of the considered function at the mean value of Y_1. Formally, the theorem states[58]

$$\lim_{n\to\infty} \frac{\sqrt{n}}{\sigma}\left(f\left(\frac{Y_n}{n}\right) - f(\mathbf{E}[Y_1])\right) \overset{i.d.}{=} f'(\mathbf{E}[Y_1])X. \tag{3.45}$$

If the function f is continuous but not differentiable at the mean value $\mathbf{E}[Y_1]$, a similar convergence relation can be shown, which uses the limits from above and below at the critical mean value $\mathbf{E}[Y_1]$. If $f'(\mathbf{E}[Y_1]+)$ denotes the limit from above and $f'(\mathbf{E}[Y_1]-)$ from below, we get

$$\lim_{n\to\infty} \frac{\sqrt{n}}{\sigma}\left(f\left(\frac{Y_n}{n}\right) - f(\mathbf{E}[Y_1])\right) \overset{i.d.}{=} f'(\mathbf{E}[Y_1]+)\{X\}^+ - f'(\mathbf{E}[Y_1]-)\{X\}^-, \tag{3.46}$$

where $\{X\}^-$ denotes the maximum of $-X$ and zero.

By means of these results the stochastic differential equation can be derived, which is needed to determine the variances and the covariances of the different processes. The variances and covariances are used to measure the randomness of the processes.

First of all, we determine the limiting stochastic differential equation for the counting process $Q(t)$ of the number of customers in the system. According to Mandelbaum et al. (1998, pp. 153-155), the limiting process we require is given by

$$\lim_{n\to\infty} \frac{Q^n(t) - nQ^F(t)}{\sqrt{n}} = \lim_{n\to\infty} Q^{\sqrt{n}}(t) \overset{i.d.}{=} Q^D(t) \tag{3.47}$$

for all $t \in \mathbb{R}_0^+$. In order to derive $Q^D(t)$ the sample-path representation $Q^n(t)$ in (3.47) is approximated[59] by a Brownian motion according to the theorem by Komlós et al. (1975). We substitute n-times the fluid process $Q^F(t)$ by n-times the left hand side of Equation (3.23) on Page 36 and dividing by square

[58] See, Mandelbaum et al. (1998) p. 154.

[59] For a proof of this strong approximation for growing n see Mandelbaum et al. (1998) on Pages 156-157 and Section 2.

root of n. Consequently, we get for the fraction in Equation 3.47[60]

$$\frac{Q^n(t) - nQ^F(t)}{\sqrt{n}}$$

$$= \frac{1}{\sqrt{n}} \left(Q^n(0) - nQ^F(0)\right) \tag{3.48a}$$

$$+ \frac{1}{\sqrt{n}} \left(n\int_0^t \lambda(s)\, ds - n\int_0^t \lambda(s)\, ds \right) \tag{3.48b}$$

$$- \frac{1}{\sqrt{n}} \left(n\int_0^t \mu(s) \min\left\{ \frac{Q^n(s)}{n}, N(s) \right\} ds - n\int_0^t \mu(s) \min\left\{Q^F(s), N(s)\right\} ds \right) \tag{3.48c}$$

$$- \frac{1}{\sqrt{n}} \left(n\int_0^t \nu(s) \left\{ \frac{Q^n(s)}{n} - N(s) \right\}^+ ds - n\int_0^t \nu(s) \left\{Q^F(s) - N(s)\right\}^+ ds \right) \tag{3.48d}$$

$$+ \frac{1}{\sqrt{n}} B_1 \left(\int_0^t n\lambda(s)\, ds \right) \tag{3.48e}$$

$$- \frac{1}{\sqrt{n}} B_2 \left(\int_0^t n\mu(s) \min\left\{ \frac{1}{n} Q^n(s), N(s) \right\} ds \right) \tag{3.48f}$$

$$- \frac{1}{\sqrt{n}} B_3 \left(\int_0^t n\nu(s) \left\{ \frac{1}{n} Q^n(s) - N(s) \right\}^+ ds \right) + o\left(\frac{\log n}{\sqrt{n}} \right). \tag{3.48g}$$

If n approaches infinity and we apply the mentioned result by Rao, we expect to get the stochastic differential equation for a centralised diffusion process. However, the minimum and maximum functions are not continuously differentiable, therefore Mandelbaum, Massey, and Reiman (1998) introduced a new form of derivative which they call the *scalable Lipschitz derivative* to circumvent these difficulties.

As an example we consider the function $f(x) = \min\{x, N\}$ for fixed N depicted in Figure 3.27. The Lipschitz derivative[61] of $f(x)$ at x for any real value y is given as the positive part of y, if $x < N$, minus the negative part of y, if $x \leq N$, i.e.,

[60] Compare Mandelbaum et al. (1998) p. 163 and Whitt (2002a) p. 360.

[61] See Mandelbaum et al. (1998) in Sections 3 and 12 for further properties of this derivative and the formal derivation.

$$\begin{aligned}\lim_{y\to 0}&\frac{\min\{x+y,N\}-\min\{x,N\}}{|y|}\\ &=\max\{y,0\}\mathbb{1}_{\{x<N\}}-\max\{-y,0\}\mathbb{1}_{\{x\le N\}} \qquad (3.49)\\ &=:\frac{d}{dx}\min\{x,N\}(y)\end{aligned}$$

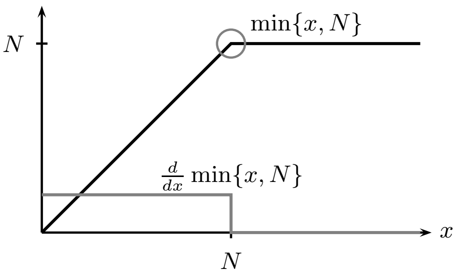

Fig. 3.27. Scalable Lipschitz derivative

In order to show how this result can be applied to Equation 3.48, we consider the term (3.48c). First of all, we transform the term, such that the structure of the functional central limit theorem becomes more obvious, i.e.,

$$\begin{aligned}\lim_{n\to\infty}-\frac{1}{\sqrt{n}}&\left(n\int_0^t \mu(s)\min\left\{\frac{Q^n(s)}{n},N(s)\right\}ds\right.\\ &\qquad\left.-\,n\int_0^t \mu(s)\min\{Q^F(s),N(s)\}\,ds\right) \qquad \text{(Eq. (3.48c))}\\ &=\lim_{n\to\infty}-\frac{n}{\sqrt{n}}\int_0^t \mu(s)\left(\min\left\{\frac{Q^n(s)}{n},N(s)\right\}-\min\{Q^F(s),N(s)\}\right)ds\\ &=\lim_{n\to\infty}-\int_0^t \mu(s)\sqrt{n}\left(\min\left\{\frac{Q^n(s)}{n},N(s)\right\}-\min\{Q^F(s),N(s)\}\right)ds\\ &=-\int_0^t \mu(s)\lim_{n\to\infty}\sqrt{n}\left(\min\left\{\frac{Q^n(s)}{n},N(s)\right\}-\min\{Q^F(s),N(s)\}\right)ds.\end{aligned}$$

Then we let n approach infinity and use Equations (3.46) and (3.49) to derive

$$\begin{aligned}\lim_{n\to\infty}\sqrt{n}&\left(\min\left\{\frac{Q^n(s)}{n},N(s)\right\}-\min\{Q^F(s),N(s)\}\right)\\ &=\{Q^D(s)\}^+\mathbb{1}_{\{Q^F(s)<N\}}-\{Q^D(s)\}^-\mathbb{1}_{\{Q^F(s)\le N\}}. \qquad (3.50)\end{aligned}$$

Although an application of the results by Mandelbaum et al. (1998) and the scalable Lipschitz derivative on the terms 3.48c and 3.48d in Equation 3.48 leads to an exact solution, the derived differential equations are neither easily solvable nor concrete.

Therefore, another assumption is made which gives rise to a much more simple and practical equation. We assume that the set of critical times $\boldsymbol{S}$ is supposed to have measure zero. The critical times are those times when the call center is critically loaded, i.e., the number of servers equals the number of customers in the system, i.e.,

$$\boldsymbol{S} = \{t | Q(t) = N(t), t \in [0, T]\} \tag{3.51}$$

A set has measure zero if it is almost empty[62]. Therefore the set of critical times has measure zero if the process $Q^D(t)$ passes the state with equal numbers of agents and customers in the system very quickly. This assumption is reasonable, as for a normally distributed process the measure of a single point is zero.

Under this assumption the limit derived in Equation (3.50) reduces to

$$\lim_{n\to\infty} \sqrt{n} \left(\min\left\{ \frac{Q^n(s)}{n}, N(s) \right\} - \min\left\{ Q^F(s), N(s) \right\} \right)$$
$$= Q^D(s) \mathbb{1}_{\{Q^F(s) \leq N\}}. \tag{3.52}$$

Similarly, the limits of the term (3.48d) related to the abandonment of customers is derived. Under the assumption just mentioned, the term (3.48d) converges to $-\int\limits_0^t \nu(s) \mathbb{1}_{\{Q^F(s) > N(s)\}} Q^D(s)\, ds$. In other words, under the assumption that the set of critical times is a null set, the limit processes including maximum or minimum functions are given by the product of the rates, the diffusion process $Q^D(t)$, and an indicator function $\mathbb{1}_{\{\}}$. The indicator function has value 1, if the associated Poisson process depends on the number of customers in the system. In the case of the service process this means that the number of customers is below the number of agents, and in the case of the abandonment process, that the number of customers exceeds the number of available agents.

The limit of the term (3.48a) is simply the diffusion process at 0, and term (3.48b) dimishes. To derive the limits of the Brownian motions (3.48e) through (3.48g), we use the so-called self-similarity scaling property[63] of Brownian motions to derive the limiting processes[64], which leads to[65]

[62] See, e.g., Bauer (2001a) and Billingsley (1999) for an introduction and further information on measure theory.

[63] See Whitt (2002a) p. 102.

[64] For overview of central limit theorems associated with Brownian motion see Whitt (2000).

[65] See also Mandelbaum et al. (1998).

$$\begin{aligned}
&\lim_{n\to\infty} -\frac{1}{\sqrt{n}} B_2\left(\int_0^t n\mu(s)\min\left\{\frac{1}{n}Q^n(s), N(s)\right\}\, ds\right)\\
&\quad= \lim_{n\to\infty} -B_2\left(\int_0^t \mu(s)\min\left\{\frac{1}{n}Q^n(s), N(s)\right\}\, ds\right)\\
&\quad= -B_2\left(\int_0^t \mu(s)\min\left\{Q^F(s), N(s)\right\}\, ds\right) \qquad (3.53)
\end{aligned}$$

for the Brownian process (3.48f) in Equation (3.48). Similarly, the other limits of the Brownian motions are calculated. Finally, we get the stochastic differential equation of the diffusion process by

$$\begin{aligned}
Q^D(t) = Q^D(0) &- \int_0^t \mu(s)\mathbb{1}_{\{Q^F(s)\le N(s)\}} Q^D(s)\, ds\\
&- \int_0^t \nu(s)\mathbb{1}_{\{Q^F(s)> N(s)\}} Q^D(s)\, ds \qquad (3.54a)\\
&+ B_1\left(\int_0^t \lambda(s)\, ds\right) + B_2\left(\int_0^t \mu(s)\min\left\{Q^F(s), N(s)\right\}\, ds\right)\\
&+ B_3\left(\int_0^t \nu(s)\left\{Q^F(s) - N(s\right\}^+\, ds\right) \qquad (3.54b)
\end{aligned}$$

Formally, the limit in Equation (3.54) derived from the functional central limit theorem in Equation (3.47) consists of two parts.
The first part (3.54a) of the limiting process describes the drift of the customers in the system towards the exit. The system wants to get the customers out of the system both by service and abandonment. This part results from the terms (3.48b) through (3.48e) if n approaches infinity.
The second part (3.54b) of the stochastic functional equation describes the variation by means of standard Brownian motions. This part results from the terms (3.48f) and (3.48g) in Equation (3.48) if n reaches infinity. The number of customers in the system varies around the mean value. This process can be interpreted as the deviation from the mean throughput process.

The next step is to derive the differential equations for the variances and covariances of the model. In this case of just one customer class we only have to determine the variance. Therefore, we have to utilise some additional properties of Brownian motions as well as the so-called *chain rule of stochastic* *calculus* commonly known as *Ito's formula*[66] to get the differential equation for the variance. The variance of the process can be calculated by the mean diffusion processes by

$$\mathbf{VAR}\big[Q^D(t)\big] = \mathbf{E}\Big[\big(Q^D(t)\big)^2\Big] - \mathbf{E}\big[Q^D(t)\big]^2 . \tag{3.55}$$

Beside the chain rule of stochastic calculus, we have to use the following property of standard Brownian motions

$$dB_i(t) \cdot dB_j(s) = \begin{cases} t, \; i = j \\ 0, \; i \neq j \end{cases} \tag{3.56}$$

to derive the differential equation of the mean value of the squared process $\mathbf{E}\Big[\big(Q^D(t)\big)^2\Big]$. First of all, we determine an expression for the differential[67]

$$d\left(Q^D(t)\right)^2 = 2 \cdot dQ^D(t) \cdot Q^D(t) + dQ^D(t) \cdot dQ^D(t). \tag{3.57}$$

Afterwards we apply the mean function on this expression. The differential of $Q^D(t)$ is simply derived from Equation 3.54 by means of the self-similarity scaling property mentioned before. We get

$$\begin{aligned} dQ^D(t) = &- \left(\mu(t)\mathbb{1}_{\{Q^F(t)\leq N(t)\}}Q^D(t) + \nu(t)\mathbb{1}_{\{Q^F(t)>N(t)\}}Q^D(t)\right) dt \\ &+ \sqrt{\lambda(t)}dB_1(t) - \sqrt{\mu(t)\min\{Q^F(t), N(t)\}}dB_2(t) \\ &- \sqrt{\nu(t)\{Q^F(t) - N(t)\}^+}dB_3(t). \end{aligned} \tag{3.58}$$

With the help of this equation and the Properties (3.56) and $\mathbf{E}[B_i(t)] = 0$, the differential equation of the mean of the squared process is given by[68]

$$\begin{aligned} \frac{d}{dt}\mathbf{E}\Big[\big(Q^D(t)\big)^2\Big] &= -2\left(\mu(t)\mathbb{1}_{\{Q^F(t)\leq N(t)\}} + \nu(t)\mathbb{1}_{\{Q^F(t)>N(t)\}}\right)\mathbf{E}\Big[\big(Q^D(t)\big)^2\Big] \\ &\quad + \lambda(t) + \mu(t)\min\big\{Q^F(t), N(t)\big\} + \nu(t)\big\{Q^F(t) - N(t)\big\}^+ \end{aligned} \tag{3.59}$$

The differential equation for the squared mean of the diffusion process $\mathbf{E}\big[Q^D(t)\big]^2$ is derived much more simply, by the mentioned properties of the standard Brownian motions and Equation 3.58 and given by

$$\frac{d}{dt}\mathbf{E}\big[Q^D(t)\big]^2 = 2\left(\mu(t)\mathbb{1}_{\{Q^F(t)\leq N(t)\}} + \nu(t)\mathbb{1}_{\{Q^F(t)>N(t)\}}\right)\mathbf{E}\big[Q^D(t)\big]^2 . \tag{3.60}$$

[66] See Karatzas and Shreve (1991, pp. 149-156) for a formal explanation of the chain rule.

[67] See Mandelbaum et al. (1998) Section 10.

[68] See Mandelbaum et al. (1998) pp. 197–198.

Subtracting Equation (3.60) from Equation (3.59) leads to[69]

$$\begin{aligned}\frac{d}{dt}\mathbf{VAR}\big[Q^D(t)\big] &= -2\left(\mu(t)\mathbb{1}_{\{Q^F(t)\leq N(t)\}} + \nu(t)\mathbb{1}_{\{Q^F(t)>N(t)\}}\right)\mathbf{VAR}\big[Q^D(t)\big] \\ &\quad + \lambda(t) + \mu(t)\min\big\{Q^F(t), N(t)\big\} + \nu(t)\big\{Q^F(t) - N(t)\big\}^+. \end{aligned} \tag{3.61}$$

This differential equation again can be solved by standard numerical methods. We used the fourth order Runge-Kutta method as the Euler method led to instabilities in the excluded critical points, when the number of customers equals the number of agents. It gives us an estimate of the variability of the queueing process.

Additionally, the probability of delay as a performance measure similar to the service level used in practice is determined by Garnett et al. (2002) from the heavy traffic diffusion approximation by Halfin and Whitt. The probability of delay is the probability that the wait of an arriving customer is greater than zero. This approximation is often used in both time-dependent and stationary contact center models for setting the staffing level[70], as this approximation gives rise to the *square-root safety-staffing rule*. Feldman et al. (2005) extended the method used by Jennings et al. (1996) to achieve time-stable performance. Garnett et al. (2002) show that the probability of delay α for the system described above can be approximated by

$$P(\text{wait} > 0) = \alpha = \left[1 + \sqrt{\frac{\mu}{\nu}}\frac{h\left(\beta\sqrt{\frac{\mu}{\nu}}\right)}{h(-\beta)}\right]^{-1} \tag{3.62}$$

with h denoting the hazard rate function of the standard normal distribution which is

$$h(x) = \frac{\phi(x)}{1 - \Phi(x)}$$

and a real-valued constant β which is determined by

$$\beta = \frac{Q^D(t) - \frac{\lambda(t)}{\mu}}{\sqrt{\frac{\lambda(t)}{\mu}}} \tag{3.63}$$

Further approximations of performance measure especially for the mean waiting time and the probability of abandoning are given in Garnett et al. (2002) depending on α, β and the other parameters of the model. More recently, Whitt (2003) and Jelenkovic et al. (2004) determined performance measures based on the diffusion process.

However, these performance measures are often limited to the case of homogeneous agents and customers, who do not retry. Therefore, we do not

[69] See Mandelbaum et al. (1998) p. 199.

[70] See, e.g., Jennings et al. (1996).

investigate these approximations any further. Our concern is to show how the variance just derived can be used for contact center analysis and how the variance is influenced by the different parameters of the contact center. Furthermore, we compare the approximated variance to simulation results in the next subsection.

3.3.3 Comparison of Approximation and Simulation Results

The main question of this subsection is whether the resulting diffusion process describes the variances of the queueing process accurately. If this is the case, we are able to calculate confidence intervals around the fluid process within which the realisation of the processes should be with a certain probability.

The simulation results for the variances were calculated over 500 repetitions. We assume that in general the mean service time μ^{-1} is one minute and the mean time to abandon ν^{-1} is thirty seconds. Furthermore, 150 agents are scheduled for the whole day. The arrival rate is given by Equation (3.13) on Page 29 with parameters

$$m_1 = 9500 \qquad t_0 = 7 \text{ am} \qquad t_2 = 4 \text{ pm}$$
$$m_2 = 8000 \qquad t_1 = 12{:}30 \text{ pm} \qquad t_3 = 8 \text{ pm}.$$

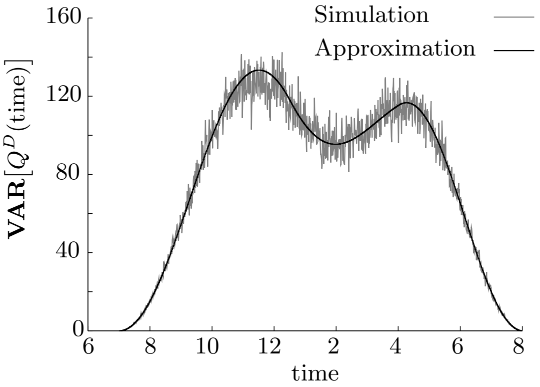

Fig. 3.28. Comparison of variances with low load

In Figures 3.28 and 3.29 the time-dependent variances of the simulated queueing process are compared to the variances of the limiting diffusion approximation. Contrary to the other examples, in the first figure we assumed a lower arrival rate function with $m_1 = 8000$ and $m_2 = 7000$. In this case the traffic intensity of the contact center is below one such that no customer abandons in the fluid model. Figure 3.28 shows that the variance of the queueing process is quite well approximated. The results of the simulation vary more strongly than the results of the initial value problem given by the differential equations for the fluid process (3.23) on Page 36, the variance process (3.61),

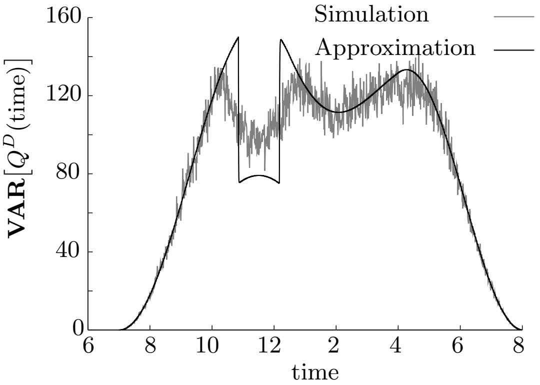

Fig. 3.29. Comparison of variances with high load

and some initial values for these processes. As in the fluid approach, the initial value problem was solved by fourth order Runge-Kutta methods.

In the second Figure 3.29 the contact center has a higher traffic intensity. The contact center is temporarily overloaded. The abandonment rate differs from the service rate. The differential equation reacts much more quickly on the change in the rate than the true variance of the simulation model. That is why the approximation differs from the variance of the queueing process in this part of the graph.

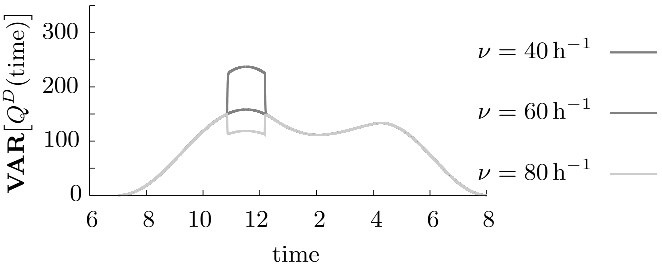

Fig. 3.30. Comparison of variances calculated by the approximation for varying abandonment rates ν

To confirm this observation we vary the abandonment rate from high patience to low patience and compare the results of the differential equations and the simulation model in Figures 3.31 and 3.30. If no customer has to wait, the curves describing the approximated variance $\mathbf{VAR}[Q^D(\text{time})]$ are identical for all abandonment rates, because no-one reneges. In the case of waiting customers, a low abandonment rate $\nu = 40\,\mathrm{h}^{-1}$ leads to a higher variance in Figure 3.31 and a high abandonment rate $\nu = 80\,\mathrm{h}^{-1}$ to a lower variance.

Comparing the solution of the differential equation in Figure 3.30 and the simulation results in Figure 3.31, we recognise the same behaviour of the variances in the differential equation solution and the simulation results. The main difference is that the simulated variances need some time to increase

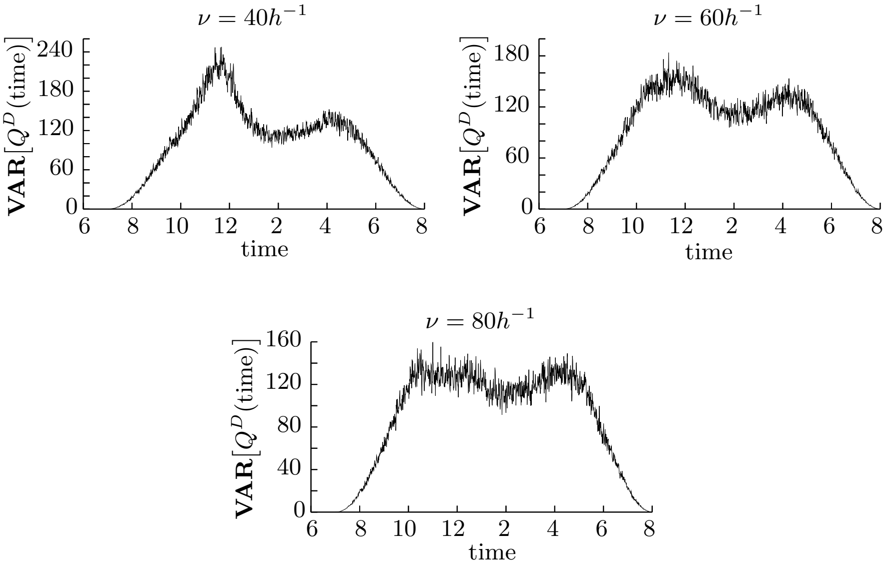

Fig. 3.31. Comparison of variances calculated by the simulation tool for varying abandonment rates ν

and decrease while the differential equations react instantaneously. The time the simulation takes to get into the new variance state seems to be equivalent to the time a system with constant rates needs to reach the steady state. It would be worthwhile studying this phenomenon in detail which could lead to a better understanding of the processes in contact centers.

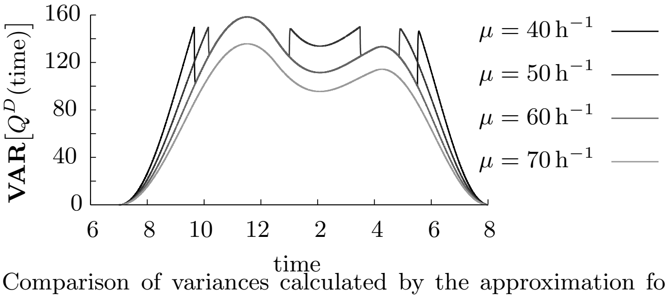

Fig. 3.32. Comparison of variances calculated by the approximation for varying service rates

We now fix ν at $60\,\text{h}^{-1}$ and vary the service rate from 40 to 70 by steps of 10 customers per hour to study a call center under very different load conditions. In the case of service rates below the abandonment rate, Figure 3.32 and 3.33 show that whenever the call center is overloaded, the variance decreases rapidly and increases as fast if the traffic intensity becomes less than one. If the abandonment rate equals the service rate, the variance does not change so strongly, as the model with equal rates can be interpreted as an infinity server queue and the customers leave by abandonment as fast as by

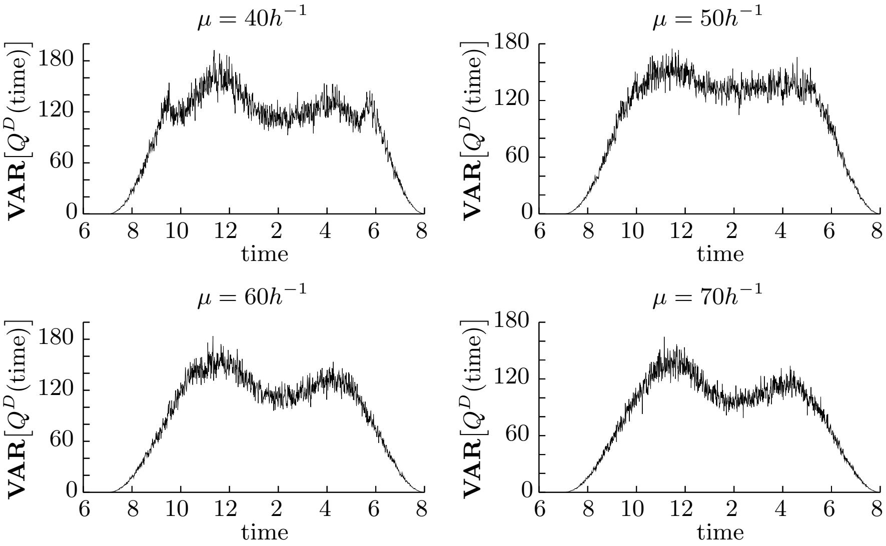

Fig. 3.33. Comparison of variances calculated by the simulation tool for varying service rates μ

service completion. In the last case with high service rates the product of the number of servers and the service rate always stays above the arrival rate such that no abandonment takes place. Consequently, the variance does not react that strongly.

Summarising the results, we find that the parameters of the contact center model have a high impact on the variance, especially when the process of the number of customers in the system passes the critical level. Furthermore, by means of the diffusion refinement we are able to make a good guess about the variance in the system. The estimated variances can be used to compute confidence intervals for the fluid approximation of the customers in the system.

3.3.4 Applicability and Limitation for Contact Center Analysis

The advantage of the diffusion model lies in the opportunity to model both time-dependencies and randomness in contact centers. However, these equations are very difficult to derive and the applicability in a staffing and shift scheduling procedure has not yet been examined.

The differential equations for the variances and covariance are linked to the fluid results. Therefore, some disadvantages of the fluid approach are carried over to the diffusion refinement. However, Mandelbaum et al. (1999a,b, 2002) show that the variances can be used to calculate variance envelopes for the fluid approach and the performance measures calculated by this approach. By both the fluid approach and the diffusion refinement the stochastic processes of the number of customers can very well be approximated.

The diffusion approximation is tied to the fluid model as the fluid approach is needed to determine the mean process. This means that the diffusion approach is a refinement or enlargement rather than a separate approach as stressed by many authors[71]. The derivation of the diffusion equations and the differential equations for the variances appears to be very complicated. If the additional assumption with respect to the times of critical loading is fulfilled, the equations become more simple.

Furthermore, in a staffing and scheduling approach the variances calculated by the diffusion refinement can be used to make staffing decisions more robust with respect to the stochasticity of the contact center. In large contact centers the results based on a fluid approximation will already be quite robust. Therefore the fluid approach will suffice. However, Feldman et al. (2005) show that the diffusion limit can be used to stabilise the performance of the contact center if staffing is done according to the square-root safety-staffing rule which results from the approximation of the delay probability.

All these advantages make the fluid and the diffusion approach a worthwhile instrument of contact center analysis, which should be extended.

3.4 Literature related to the Fluid and Diffusion Approach

The literature on the approximation of contact center models by means of fluid and diffusion approaches is extensive. Especially since 2000 it has been rapidly growing. Therefore, the overview given here cannot be all-embracing. We aim to give some suggestions for further reading.

The literature on fluid approximation can be subdivided into the literature based on the stochastic theory of deriving fluid limits for service systems by Newell[72] and the fluid approximations based on the scaling approach by Halfin and Whitt (1981) which is used in this thesis. Therefore, the overview is restricted to the recent literature based on this approach.

The growing interest in the development of fluid limits for contact center models based on the scaling approach by Halfin and Whitt started in 1995 with a paper by Mandelbaum and Massey. They used the theory to analyse a Markovian queueing system with time-dependent arrival rates and service rate. In 1998 Mandelbaum, Massey, and Reiman extended this theory by adding the fluid refinement to the theory of so-called strong approximations of Markovian service networks. Furthermore, they showed that the fluid limit derived by Halfin and Whitt ca n be more useful than the one by Newell and differs significantly for multiserver queues. The paper by Mandelbaum et al. (1998) covers many features found in real-world call centers such as abandonment, retrials and skills-based routing.

[71] See Whitt (2002a); Mandelbaum et al. (1998) and references therein.

[72] See Newell (1982) and Jiménez and Koole (2004).

Within the theoretical framework of strong approximations for Markovian service networks Mandelbaum et al. (1999a,b), and Mandelbaum et al. (2002) present an extensive analysis of a special fluid queue with retrials. Additionally, they analyse and derive diffusion limits for the theoretical queue length and virtual waiting time.

Altman, Jiménez, and Koole (2001) and Jiménez and Koole (2004) show that the fluid limits are worthwhile if the system is partially overloaded. Jiménez and Koole (2004) prove that the fluid limits are lower bounds for the actual queue length.

More recently, the fluid approximation is used to analyse even more complex contact center models, such as models with retrials or skills-based routing. But not only for analysis but also for other concerns, e.g. staffing and control, are these approximations used. Aguir et al. (2004) developed a model to separate the retrials arriving in a call center from the primary calls of an empirical arrival rate.

Feldman et al. (2005), Harrison and Zeevi (2004, 2005), Whitt (2006a), and Hampshire and Massey (2005) developed methods and rules of thumb for staffing call centers based on fluid approximation with different objective targets. In the paper by Feldman et al. (2005) the target is stabilising the time-dependent performance, while Hampshire and Massey (2005) aim to maximise a profit function. Furthermore, they present an extensive theoretical analysis of the influences of the parameters on the profit function and the optimal solution. Harrison and Zeevi (2004, 2005) investigate the consequences of random arrival rate functions and Whitt (2006a) the influence of so-called absenteeism, i.e., the random events that cause scheduled agents not to appear for working.

An important sector in contact centers with heterogeneous structures, where fluid models and diffusion refinement play a major role, is the control or routing of customers and agents to each other. Atar (2005a,b) obtains asymptotically optimal routing policies for scheduling control and design of contact centers with heterogeneous structures. The analysis is deepened and extended in the papers by Atar et al. (2004a,b), Armony and Mandelbaum (2004), and Gurvich et al. (2004), in which Armony and Mandelbaum (2004) and Gurvich et al. (2004) consider the staffing problem as well. In contrast to these authors, Chang et al. (2004) use a fluid model to develop routing policies.

Besides a few examples of papers that deal with diffusion approximation of contact centers with different customer and agent classes, in most papers only single server systems are analysed. Mainly the diffusion approach is used to circumvent the difficulties associated with other than exponential distributions of service and interarrival times, see Anisimov and Atadzhanov (1994); Glynn and Whitt (1995); Mitzlaff (1997) for earlier results. In addition to the these papers, Ward and Glynn (2003b, 2005) consider balking and reneging.

This overview shows that the research on contact centers by means of fluid and diffusion approaches is a huge and vivid research field with various directions. Our main focus lies on the analysis and shift scheduling in contact centers with retrials.

4

Analysis of Time-Dependent Contact Centers with Retrials

4.1 Contact Centers with Homogeneous Customers and Agents

4.1.1 Description of a Contact Center Model with Retrials

This chapter is dedicated to the analysis of contact center models with retrial behaviour of impatient customers by means of the fluid approach and diffusion refinement introduced in the previous chapter. We aim to show how retrials of customers influence the time-dependent and aggregated performance of the contact center as well as the profitability.

For the analysis of the model with retrials we use the framework of Markovian Service Networks developed in Mandelbaum et al. (1998) and used in Sections 3.2 and 3.3. We want to consider both the time-dependent behaviour and the stochasticity of the queueing processes.

Figure 4.1 is a schematic presentation of the contact center considered in this section. This model has been analysed by Mandelbaum et al. (1998) before[1]. We assume homogeneous customers and agents where the number of agents $N(t)$ may vary over the day. We assume that customers arrive according to a Poisson process with time-dependent rate $\lambda(t)$.

If an agent is available, an arriving customer is served immediately, otherwise the caller has to wait for service in an infinite waiting room. Because customers are supposed to be impatient, this is not a very restrictive assumption in a large service system as shown by Garnett et al. (2002) and Figure 3.11 on Page 45. If a waiting customer in the queue is not willing to wait any longer, i.e., he reaches his individual waiting time limit, this customer hangs up and may call again later. On average customers abandon with rate $\nu(t)$, i.e., the waiting time limits are exponentially distributed with parameter $\nu(t)$. The percentage of callers who will retry after an exponentially distributed time in the so-called orbit with rate $\gamma(t)$, is denoted by p. The orbit is modeled as an

[1] See, also Mandelbaum et al. (1999a,b), and Mandelbaum et al. (2002).

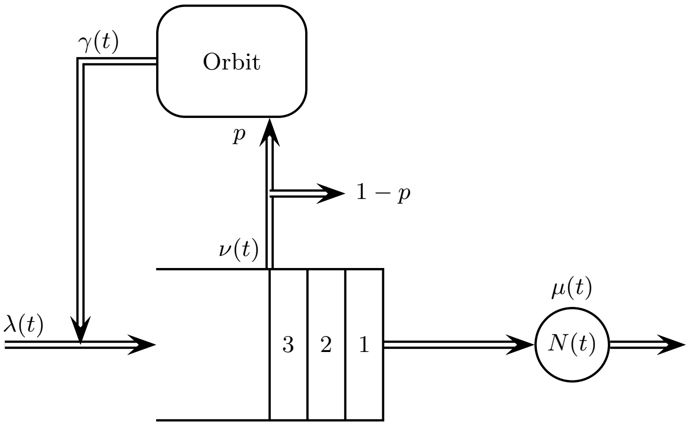

Fig. 4.1. A contact center model with a single class of customers, a single group of agents, and retrials of impatient customers

infinite server queue as each customer in the orbit is his own server. Both the times to abandonment and times to retrial are allowed to be time-dependent. Furthermore, the agents serve customers with exponentially distributed service times with time-varying parameter $\mu(t)$.

4.1.2 Determination of the Fluid Processes

To derive the two fluid processes, we describe the stochastic system by a vector $\boldsymbol{Q}(t) = (Q_S(t), Q_{\mathcal{O}}(t))$. This vector consists of the processes $Q_S(t)$ for the number of customers in the system and $Q_{\mathcal{O}}(t)$ for the number of customers in the orbit at some time $t \in \mathbb{R}_0^+$.

As shown in Subsection 3.2.2 we can model these processes by a sum of independent, non-homogeneous Poisson processes which describe the number of customers moving into and out of the system and the orbit, respectively. If we scale the processes according to the well-known heavy traffic scaling presented in Halfin and Whitt (1981) with scaling parameter n, we arrive at the so-called fluid limits justified by the functional strong law of large numbers[2].

To determine the system of differential equations of the fluid processes approximating the stochastic processes, these stochastic processes $Q_S(t)$ and $Q_{\mathcal{O}}(t)$ are substituted by the deterministic processes $Q_S^F(t)$ and $Q_{\mathcal{O}}^F(t)$. Then the change in the amount of customers in the system and the orbit given by the derivative of the processes can be explained by the rates of changes.

The change in the number of customers in the system $dQ_S^F(t)$ during a small time interval dt is given by the number of customers who enter the

[2] Mandelbaum et al. (1998) determine the fluid differential equations analytically as shown in Section 3.2.2. They also show a more general case with respect to the time-dependent rates for the model investigated in this section.

system $(\lambda(t)+\gamma(t)Q_{\mathcal{O}}^F(t))dt$ during the time interval minus the number of customers who leave the system $\mu(t)\min\{Q_S^F(t),N(t)\}\,dt+\nu(t)\{Q_S^F(t)-N(t)\}^+dt$. As in Equation (3.23) on Page 36, customers enter as primary calls with rate $\lambda(t)$ but additionally customers from the orbit retry with an individual rate $\gamma(t)$. The rate by which customers leave the contact center in the model with retrials is equal to the rate in the model without retrials in Equation (3.23).

The number of customers in the orbit $Q_{\mathcal{O}}^F(t)$ increases by the customers who have left the contact center because of impatience and are willing to call again with rate $p\nu(t)\{Q_S^F(t)-N(t)\}^+$. During a small time interval dt the number in orbit decreases by the number of customers who retry $\gamma(t)Q_{\mathcal{O}}^F(t)dt$. Consequently, this gives rise to[3]

$$\begin{aligned}\frac{d}{dt}Q_S^F(t) &= \lambda(t)+\gamma(t)Q_{\mathcal{O}}^F(t)-\mu(t)\min\{Q_S^F(t),N(t)\}\\ &\quad -\nu(t)\{Q_S^F(t)-N(t)\}^+ \end{aligned} \tag{4.1a}$$

$$\frac{d}{dt}Q_{\mathcal{O}}^F(t) = p\nu(t)\{Q_S^F(t)-N(t)\}^+-\gamma(t)Q_{\mathcal{O}}^F(t) \tag{4.1b}$$

for all $t\in\mathbb{R}_0^+$. As mentioned in Section 3.2 we solved the initial value problems linked to differential equations numerically by means of Euler and Runge-Kutta methods. Afterwards the results are used to determine performance measures for the underlying contact center model, so that the performance can be analysed.

4.1.3 Refinement to a Diffusion Model

The scaled processes and the fluid limit just determined can now be utilised to derive the diffusion limits[4]. Applying the functional central limit theorem leads to stochastic functional equations for the diffusion processes which are needed to determine the variances and covariances of the processes as described in Section 3.3.

Therefore, we assume that the set of critical times has measure zero. Similar to Section 3.3 on Page 62, the set of critical times $\boldsymbol{S}$ contains all moments when the number of customers in the system equals the number of agents on duty. Then the stochastic functional equations[5] and their interpretation are much more transparent.

Based on the stochastic functional equations the differential equations of the variances and covariance are derived as described in Section 3.3 on

[3] Compare, Mandelbaum et al. (1999a,b), and Mandelbaum et al. (2002). They derive the fluid limits by means of the strong law of large number as presented in the previous chapter.

[4] See Mandelbaum et al. (1998) Section 5 or Mandelbaum et al. (1999a).

[5] The stochastic functional equations with and without this assumption can also be found, e.g., in Mandelbaum et al. (1999a) and Mandelbaum et al. (2002).

Pages 64f. These differential equations for the variances and covariance are given by[6]

$$\begin{aligned}
&\frac{d}{dt}\mathbf{VAR}\big[Q_S^D(t)\big] \qquad (4.2a)\\
&\quad = -2\left(\nu(t)\mathbb{1}_{\{Q_S^F(t)>N(t)\}} + \mu(t)\mathbb{1}_{\{Q_S^F(t)\le N(t)\}}\right)\mathbf{VAR}\big[Q_S^D(t)\big]\\
&\qquad + 2\gamma(t)\mathbf{COV}\big[Q_S^D(t), Q_{\mathcal{O}}^D(t)\big] + \lambda(t) + \gamma(t)Q_{\mathcal{O}}^F(t)\\
&\qquad + \mu(t)\min\big\{Q_S^F(t), N(t)\big\} + \nu(t)\big\{Q_S^F(t) - N(t)\big\}^+\\
&\frac{d}{dt}\mathbf{VAR}\big[Q_{\mathcal{O}}^D(t)\big] \qquad (4.2b)\\
&\quad = 2p\nu(t)\mathbb{1}_{\{Q_S^F(t)>N(t)\}}\mathbf{COV}\big[Q_S^D(t), Q_{\mathcal{O}}^D(t)\big] - 2\gamma(t)\mathbf{VAR}\big[Q_{\mathcal{O}}^D(t)\big]\\
&\qquad + p\nu(t)\big\{Q_S^F(t) - N(t)\big\}^+ + \gamma(t)Q_{\mathcal{O}}^F(t)\\
&\frac{d}{dt}\mathbf{COV}\big[Q_S^D(t), Q_{\mathcal{O}}^D(t)\big] \qquad (4.2c)\\
&\quad = -\left(\nu(t)\mathbb{1}_{\{Q_S^F(t)>N(t)\}} + \mu(t)\mathbb{1}_{\{Q_S^F(t)\le N(t)\}}\right)\mathbf{COV}\big[Q_S^D(t), Q_{\mathcal{O}}^D(t)\big]\\
&\qquad + \gamma(t)\left(\mathbf{VAR}\big[Q_{\mathcal{O}}^D(t)\big] - \mathbf{COV}\big[Q_S^D(t), Q_{\mathcal{O}}^D(t)\big]\right) + p\nu(t)\mathbf{VAR}\big[Q_S^D(t)\big]\\
&\qquad - p\nu(t)\big\{Q_S^F(t) - N(t)\big\}^+ - \gamma(t)Q_{\mathcal{O}}^F(t)
\end{aligned}$$

for all $t \in \mathbb{R}_0^+$. The meaning of these equations as well as the influence of the various parameters are explained by means of numerical examples in Subsection 4.1.5.

4.1.4 Performance Measures

Based on the fluid results we calculate some performance measures which are quite similar to the performance measures derived for the simpler time-dependent Erlang-A model discussed in Chapter 3. First, we have to distinguish between the waiting times in the queue and the times in the orbit.

The time-dependent waiting time $W_S^F(t)$ of a customer arriving in the system is derived equivalently to the waiting time in a system without retrials by dividing the number of waiting customers by the departure rate $d(t)$ in Equation (3.26) on Page 38. Therefore, the Equations (3.27) on Page 39 and Equation (4.3) are the same. The time-dependent waiting time is given by

$$W_S^F(t) = \frac{\max\{0, Q^F(t) - N(t)\}}{\mu(t)\min\big\{Q_S^F(t), N(t)\big\} + \nu(t)\big\{Q_S^F(t) - N(t)\big\}^+}. \qquad (4.3)$$

As the orbit is modelled as an infinite server queue, the time-dependent waiting time $W_{\mathcal{O}}^F(t)$ of customers in the orbit equals the average time the

[6] For the derivation of the differential equations see Appendix A.1 and Section 5 of Mandelbaum et al. (1998).

caller spends in the orbit, which is the mean time to redial $\gamma(t)^{-1}$. The aggregated waiting time of customers $W^F_{S,agg}(T)$ in the system during the time interval $[0, T]$ is derived as shown in Equation (3.29) on Page 40.

In this model the time-dependent probabilities of being served or abandoning are similar to the probabilities determined on Page 40. For a customer entering the system at time t the probability of being served is given by the fraction of customers that are served out of all customers who leave the system. The number of customers leaving at time t is simply the departure rate $d(t)$ and the number of customers being served at time t is the service rate multiplied by the number of agents, if the number of customers in the system exceeds the number of agents on duty. Otherwise the probability of being served would be one. Therefore, the probability of being served is given by

$$P^F(\text{served}, t) = \frac{\mu(t) \min\bigl\{Q^F_S(t), N(t)\bigr\}}{\mu(t) \min\bigl\{Q^F_S(t), N(t)\bigr\} + \nu(t)\bigl\{Q^F_S(t) - N(t)\bigr\}^+}. \tag{4.4}$$

Similarly, the time-dependent percentage of customers abandoning without retrying can be determined by multiplying the probability of abandoning derived in Equation 3.31 on Page 40 by the probability of not retrying $(1-p)$, i. e.,

$$P^F(\text{abandon}, t) = \frac{(1-p)\nu(t)\bigl\{Q^F(t) - N(t)\bigr\}^+}{\mu(t) \min\bigl\{Q^F_S(t), N(t)\bigr\} + \nu(t)\bigl\{Q^F_S(t) - N(t)\bigr\}^+}. \tag{4.5}$$

As the probability of being served, abandoning without retrial, and moving into the orbit must add up to one, the probability of abandoning and moving into the orbit is given by the probability of abandoning determined in Equation (3.31) weighted by the probability of retrial p, i.e.,

$$P^F(\text{into orbit}, t) = \frac{p\nu(t)\bigl\{Q^F(t) - N(t)\bigr\}^+}{\mu(t) \min\bigl\{Q^F_S(t), N(t)\bigr\} + \nu(t)\bigl\{Q^F_S(t) - N(t)\bigr\}^+}. \tag{4.6}$$

In Section 3.2 on Page 40 two different aggregated probabilities of being served were derived, which numerically do not differ much in the case of the contact center considered there. However, for the contact center model in this section, these probabilities can be used to determine on the one hand the percentage of customers who have been served out of all primary calls during the time interval $[0, T]$ $P^F_\lambda(\text{served}, T)$ and on the other hand the percentage of customers who have been served out of all departures $P^F_{agg}(\text{served}, T)$. The second probability does not distinguish between primary calls and retrials, therefore it judges the actual performance of the contact center in the considered time interval more accurately. Furthermore, this performance measure is associated with the empirical performance measure.

Although agents in the system do not know whether an arriving customer calls for the first time or retries, the probability that an arriving customer is a recaller can be calculated by means of the retrial rate $\gamma(t)$, the arrival rate $\lambda(t)$ and the process describing the number of customers in the orbit $Q_{\mathcal{O}}(t)$. Customers from the orbit retry on average with rate $\gamma(t)$, which leads to a total arrival rate of recalls to the system of $\gamma(t)Q_{\mathcal{O}}^F(t)$. The sum of the primary arrival rate and the retrial arrival rate is the total arrival rate in the system. Consequently, the probability that a customer arriving at time t is a recaller is given by the fraction of the retrial arrival rate and the total arrival rate in the system, i.e.,

$$P^F(\text{recaller}, t) = \frac{\gamma Q_{\mathcal{O}}^F(t)}{\lambda(t) + \gamma Q_{\mathcal{O}}^F(t)}. \tag{4.7}$$

Equivalently, the probability that a customer is a primary caller is given by

$$P^F(\text{primary call}, t) = \frac{\lambda(t)}{\lambda(t) + \gamma Q_{\mathcal{O}}^F(t)}. \tag{4.8}$$

These probabilities can be used in an empirical analysis to separate the primary calls from the retrial, but they are of almost no benefit for a performance analysis. Therefore, these probabilities are not investigated any further[7].

Finally the utilisation of the agents at time t is derived by dividing the number of busy agents by the number of agents scheduled:

$$U^F(t) = \frac{\min\{Q_S^F(t), N(t)\}}{N(t)}. \tag{4.9}$$

The aggregated utilisation $U_{agg}^F(T)$ is given by Equation 3.36 on Page 42.

The formal similarity of all performance measures of this section and section 3.2 could lead to the conclusion that the retrial behaviour of customers does not influence the performance of the system. However, the number of customers in the system calculated by solving the initial value problem defined by Equation (4.1) and some initial conditions depends on the retrial parameters. Therefore, the retrial behaviour influences the performance measures implicitly.

In addition to the technical performance measures, the profit gained in a time period $[0, T]$ is a measure of the economic performance of the contact center. It is derived in the same way as in Section 3.2.3.2 on Pages 42f. and given there in Equation (3.38), i.e.,

$$\text{profit}(T) = \int_0^T r\mu(t) \min\{Q_S^F(t), N(t)\} - \ell Q_S^F(t) - wN(t)\, dt. \tag{4.10}$$

Equivalently to the technical performance measures, the profit is implicitly influenced by the retrial parameters. Similar to the profit function (3.38) on Page 42, the profit function is closely related to the aggregated probability of being served $P_{agg}^F(T)$.

[7] See Aguir et al. (2004) and Aguir et al. (2005) for an application of these probabilities in staffing.

4.1.5 Numerical Analysis

4.1.5.1 The Number of Customers in the System and in the Orbit

In this section we investigate whether the numerical solution to the initial value problem given by the Differential Equations (4.1) and and the initial conditions

$$Q_S(t_0) = 0 \quad \text{and} \quad Q_{\mathcal{O}}(t_0) = 0 \tag{4.11}$$

approximates the simulation results accurately. To solve the initial value problem we use again the fourth order Runge-Kutta methods. We assume that all processes start at zero, i.e., customers are neither in the system nor in the orbit.

For the simulation results we have extended the simulation tool developed by Feldman (2004), so that we can simulate retrials as well. We generate 500 repetitions and compare the average processes to the results of the initial value problem.

The primary arrival rate is defined by Equation (3.13) on Page 29 with parameters

$$\begin{aligned} m_1 &= 9500 & t_0 &= 7 \text{ am} & t_2 &= 4 \text{ pm} \\ m_2 &= 8000 & t_1 &= 12{:}30 \text{ pm} & t_3 &= 8 \text{ pm.} \end{aligned} \tag{4.12}$$

In Figure 4.2 the graph of the arrival rate function is depicted. In Table 4.1 we summarise the parameters for the investigation.

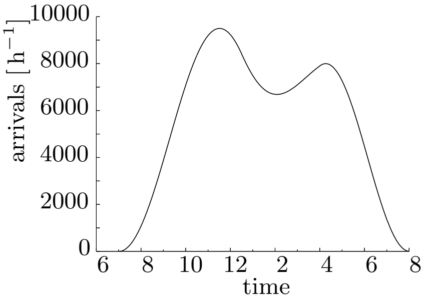

Fig. 4.2. Time-Dependent primary arrival rate function $\lambda(t)$

Figures	$\mu(t)$	$N(t)$	$\nu(t)$	$\gamma(t)$	p
4.3, 4.4	variable	150	120 h^{-1}	15 h^{-1}	0.5
4.5, 4.6	60 h^{-1}	variable	120 h^{-1}	0.5 h^{-1}	0.5
4.7, 4.8	60 h^{-1}	150	variable	15 h^{-1}	0.5
4.9, 4.10	60 h^{-1}	100	120 h^{-1}	variable	0.5
4.11, 4.12	60 h^{-1}	100	120 h^{-1}	15 h^{-1}	variable

Table 4.1. Parameters for the numerical analysis

First of all, the approximation and simulation results are compared for different service rates to show the influence of service rates on the number of customers both in the system and the orbit.

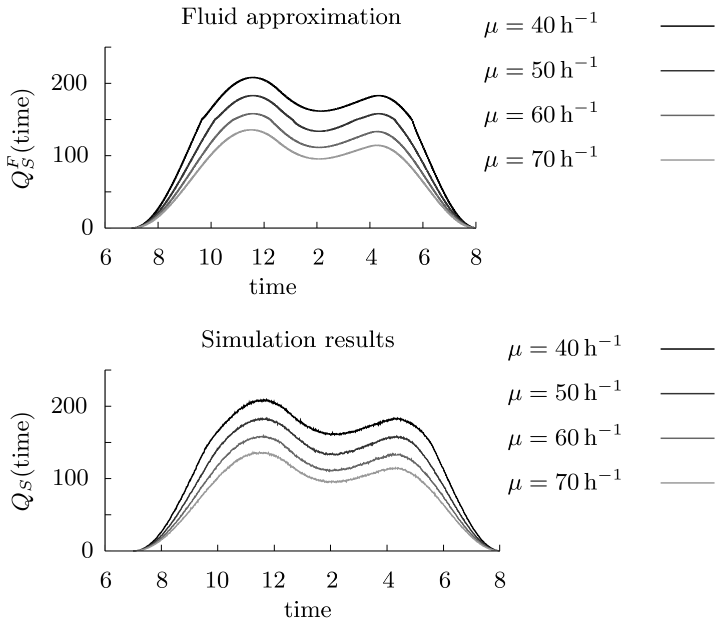

Fig. 4.3. Comparison of the results for the number of customers in system for different service rates μ

We vary the service rate from 40 customers per hour to 70 customers per hour. In Figure 4.3 the approximation fits the simulation remarkably well. The service rate has a huge influence on the number of customers in the system. Although the abandonment rate is high, a lot of customers have to wait in the system with low service rate $\mu = 40\,\mathrm{h}^{-1}$ or $\mu = 50\,\mathrm{h}^{-1}$. A similar result had already been observed for the Erlang-A model in Figures 3.11 on Page 45.

In Figure 4.4 the number of customers in the orbit calculated by the fluid approach is compared to the simulation results. Similar to the approximation of the number of customers in the system, the approximation of the number of customers in the orbit works well and the influence of the service rate on the number of customers in the orbit is quite strong. If the average service times are short, i.e., $\mu = 70\,\mathrm{h}^{-1}$, no customer abandons. Consequently, the orbit remains empty both in the simulation and in the fluid approximation during the whole day. If the contact center is critically loaded or slightly overloaded, the simulation predicts a few more customers in the orbit than the fluid approximation. This result is due to the neglected randomness of the fluid approach. In a real-world contact center and in the simulation some customers will have to wait and already abandon, if the contact center is nearly critically loaded, whereas in the fluid model all customers are still served immediately.

In both Figures 4.3 and 4.4 the maximum numbers of customers in the system and the orbit during the day are reached almost at the same time. If

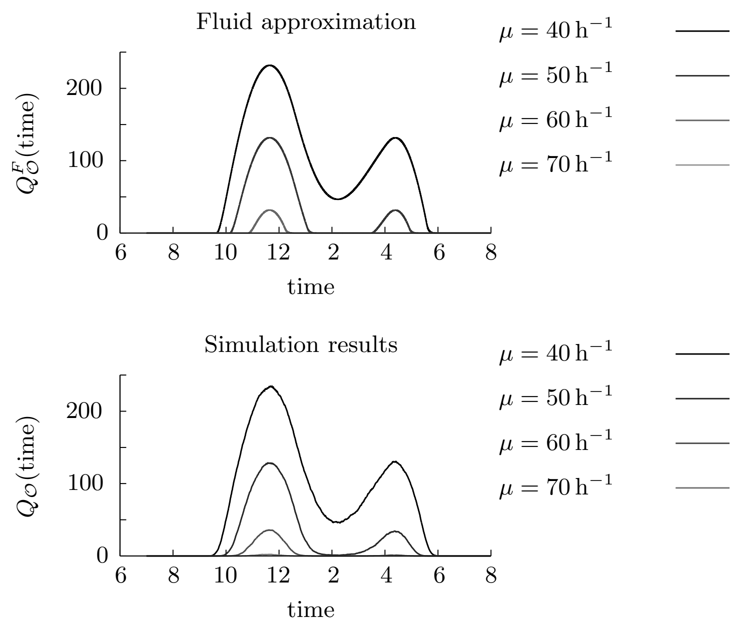

Fig. 4.4. Comparison of the results for the number of customers in orbit for different service rates μ

the contact center is heavily loaded, a lot of customers will wait in the orbit to recall as well. If the number of customers in the system falls below the number of available servers, the number of customers in the orbit decreases as retrying customers are served and no-one has to abandon.

Next the fluid approximation and simulation results for different numbers of agents on duty are compared in Figure 4.5. For this example we use a mean time to redial of two hours, i.e., $\gamma = 0.5\,\text{h}^{-1}$, to illustrate the shifting of work during the day. In Figure 4.5 the results of the fluid approximation and simulation are very similar. During the first half of the day a low number of agents (e.g. $N(t) = 80$) on duty leads to fewer customers in the system than a high number of agents (e.g. $N(t) = 140$). If few agents are staffed, a lot of customers abandon because the abandonment rate is quite high. This observation corresponds to the results of Figure 3.11 on Page 45 for the Erlang-A model. In the late afternoon fewer working agents lead to more customers in the system. These customers have abandoned during the day and return on average two hours later, such that they arrive when the primary workload has decreased.

The results depicted in Figure 4.6 stress that a lot of customers have to wait in the orbit if only few agents are staffed. As the mean time to retrial γ^{-1} is assumed to be two hours, some customers who have abandoned during the period of high load in the afternoon remain in the orbit at the end of the

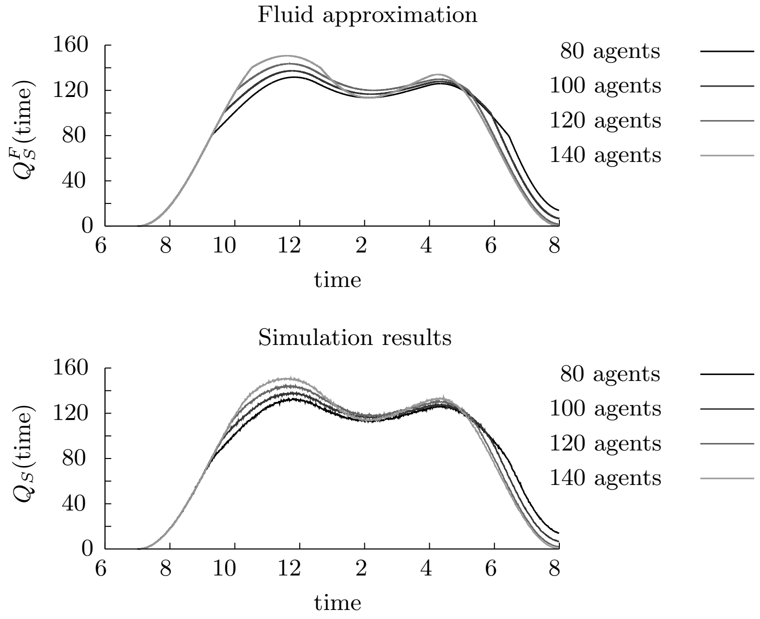

Fig. 4.5. Comparison of the results for the number of customers in the system with different numbers of agents on duty $N(t)$

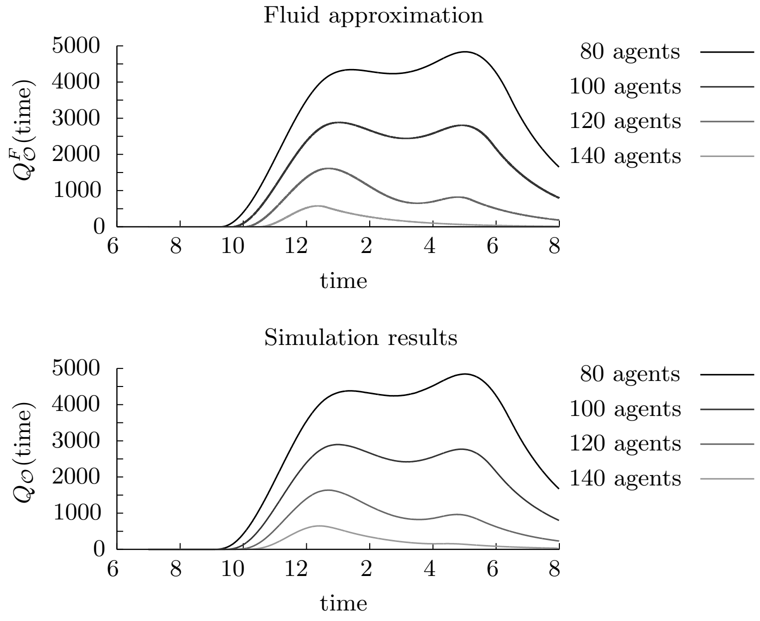

Fig. 4.6. Comparison of the results for the number of customers in the orbit with different numbers of agents on duty $N(t)$

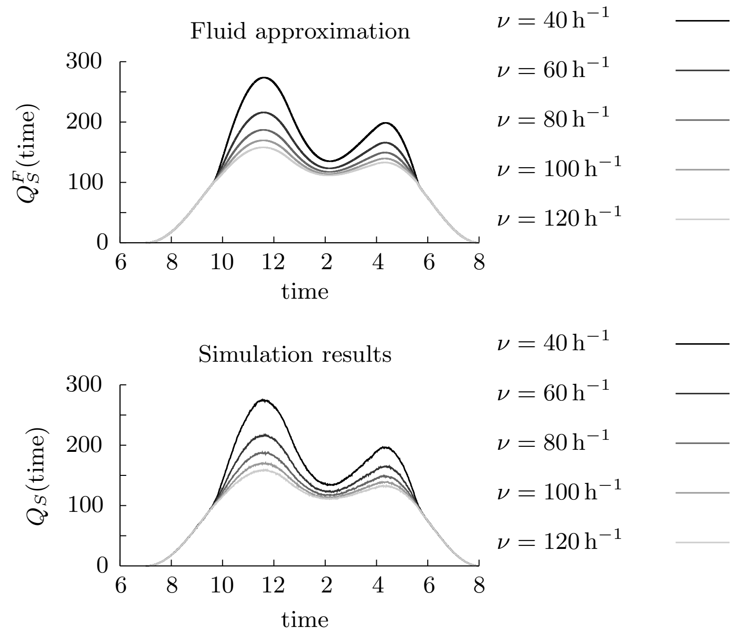

Fig. 4.7. Comparison of fluid approximation and the simulation results for the number of customers in the system with different abandonment rates ν

day. Contrary to Figures 4.3 and 4.4 the times when the maximum number of customers is reached differ for the orbit and the system.

In Figure 4.7 the fluid approximation for the number of customers in the system is compared to simulation results for different abandonment rates ν. The results almost equal the results for the Erlang-A model presented in Figure 3.11 on Page 45. If customers are more patient, i.e., the abandonment rate ν decreases, more customers have to wait in the system. Consequently, the number of customers in the system increases.

Remarkably, in Figure 4.8 the abandonment rate has no influence on the number of customers in the orbit in the approximation as well as in the simulation, as the curves are indistinguishable. This empirical observation is in line with the result of Aguir et al. (2004) who show that the arrival rate of retrials is independent of the abandonment rate.

To complete the comparison of fluid approximation and simulation results the retrial behaviour of customers is investigated. Figures 4.9 and 4.10 present the influence of the mean time to retry γ^{-1} on the number of customers in the system and the orbit.

Contrary to the previous examples the number of agents $N(t)$ is assumed to be 100 in order to cause more abandonment so that the effect of retrial is amplified. In Figure 4.8 we show that the abandonment rate ν does not influence the number of customers in the orbit $Q_{\mathcal{O}}(t)$. Therefore, one would

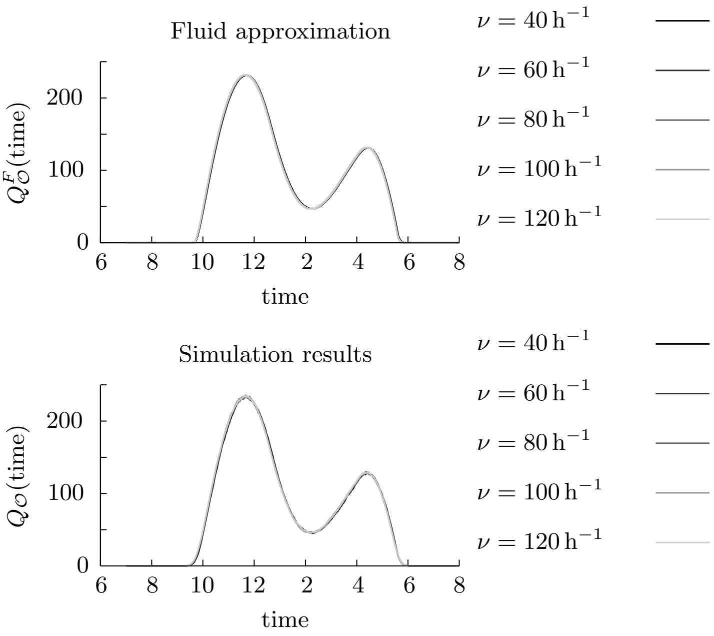

Fig. 4.8. Comparison of the results for the number of customers in the orbit with different abandonment rates ν

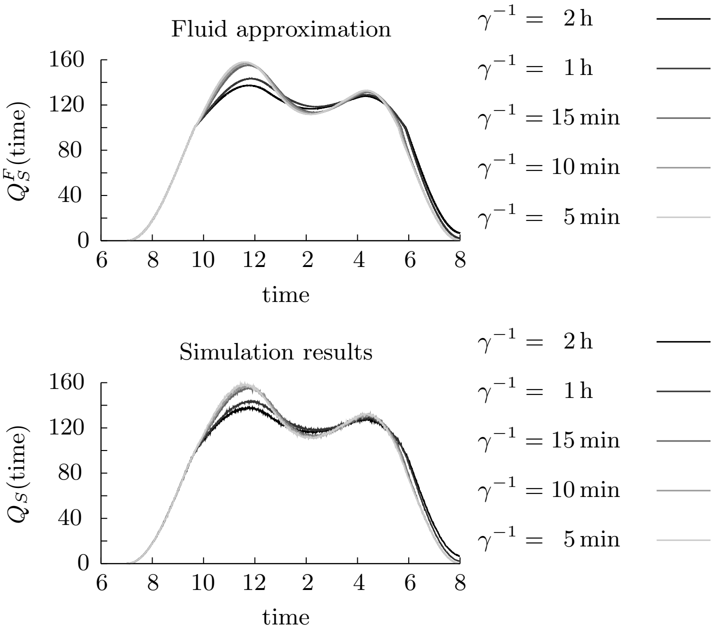

Fig. 4.9. Comparison of the results for the number of customers in the system with different redialling rates γ

assume that the redialling rate might not influence the number of customers in the system $Q_S(t)$ as well.

However, in Figure 4.9 the redialling rate has an effect on the number of customers in the system. If the mean time to redial is low, e.g. $\gamma^{-1} = 5\,\text{min}$, and the contact center is heavily loaded, more customers are in the system, although a lot of customers abandon. If the load of the contact center decreases, a short mean time to retrial leads to a quicker reduction of the number of customers in the system, because the arrival rate from the orbit decreases earlier. If the mean time to retrial γ^{-1} increases, the number of customers in the system decreases, because customers wait in the orbit for a longer time as depicted in Figure 4.10, i.e., low retrial rates cause a shifting of work into later periods when the workload of the system is lower. In this case more customers are stored in the orbit, because a retrial is done quite rarely.

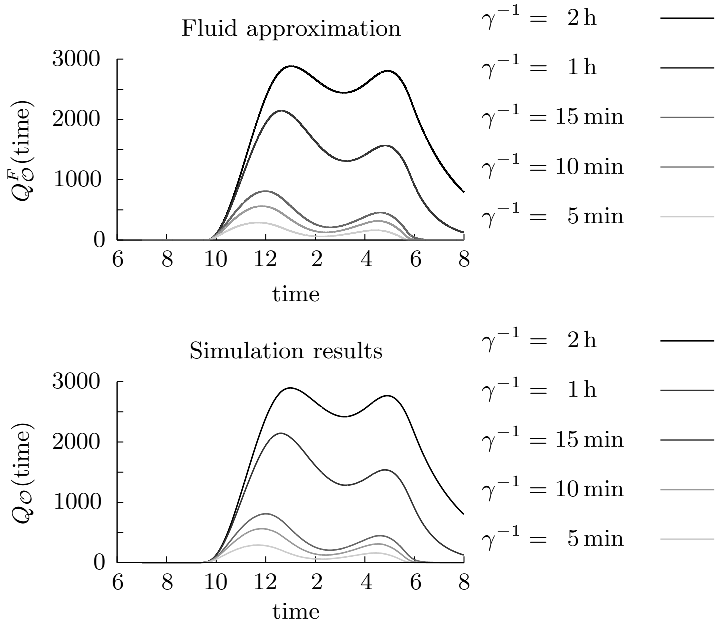

Fig. 4.10. Comparison of the results for the number of customers in the orbit with different redialling rates γ

Consequently, low retrial rates, i.e., high mean times to retrial, help to improve the service for each customer and high retrial rates make the service worse and cause congestion. Similar to the previous example, the approximation works fine.

Finally, the results of the fluid approximation and the simulation for different redial probabilities should be compared. In Figure 4.11 the number of customers in the system is depicted. Similar to the previous investigation of the redialling rates the number of agents on duty is 100. The results of the approximation and the simulations are again alike. Unlike the retrial rate the

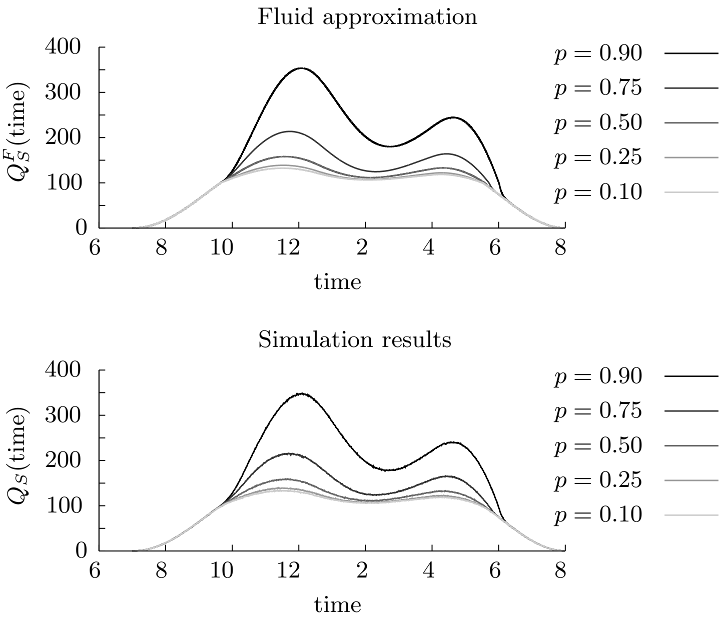

Fig. 4.11. Comparison of the results for the number of customers in the system with different probabilities of retrial p

impact of the retrial probability on the number of customers in the system and the orbit is much stronger. If a lot of customers are willing to retry, the number of customers in the system increases drastically. Consequently, more customers have to wait and abandon. If the probability of retrial decreases, the model will become more and more like the Erlang-A model investigated in the previous Chapter 3.

In Figure 4.12 on Page 87 the influence of the probability of retrial on the number of customers in the orbit is similar to the influence on the number of customers in the system. Hence, it is more important to estimate the probability of retrial correctly from empirical data than to determine the retrial rate exactly.

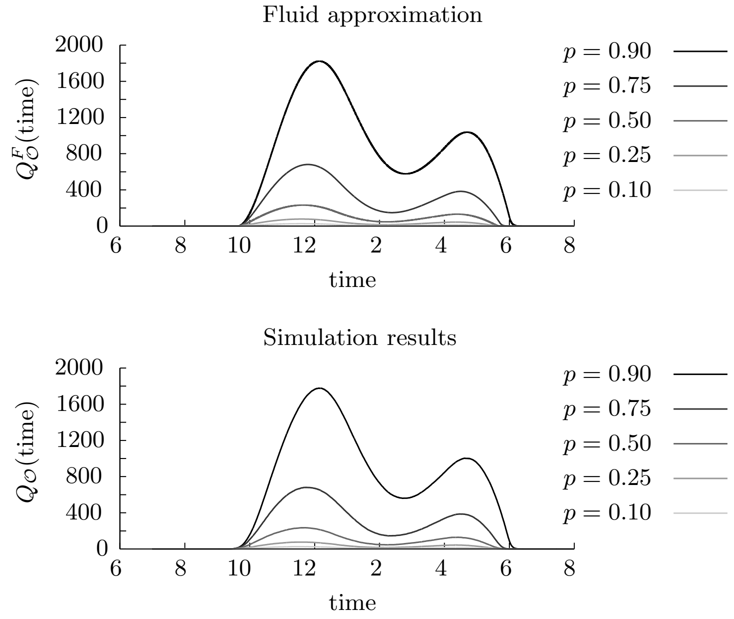

Fig. 4.12. Comparison of the results for the number of customers in the orbit with different probabilities of retrial p

4.1.5.2 Influence of the Parameters on the Time-Dependent Waiting Time

Next the influence of the parameters on time-dependent technical performance measures is studied. We again compare the approximated performance measures with the results from the simulation model. As mentioned before we averaged the results over 500 simulation runs to determine the performance measures.

The arrival rate is time-dependent as shown in Figure 4.2 on Page 79. It is defined by Equation (3.13) with parameters given in (4.12) on Page 79. The other parameters do not depend on time. We assume the same configurations as in the previous section given in Table 4.1 on Page 79 for the waiting time and the probability of being served. When we analyse the influence of the retrial parameters γ and p on the utilisation, we suppose that 150 agents are on duty because otherwise all the agents would be busy nearly all the time. Therefore, different retrial parameters would have almost no influence on the time-dependent utilisation.

We start by studying the influence of the service rate on the time-dependent waiting time in the system. The time-dependent waiting time is given in Equation (4.3) on Page 76. The service rate μ is varied from 40 customers per hour to 70 customers per hour. However, in Figure 4.13 only three curves are visible in the picture referring to the fluid approximation. In the

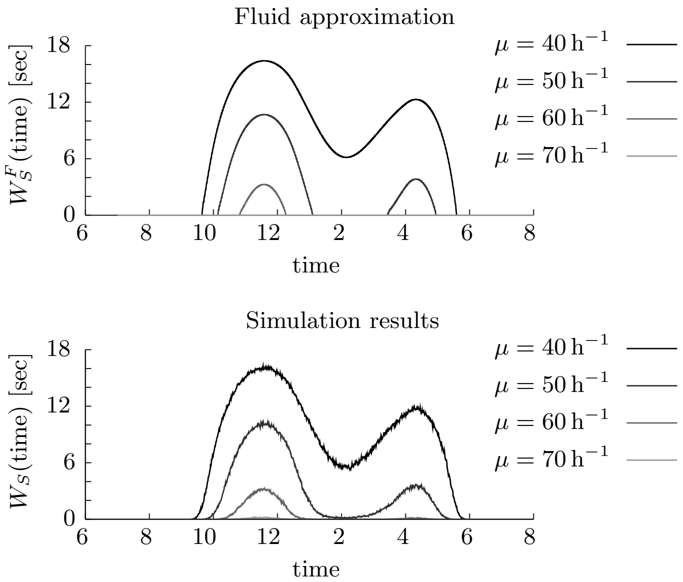

Fig. 4.13. Influence of the service rate μ on the time-dependent waiting time of customers in the system

picture referring to the simulation results the curve associated with $\mu = 70\,\mathrm{h}^{-1}$ is nearly invisible, because the waiting time is almost zero. If the agents work very fast, i.e. $\mu = 70\,\mathrm{h}^{-1}$, no customer has to wait in the fluid model. Therefore, the waiting time is zero all the time. The comparison of the simulation results and the fluid approximation for the other service rates shows that the waiting times are well approximated by the fluid approach. Only if the waiting time is almost zero, the waiting times calculated by the fluid approximation differs somewhat from the waiting times in the simulation. The reason for this difference is that the fluid approach neglects randomness.

In Figure 4.14 the fluid approximations of the waiting time of customers is compared to the waiting times calculated by means of the simulation model for different numbers of agents on duty. As for Figures 4.5 on Page 82 and 4.6 on Page 82, the mean time to redial was assumed to be two hours, i.e., $\gamma = 0.5\,\mathrm{h}^{-1}$. The approximation is accurate except in the surrounding of zero, i.e., if the workload of the contact center is approximately one. A direct comparison of the simulation and approximation results would show that the waiting time calculated by the approximation is a lower bound for the waiting time in the simulation model. The more agents are staffed, the shorter the waiting time becomes. As the mean time to abandonment is assumed to be 30 seconds, the waiting time in the fluid approach will never exceed 30 seconds, even if no agents are staffed at all.

In order to point out the impact of varying abandonment rates ν, 100 agents are assumed to be staffed. In Figure 4.15 the simulation results and the ap-

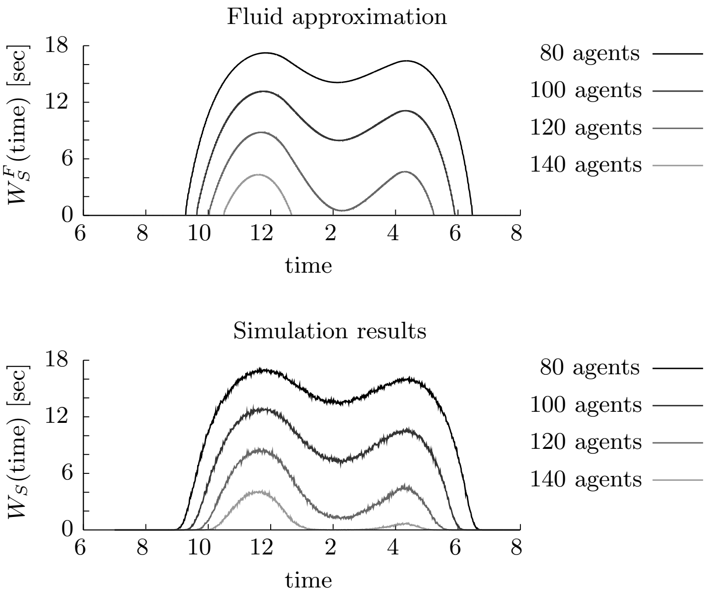

Fig. 4.14. Influence of the number of agents $N(t)$ on duty on the time-dependent mean waiting time of customers in the system

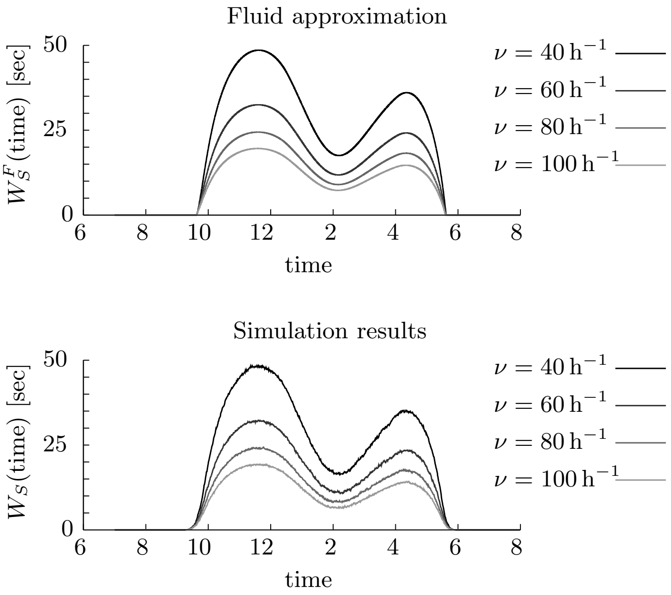

Fig. 4.15. Influence of the abandonment rate ν on the time-dependent waiting time of customers in the system

proximation are compared. The comparison emphasises previous observations. The waiting times enlarge if the customers are more patient. The mean waiting times are far below the waiting time limit, because a lot of customers are served even if the contact center is heavily loaded. If customers are more patient, e.g., $\nu = 40\,\mathrm{h}^{-1}$ or $\nu = 60\,\mathrm{h}^{-1}$, fewer customers abandon and more customers remain in the system in order to wait for attendance. Therefore the waiting time of all customers increases.

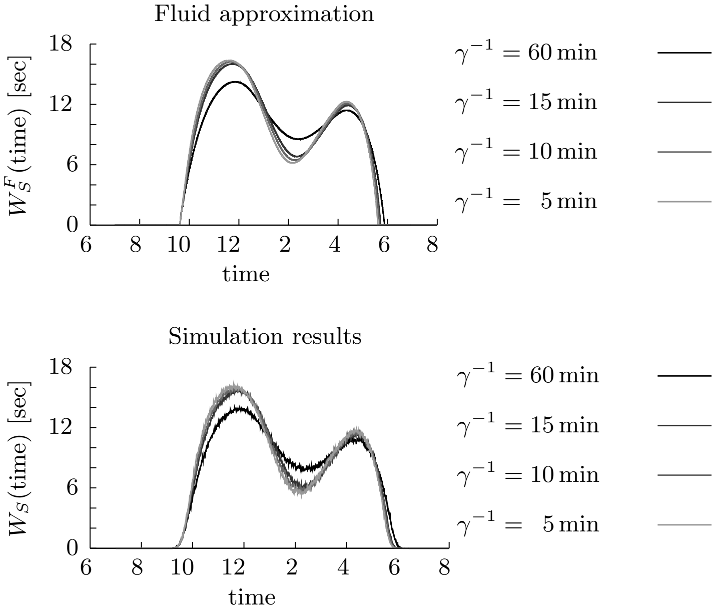

Fig. 4.16. Influence of the mean time to redial γ^{-1} on time-dependent mean waiting time of customers in the system

The influence of the mean time to redial on the time-dependent waiting time is displayed in Figure 4.16. In order to amplify the effects of redialling, we assume that 100 agents are on duty. The waiting time calculated by the fluid approach approximates the waiting time in the simulation model well. If the mean time to redial decreases from 60 minutes to 5 minutes and the contact center is overloaded, the waiting time increases, because more customers retry quickly such that more customers are in the system[8]. Long mean times to redial lead to a shift of workload into later hours of the day, when the traffic intensity is lower. Therefore, the waiting times increase, if the contact center is underloaded during the early afternoon (2 pm) and the mean time to retrial increases. If the mean times to redial are quite short, the difference in the waiting times diminishes.

[8] Compare Figures 4.9 on Page 84 and 4.10 on Page 85.

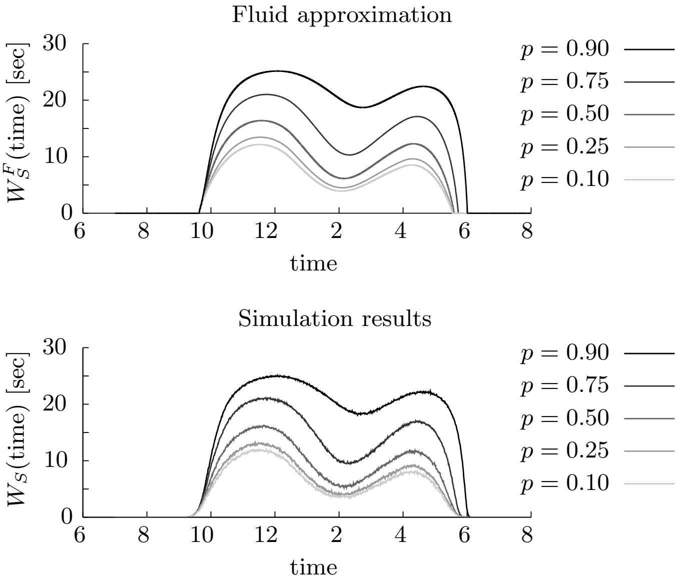

Fig. 4.17. Influence of the probability of retrial p on the time-dependent waiting time of customers in the system

Finally, the results of the fluid approximation for the time-dependent waiting times calculated by Equation (4.3) on Page 76 and the simulation results are compared for different probabilities of retrial p in Figure 4.17. As in the previous example with different mean times to retrial, the number of agents is assumed to be 100.

If the probability of retrial decreases, the waiting time of customers in the system shortens, because more customers will never return and are lost. The remaining customers can be attended to more quickly. If the probability of retrial increases, the local minimum of the waiting time function during the lunch time is shifted into later hours of the day. The reason for this shift is that more customers wait in the orbit, such that some work is stored in the orbit whenever the contact center is overloaded. This work has to be done when the primary arrival rate decreases as shown in Figure 4.2 on Page 79. Then the percentage of arrivals due to retrials $P(\text{recaller}, t)$ given by Equation (4.7) on Page 78 increases.

Therefore, with high probabilities of retrial the total arrival rate stays high for a longer time before decreasing. This leads to the shift of the local minimum of the number of customers in the system depicted in Figure 4.11 on Page 86 and the minimum of the waiting time function shown in Figure 4.17.

4.1.5.3 Influence of the Parameters on the Time-Dependent Probability of Being Served

Next the probability of being served calculated in Equation (4.4) on Page 77 is investigated. As the sum of the probability of reneging (4.5) and the probability of moving into the orbit (4.6) is complementary to the probability of being served, these probabilities are not presented.

Again we start by studying the influence of the service rates followed by the influence of the number of agents. Afterwards we investigate the impact of the retrial parameters. The probability of being served is independent of the abandonment rate ν[9]. Therefore the influence of the abandonment rates is not studied. We assume the parameters of Table 4.1 on Page 79.

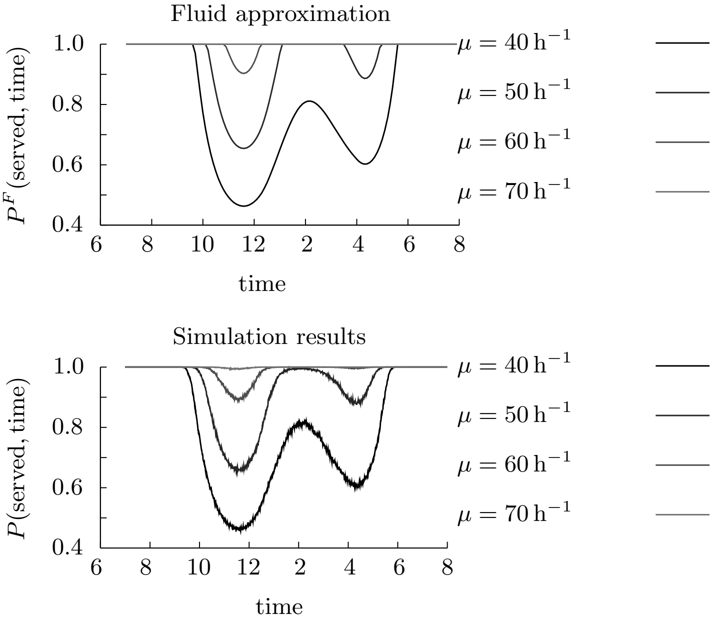

Fig. 4.18. Influence of the service rate μ on the time-dependent probability of being served

In Figure 4.18 the probability of being served for different mean service times μ^{-1} is depicted. If the service rate is high, i.e. $\mu = 70\,\text{h}^{-1}$, all customers are served. Therefore the probability of being served is one both in the simulation and the approximation. Lower service rates lead to lower probabilities of being served. Additionally, if a lot of customers retry after a short sojourn time in the orbit, the probability of being served decreases.

The results of the approximation fit the simulation results well. Furthermore, the graphs of Figure 4.18 look very similar to the horizontally mirrored

[9] See Figure 3.18 on Page 50.

graphs of the time-dependent waiting time in Figure 4.13 on Page 88. Therefore it might suffice to investigate just the waiting time to measure the performance experienced by customers, if the influence of the other parameters on the probability of being served is similar to the influence on the waiting time or the parameters have no influence on the probability of being served. In the first case the other graphs will also look like horizontally mirrored graphs of the waiting time. Consequently, if the waiting times are low a lot of customers are served and the probability of being served is high.

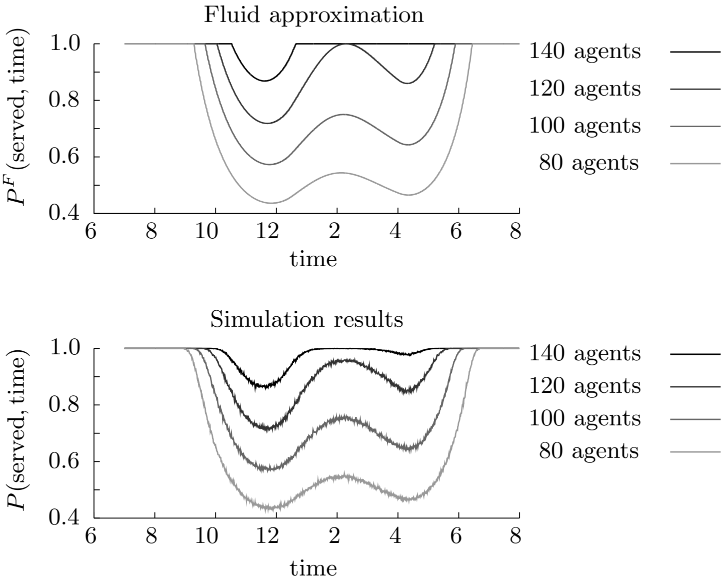

Fig. 4.19. Influence of the number of active agents $N(t)$ on the time-dependent probability of being served

A comparison of the simulation and fluid results for a varying number of agents in Figure 4.19 underlines that the approximation is accurate. The number of agents is constant for the whole day. We compare the results for 140, 120, 100, and 80 agents on duty. The other parameters are given in Table 4.1 on Page 79.

If the probability of being served is almost one, e.g. at 2 pm with 120 agents on duty, the approximation overestimates the probability of being served, because the fluid approximation neglects randomness. This result is consistent with the underestimation of the waiting time and number of customers in the system, when the traffic intensity is almost one. The probability of being served increases, if more agents are on duty, because fewer customers have to wait at all.

Equivalent to the Figure 4.18 on Page 92, the graphs of Figure 4.19 seem to be structurally equal to the horizontally mirrored graphs of Figure 4.14

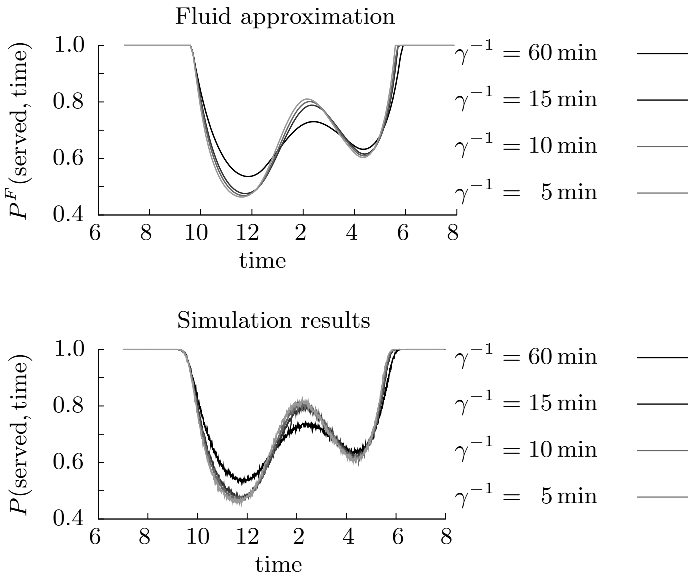

Fig. 4.20. Influence of the mean time to redial γ^{-1} on the probability of being served

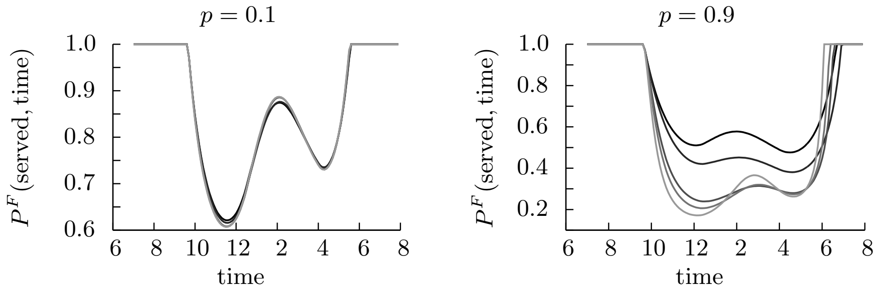

Fig. 4.21. Influence of the mean time to redial γ^{-1} and the probability of retrial on the probability of being served for different retrial probabilities p

on Page 89. This observation stresses the implications for the performance analysis which were already formulated as the influence of the service rates on the probability of being served has been investigated.

The influence of the mean time to recalling on the probability of being served presented in Figures 4.20 and 4.21 is closely connected to the probability of redialling p. Therefore, two figures are presented. The legends of the first figure and the second figure are the same.

If the probability of redialling approaches zero, the probability of being served seems to be nearly independent of the mean time to redial as shown in the first graph of Figure 4.21. If the probability of redialling approaches one in the second graph of Figure 4.21 the influence becomes remarkably strong.

The reason for this phenomenon is that with a redialling probability of one the customers will circulate from system to orbit and back into the system until they are finally served. Therefore, the faster the customers recall the more customers are in the system. This causes more customers to abandon after waiting a certain time. All customers who abandon will move into the orbit and retry and so on and so forth. If the customers remain in the orbit for a longer time, the congestion of the system will be spread into later periods of the day. During these later periods of the working day, fewer primary customers arrive, so that the workload decreases. Consequently, fewer customers leave the system because of impatience, so that the fraction of customers who are served later on from those who abandon increases.

As for the previous graphs, the graphs of the time-dependent probability of being served in Figure 4.20 are similar to the horizontally mirrored graphs in Figure 4.16. Therefore, the effects of different mean times to retrial can be explained in an equivalent way.

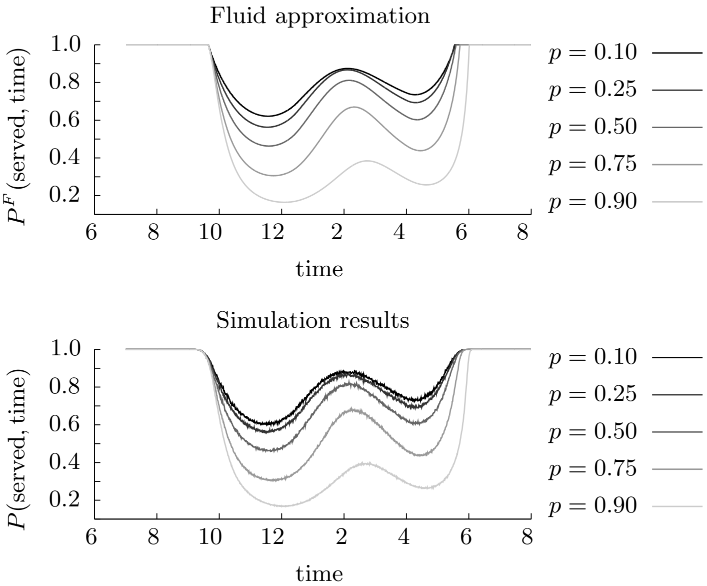

Fig. 4.22. Influence of the probabilities of retrial on the time-dependent of being served

Finally, in Figure 4.22 the influence of the probability of retrial on the probability of being served is presented. As in Figures 4.20 and 4.21 we assume that 100 agents are working.

The influence of the probability of retrial on the probability of being served is strong. This is in line with the observations of Figure 4.21. If the probability of retrial increases and the system is temporarily overloaded the performance of the system may become worse[10]. Therefore, it is important to estimate the retrial parameters, particularly the probability of retrial, from historical data before staffing decisions are made.

4.1.5.4 Influence of the Parameters on the Time-Dependent Utilisation

The time-dependent utilisation of the agents calculated by Equation (4.9) on Page 78 leads to a different point of view on the system. If agents are utilised too highly for long, this might lead to so-called burn-out syndromes, which might cause inefficiencies and more agents slow down[11].

Firstly, we study the influence of the service rate and the number of agents on the utilisation. We assume the parameters mentioned in Table 4.1 on Page 79. Because the utilisation of agents is independent of the abandonment rate[12], no pictures are presented. Finally, we analyse the impact of the retrial parameters on the utilisation. Differing from Table 4.1 we assume 150 agents on duty to stress the effect of varying parameters, because otherwise all agents would be fully utilised nearly all the time, i.e., the time-dependent utilisation would be one.

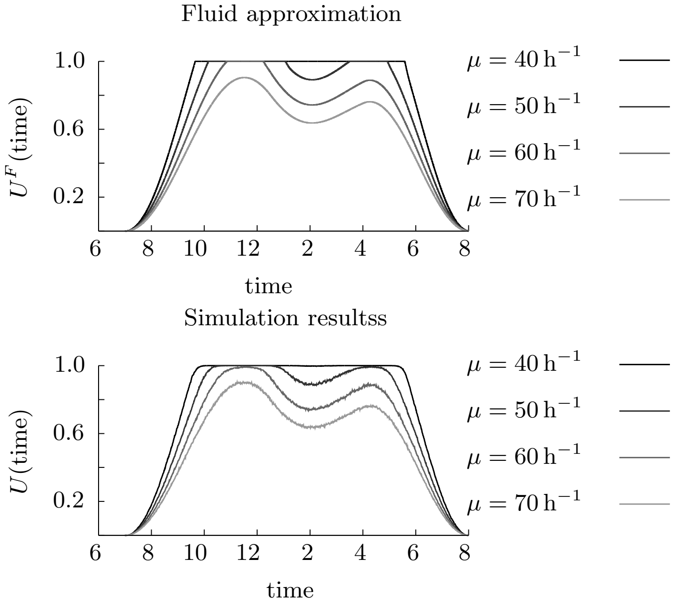

Fig. 4.23. Influence of the service rate μ on the utilisation of agents

[10] See also Figure 4.17 on Page 91.

[11] See Lüde and Nerlich (2002); Taylor et al. (2002); Witness Systems (2004), and references therein.

[12] See Figure 3.22 on Page 53.

In Figure 4.23 the fluid approximation overestimates the utilisation of agents if the workload approaches one, e.g., for $\mu = 60\,\mathrm{h}^{-1}$ during lunch time. This observation is in line with earlier results. If the contact center is over- or underloaded the approximation is accurate. Higher service rates give rise to lower utilisation. Consequently, more customers are served and fewer retry. The utilisation cannot exceed one. Therefore, a higher service rate than $\mu = 40\,\mathrm{h}^{-1}$ would lead in this example to minor changes in the utilisation.

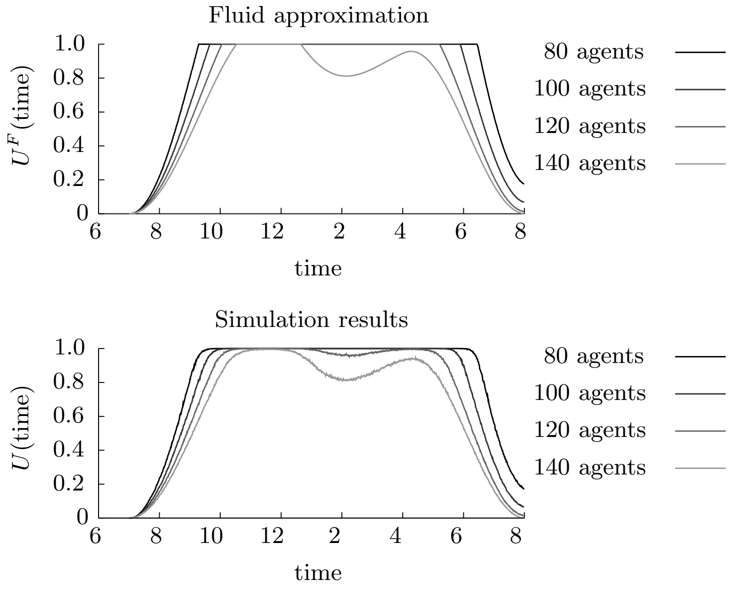

Fig. 4.24. Influence of the number of active servers on the utilisation of agents

Similarly to the service rate, more working agents lead to lower utilisation as shown in Figure 4.24. If the workload approaches one, the approximation again overestimates the utilisation obtained by the simulation for the same parameters of the system. The graphs underline earlier observations about the impact of the number of agents on the performance.

Finally, the influence of the retrial parameter on the utilisation of agents is analysed. Amazingly, the mean time to retrial seems to have no effect on the utilisation of agents as shown in Figure 4.25. If the utilisation approaches one, i.e., all agents are working, the observation is clear. If otherwise some agents are not busy, all customers are served and no-one will abandon. Therefore, no customer has to redial. Consequently, the mean time to retrial has almost no influence on the utilisation.

A little effect of the mean time to retrial on the utilisation is observable, if the contact center was overloaded for some time, such that a lot of customers wait in the orbit in order to retry. If the mean time to retrial is quite long, e.g., $\gamma^{-1} \geq 60$ min, some customers will retry when some agents are already free. Then the utilisation increases a little. The same argumentation holds for the

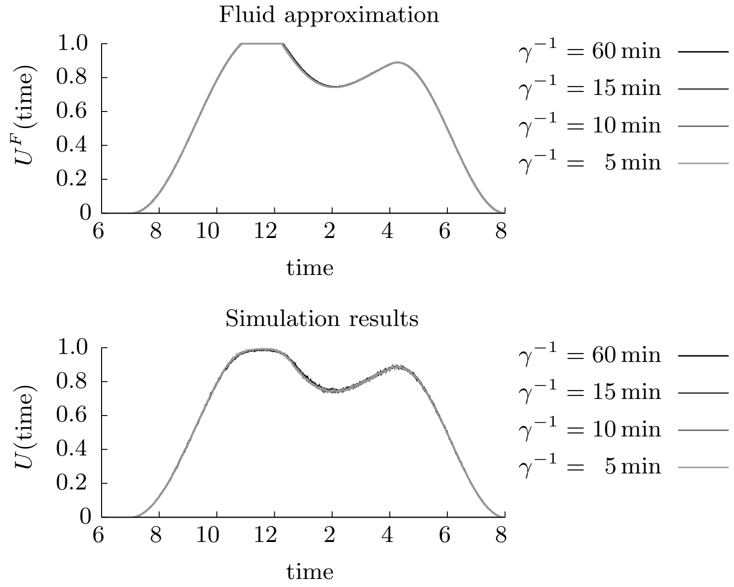

Fig. 4.25. Influence of the mean time to redialling γ^{-1} on the utilisation of agents

impact of the probability of retrial. That is why we do not present a picture of the impact on the probability of retrial.

After having investigated all these performance measures it becomes clear that only few of them are needed to analyse and evaluate a contact center model. As the waiting times and the probabilities of being served have nearly the same shape, it suffices to compare the effects of several parameters on the waiting time. Apart from the effects of different abandonment and redialling rates, the influence of the other parameters on the waiting times results in an influence of the parameters on the other performance measures. The same holds for the utilisation of agents. Consequently, in an optimisation of server staffing, considering one performance measure will do. For this reason, in Section 4.2 fewer performance measures will be compared.

4.1.5.5 Aggregated Technical Performance Measures

To complete the analysis of the fluid model, aggregated performance measures should also be considered. The waiting time, the probability of being served and the utilisation are aggregated over one working day starting at 7 am and ending at 8 pm in the same manner as in Subsection 3.2.3 on Pages 40–42.

As the time-dependent performance measure except for the waiting time have been insensitive of the abandonment rate, the abandonment rate will not have any effect on the aggregated performance measures. Therefore, effects due to changes in the abandonment rate on the probability of being served and the utilisation are not considered. For the same reason no figures on the influence of the retrial parameters on the aggregated utilisation are presented. The arrival rate is assumed as shown in Figure 4.2 with parameters given in (4.12). Furthermore, the parameter combinations is given in Table 4.2.

$\mu(t)$	$N(t)$	$\nu(t)$	$\gamma(t)$	p
variable	variable	$120\,\mathrm{h}^{-1}$	$0.5\,\mathrm{h}^{-1}$	0.5
$60\,\mathrm{h}^{-1}$	variable	variable	$0.5\,\mathrm{h}^{-1}$	0.5
$60\,\mathrm{h}^{-1}$	100	$120\,\mathrm{h}^{-1}$	variable	variable

Table 4.2. Parameters for the analysis of aggregated performance measures

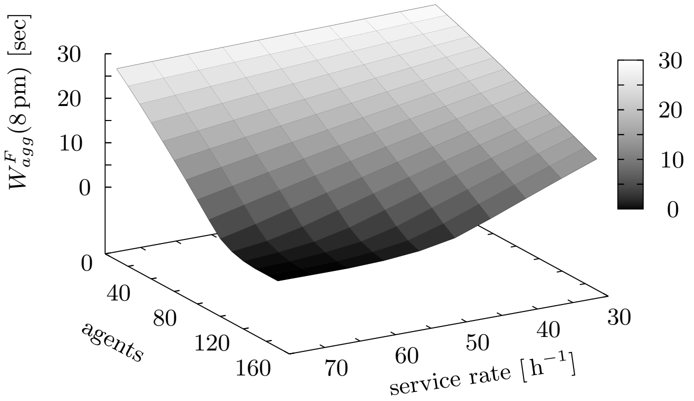

Fig. 4.26. Influence of the service rate μ and the number of active agents $N(t)$ on the aggregated waiting time $W^F_{agg}(T)$

In Figure 4.26 the aggregated mean waiting time $W^F_{agg}(T)$ as a function of the number of agents $N(t)$ staffed for the whole day and the service rate is depicted. The more agents who are on duty and the faster these agents work, the smaller the aggregated waiting time of customers is. Whenever few agents are working many customers will abandon. Therefore, the aggregated waiting time does not exceed thirty seconds. The influence of the number of agents and service rate on the aggregated waiting time corresponds to the influence on the time-dependent waiting time.

If we compare Figures 4.13 and 4.14 on Pages 88f. with Figure 4.26, we observe that the aggregated waiting time averages the time-dependent waiting times. Although the time-dependent waiting time is temporarily high, e.g., $\mu = 60\,\mathrm{h}^{-1}$ in Figure 4.13 and $N(t) = 80$ in Figure 4.14, the aggregated waiting time is smaller than 20 seconds.

In Figure 4.27 the number of agents $N(t)$ and the abandonment rate ν are varied. If only few agents are staffed and the customers are very patient, the waiting time is almost six minutes. If the abandonment rate ν grows, i.e., the customers are more impatient, the waiting time decreases very quickly. The same holds for a growing number of agents $N(t)$ on duty. These effects

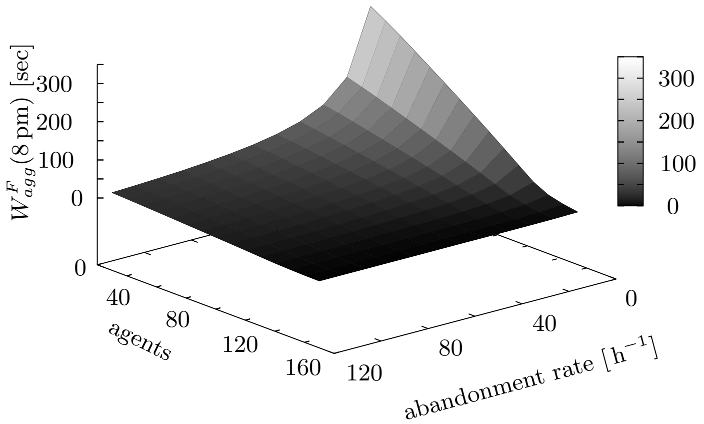

Fig. 4.27. Influence of the abandonment rate ν and the number of active agents $N(t)$ on the aggregated waiting time $W^F_{agg}(T)$

are in line with the effects on the time-dependent waiting time presented in Figures 4.14 and 4.15 on Page 89. Although the aggregated waiting time is quite long, if customers are very patient, the aggregated waiting time in the fluid approach will never exceed the waiting time limit ν^{-1} equivalently to the time-dependent waiting time $W^F_S(t)$.

The reason for this observation is that in the fluid model the amount of waiting customers is processed by the pumps representing the abandonment of customers. These pumps work with rate ν for each waiting customer. Therefore, every amount of customers in the system will be pumped out at once, if the waiting time limit ν is reached. In other words, in the fluid model the waiting time limit determines the maximum waiting time.

The influence of the retrial parameters γ and p on the aggregated waiting time is depicted in Figure 4.28. Obviously, higher probabilities of retrial p lead to a stronger influence of the retrial rate γ on the aggregated waiting time $W^F_{agg}(t)$. If the retrial rate increases from zero to four customers per hour, i.e., the mean time in orbit decreases from infinity to fifteen minutes, the aggregated waiting time increases in an almost concave manner.

Furthermore, the aggregated waiting time seems to be a convex function of the probability of retrial, if the retrial rate is greater than one. This observation stresses earlier results, that the influence of the probability of retrial on the performance of the contact center is much stronger[13] than the influence of the retrial rate. Therefore, the probability of retrial should be carefully estimated from historical data. The influence of the retrial rate grows, if the probability of retrial increases.

[13] Compare the Figures 4.16 and 4.17 on Page 91.

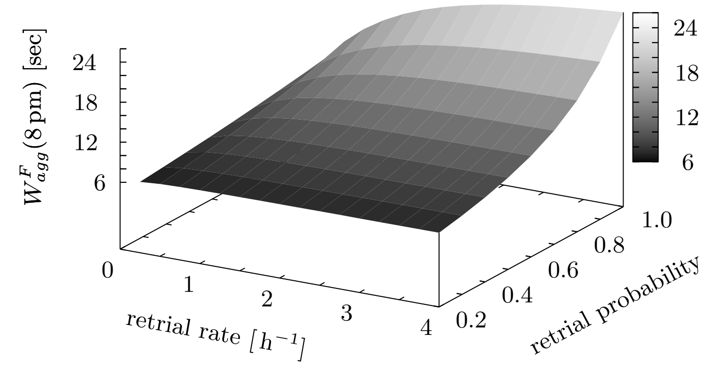

Fig. 4.28. Influence of the retrial rate γ and the probability of retrial p on the aggregated waiting time $W_{agg}^F(T)$

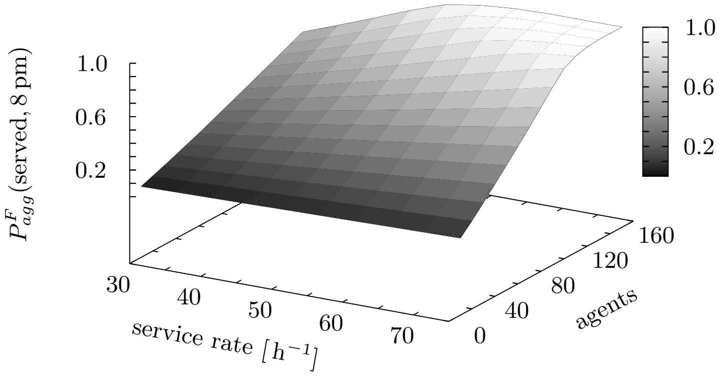

Fig. 4.29. Influence of the service rate μ and the number of active agents $N(t)$ on the aggregated probability of being served $P_{agg}^F(\text{served}, T)$

Next, the aggregated probability of being served $P_{agg}^F(\text{served}, T)$, calculated by aggregating the time-dependent probabilities given by Equation (4.4), is investigated. The time-dependent probabilities were aggregated according to Equation (3.32) on Page 40. The parameters of the model are as before.

In Figure 4.29 the aggregated probability of being served for all customers having left is shown as a function of the number of staffed agents $N(t)$ and the service rate μ. Equivalently to Figures 4.18 and 4.19 on Pages 92f. the probability of being served increases if the service rate or the number of agents on duty increases. The reaction on a change in the number of agents seems to be stronger than the response on a growing service rate.

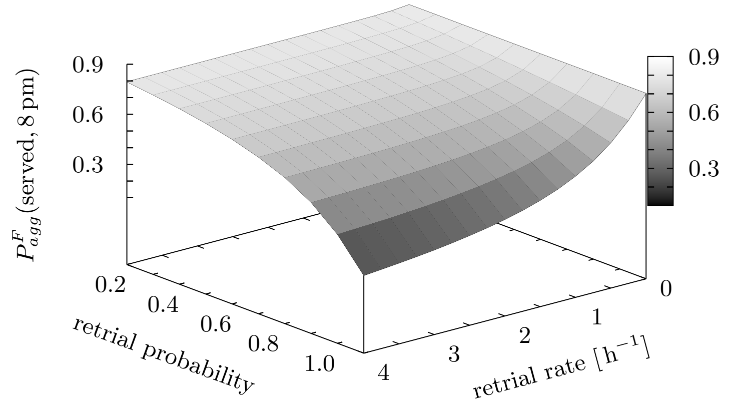

Fig. 4.30. Influence of the retrial rate γ and the probability of retrial p on the aggregated probability of being served $P_{agg}^F(\text{served}, T)$

The impact of the retrial parameters on the probability of being served for all customers entering the system is presented in Figure 4.30. If the probability of retrial p increases, the aggregated probability of being served of all customers who have left decreases. This observation conforms to the observation of Figure 4.22 on Page 95. The reason for this decrease is that primary calls and recalls cannot be distinguished after having entered the system. A retrial increases the number of customers in the system. If the number of customers in the system exceeds the number of the agents, some customers will abandon. These customers are not served in the current attempt. For this performance measure it does not matter whether the customer is finally served in a later attempt. This aggregated probability of being served just shows whether an arriving customer is served in his current attempt.

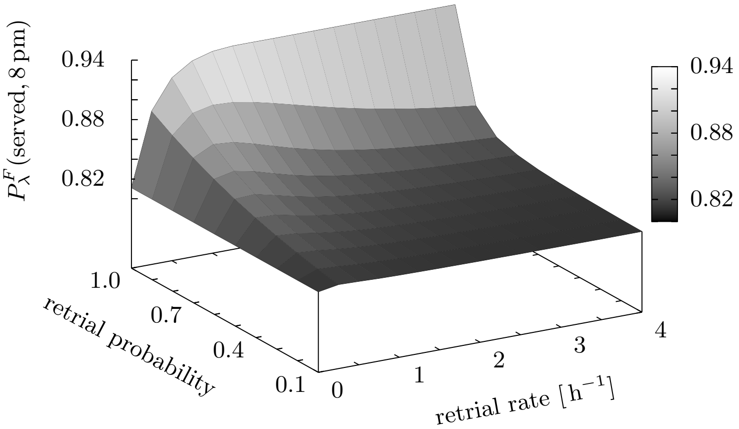

Fig. 4.31. Influence of the retrial rate γ and the probability of retrial p on the aggregated probability of finally being served $P_{\lambda}^F(\text{served}, T)$

Besides the probability of all customers leaving the system being served, we have also derived a performance measure, which represents the percentage of customers served finally $P_\lambda^F(T)$, i.e., the probability of all primary attempts being served independent of the number of retrials needed to get service. This performance measure is given in Equation (3.34) on Page 41. In Figure 4.31 the probability of finally being served increases if the probability of retrial increases, because customers who have abandoned can be served in a later attempt.

This performance measure is useful if primary calls and recalls can be distinguished. If we assume that customers get annoyed by their unavailing attempts, the first aggregated probability of being served is more reasonable to measure the performance of the system. For most contact centers the second probability of finally being served is indeterminable, because mostly primary calls and retrials cannot be distinguished by the ACD unit.

Amazingly, the minimum of the first probability corresponds to the maximum of the second probability, if the probability of retrial p is one and the retrial rate γ increases. In this case all customers who have abandoned will retry, i.e., all customers will be served at some time. Consequently the second probability of finally being served will be one, if both the system and the orbit become empty. Contrarily, the aggregated probability of being served is the ratio of the number of customers being served and the sum of the arrival rates from outside and the orbit. If all customers who abandon will retry after a short sojourn time in the orbit, a lot of customers will recall when the primary arrival rate is also high. Therefore the congestion in the system rises, more customers will abandon and fewer customers will be served. As long as the primary arrival rate increases, the congestion will build up and it will be carried on into later periods of the day. Therefore, the aggregated probability of being served decreases because of the high number of useless attempts.

Finally, the impact of the number of agents $N(t)$ and the service rate μ on the aggregated utilisation $U_{agg}^F(t)$ is presented in Figure 4.32. The observations of this Figure underline earlier results in Figures 4.23 and 4.24 on Pages 96f. Although the temporary utilisation might be high, e.g., $N(t) = 80$ in Figure 4.24 and $\mu(t) = 60$ in Figure 4.23, the aggregated utilisation remains acceptable below 90%. If the number of agents is increased by 20 agents the utilisation reduces to 84%.

In this section we have observed the averaging effect of aggregation as well as some special properties of the fluid approach with respect to the waiting time and abandonment. Furthermore, in retrial queues a higher percentage of customers might actually be served than might be predicted by the aggregated probability of all customers who have left the system being served associated with the X/Y service level performance measure which appears to be used occasionally in reality.

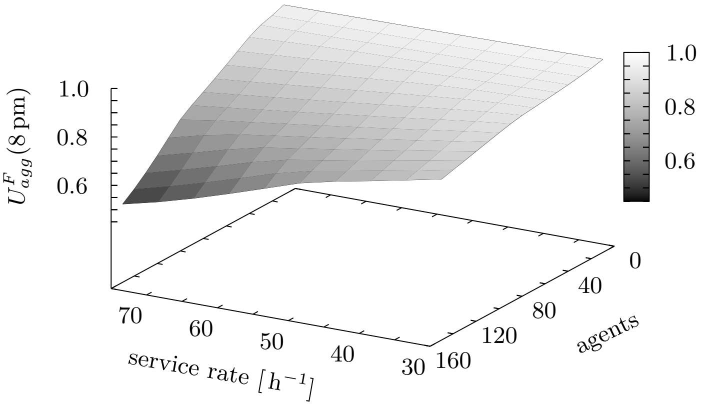

Fig. 4.32. Influence of the service rate μ and the number of active agents $N(t)$ on the aggregated utilisation $U^F_{agg}(T)$

4.1.5.6 Economical Performance Measures

After considering the technical performance measures, the economical aspect should also be investigated. We study the impact of the parameters on the profit function given in Equation (4.10) on Page 78. We assume the same arrival rate function as in Figure 4.2 on Page 79. The parameters of the contact center are given in Table 4.2 on Page 99. Furthermore, the hourly wage w of an agent is 10€ and the costs for an occupied trunk ℓ is 6€ per hour. The average revenue r gained by serving a customer is 0.50€.

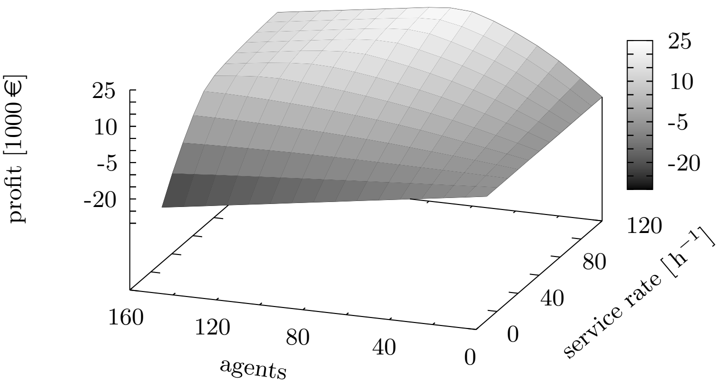

Fig. 4.33. Influence of the service rate μ and the number of active agents $N(t)$ on the profit

In Figure 4.33 the influence of the service rate μ and the number of staffed agents $N(t)$ is presented, where the number of agents is supposed to be constant for the whole day. The profit of the contact center increases, if the number of agents or the service rate grows until a maximum level is reached. Afterwards the profit decreases again. The maximum profit can be achieved for different combinations of the number of agents and the service rate.

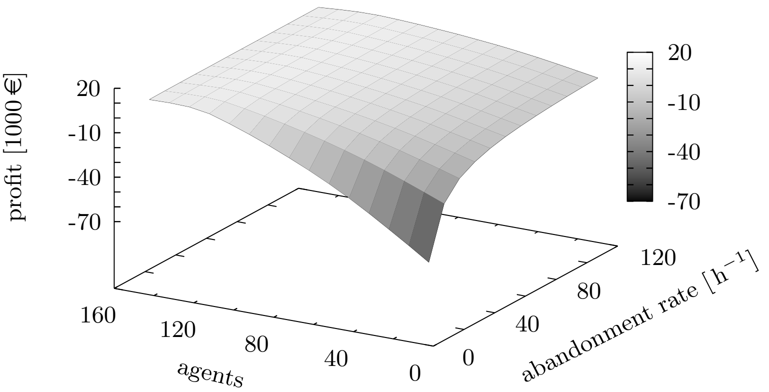

Fig. 4.34. Influence of the abandonment rate ν and the number of active agents $N(t)$ on the profit

In Figure 4.34 the profit is a function of the number of agents and the abandonment rate. Because of the special revenue and cost parameters the daily profit is quite small and even a loss is possible. Amazingly, the profit function has a minimum if almost no agents are staffed and the customers are very patient. If few agents are staffed and the mean abandonment rate is low, many customers wait in the system for attendance. Each customer occupies a telephone line. Consequently, the cost of occupied trunks will increase and finally exceed the revenue gained by serving customers. The more agents who are staffed and the more impatient customers become, the more money is made out of service, because the costs for occupied trunks decrease more quickly than the costs for staffed agents increase. If the abandonment increases further, the profit becomes independent of the abandonment rate. This observation underlines the relationship between the aggregated probability of being served and the profit function.

The retrial rate γ and the retrial probability p seem to have a minor influence on the profit, as shown in Figure 4.35 with the parameters given in Table 4.2 on Page 99. However, in the case of very high probabilities of retrial and high retrial rates the profit decreases. Then a high percentage of customers recalls after a very short time in orbit. This behaviour raises the number of customers in the system[14]. Similar to the previous example about

[14] See Figures 4.9 on Page 84 and 4.11 on Page 86.

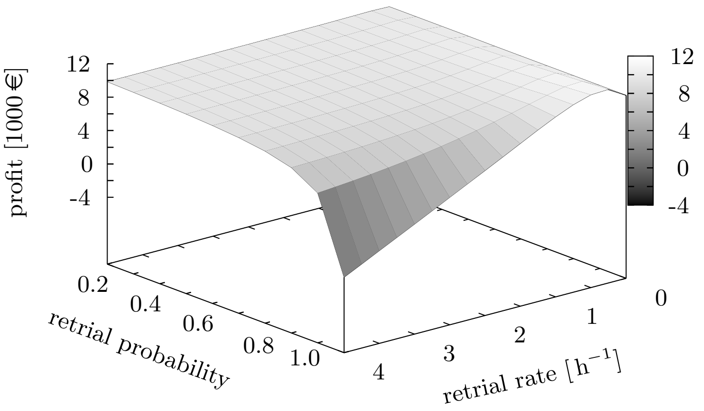

Fig. 4.35. Influence of the retrial rate γ and the probability of retrial p on the profit

the abandonment rate, the number of occupied trunks increases and so do the costs for these trunks. Furthermore, some customers may even get lost after starting several attempts, so that these customers caused a lot of costs but no revenue.

If the probability of retrial decreases, more customers abandon and do not retry. Then fewer trunks are occupied and the costs for these trunks decrease. If the retrial rate decreases, customers stay in the orbit a longer time before starting a new attempt. In this case some customers recall when the workload of the contact center has already lessened, so that these retrials can be served immediately. The number of customers in the system during periods of high load decreases while the number of customers in the system during periods of low load increases a little as depicted in Figure 4.9 on Page 84. Therefore the revenue gained by served customers grows and the costs for occupied trunks during periods of high load shrinks.

4.1.5.7 The Variance of the Number of Customers in the System

In this section we compare the variances of the diffusion approximation to simulation results. Furthermore, we investigate the impact of the different parameters on the variances of the diffusion processes. We assume that the arrival rate is given as depicted in Figure 4.2 on Page 79 with parameters (4.12). The other parameters for the model are given in Table 4.3.

The simulation results were generated by the extended simulation tool mentioned before[15]. The simulated variances for the processes of the number of customers in the system were calculated based on the results of 500 repetitions.

The analytical variances and covariances were determined by numerically solving the extended initial value problem consisting of the initial value prob-

[15] See Page 79

Figures	$\mu(t)$	$N(t)$	$\nu(t)$	$\gamma(t)$	p
4.36, 4.41	variable	150	$120\,\mathrm{h}^{-1}$	$15\,\mathrm{h}^{-1}$	0.5
4.37, 4.42	$60\,\mathrm{h}^{-1}$	variable	$120\,\mathrm{h}^{-1}$	$0.5\,\mathrm{h}^{-1}$	0.5
4.38, 4.43	$60\,\mathrm{h}^{-1}$	150	variable	$15\,\mathrm{h}^{-1}$	0.5
4.39, 4.44	$60\,\mathrm{h}^{-1}$	100	$120\,\mathrm{h}^{-1}$	variable	0.5
4.40, 4.45	$60\,\mathrm{h}^{-1}$	100	$120\,\mathrm{h}^{-1}$	$15\,\mathrm{h}^{-1}$	variable

Table 4.3. Parameters for the numerical analysis and comparison of the diffusion approximation and simulation results

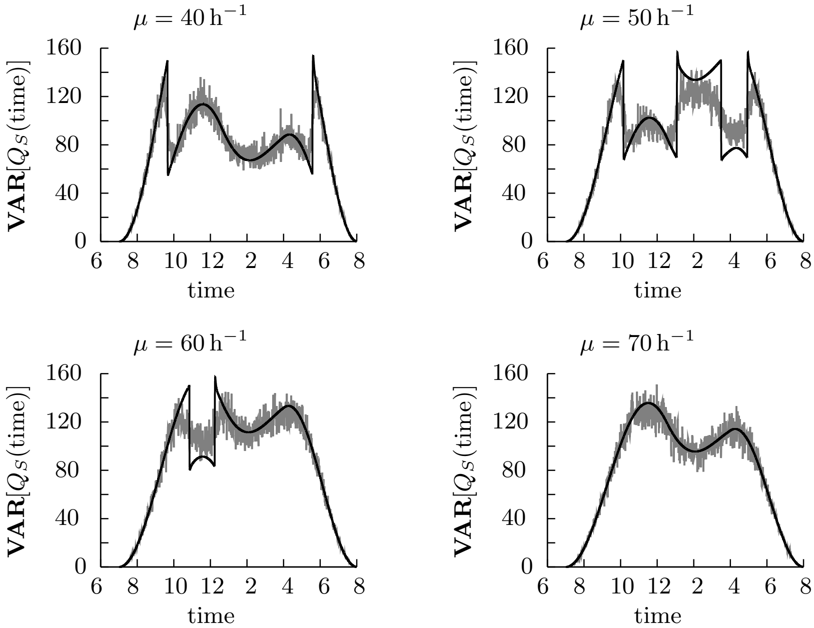

Fig. 4.36. Comparison of the approximation (black line) for the variance of the number of customers in the system to the simulation results (gray line) for different service rates μ

lem of the fluid approach, the differential Equations 4.2 on Page 76 for the variances and covariances and the initial conditions

$$\begin{gathered}\mathbf{VAR}\big[Q_S^D(t_0)\big] = 0,\ \mathbf{VAR}\big[Q_O^D(t_0)\big] = 0,\\ \text{and } \mathbf{COV}\big[Q_S^D(t_0), Q_O^D(t_0)\big] = 0.\end{gathered} \tag{4.13}$$

First of all, in Figure 4.36 the diffusion approximation of the variance of the number of customers (black line) is compared to the simulation results of these variances (gray line) for four different values of the service rate μ. The variance changes drastically if the number of customers in the system exceeds the capacity of the contact center, i.e., the number of available agents[16]. If

[16] Compare Figure 4.3 on Page 80.

the number of customers in the system remains below the maximum capacity of the contact center in the fourth picture of Figure 4.36, such that all customers are served and no-one has to abandon, the variance does not change so drastically.

In the first picture the variance reduces from 150 to about 50 which is one third. In the second picture the variance reduces from 150 to about 60, and in the third from 150 to about 75. If the number of customers decreases, the variance raises again. If the service rate is small, many customers abandon and the number of customers in the system stays above the number of available servers nearly all the time. Then the variance is mainly driven by the arrivals and the abandonments. In this example customers are very impatient. Their mean waiting time limit ν^{-1} is thirty seconds. Therefore, more customers abandon during the same short time interval than are served. Furthermore, half of all abandoning customers retry after a very short time, which increases the total arrivals. The process of the number of customers in the system becomes less variable. The variability is transferred to the orbit[17]. Amazingly, the reduction of the variances always approximately equals the ratio of the service and the abandonment rate $\frac{\mu}{\nu}$.

If the contact center is underloaded, e.g. in the fourth picture, the variance is governed by the primary arrival and the service process. The other processes do not influence the variance. Hence, no jump discontinuities arise.

Next we study the influence of the number of active agents in Figure 4.37. In Figure 4.37 the same phenomenon is visible as in the previous Figure 4.36. Therefore, a higher number of agents leads to more variability, but if more agents are staffed such that the number of customers does not cross the level of staffed agents[18], the smoother the curve is. Furthermore the simulation results for the variances are well approximated by the numerical solution of the differential equations for the variances generated by the diffusion approach.

In Figure 4.38 the effect observed in Figure 4.36 if the system becomes overloaded is investigated more deeply. In Figure 4.36 the decrease in the variance is approximately $1 - \frac{\mu}{\nu}$. Therefore, the question arises whether this can be considered as a rule or just a coincidence.

In the first picture the abandonment rate is less than the service rate. The ratio $\frac{\mu}{\nu}$ is 1.5. If the number of customers exceeds one hundred, the variance should increase instantaneously to 150, if the proposal holds. In the second picture the abandonment rate and the service rate are equal. Therefore, the variance should not change drastically at one hundred customers. In the third and fourth picture the abandonment rate exceeds the service rate with $\frac{\mu}{\nu}$ being $\frac{3}{4}$ and $\frac{3}{5}$, respectively. That is why the variance decreases after reaching the number of agents. Obviously, the proposed rule applies to all four pictures.

[17] See Figure 4.41 on Page 112.

[18] Compare Figure 4.5 on Page 82.

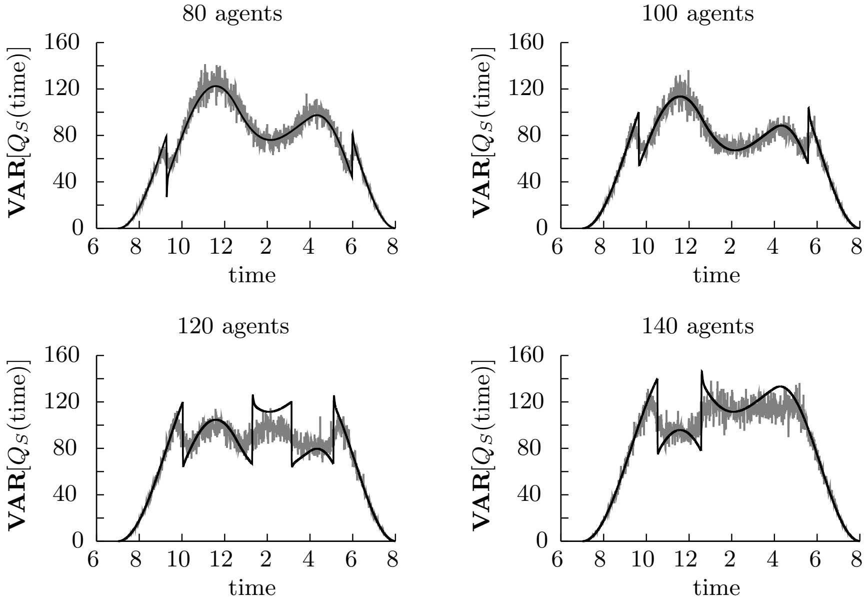

Fig. 4.37. Comparison of the approximation (black line) for the variances of the number of customers in the system to the simulation results (gray line) for different numbers of active agents $N(t)$

Hence, the variance depends on all parameters. The number of agents determines the point where the variance may change drastically and the ratio of the service and abandonment rate estimates the magnitude of the change.

In Figure 4.39 the influence of different mean times to redial γ^{-1} is depicted. Contrary to the other parameters, the mean time to redial has a very small influence. If customers wait for a long time in the orbit the variance is smaller than in the case of quickly recalling customers. The influence of the retrial rate is limited to the period when the contact center is overloaded. The approximation of the simulated variances is remarkably good.

Finally, the influence of the probability of retrial is shown in Figure 4.40. Obviously, this probability has an enormous effect in the cases depicted[19]. If 10% of all customers are willing to recall, the graph almost equals the graph of the previous Figures 4.39. If 90% of all customers redial, the variance increases very strongly. Comparing the second picture of Figure 4.40 to the last picture of Figure 4.39 and the picture referring to $p = 0.1$, the variance for $p = 0.9$ is about three to four times as high as the other variances. That means, in a contact center with many redials, the stochasticity has a higher impact and should be considered carefully.

[19] Consider the different scalings in Figure 4.40.

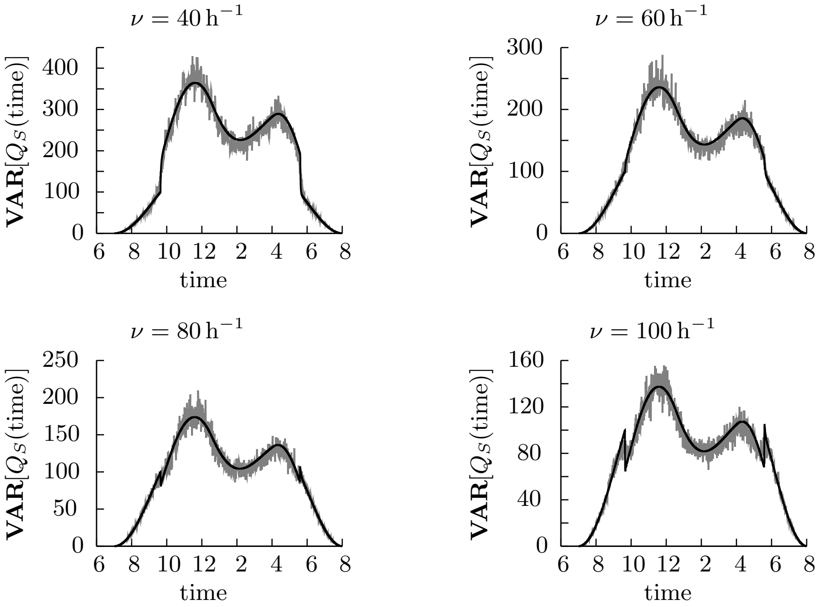

Fig. 4.38. Comparison of the approximation (black line) for the variance of the number of customers in the system to the simulation results (gray line) for different abandonment rates ν

4.1.5.8 The Variance of the Number of Customers in the Orbit

Next, the influence of the diverse parameters given in Table 4.3 on Page 107 on the variance of the process describing the amount of customers in the orbit is investigated. As in the previous section we assume time-dependent arrival rates and the initial conditions given in Equation (4.13) on Page 107. We use the same order, i.e., we start with investigating the influence of the service rates and end with studying the influence of the probabilities of retrial.

In Figure 4.41 the variance of the diffusion process describing the number of customers in the orbit is presented. Unlike Figure 4.36, in Figure 4.41 the variance of the number of customers in the orbit does not equal the average number of customers in the orbit depicted in Figure 4.4 on Page 81.

The variance decreases as the service rate increases, because fewer customers abandon and wait in the orbit to start another attempt. The diffusion approximation is quite accurate in the case of low service rates, $\mu = 40\,\mathrm{h}^{-1}$, in the first picture of Figure 4.41. If the service rate is $\mu = 60\,\mathrm{h}^{-1}$ in the third picture, the variance of the simulation is overestimated by the numerical results of the diffusion approximation. If the service rate increases further, as shown in the fourth picture, only the simulation determines a variance for the number of customers in the system, although this variance is quite small.

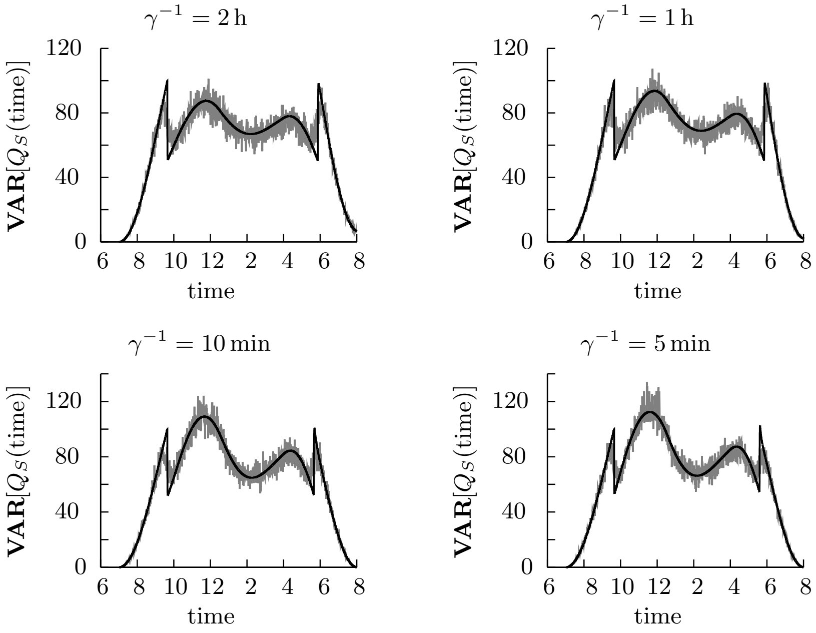

Fig. 4.39. Comparison of the approximation (black line) for the variance of the number of customers in the system to simulation results (gray line) for different mean times to redial γ^{-1}

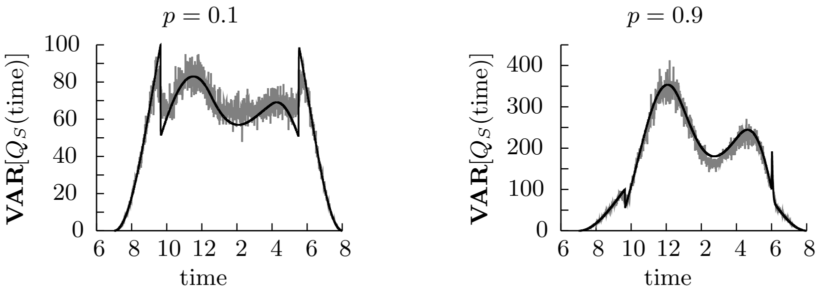

Fig. 4.40. Comparison of the approximation (black) for the variances of the number of customers in the system to the simulation results (gray) for different redialling probabilities p

These two effects just mentioned are due to the fact that the diffusion approximation is a refinement of the fluid model. In the fluid model we showed in Figure 4.4 on Page 81 that in the case of a service rate of $70\,\mathrm{h}^{-1}$ no customer abandons and moves into the orbit. Therefore, the variance of the diffusion process becomes zero as well. However, some customers will abandon and move into the orbit because of randomness in the underlying system. In the case of a service rate of $60\,\mathrm{h}^{-1}$, few customers move into the orbit and the

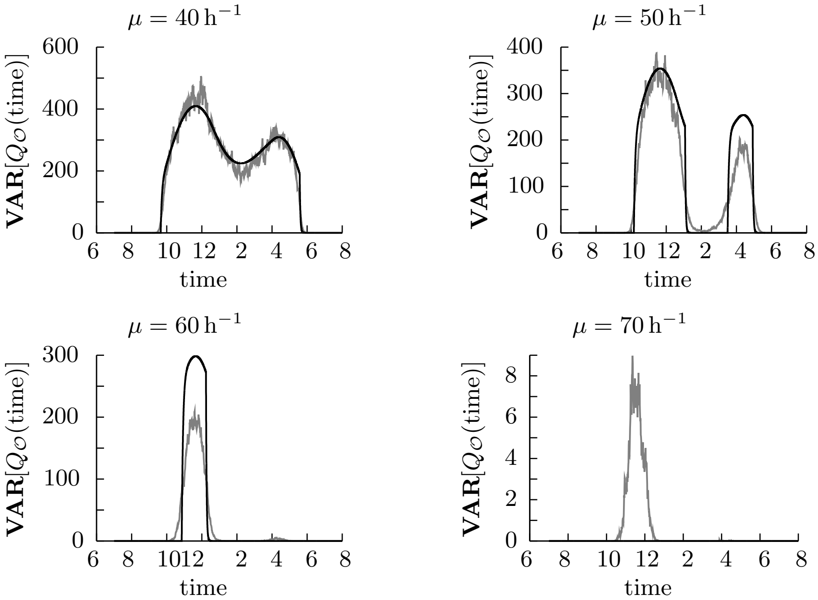

Fig. 4.41. Comparison of the approximation (black line) for the variance of the number of customers in the orbit to the simulation results (gray line) for different service rates μ

solution of the differential equation reacts much more strongly to changes in the parameters than the simulations, as already observed in earlier figures.

In Figure 4.42 the influence of the number of agents is presented. Unlike the previous Figure 4.41, the amplitude of the variance during the period of maximum load of the contact center remains almost unchanged if the number of agents on duty increases. In general, the simulation results for the variances are approximated accurately by the diffusion approach. The under- and overestimation of the variance during some periods can be explained similar to the varying service rates in the previous Figure 4.41.

As the abandonment rate has no influence on the number of customers in the orbit, one would suspect that the abandonment rate has no impact on the variance as well. This statement is confirmed in Figure 4.43. The approximation shows no influence of the abandonment rate. In the simulation results the variances changes a little in the fourth picture. However, these changes cannot be explained by the different abandonment rates but by the different simulation runs.

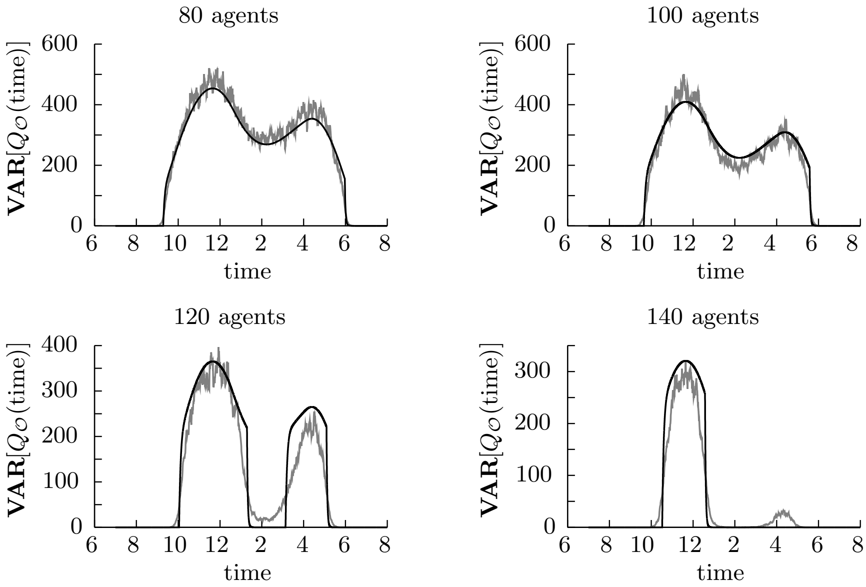

Fig. 4.42. Comparison of the approximation (black line) for the variance of the number of customers in the orbit to the simulation results (gray line) for different numbers of active agents $N(t)$

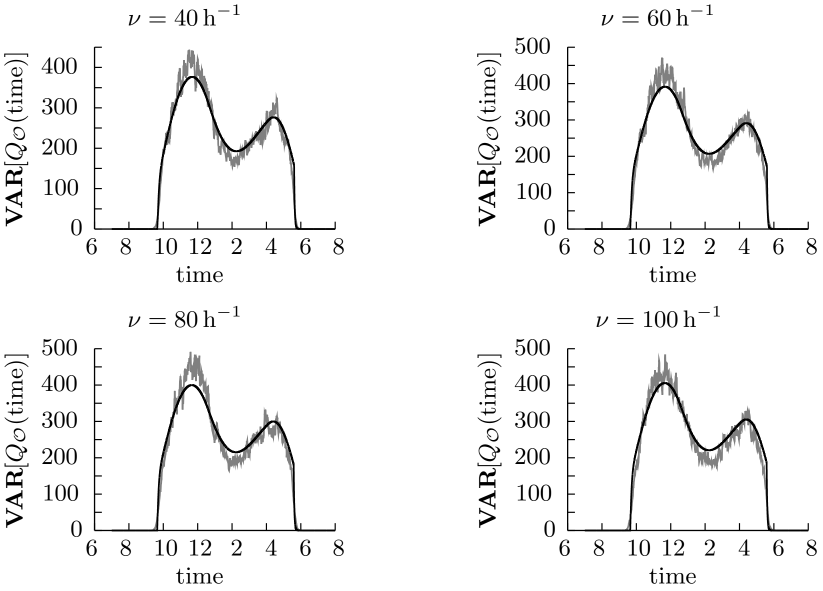

Fig. 4.43. Comparison of the approximation (black line) for the variance of the number of customers in the orbit to the simulation results (gray line) and the missing influence of the abandonment rates ν

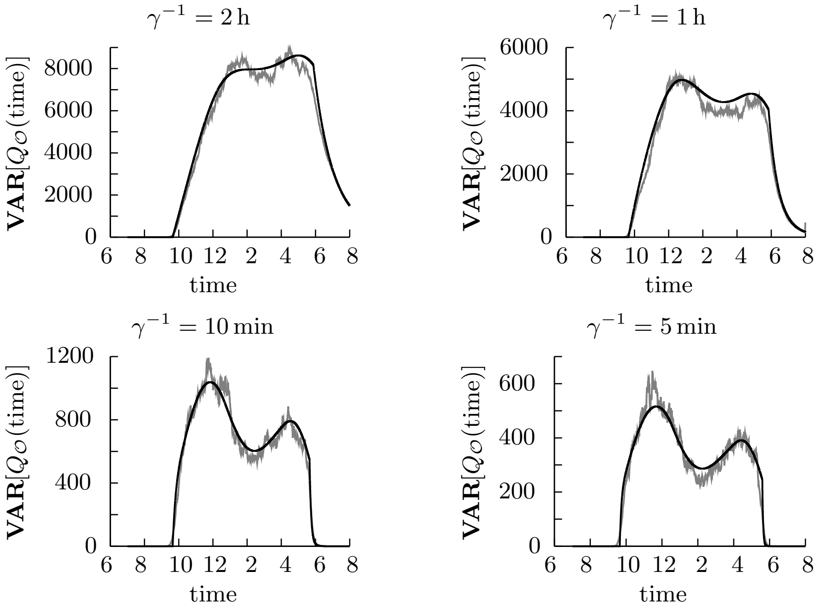

Fig. 4.44. Comparison of the approximation (black line) for the variance of the number of customers in the orbit to the simulation results (gray line) for different mean times to redial γ^{-1}

In Figure 4.44 the effects of different mean times to recall γ^{-1} are shown. Obviously, the time the customers are willing to wait in the orbit has an enormous effect on the variance of the process describing the number of customers in the orbit. If the mean time to redial γ^{-1} is two hours on average, the variance is almost eight times as high as in the case of a mean time to redial of ten minutes and almost 15 times as high as in the case of five minutes. The variances of the simulation are approximated well.

In the first picture of Figure 4.44 the maximum of the variance is reached during the late afternoon, while in the other three pictures the maximum is located in the time interval, when the maximum arrival rate is achieved[20]. The reason for this shift of the maximum is due to the shift of work into later periods because of the long mean time to redial.

Finally, the influence of the probability of retrial is presented in Figure 4.45. This probability has a huge influence on the variance as well. High retrial probabilities lead to extremely high variances. If the probability of retrial is very small, as depicted in the first picture of Figure 4.45, the variance is several times smaller than in the case of high probabilities. The approximated variance follows the simulated variance very well.

[20] Compare Figure 4.2 on Page 79.

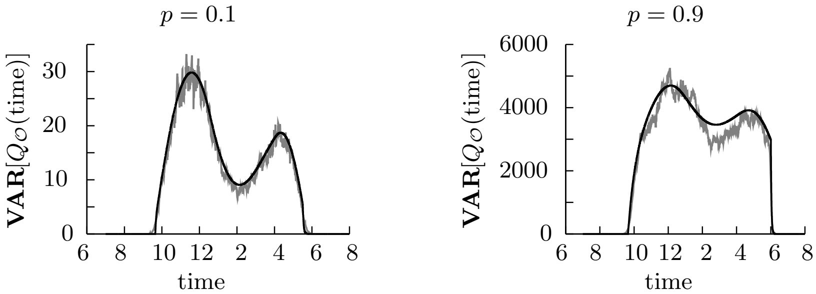

Fig. 4.45. Comparison of the approximation (black line) for the variance of the number of customers in the orbit to the simulation results (gray line) for different redialling probabilities p

4.1.5.9 The Correlation between the Number of Customers in the System and the Orbit

In order to understand the influence between the different queueing processes in the orbit and the system we investigate the correlation of these two processes. The correlation is a measure of the strength of the interdependencies of processes. If it is greater than zero, both processes are said to be positively correlated, i.e., an increase in the first process will lead to an increase in the second process by a proportional amount given by the correlation. If otherwise the correlation is smaller than zero, the processes are negatively correlated. In this case an increase of the first process will lead to a decrease of the second process. If the correlation is 1 or -1 the relationship between the processes is linear. The correlation of the diffusion processes of the number of customers in the system and the orbit is calculated by dividing the covariance by the square root of the product of the variances, i.e.,

$$\mathbf{Corr}\big[Q_S^D(t), Q_O^D(t)\big] = \frac{\mathbf{COV}\big[Q_S^D(t), Q_O^D(t)\big]}{\sqrt{\mathbf{VAR}\big[Q_S^D(t)\big]\,\mathbf{VAR}\big[Q_O^D(t)\big]}}. \tag{4.14}$$

As before, we analyse the influence of the different parameters on the correlation of these two processes. The covariances are calculated by numerically solving the initial value problem described on Page 107 for the variances containing the differential equation systems in (4.2) on Page 76 and (4.1) on Page 75 with the initial condition that all processes start at zero. For the different examples presented in the next figures, we assume the arrival rate function in Figure 4.2 on Page 79 and the parameters in Table 4.3 on Page 107.

At first glance Figures 4.46 through 4.50 may look a little strange. Especially the spikes at the beginning of each curve may cause astonishment. These spikes are due to the fact that in the fluid model the number of customers suddenly exceeds the number of available agents. At this time a small amount of customers abandons instantaneously and moves to the orbit. This is an effect resulting from the fluid approach, which assumes that the customers are infinitely divisible. Therefore, the number of customers increases proportionally to the number of customers in the system and the correlation of the number of customers in the system and the orbit is one.

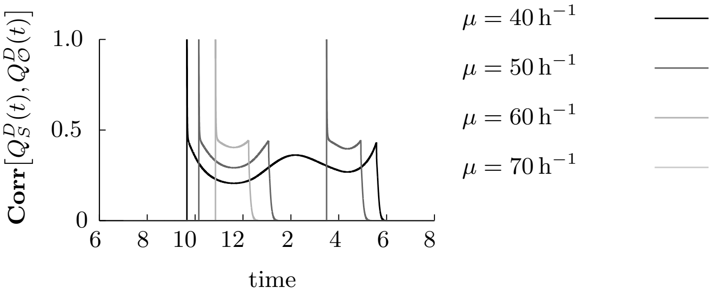

Fig. 4.46. Influence of the service rate μ on the correlation $\mathbf{Corr}\big[Q_S^D(t), Q_O^D(t)\big]$ of the processes of the number of customers in the system and in the orbit

First of all, in Figure 4.46 the influence of the service rate on the correlation of the orbit and system processes is depicted. If the agents work fast enough to serve all arriving customers, the correlation is zero. Smaller service rates, i.e. long average service times, lead to weaker correlations. Furthermore, the correlation decreases if the number of customers in the system and in the orbit increases. In other words, the more customers arrive in the overloaded system, the more independent are the processes which describe the number of customers in the system and the orbit. The reasons for this growing independency are similar to the reasons for the independency of the number of customers in the orbit from the abandonment parameter.

If the number of customers in the system exceeds the number of agents, the arrival process to the orbit is a non-homogeneous Poisson process with parameter $p(\lambda(t) - \mu N(t))$[21]. This process is independent of the process describing the number of customers in the system. The more customers try to enter the system, the more customers are directly[22] transferred to the orbit not influencing the congestion in the system. Hence, the correlation decreases. If the service rate decreases, the state where the number of customers in the system exceeds the number of active servers is reached earlier and lasts longer.

Nearly the same statements on the correlation hold for a varying number of agents as displayed in Figure 4.47. The correlation of the two processes increases, if the number of agents on duty increases. Consequently, the more agents are staffed, the stronger is the influence of the number of customers in the system on the number of customers in the orbit and the other way

[21] See the explanation on Page 12.

[22] In the fluid model the customers are assumed to be like liquid, which is processed by drains. Therefore, the liquid flows out of the basin immediately after entering but is limited by the drains.

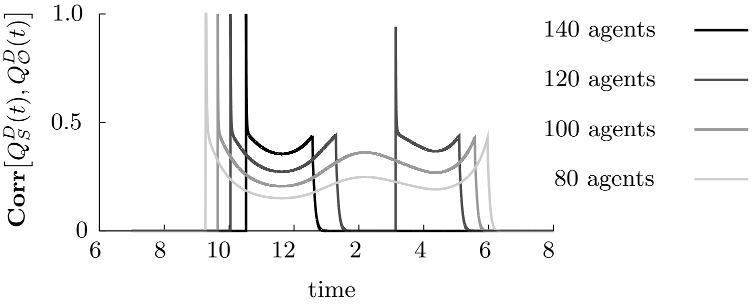

Fig. 4.47. Influence of the number of active agents $N(t)$ on the correlation $\mathbf{Corr}\big[Q_S^D(t), Q_\mathcal{O}^D(t)\big]$ of the processes of the number of customers in the system and in the orbit

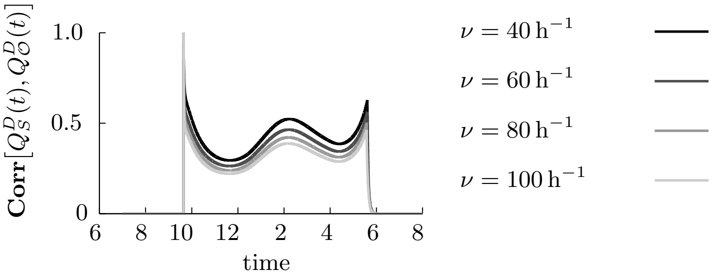

Fig. 4.48. Influence of the abandonment rate ν on the correlation $\mathbf{Corr}\big[Q_S^D(t), Q_\mathcal{O}^D(t)\big]$ of the processes of the number of customers in system and in the orbit

round. If the system is underloaded and no customers remain in the orbit the correlation of the processes is zero. Obviously, both processes are always positively correlated, if the number of customers in the system exceeds the number of agents, i.e., if the number of customers in the system grows, the number of customers in the orbit increases as well and vice versa.

In Figure 4.48 the influence of the abandonment rate ν on the correlation $\mathbf{Corr}\big[Q_S^D(t), Q_\mathcal{O}^D(t)\big]$ of the two processes is presented. In Figures 4.8 on Page 84 and 4.43 on Page 113 the abandonment rate had no impact on either the number of customers in the orbit or on the variance of the number of customers in the system. Therefore, the influence of the abandonment rate on the correlation is based on the influence on the number of customers in the system shown in Figure 4.7 on Page 83 and the variance of the number of customers in the system shown in Figure 4.38 on Page 110. The impact on the correlation is very small. The more impatient the customers become, the more independent the processes are.

Finally, the influence of the mean times to retrial γ^{-1} and the probability of retrial p are inspected in Figures 4.49 and 4.50. If the mean time to retrial decreases, the correlation of the two processes of the number of customers in

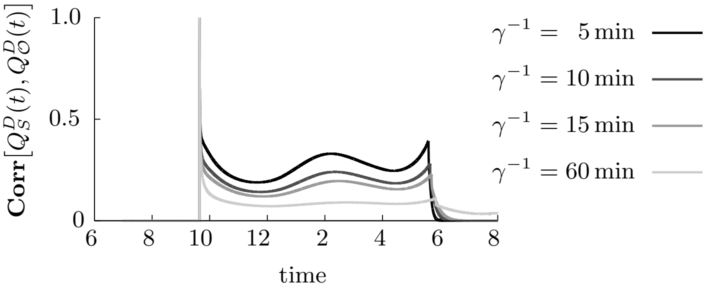

Fig. 4.49. Influence of the mean time to retrial γ^{-1} on the correlation $\mathbf{Corr}\big[Q_S^D(t), Q_O^D(t)\big]$ of the processes of the number of customers in system in the orbit

the system and in the orbit increases, i.e., the quicker customers start a new attempt, the higher is the influence of the processes on each other. The longer customers wait in the orbit, the more independent the queueing processes in the system and the orbit become. Therefore, the performance of the system is less influenced by retrials with long mean times to retrial, i.e, long sojourn times of customers in the orbit.

Because many of the performance measures improve as well if the mean time in orbit increases, it would be worthwhile asking customers to recall after some time, e.g., an hour, if the contact center is currently overloaded. If the service is not too urgent for the customers, this might be an alternative to letting customers wait for a long time.

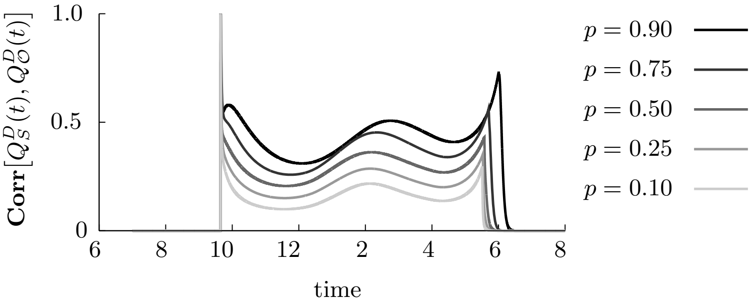

Fig. 4.50. Influence of the probability of retrial p on the correlation $\mathbf{Corr}\big[Q_S^D(t), Q_O^D(t)\big]$ of the processes of the number of customers in system and in the orbit

However, if many customers are willing to retry after some time in the orbit, the correlation of the processes is higher than if only few retry. This effect is shown in Figure 4.50. This effect is reasonable, because the more customers are willing to recall, the more the orbit becomes an additional waiting room. If all customers are willing to retry the customers circulate between the system and the orbit until they are finally served.

In the sections about the variance and the correlation of the processes of the number of customers in the system and the orbit, we have shown that stochasticity has some impact on the contact center and how the processes influence each other. By means of the variance we are able to calculate confidence intervals around the mean number of customers in the system, such that the variability of the processes can be considered in a staffing or scheduling approach. In the next section we will see how these confidence intervals improve the approximation of the simulation results.

4.2 Contact Centers with Heterogeneous Customers and Agents

4.2.1 Description of the Model

The structure of the contact center considered in this section is depicted in Figure 4.51. The contact center presented in Figure 2.5 on Page 13 and analysed in Mandelbaum et al. (1998) is a special case of this model. We assume two types of customers who arrive at the contact center according to Poisson processes with rates $\lambda_1(t)$ and $\lambda_2(t)$ for type-1 and type-2 customers, respectively. These rates are assumed to be time-dependent as in the previous section. This model combines the problems associated with priority and with retrial queues which have so far been considered solely in separation.

If no-one is waiting in each queue the arriving customer is served immediately. Customers of the first type, called type-1 customers, are either served by a type-1 specialist with mean service time μ_1^{-1} or by a generalist with mean service time $\overline{\mu_1}^{-1}$, if all specialists are busy. Otherwise they have to wait.

A type-2 customer is served by a type-2 specialist with mean service time μ_2^{-1}. If no specialist of type 2 is available and there are generalists who are not attending type-1 customers, the type-2 customer is served by a generalist with mean service time $\overline{\mu_2}^{-1}$. If all servers are busy, arriving customers have to wait. If customers of both types are waiting and a generalist finishes a service, he serves a type-1 customer next. Otherwise he serves the next waiting type-2 customer. That means type-1 customers have non-preemptive priority.

We assume that the customers are impatient and abandon if an exponentially distributed waiting time limit ν_1 or ν_2 depending on their type is reached. Some of these customers leave the system and are supposed to never call again. Such customers are lost. A fraction of p_1 and p_2 of all abandoning customers is assumed to try again during the day. These customers move into the orbits, which are distinguished according to the different kinds of customers. After an exponentially distributed time type-1 customers retry with mean rate $\gamma_1(t)$ and type-2 customers with mean rate $\gamma_2(t)$.

Like the customers who enter the system from outside, these retrying customers either are served immediately or have to wait. The agents do not know whether an arriving customers is a primary caller or a retrial.

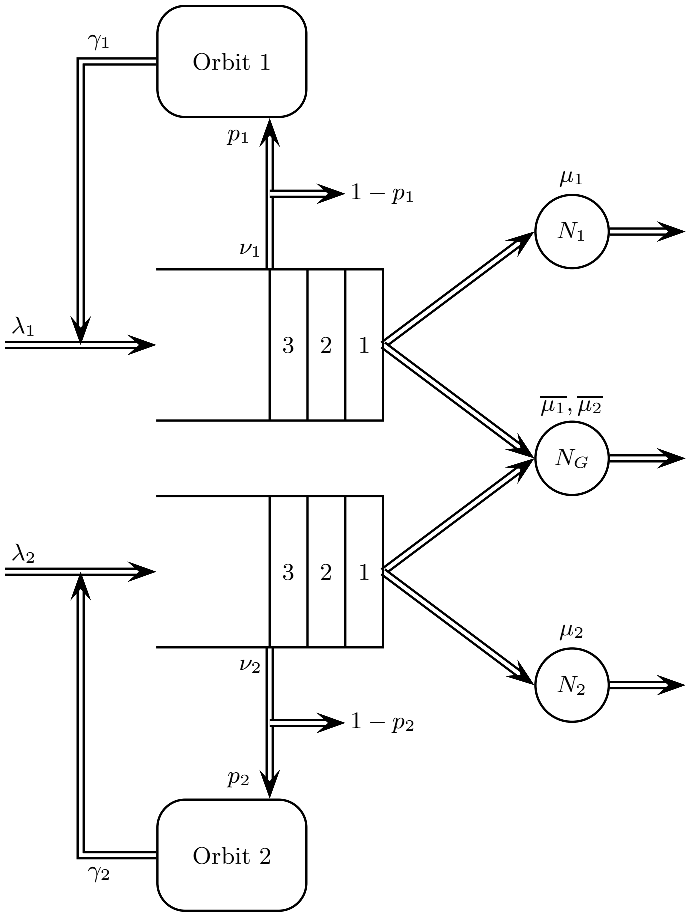

Fig. 4.51. A contact center model with two types of customer classes, three kinds of agent groups and retrials of impatient customers

4.2.2 Determination of the Fluid Processes

Firstly, we want to develop and analyse a fluid approximation of this model. Therefore, we denote by

$$\big(\boldsymbol{Q}(t)\big)_{t\in\mathbb{R}_0^+} = (Q_1(t), Q_2(t), Q_{\mathcal{O}1}(t), Q_{\mathcal{O}2}(t))_{t\in\mathbb{R}_0^+} \tag{4.15}$$

the stochastic process which describes the number of customers in the different queues, where $Q_1(t)$ represents the number of type-1 customers in the system waiting or being served. $Q_2(t)$ denotes the number of type-2 customers who are served or waiting for an available server. The third and fourth processes describe the number of type-1 and type-2 customers in orbit 1 and 2 who are going to retry. As we want to get insights into the time-dependent behaviour of the queues, we will look at the change of the process at time t.

The change of the amount of customers in the system or in the orbits in a small time interval can be expressed according to Section 3.2 on Page 36 by the derivative of the four processes of the vector $\big(\boldsymbol{Q}(t)\big)_{t\in\mathbb{R}_0^+}$ with respect to the time. Therefore differential equations can be used to describe how many customers enter and leave the system on average during a very small time interval.

The differential equation (4.23a) for the first process describing the number of type-1 customers in the system $Q_1^F(t)$ is equivalently derived just as the process $Q_S^F(t)$ in Section 4.1.2 on Page 75 except for the terms containing the number of generalists $N_G(t)$. The generalists become important if the number of type-1 customers is greater than the number of their specialists. If the amount of type-1 customers exceeds $N_1(t)$ but remains below the sum of the number of type-1 specialists and generalists $N_1(t) + N_G(t)$, all type-1 customers not served by type-1 specialists can be served by generalists with rate $\overline{\mu_1}$. We denote the number of generalists serving type-1 customers in the fluid model by $B_1^F(t)$. Generalists give priority to type-1 customers. Therefore the number of busy generalists serving type-1 customers is given by the minimum of the number of generalists on duty $N_G(t)$ and the number of type-1 customers that exceeds the number of their specialists, i.e.,

$$B_1^F(t) = \min\Big\{N_G(t), \big\{Q_1^F(t) - N_1(t)\big\}^+\Big\}. \tag{4.16}$$

The departure rate of type-1 customers from the system being served by their specialists is $\mu_1 N_1(t)$ and the departure rate after being served by a generalist is $\overline{\mu_1} B_1^F(t)$. Therefore in this case the overall departure rate of type-1 customers $d_1(t)$ is the sum of these two. If the amount of customers exceeds $N_1(t)+N_G(t)$, where $N_G(t)$ is the current number of generalists, type-1 customers waiting for service may leave because of impatience. The number of waiting type-1 customers $L_1^F(t)$ in the fluid model is given by

$$L_1^F(t) = \big\{Q_1^F(t) - N_1(t) - N_G(t)\big\}^+. \tag{4.17}$$

Each waiting customer abandons with rate ν_1, such that the departure rate $d_1(t)$ becomes $\mu_1 N_1(t) + \overline{\mu_1} N_G(t)$ plus $\nu_1 L_1^F(t)$. Consequently the total departure rate is

$$d_1(t) = \begin{cases} \mu_1 Q_1^F(t), & Q_1^F(t) \leq N_1(t) \\ \mu_1 N_1(t) + \overline{\mu_1}(Q_1^F(t) - N_1(t)), & N_1(t) < Q_1^F(t) \leq N_1(t) + N_G(t) \\ \mu_1 Q_1^F(t) + \overline{\mu_1} N_G(t) & \\ \quad +\nu_1(Q_1^F(t) - N_1(t) - N_G(t)), & N_1(t) + N_G(t) < Q_1^F(t). \end{cases} \tag{4.18}$$

Taking into account that abandonment does not occur until the number of type-1 customers exceeds the sum of the number of their specialists and the generalists, the Differential Equation (4.23c) is similar to Equation (4.1b) on Page 75.

Similar to the process of type-1 customers, the rate of change of type-2 customers is equal to the primary arrival rate $\lambda_2(t)$ plus the arrival rate from the orbit, given by the product of the number of type-2 customers in the orbit $Q^F_{\mathcal{O}2}(t)$ and the retrial rate γ_2, minus the departure rate $d_2(t)$.

For the departure rate of type-2 customers in the system we have to keep in mind that type-1 customers are prioritised by generalists. If the number of type-2 customers is below the number of their specialists, the departure rate is given by product of the service rate of specialists μ_2 and the number of type-2 customers in the system, i.e.

$$d_2(t) = \mu_2 Q^F_2(t). \tag{4.19a}$$

If the number of type-2 customers exceeds the number of specialists $N_2(t)$, the number of type-1 customers in the system influences the departure rate of type-2 customers. If more than $N_1(t) + N_G(t)$ type-1 customers are in the system, generalists serve only type-1 customers, i.e., the number of busy generalists serving type-2 customers $B_2(t)$ is zero. Type-2 customers have to wait and may abandon, i.e., the number of type-2 customers in the system changes by the rate customers are served by specialists and by the rate the waiting customers $L^F_2(t)$ abandon.

$$d_2(t) = \mu_2 N_2(t) + \nu_2 L^F_2(t). \tag{4.19b}$$

In this case the number of waiting type-2 customers $L_2(t)$ is given by the difference of the number of customers in the system and the number of type-2 specialists $Q^F_2(t) - N_2(t)$. If the number of type-1 customers in the system is between $N_1(t)$ and $N_G(t)$ and the number of type-2 customers is greater than the number of type-2 specialists, generalists additionally serve type-2 customers with mean rate $\overline{\mu_2}$, i.e., the number of generalists serving type-2 customers $B^F_2(t)$ is the minimum of the number of type-2 customers who are not served by specialists and the number of generalists not occupied by type-1 customers.

$$d_2(t) = \mu_2 N_2(t) + \overline{\mu_2} B^F_2(t) + \nu_2 L^F_2(t), \tag{4.19c}$$

Therefore, the number of busy generalists serving type-2 customers at time t is

$$B^F_2(t) = \min\left\{ \big\{Q^F_2(t) - N_2(t)\big\}^+, \left\{ N_G(t) - \big\{Q^F_1(t) - N_1(t)\big\}^+ \right\}^+ \right\}. \tag{4.20}$$

In this last case the departure rate is the sum of three different kinds of departure rates (4.19) due to service by specialists, due to service by generalists, and due to abandonment, i.e.,

$$d_2(t) = \mu_2 \min\big\{Q^F_2(t), N_2(t)\big\} + \overline{\mu_2} B^F_2(t) + \nu_2 L^F_2(t), \tag{4.21}$$

where $L_2^F(t)$ is the number of type-2 customers waiting given by

$$L_2^F(t) = \Big\{Q_2^F(t) - N_2(t) - \Big\{N_G(t) - \big\{Q_1^F(t) - N_1(t)\big\}^+\Big\}^+\Big\}^+. \tag{4.22}$$

The priority rule[23] has an influence on the term describing the abandonment of customers in Equations (4.23b) and (4.23d), as well. Therefore, we get the following system of differential equations.

$$\begin{aligned} \frac{d}{dt}Q_1^F(t) &= \lambda_1(t) + \gamma_1 Q_{\mathcal{O}1}^F(t) - \mu_1 \min\big\{Q_1^F(t), N_1(t)\big\} && \text{(4.23a)}\\ &\quad - \overline{\mu_1} B_1^F(t) - \nu_1 L_1^F(t) \\ \frac{d}{dt}Q_2^F(t) &= \lambda_2(t) + \gamma_2 Q_{\mathcal{O}2}^F(t) - \mu_2 \min\big\{Q_2^F(t), N_2(t)\big\} && \text{(4.23b)}\\ &\quad - \overline{\mu_2} B_2^F(t) - \nu_2 L_2^F(t) \\ \frac{d}{dt}Q_{\mathcal{O}1}^F(t) &= p_1\nu_1 L_1^F(t) - \gamma_1 Q_{\mathcal{O}1}^F(t) && \text{(4.23c)}\\ \frac{d}{dt}Q_{\mathcal{O}2}^F(t) &= p_2\nu_2 L_2^F(t) - \gamma_2 Q_{\mathcal{O}2}^F(t), && \text{(4.23d)} \end{aligned}$$

for all $t \in \mathbb{R}_0^+$. The initial value problem associated with these equations can be solved easily by using standard numerical methods, e.g., by the fourth order Runge-Kutta method, and give us a good approximation of the mean amount of customers in both the system and the orbit as described before.

Next the differential equations for the variances and covariances have to be derived in order to make statements on the variability of these processes. Therefore the fluid model has to be refined to a diffusion model as done in Section 3.3.

4.2.3 Refinement to a Diffusion Model

Now we derive the stochastic differential equations for the diffusion process

$$\big(\boldsymbol{Q^D}(t)\big)_{t\in\mathbb{R}_0^+} = \big(Q_1^D(t), Q_2^D(t), Q_{\mathcal{O}1}^D(t), Q_{\mathcal{O}2}^D(t)\big)_{t\in\mathbb{R}_0^+} \tag{4.24}$$

as described in Section 3.3 on Pages 57 ff. Again, the sets of critical times[24] are supposed to have measure zero. These sets are defined by the turning points of the minimum and maximum functions in the differential equations in (4.23). Similar to the previous Section 3.2, two of the sets, $\boldsymbol{S}_1$ and $\boldsymbol{S}_3$, contain all

[23] In the fluid approximation the priority rule changes from non-preemptive to preemptive priority, i.e., whenever a type-1 customer arrives and has to be served by a generalist, the service of a type-2 customer is disrupted. Consequently, in the fluid model non-preemptive and preemptive priority rules are modelled in the same way. See Mandelbaum et al. (1998) Section 7 and Ridley et al. (2004).

[24] See Section 4.1 on Page 62 for the meaning of these sets.

times when the numbers of customers equals the number of their specialists. These sets are associated with the service of type-1 and type-2 specialists. For the other two sets $\boldsymbol{S}_2$ and $\boldsymbol{S}_4$, the number of generalists working plays a central role. The set $\boldsymbol{S}_2$ consists of all moments, when the number of type-1 customers equals the sum of the number of their specialists and the number of generalists, i.e., the function $\min\{N_G(t),\{Q_1^F(t)-N_1(t)\}^+\}$ specifies this set. The last set $\boldsymbol{S}_4$ is determined by the turning points of the function $B_2^F(t)$. This set is most critical with respect to the null set assumption because it depends on both processes. We get the following sets:

$$\boldsymbol{S}_1 = \{t|Q_1(t) = N_1(t)\}\,, \tag{4.25a}$$

$$\boldsymbol{S}_2 = \{t|Q_1(t) = N_1(t) + N_G(t)\}\,, \tag{4.25b}$$

$$\boldsymbol{S}_3 = \{t|Q_2(t) = N_2(t)\}\,, \tag{4.25c}$$

and

$$\boldsymbol{S}_4 = \{t|Q_1(t) + Q_2(t) = N_1(t) + N_2(t) + N_G(t)\}\,. \tag{4.25d}$$

If these sets are almost empty, the diffusion processes are Gaussian[25] and simpler functional equations can be derived than in case of non-Gaussian processes as shown in Section 3.3. These sets do not have measure zero if the processes $Q_1^D(t)$ and $Q_2^D(t)$ do not pass the level where the number of customers equals the different numbers of agents quickly. Then the non-differentiability of the maximum and minimum functions at these points become important, such that we need the scalable Lipschitz derivative[26] to calculate the functional equations of the diffusion processes.

However, if the sets in (4.25) have measure zero, the stochastic functional equations consist of three parts as described in Section 3.3. In order to make the notation shorter and to present the results more clearly, we use matrix forms[27]. The first part of the diffusion process is the row vector $\boldsymbol{Q^D}(t_0)$ which describes the number of customers in the considered system or orbit at the beginning of the observation at time t_0. The second part is an integral expression which represents the cumulative drift of the process beginning at time t_0 and ending at time t. This second part can be represented by integrating the product of the row vector $\boldsymbol{Q^D}(s)$ and a quadratic matrix $\boldsymbol{A}(s)$, which formally is the Jacobian matrix of the right hand side of the differential equations in (4.23) differentiated with respect to the fluid processes. The matrix $\boldsymbol{A}(s)$ is given by

[25] See Whitt (2000) and Ward and Glynn (2003b,a) for an explanation of the difficulties.

[26] See Mandelbaum et al. (1998) and Page 60 for an explanation.

[27] The detailed functional equations of the diffusion processes are presented in Appendix A.2.

$$\begin{pmatrix} \alpha_1 & -\nu_2\mathbb{1}_{\{c_4\}} + \overline{\mu_2}\mathbb{1}_{\{c_5\}} & p_1\nu_1\mathbb{1}_{\{c_3\}} & p_2\nu_2\mathbb{1}_{\{c_4\}} \\ 0 & \alpha_2 & 0 & p_2\nu_2\mathbb{1}_{\{c_7\}} \\ \gamma_1 & 0 & -\gamma_1 & 0 \\ 0 & \gamma_2 & 0 & -\gamma_2 \end{pmatrix} \tag{4.26}$$

with

$$\begin{aligned} \alpha_1 &= -\mu_1\mathbb{1}_{\{c_1\}} - \overline{\mu_1}\mathbb{1}_{\{c_2\}} - \nu_1\mathbb{1}_{\{c_3\}}, \\ \alpha_2 &= -\mu_2\mathbb{1}_{\{c_6\}} - \nu_2\mathbb{1}_{\{c_7\}} - \overline{\mu_2}\mathbb{1}_{\{c_8\}} \end{aligned}$$

and the following conditions for the indicator functions

$$c_1 : \; Q_1^F(s) \le N_1(s) \tag{4.27a}$$

$$c_2 : \; N_1(s) < Q_1^F(s) \le N_1(s) + N_G(s) \tag{4.27b}$$

$$c_3 : \; Q_1^F(s) > N_1(s) + N_G(s) \tag{4.27c}$$

$$\begin{aligned} c_4 : \; & N_1(s) < Q_1^F(s) \le N_1(s) + N_G(s), \\ & Q_2^F(s) > N_1(s) + N_2(s) + N_G(s) - Q_1^F(s) \end{aligned} \tag{4.27d}$$

$$\begin{aligned} c_5 : \; & N_1(s) < Q_1^F(s) \le N_1(s) + N_G(s), \\ & N_2(s) < Q_2^F(s) \le N_1(s) + N_2(s) + N_G(s) - Q_1^F(s) \end{aligned} \tag{4.27e}$$

$$c_6 : \; Q_2^F(s) \le N_2(s) \tag{4.27f}$$

$$c_7 : \; Q_2^F(s) > N_2(s) + \Big\{N_G(s) - \big\{Q_1^F(s) - N_1(s)\big\}^+\Big\}^+ \tag{4.27g}$$

$$c_8 : \; N_2(s) < Q_2^F(s) \le N_2(s) + \Big\{N_G(s) - \big\{Q_1^F(s) - N_1(s)\big\}^+\Big\}^+. \tag{4.27h}$$

The first column contains the deviation of the first differential equation in (4.23a) with respect to $Q_1^F(t)$ in the first line and with respect to $Q_{\mathcal{O}1}^F(t)$ in the third line. As the process of the number of type-1 customers is independent of the behaviour of type-2 customers the entries in the second and fourth line of the first column are zero. The entries in other columns are equivalently derived.

The last part of the functional equation is a row vector of Brownian motions $\mathcal{B}(t) = (B_1(t), B_2(t), B_1^{\mathcal{O}}(t), B_2^{\mathcal{O}}(t))$ where each entry of this vector is the sum of different standard Brownian motions with mean zero. The Brownian motions describe the deviation of the process from its average. The deviation is also named drift and shows how far the diffusion process has moved from his starting average. These notations give rise to the following stochastic functional equation of the diffusion process:

$$\boldsymbol{Q^D}(t) = \boldsymbol{Q^D}(t_0) + \int_0^t \boldsymbol{Q^D}(s)\boldsymbol{A}(s)\,ds + \mathcal{B}(t). \tag{4.28}$$

From the stochastic functional equations in (4.28) the differential equations for variances and covariances are deduced. The complete derivation of the differential equations for the variances and covariances is given in Appendix A.2. The differential equations for the covariances are presented in matrix form as well. The covariance matrix $\mathbf{COV}\big[\boldsymbol{Q^D}(t)\big]$ of the considered diffusion process is a four-dimensional, quadratic, and symmetric matrix, i.e., the entry in row i and column j equals the entry in row j and column i for $i \neq j$. The entries of the principal diagonal are the variances of each process of the vector $\boldsymbol{Q^D}(t)$. Mandelbaum et al. (1998) showed that the differential matrix equation for the covariance matrix of the diffusion processes for all queueing models can be written as

$$\frac{d}{dt}\mathbf{COV}\big[\boldsymbol{Q^D}(t)\big] = \mathbf{COV}\big[\boldsymbol{Q^D}(t)\big]\,\boldsymbol{A}(t) + (\boldsymbol{A}(t))^T\mathbf{COV}\big[\boldsymbol{Q^D}(t)\big] + \boldsymbol{B}(t). \tag{4.29}$$

In this matrix equation the matrix $\boldsymbol{A}(t)$ is given by Equation (4.26).

The four-dimensional and symmetric matrix $\boldsymbol{B}(t)$ results from the Brownian motions. The entries of this matrix are derived by applying the chain rule of stochastic calculus and the property given in Equation (3.56) on Page 64 for standard Brownian motions as shown in Section 3.3 on Page 64. This matrix has entries zero whenever the considered processes have no Brownian motion in common. The entries of the matrix are negative, if the Brownian motions in the functional equations have opposite signs, otherwise the entry is positive. The matrix $\boldsymbol{B}(t)$ contains the following positive rate functions $b_i(t)$, $i = 1, 2, 3, 4$, describing the change in the derivatives of the processes $\boldsymbol{Q}(t)$ which are

$$\begin{aligned} b_1(t) &= \lambda_1(t) + \gamma_1(t)Q^F_{\mathcal{O}1}(t) + \mu_1 \min\big\{Q^F_1(t), N_1(t)\big\} \\ &\quad + \overline{\mu_1}B^F_1(t) + \nu_1 L^F_1(t) \end{aligned} \tag{4.30a}$$

$$\begin{aligned} b_2(t) &= \lambda_2(t) + \gamma_2(t)Q^F_{\mathcal{O}2}(t) + \mu_2 \min\big\{Q^F_2(t), N_2(t)\big\} \\ &\quad + \overline{\mu_2}B^F_2(t) + \nu_2 L^F_2(t) \end{aligned} \tag{4.30b}$$

$$b_3(t) = \gamma_1(t)Q^F_{\mathcal{O}1}(t) + p_1\nu_1 L^F_1(t) \tag{4.30c}$$

$$b_4(t) = \gamma_2(t)Q^F_{\mathcal{O}2}(t) + p_2\nu_2 L^F_2(t). \tag{4.30d}$$

Then the matrix $\boldsymbol{B}(t)$ is

$$\boldsymbol{B}(t) = \begin{pmatrix} b_1(t) & 0 & -b_3(t) & 0 \\ 0 & b_2(t) & 0 & -b_4(t) \\ -b_3(t) & 0 & b_3(t) & 0 \\ 0 & -b_4(t) & 0 & b_4(t) \end{pmatrix}. \tag{4.31}$$

This system of differential equations resulting form the matrix equation in (4.29) can again be solved numerically for any initial condition. In general, we will assume that the variances and the covariances are zero at the beginning of the observation, i.e.,

$$\mathbf{VAR}\big[Q_i^D(t_0)\big] = 0 \quad \text{for } i = 1, 2, \mathcal{O}1, \mathcal{O}2, \tag{4.32}$$

$$\mathbf{COV}\big[Q_i^D(t_0), Q_j^D(t_0)\big] = 0 \quad \text{for } i, j = 1, 2, \mathcal{O}1, \mathcal{O}2 \text{ and } i \neq j. \tag{4.33}$$

The solution gives us a good approximation for the variances and covariances of the processes[28].

4.2.4 Performance Measures

Some new performance measures must be explained to compare the different systems and to estimate their performance. These performance measures are based on the results we get from numerically solving the initial value problem given by the system of Differential Equations (4.23) on Page 123 and some initial conditions for the processes of the number of customers in the system and the orbit $(\boldsymbol{Q^F}(t))_{t\in\mathbb{R}_0^+}$. In general, we will assume that the processes start with zero customers in the system and the orbit.

First of all, the number of customers in the system is directly given by the results for the fluid processes as described in Section 3.2.3 on Page 38 and 4.1.4 on Page 76. Subtracting the number of active agents gives rise to the number of customers waiting $L_1^F(t)$ and $L_2^F(t)$ in each queue, which were derived in Equations (4.17) and (4.22) on Pages 121f. Without the term $N_G(t)$ in Equation (4.17) this equation would equal the queue length determined in the previous sections.

Much more important is the waiting time of the customers at any time t. If this waiting time becomes too long most customers will abandon and only few of them will recall, if the probability of retrial is small. Especially the influence of type-1 customer on type-2 customers in the system is interesting and should be studied in detail.

The mean waiting time of type-1 customers is easily derived as these customers have preemptive priority in the fluid approximation. Consequently, the instantaneous waiting time of these customers is not influenced by the number of customers in the second queue. The waiting time of a typical type-1 customer $W_1^F(t)$ at time t is given equivalently to the waiting time of customers $W_S^F(t)$ derived in Equation (4.3) on Page 76 by the fraction of the number of customers in front of this customer divided by the departure rate $d_1(t)$ in Equation (4.18) of type-1 customers, i. e.,

$$W_1^F(t) = \frac{\big\{Q_1^F(t) - N_1(t) - N_G(t)\big\}^+}{\mu N_1(t) + \overline{\mu_1} N_G(t) + \nu_1(Q_1^F(t) - N_1(t) - N_G(t))}. \tag{4.34}$$

Contrarily, the waiting time of type-2 customers depends on the number of type-1 customers in the system because of the priority rule. Type-2-customers have to wait until either a type-2 specialist becomes available or no type-1 customers are waiting and a generalist becomes available. In general, the

[28] See also Mandelbaum et al. (1999a,b), and Mandelbaum et al. (2002).

mean waiting time is the number of type-2 customers waiting divided by the departure rate $d_2(t)$ of type-2 customers.

Unfortunately, a customer of class 2 arriving at time t as well as an observer from outside are not able to predict how the number of type-1 customers in the system during the waiting time of the type-2 customer will change. If this special type-2 customer arrives while no type-1 customer is waiting but some are served by generalists, his waiting time might become shorter, if no further type-1 customers arrive during his waiting time. However, if many type-1 customers arrive during the waiting time, the waiting time of the type-2 customer might even become longer because of the preemptive priority of type-1 customers. In this case type-1 customers expel type-2 customers from service by generalists. These type-2 customers will be put to the front of the queue until they are finally served. Therefore, assumptions on the development of the queue length of type-1 customers, in particular, on the number of generalists serving type-1 and type-2 customers, have to be made to estimate the waiting time of an arriving type-2 customer.

We assume that the number of type-1 customers in the system stays on average at the same level during the waiting time of type-2 customers. Under this assumption we solely have to consider the departure rate $d_2(t)$ of type-2 customers. If the number of type-1 customers increases during this time period, the waiting time of type-2 customers will become longer, whereas the waiting time decreases if the number of type-1 customers decreases.

The number of generalists currently serving type-2 customers

$$B_2^F(t) = \min\Big\{\big\{Q_2^F(t) - N_2(t)\big\}^+, \big\{N_G(t) - B_1^F(t)\big\}^+\Big\}.$$

is the minimum of the amount of type-2 customers that exceeds the number of type-2 specialists and the number of generalists not occupied by type-1 customers. This number $B_2^F(t)$ has been determined in Equation (4.20) on Page 122.

Using this notation the departure rate $d_2(t)$ of type-2 customers is the combination of three terms (4.19) on Page 122 as shown in Equation (4.21). By means of this departure rate and the fact that the waiting time is zero if no number of type-2 customer waits, we get for the waiting time of type-2 customers in the system

$$W_2^F(t) = \frac{L_2^F(t)}{\mu_2 N_2(t) + \overline{\mu_2} B_2^F(t) + \nu_2 L_2^F(t)}. \tag{4.35}$$

In this formula the waiting time is implicitely influenced by the number of type-1 customers at the arrival time of the type-2 customers. Therefore, the time-dependent and the aggregated waiting times of type-2 customers have to be thoroughly investigated. Contrarily, the waiting time of type-1 customers is not influenced by the other customers, so the results from the previous Section 4.1 can directly be conferred.

The probability of being served for an arriving type-1 customer is equivalently derived as in the previous Section 4.1.4 despite the fact that the generalists have to be considered as well, i.e.,

$$P_1^F(\text{served}, t) = \frac{\mu_1 \min\{Q_1^F(t), N(t)\} + \overline{\mu_1} B_1^F(t)}{\mu_1 \min\{Q_1^F(t), N(t)\} + \overline{\mu_1} B_1^F(t) + \nu_1 L_1^F(t)}. \tag{4.36}$$

The probability of being served for type-2 customers depends on the number of customers in queue one. The number of served customers is given by the number served by type-2 specialists plus the number of customers served by generalists if a generalist is available for serving type-2 customers. The amount of type-2 customers leaving the system at time t is given by the amount served as explained before plus the amount abandoning. Therefore the probability of being served is

$$P_2^F(\text{served}, t) = \frac{\mu_2 \min\{Q_2^F(t), N_2(t)\} + \overline{\mu_2} B_2^F(t)}{\mu_2 \min\{Q_2^F(t), N_2(t)\} + \overline{\mu_2} B_2^F(t) + \nu_2 L_2^F(t)}. \tag{4.37}$$

The probability of abandoning and the probability of moving into the orbit are determined similarly to the derivation of the previous Section 4.1 on Page 40.

As mentioned before[29], we can calculate two different aggregated probabilities of being served. The first one is the aggregated probability of being served

$$P_{agg}^F(\text{served}, T) = \frac{\int_0^T \mu(t) \min\{Q(t), N(t)\}\ dt}{\int_0^T d(t)\, dt}.$$

for all customers who have left the system either as primary calls or as recalls in the time interval $[0, T]$, the other aggregated probability of being served is the probability of being finally served

$$P_\lambda^F(\text{served}, T) = \frac{\int\limits_0^T \mu(t) \min\{Q(t), N(t)\}\ dt}{\int_0^T \lambda(t)\, dt}$$

for all primary calls of the time interval $[0, T]$. On Page 40 we have argued why the first one is more useful for the purpose of analysis. The first aggregated probability for both customer classes is derived equivalently to Equation (3.32) on Page 40 by weighting the probabilities of being served by the departure rates of the considered customer type and dividing by the aggregated departures. The second probability is determined as in Equation (3.34) on Page 41 by aggregating the number of served customers over the time horizon and dividing by the aggregated number of arrivals.

[29] See Pages 40f.

Furthermore, the utilisation of the different agent groups has to be determined to estimate the burden of work placed on the agents. Similarly to Equation (4.9) on Page 78, we get

$$U_1^F(t) = \frac{\min\{Q_1^F(t), N_1(t)\}}{N_1(t)}, \tag{4.38}$$

$$U_2^F(t) = \frac{\min\{Q_2^F(t), N_2(t)\}}{N_2(t)}, \tag{4.39}$$

and

$$U_G^F(t) = \frac{B_1^F(t) + B_2^F(t)}{N_G(t)} \tag{4.40}$$

for the utilisation of type-1 specialists, type-2 specialists and generalists, respectively. The utilisation is aggregated according to Equation (3.36) on Page 42.

Finally, the profit function for this complex model aggregates the economical performance of the whole system for the period $[0, T]$. While the technical performance measures focus on special perspectives, the economical performance measures represent the position of the management, which is interested in the revenue gained from served customers and in the costs for paying agents and for occupied trunks. If we assume each served customer of class 1 leads on average to a revenue of r_1 and each served customer of class 2 to a revenue of r_2, the revenue gained over the considered length of the time period $[0, T]$ is

$$\begin{aligned} \text{rev}(T) = &\int_0^T r_1\mu_1 \min\{Q_1^F(t), N_1(t)\} + r_1\overline{\mu_1}B_1^F(t)dt \\ &+ \int_0^T r_2\mu_2 \min\{Q_2^F(t), N_2(t)\} + r_2\overline{\mu_2}B_2^F(t)dt. \end{aligned} \tag{4.41}$$

If a type-1 specialist has a hourly wage w_1, a type-2 specialist a hourly wage of w_2, and a generalist of w_G, the costs for agents are

$$\text{cost}_{\text{Agents}}(T) = \int\limits_0^T w_1 N_1(t) + w_2 N_2(t) + w_G N_G(t)\, dt. \tag{4.42}$$

Furthermore, we assume that an occupied line costs ℓ_1 or ℓ_2 per hour in the case of a call center[30]. Therefore the profit of the call center during a time period of legnth T is given by

[30] For e-mails and letters these parameters are zero. In this case the redialling and abandonment parameters will be zero as well.

$$\text{profit}(T) = \text{rev}(T) - \int_0^T \left(\ell_1 Q_1^F(t) + \ell_2 Q_2^F(t)\right) \, dt - \text{cost}_{\text{Agents}}(T). \qquad (4.43)$$

Comparing Equations (4.10) on Page 78 and (4.43), we observe that the structures of these equations are similar. This result holds for all performance measures calculated in this section and underlines the advantage of the fluid approach to be easily adaptable to more complex contact centers. Furthermore, in the model discussed here the type-1 customers are not influenced by the behaviour of type-2 customers, therefore the performance measures for type-1 customers and type-1 specialists are the same as in the previous Section 4.1.

4.2.5 Numerical Results

4.2.5.1 The Influence of the Priority Rule on the Number of Customers in the System and in the Orbit

Before we investigate the influence of the different parameters on the number of customers and the performance measures, we will have a closer look at the difference between preemptive and non-preemptive priority. For this purpose, we simulate the contact center in consideration with preemptive and non-preemptive priority of type-1 customers. By means of this investigation we want to find out whether these priority rules have a high influence of the number of customers in the system because in the fluid model the priority rules make no difference. The fluid model always leads to a preemptive priority, i.e., the prioritised customers disturb the service of the other customers by generalists and displace them. Contrarily, non-preemptive priority means that the prioritised customers have to wait until a current service is finished. In order to show the influence of the priority rules we use the simulation and compare the results of the number of customers of each type both in the system and the orbit for the different rules.

Three different cases with respect to the arrival rate functions $\lambda_1(t)$ and $\lambda_2(t)$ are distinguished. In all cases the arrival rate functions $\lambda_1(t)$ and $\lambda_2(t)$ have a sinusoidal form and are given by the same Equation (3.13) on Page 29 with different parameters.

In the first case the arrival rate functions of type-1 and type-2 customers shown in Figure 4.52 coincide, i.e., $\lambda_1(t) = \lambda_2(t)$ for all times t. In this case the parameters of the arrival rate function are given by

$$\begin{aligned} m_1 &= 9000 \qquad & t_0 &= 7 \text{ am} \qquad & t_2 &= 4 \text{ pm} \\ m_2 &= 7500 \qquad & t_1 &= 12{:}30 \text{ pm} \qquad & t_3 &= 8 \text{ pm}. \end{aligned} \qquad (4.44)$$

In the second case the maximum of the arrival rate function for type-1 customers $\lambda_1(t)$ is reached later than the one for type-2 customers $\lambda_2(t)$. In the third case the arrival rate function of type-1 customers $\lambda_1(t)$ approaches its maximum earlier. The second and the third cases are presented in Figure 4.53.

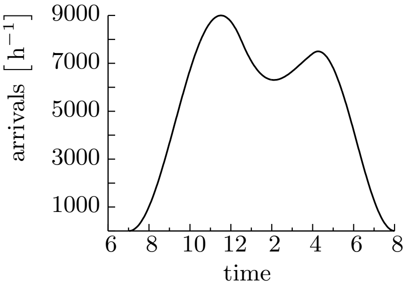

Fig. 4.52. Equal arrival rate functions $\lambda_1(t) = \lambda_2(t)$ for type-1 and type-2 customers

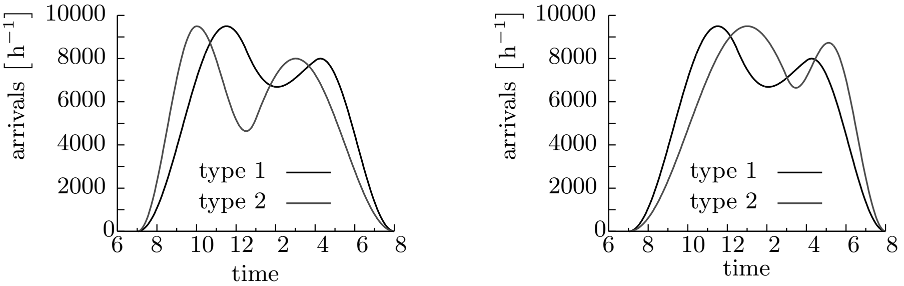

Fig. 4.53. Arrival rate functions of type-1 and type-2 customers with different positions of the maximum arrival rate

In the first picture of Figure 4.53 the parameters of the arrival rate function for type-1 customers $\lambda_1(t)$ are

$$\begin{aligned} m_1^{(1)} &= 9500 & t_0^{(1)} &= 7 \text{ am} & t_2^{(1)} &= 4 \text{ pm} \\ m_2^{(1)} &= 8000 & t_1^{(1)} &= 12{:}30 \text{ pm} & t_3^{(1)} &= 8 \text{ pm.} \end{aligned} \tag{4.45}$$

and for type-2 customers $\lambda_2(t)$

$$\begin{aligned} m_1^{(2)} &= 9500 & t_0^{(2)} &= 7 \text{ am} & t_2^{(2)} &= 1 \text{ pm} \\ m_2^{(2)} &= 8000 & t_1^{(2)} &= 10 \text{ am} & t_3^{(2)} &= 8 \text{ pm.} \end{aligned} \tag{4.46}$$

In the second picture the parameters of the arrival rate function of type-1 customers $\lambda_1(t)$ are the same while the parameters of the second arrival rate function $\lambda_2(t)$ are

$$\begin{aligned} m_1^{(2)} &= 9500 & t_0^{(2)} &= 7 \text{ am} & t_2^{(2)} &= 7 \text{ pm} \\ m_2^{(2)} &= 7000 & t_1^{(2)} &= 3 \text{ pm} & t_3^{(2)} &= 8 \text{ pm.} \end{aligned} \tag{4.47}$$

For the comparison of the results we assume the parameters given in Table 4.4. We distinguish two cases with respect to the retrial rates γ_1 and γ_2 to study the influence of the retrial rate on the priority rule and the number of type-2 customers in the system later on.

Service rate				Number of agents		
μ_1	μ_2	$\overline{\mu_1}$	$\overline{\mu_2}$	$N_1(t)$	$N_2(t)$	$N_G(t)$
$60\,\mathrm{h}^{-1}$	$60\,\mathrm{h}^{-1}$	$40\,\mathrm{h}^{-1}$	$40\,\mathrm{h}^{-1}$	100	100	50

Abandonment rate		Retrial parameters			
ν_1	ν_2	γ_1	γ_2	p_1	p_2
$120\,\mathrm{h}^{-1}$	$120\,\mathrm{h}^{-1}$	$0.5\,\mathrm{h}^{-1}$	$0.5\,\mathrm{h}^{-1}$	0.5	0.5
		$12\,\mathrm{h}^{-1}$	$12\,\mathrm{h}^{-1}$		

Table 4.4. Parameters of the contact center model with heterogeneous agents and customers with retrials

The simulation results were produced by the simulation tool introduced earlier. We carried out 500 repetitions and calculated the mean values for the number of customers in the system and in the orbit. In order to compare the influence of the two priority rules on the number of customers in the system, we determine the differences $\varDelta \boldsymbol{Q}(t)$ between the simulation results. We subtract the number of customers resulting from applying the non-preemptive priority rule $\boldsymbol{Q}^{\text{non}}(t)$ from the number of customers resulting from applying the preemptive priority rule $\boldsymbol{Q}^{\text{pre}}(t)$, i.e.,

$$\varDelta Q_1(t) = Q_1^{\text{pre}}(t) - Q_1^{\text{non}}(t) \tag{4.48a}$$

$$\varDelta Q_2(t) = Q_2^{\text{pre}}(t) - Q_2^{\text{non}}(t) \tag{4.48b}$$

$$\varDelta Q_{\mathcal{O}1}(t) = Q_{\mathcal{O}1}^{\text{pre}}(t) - Q_{\mathcal{O}1}^{\text{non}}(t) \tag{4.48c}$$

$$\varDelta Q_{\mathcal{O}2}(t) = Q_{\mathcal{O}2}^{\text{pre}}(t) - Q_{\mathcal{O}2}^{\text{non}}(t) \tag{4.48d}$$

In the first Figure 4.54 the influence of the priority rule on the number of customers for a low and a high retrial rate and equal arrival rates of both customer classes is shown. The influence of the priority is small as the difference is at most six customers. If type-1 customers have preemptive priority, i.e., they displace type-2 customers from service by generalists, the number of type-1 customers is higher than in the case of non-preemptive priority. The number of type-2 customers acts the other way. The reason for this behaviour of the number of customers is due to the high abandonment rate in comparison to the service rate. If type-1 customers have preemptive priority, more type-1 customers are served. These customers stay on average for a longer time in the system than the customers who abandon. In the case of non-preemptive priority more type-1 customers have to wait. Consequently, more type-1 customers abandon and leave the system more quickly. Therefore, fewer customers of this type are in the system.

If the retrial rate increases, i.e., the mean sojourn time in orbit γ_i^{-1}, $i = 1, 2$, decreases, the influence of the priority rule reduces. In general, the graph of the difference in the number of type-2 customers is almost the mirror image of the graph for type-1 customers. However, the influence on the number of type-2 customers is slightly higher than on the number of type-1 customers. If

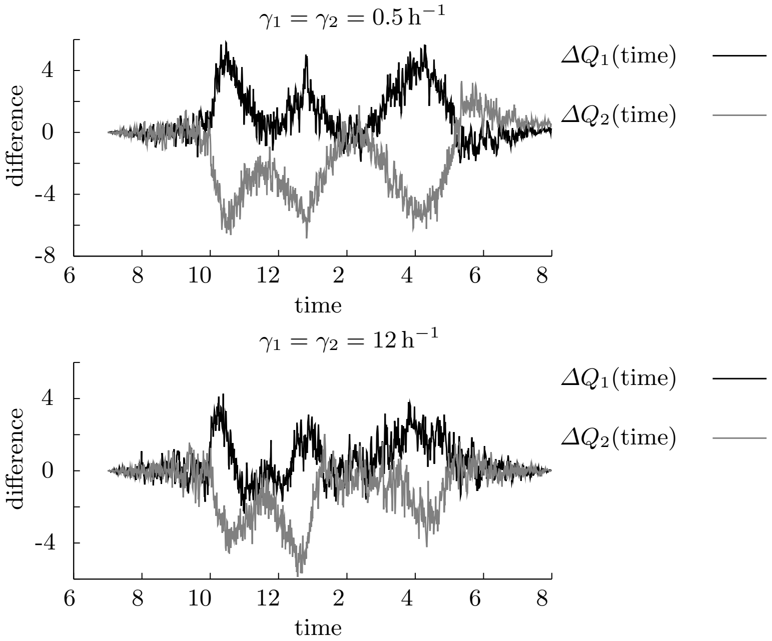

Fig. 4.54. Deviation of the number of customers in the system for different retrial rates and equal arrival rate functions

we compare the developement of the arrival rates in Figure 4.52 to the course of the graphs in Figure 4.54, we observe that the priority rule has solely an influence if the arrival rate for type-1 customers is below the serving capacity of the type-1 specialists and generalists. If the arrival rate increases onward, the difference in the number of customers decreases according to the amount.

Contrary to the difference in the number of customers in the system, the graphs of the difference in the number of customers in the orbit behave as shown in Figure 4.55. If type-1 customers have preemptive priority and dispose type-2 customers from service by generalists, fewer type-1 customers are in the orbit, because less type-1 customers abandon and have to retry. In this case the number of type-2 customers in the orbit is for the same reasons higher than in the case of a non-preemptive priority rule. Amazingly, the influence of the priority rule on the number of customers in the orbit is much stronger than the influence on the number of customers in the system. As observed in Section 4.1.5.1, the number of customers in the orbit increases if the retrial rate decreases, i.e., the customers stay for a longer time in the orbit. The impact of the priority rule is proportional to the absolute number of customers in the orbit, i.e., if more customers are in the orbit, the difference in the number of customers between the two priority rules increases. Unlike the graphs for the difference in the number of customers in the system, the graphs shown in Figure 4.55 are no reflections of each other.

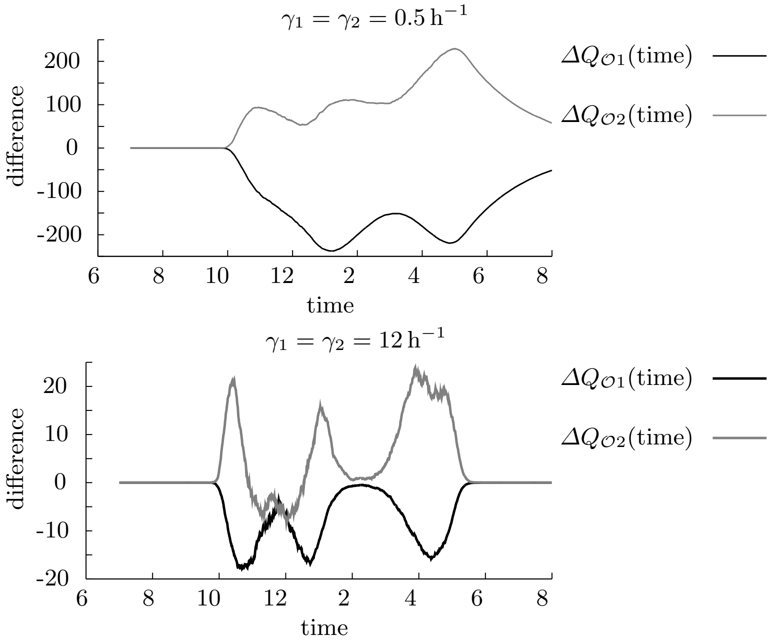

Fig. 4.55. Deviation of the number of customers in the orbit for different retrial rates and equal arrival rate functions

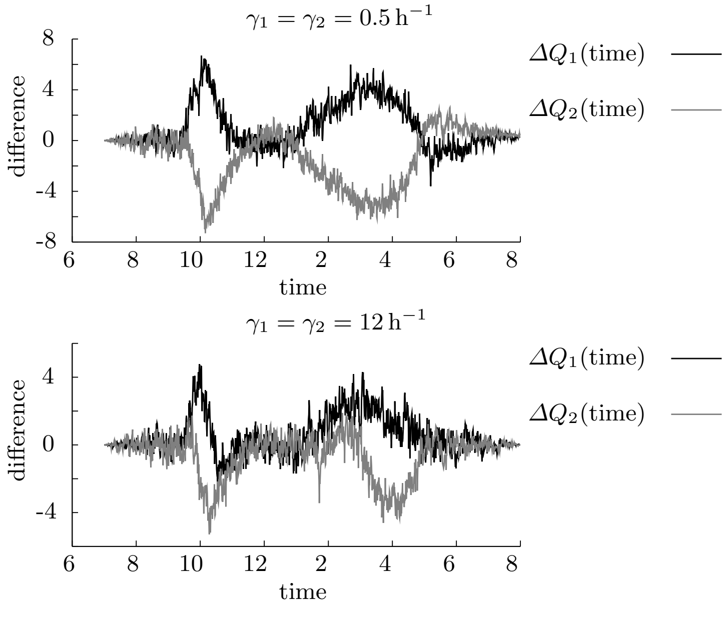

Fig. 4.56. Deviation of the number of customers in the system for different retrial rates and the arrival rate function in the first picture of Figure 4.53

If the arrival rates of the two customer classes differ, such that more type-2 customers arrive before type-1 customers the influence of the priority rule changes slightly as depicted in Figure 4.56. As before, the influence of the priority is quite small. If the retrial rate is very small in the first picture the same conclusions as for the previous figure can be drawn. However, if the retrial rate increases, i.e., the customers retry very quickly, the interaction between the priority rule and the shifted arrival rate functions becomes more obvious.

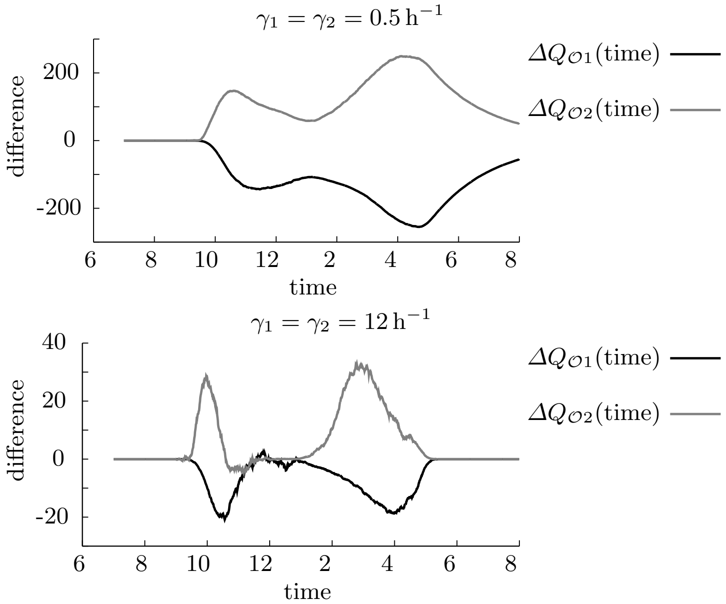

Fig. 4.57. Deviation of the number of customers in the orbit for different retrial rates and the arrival rate function in the first picture of Figure 4.53

From Figure 4.57 similar conclusions can be drawn as from Figure 4.55. The graphs of the difference in the number of customers in the orbit take course in the other direction than the graphs of the difference in the number of customers in the system. The impact of the priority rule on the number of customers in the orbit is much stronger than on the number of customers in the system.

From the last Figure 4.58 the same conclusions with respect to the influence of the priority rule can be drawn. Therefore, the graphs of the difference in the number of customers in the orbit are not shown. As observed in the previous figures the influence decreases if the customers retry more quickly. If the retrial rate is high the difference in the number of customers is linked to the arrival rate function to a larger extent.

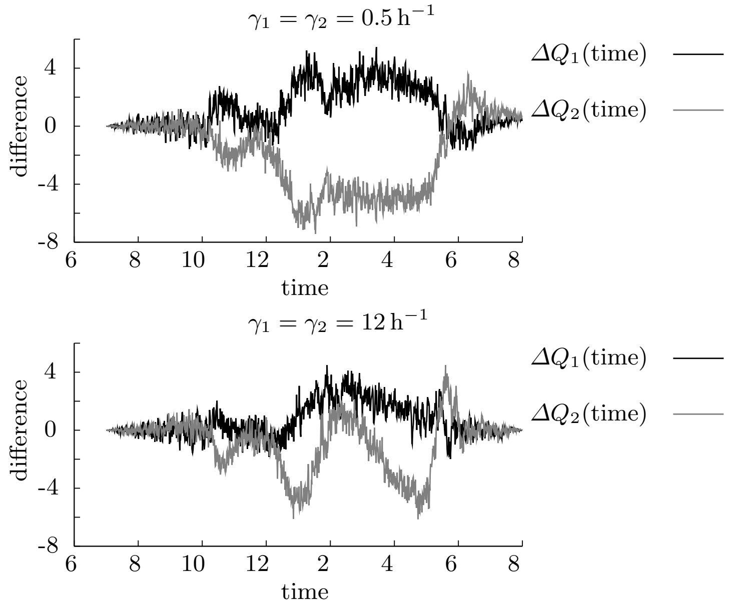

Fig. 4.58. Deviation of the number of customers in the system for different retrial rates and the arrival rate functions in the second picture of Figure 4.53

4.2.5.2 The Number of Type-2 Customers in the System and in the Orbit

In this section we conpare the simulation results and the numerical results of the fluid approximation. As the type-1 customers have preemptive priority in the fluid approximation, they are not influenced by the type-2 customers. Consequently, our main viewpoint is the change in the number of type-2 customers both in the system and the orbit. As shown in the previous section, the priority rule has at least a very small influence on the number of customers in the system. Therefore, in the simulation a non-preemptive priority was assumed as this priority rule is more reasonable for calls. The results from Section 4.1.5 apply to the results of the fluid approximation for the number of type-1 customers in this section.

In order to compare the simulation results to the numerical solution of the initial value problem given by the system of Differential Equations (4.23) for the fluid processes and the initial conditions

$$Q_1^F(t_0) = 0 \tag{4.49a}$$

$$Q_2^F(t_0) = 0 \tag{4.49b}$$

$$Q_{\mathcal{O}1}^F(t_0) = 0 \tag{4.49c}$$

$$Q_{\mathcal{O}2}^F(t_0) = 0 \tag{4.49d}$$

Furthermore, we assume the parameters given in Table 4.4 and the different arrival rates given on Pages 132f, i.e., we distinguish three cases according to the arrival rate functions of type-2 customers. For each of these cases we investigate two different retrial rates for both customer classes.

The confidence belt is calculated by means of the diffusion approximation. Therefore, we solve the extended initial value problem containing the differential equations for the variances and covariance (4.29) on Page 126 and the initial value problem for the fluid processes given by Equations (4.23) on Page 123 and (4.49) numerically. We assume initial values of zero for all variances and covariances at time t_0. Then the upper 95% confidence bound for the number of customers in the system is given by

$$\text{upper bound} = Q_S^F(t) + 1.96\sqrt{\mathbf{VAR}\big[Q_S^D(t)\big]} \tag{4.50}$$

and the lower bound by

$$\text{lower bound} = Q_S^F(t) - 1.96\sqrt{\mathbf{VAR}\big[Q_S^D(t)\big]}. \tag{4.51}$$

The confidence belts for the number of type-2 customers in the orbit are calculated equivalently.

In Figure 4.59 simulation results for the number of type-2 customers in the system and in the orbit for the case of equal arrival rate functions for both

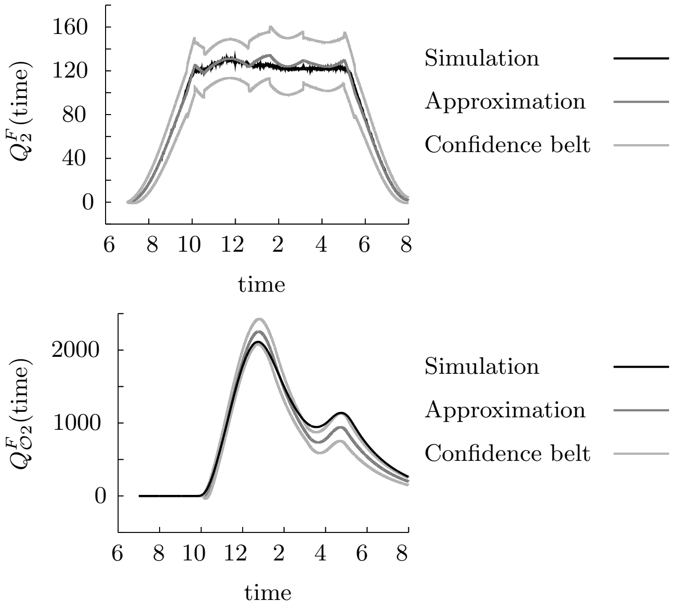

Fig. 4.59. Comparison of the number of type-2 customers in the system $Q_2(t)$ and in the orbit $Q_{\mathcal{O}2}(t)$ in case of identical arrival rate functions

customers classes are compared to the numerical solution of the fluid approximation with an 95% confidence envelope. Obviously, the approximation works well to estimate the number of type-2 customers in the system. The deviation of the results from the simulation during the period from 12 pm to 6 pm has various reasons. First of all, the deviation is due to the different priority rules as shown in the previous section. During this time interval the set S_4 described in Equation 4.25 on Page 124 is not a null set, i.e., the set might not have a measure zero. Furthermore, the differential equations react much faster to changes in the departure rate than the simulation. The simulation results are slightly overestimated by the fluid approach, but the simulation results are within the confidence belt of the diffusion approach.

During the same time interval the approximated number of type-2 customers in the orbit differs from the simulation results. Contrary to the number of customers in the system the number of customers in the orbit is underestimated, i.e., more customers might actually be waiting in the orbit. If we consider Figure 4.55 on Page 135 the difference in the simulation and approximation results can easily be explained by the different priority rules.

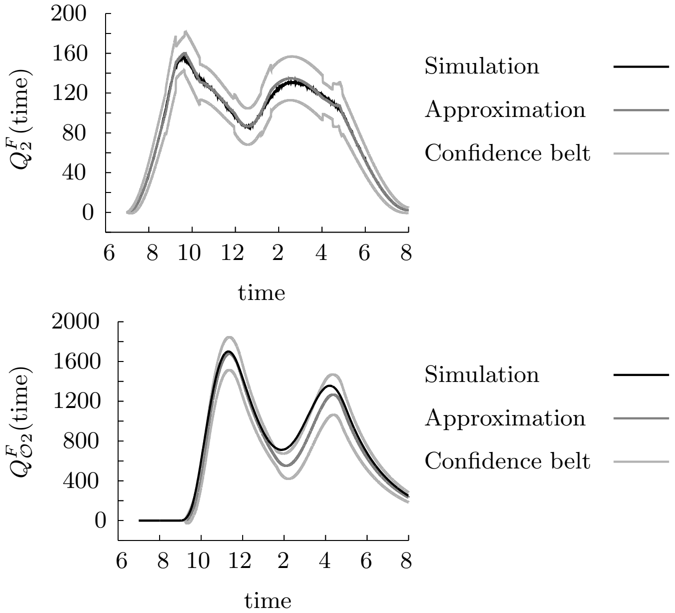

Fig. 4.60. Comparison of the number of type-2 customers in the system $Q_2(t)$ and the orbit $Q_{\mathcal{O}2}(t)$ in the case of the different arrival rate functions shown in the first picture of Figure 4.53 on Page 132

In the case of different arrival rate functions the approximation works as well as in the case of identical arrival rate functions. In Figure 4.60 the simulation results and the fluid approximation for the number of type-2 customers in the system and in the orbit are compared. The number of customers in the system is slightly overestimated during the time interval from 1 pm to 4 pm. During this interval the number of customers in the orbit is underestimated. The deviation of the simulation and approximation results is stronger for the number of customers in the orbit. This effect can be explained by critical loading and the different priority rules used in the simulation and the approximation. If we consider the first picture of Figure 4.57 on Page 136, we observe that during the time period from 1 pm to 4 pm the difference in the number of type-2 customers in the orbit rises quickly. The difference between the graphs in the second picture of Figure 4.60 is about 200 customers which corresponds to the results of the first picture of Figure 4.57.

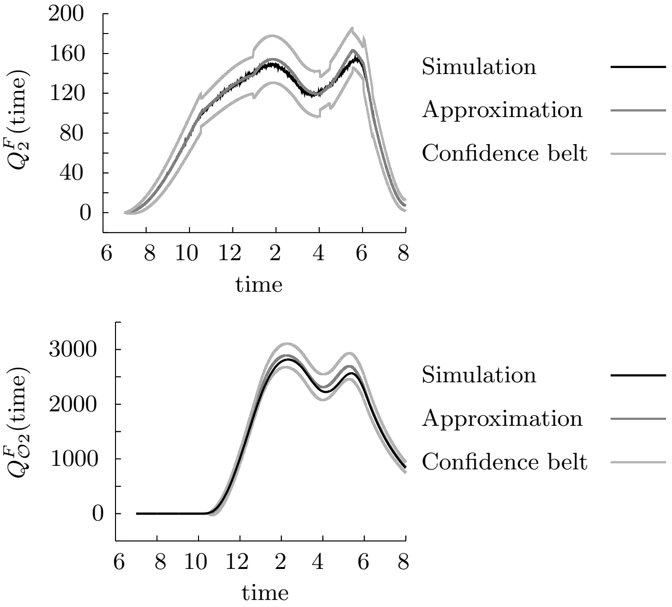

Fig. 4.61. Comparison of the number of type-2 customers in the system $Q_2(t)$ and the orbit $Q_{\mathcal{O}2}(t)$ in the case of the different arrival rate functions shown in the second picture of Figure 4.53 on Page 132

Contrary to the previous cases, the number of customers in the orbit is overestimated during the time intervals from 1:30 pm to 2:30 pm and from 4 pm to 6 pm in Figure 4.61, i.e., the fluid approach predicts more customers in the orbit than the simulation. In this case the priority rule has a minor influence, because more type-2 customers arrive after the maximum arrival rate of type-1 customers has been reached. Therefore, in the fluid approach the service of fewer type-2 customers by generalists is disrupted as in the previous cases. The generalists are already busy serving type-1 customers when most type-2 customers arrive. Consequently, they queue at the end of the queue like regular customers. Therefore, the number of customers in the orbit is mainly influenced by the arrival and retrial rate. The time intervals in which

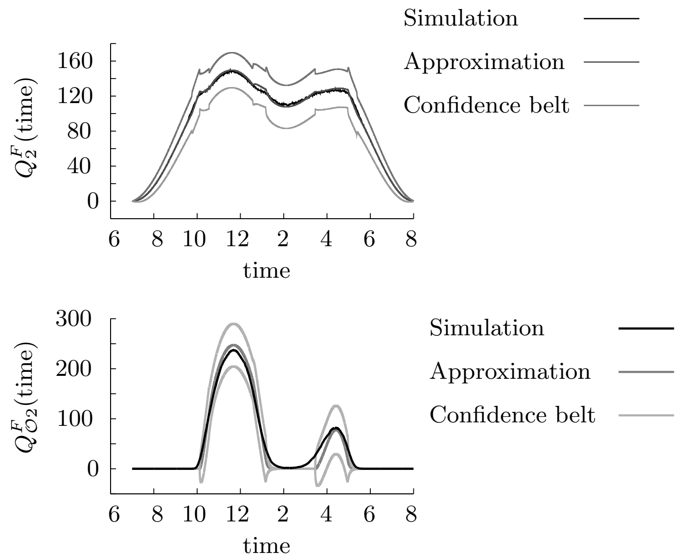

Fig. 4.62. Comparison of the number of type-2 customers in the system $Q_2(t)$and the orbit $Q_{\mathcal{O}2}(t)$ with identical arrival rate functions shown in Figure 4.52 on Page 132 and short mean times to retrial $\gamma_1^{-1} = \gamma_2^{-1} = 5\,\text{min}$

the simulation processes of the two different processes $Q_2^F(t)$ and $Q_{\mathcal{O}2}^F(t)$ are overestimated no longer coincide. As the deviation from the simulation results is small, the approximation can be judged as accurate.

Additionally, the simulation and approximation results for short mean times to retrial should be compared. Therefore, the retrial rates are assumed to be $\gamma_1 = \gamma_2 = 12\,\text{h}^{-1}$. This means that customers of both classes retry on average after five minutes in the orbit. In this case similar statements on the accuracy of the approximations can be made. The number of customers in the system $Q_2(t)$ is approximated almost exactly while the approximation of the number of customers in the orbit $Q_{\mathcal{O}2}(t)$ differs slightly from the simulation results, as shown in Figure 4.62.

However, the simulation results almost always stay within the confidence envelopes given by the diffusion refinement. During the time interval from 10 am to 12 pm the number of customers in the orbit is overestimated, while it is underestimated during the interval from 3 pm to 4 pm in Figure 4.62. The deviation again can be reasoned by the results of the previous section and by critical loading. As the calculation of the variances and covariances relies on the fluid approach, the variance of the process of the number of customers in the orbit is zero if this process is zero itself. That is why no confidence envelope is calculated during the interval from 1 pm to 3 pm.

The same effects are visible if the arrival rate functions differ. In Figure 4.63 the fluid approximation almost coincides with the simulation results for the number of customers in the system $Q_2(t)$. The approximated num-

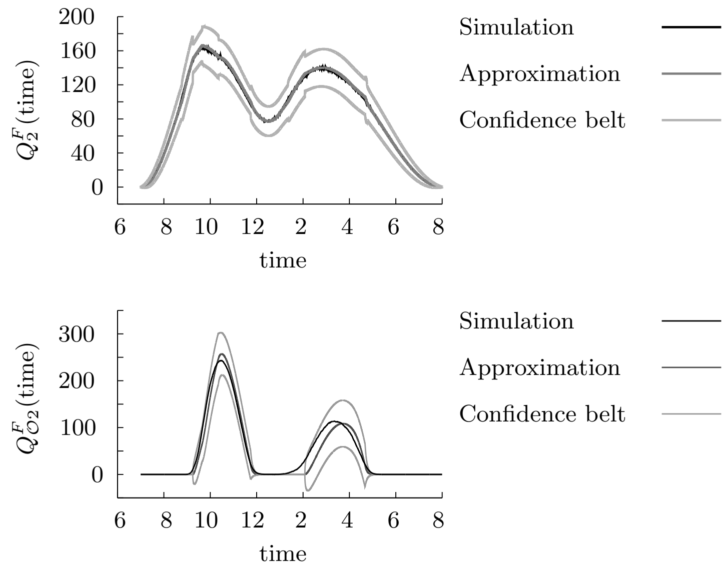

Fig. 4.63. Comparison of the number of type-2 customers in the system $Q_2(t)$ and the orbit $Q_{\mathcal{O}2}(t)$ in the case of different arrival rate functions shown in the first picture of Figure 4.53 on Page 132 and short mean times to retrial $\gamma_1^{-1} = \gamma_2^{-1} = 5\,\text{min}$

ber of customers in the orbit $Q_{\mathcal{O}2}^F(t)$ is higher than the simulated number of customers $Q_{\mathcal{O}2}(t)$ at the peak arrival rate and during the late afternoon from 3 pm to 5 pm. During the early afternoon the approximated number is smaller than the simulated values. The reason for the deviation is that the system is critically loaded.

In particular, in Figure 4.64 during the period from 1 pm to 3 pm the violation of the set conditions becomes visible for the process of the number of customers in the orbit $Q_{\mathcal{O}2}(t)$. During this time interval the simulated number of customers in the orbit is underestimated by the fluid approach. The number of customers in the system calculated by the fluid approach almost coincides with the results of the simulation. This result stresses the minor influence of the priority rule on the number of customers in the system.

Comparing the Figures 4.59, 4.60, and 4.61 on Pages 138-140 referring to the small retrial rates $\gamma_1 = \gamma_2 = 0.5\,\text{h}^{-1}$ and the Figures 4.62, 4.63, and 4.64 on Pages 141-143 referring to the high retrial $\gamma_1 = \gamma_2 = 12\,\text{h}^{-1}$ rate, the main difference can be seen in the number of customers in the orbit. In the case of small retrial rates, the number is almost ten times as high as in the case of high retrial rates. Amazingly, the number of customers in the system does not seem to differ much. If the customers recall after a short mean time in orbit the approximation of the number of customers in the system appears to be more accurate because in this case the priority rule has a minor impact on the number of customers in the system and the orbit

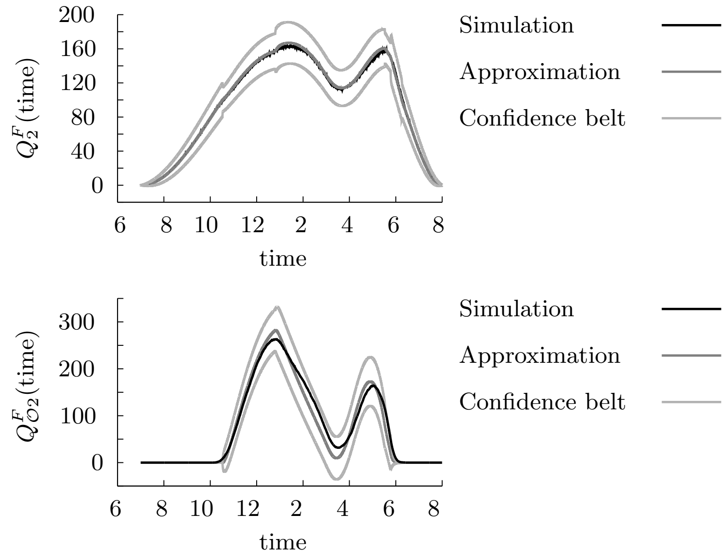

Fig. 4.64. Comparison of the number of type-2 customers in the system $Q_2(t)$ and the orbit $Q_{\mathcal{O}2}(t)$ in the case of different arrival rate functions shown in the second picture of Figure 4.53 on Page 132 and short mean times to retrial $\gamma_1^{-1} = \gamma_2^{-1} = 5\,\text{min}$

4.2.5.3 The Time-Dependent Waiting Time of Type-2 Customers

In addition to the comparison of the number of type-2 customers in the system $Q_2(t)$ and the orbit $Q_{\mathcal{O}2}(t)$, we compare the result for the waiting time of type-2 customers calculated by the fluid approach $W_2^F(t)$ and by the simulation tool $W_2(t)$. For this purpose, we use the three different arrival functions presented in Figures 4.52 and 4.53 on Page 132 and the parameters given in Table 4.4 on Page 133. The fluid approximation of the time-dependent waiting time of type-2 customers was calculated by means of Equation (4.35) on Page 128.

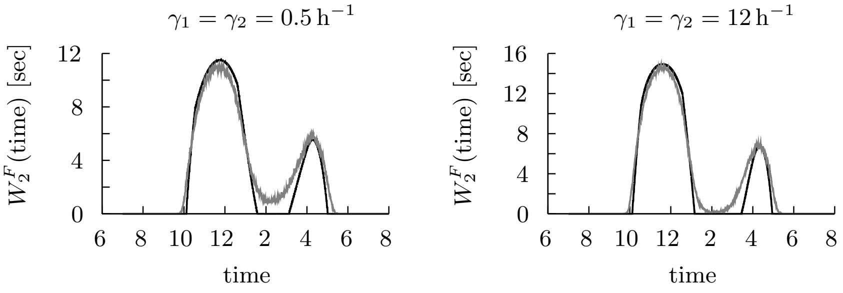

Fig. 4.65. Comparison of approximated time-dependent waiting time of type-2 customers (black line) to simulation results (gray line) with different retrial rates and identical arrival rate functions

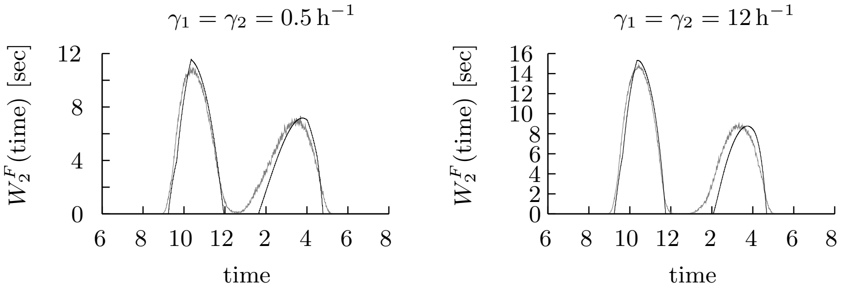

Fig. 4.66. Comparison of approximated time-dependent waiting time of type-2 customers (black line) to simulation results (gray line) with different retrial rates and different arrival rate functions shown in the first picture of Figure 4.53 on Page 132

In Figure 4.65 the waiting time of type-2 customers $W_2(t)$ calculated by the simulation tool is overestimated during the time interval from 10:30 am to 1 pm. During this time interval the maximum arrival rate of both customer classes is reached. The number of type-1 customers exceeds the number of type-1 specialists and generalists. Consequently, type-2 customers are served by their specialists only. The overestimation is due to the different priority rules. During the afternoon the simulated waiting time is underestimated by the numerical results. This underestimation is caused by the fact that the customers in the simulation model are discrete, while they are assumed to be infinitely divisible in the fluid model. However, the approximation is not far away from the simulation results. In the case of small retrial rates $\gamma = 0.5\,\mathrm{h}^{-1}$, i.e. long mean times to retrial, the deviation is slightly higher than in the case of high retrial rates.

If the arrival rate functions of type-1 and type-2 customers differ as shown in the first picture of Figure 4.53 on Page 132 the curve of the waiting time of type-2 customers has a different shape. In Figure 4.66 the curve of the waiting time has sharp bends at 10 am and in the case of small retrial rates $\gamma_1 = \gamma_2 = 0.5\,\mathrm{h}^{-1}$ additionally at 4 pm. These sharp bends result from the influence of type-1 customers on type-2 customers. The number of type-1 customers exceeds the number of type-1 specialists and generalists, such that type-2 customers are crowded out of the service by generalists. As the customers are assumed to be fluid in the approximation, the waiting time of type-2 customers increases instantaneously. During the time interval from 12:30 pm to about 4 pm the waiting time of type-2 customers is underestimated, because customers are assumed to be a continuous fluid and the traffic intensity is almost one.

Finally, in Figure 4.67 the approximated waiting times of type-2 customers $W_2^F(t)$ are compared to the simulation results $W_2(t)$, if the arrival rate functions are given according to the second picture of Figure 4.53 on Page 132. As in the previous figure some sharp bends in the curve of the

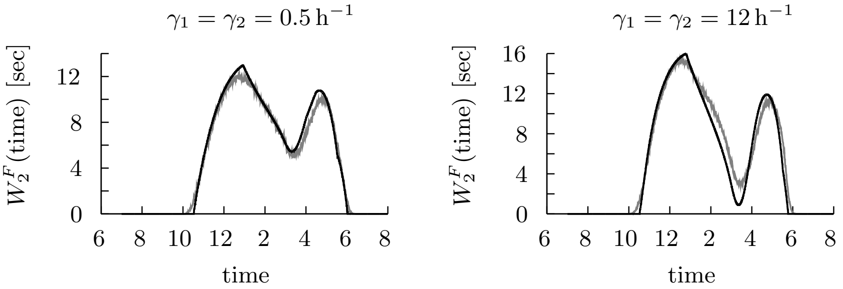

Fig. 4.67. Comparison of approximated time-dependent waiting time of type-2 customers (black line) to simulation results (gray line) with different retrial rates and the different arrival rate function shown in the second picture of Figure 4.53 on Page 132

approximated waiting time appear. But unlike the previous figure, the waiting time of type-2 customers decreases instantaneously at these bends. The reasons for the drastic change in the waiting time are due to the assumed continuity as mentioned before.

4.2.5.4 Aggregated Technical Performance Measures

The numerical investigation of the influence of the different parameters focuses on the aggregated waiting time of type-2 customers[31] and the aggregated utilisation of type-2 specialists and of generalists[32]. The performance measures of type-1 customers are not influenced by the behaviour of type-2 customers. As the influence of the other parameters on the technical performance measures of type-1 customers has been studied in the previous Section 4.1, these performance measures are not reported again[33].

We assume that the arrival rate functions are given by Figures 4.52 and 4.53 on Page 132. The default parameters of the different examples are given in Table 4.4 on Page 133, whereas we consider solely the small retrial rates $\gamma_1 = \gamma_2 = 0.5\,\mathrm{h}^{-1}$. The performance measures are aggregated over the time interval from 7 am to $T = 8\,\mathrm{pm}$.

We start to investigate the impact of the service rate of type-2 specialists μ_2 and the number of these specialists $N_2(t)$ on the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ in Figures 4.68 and 4.69. In the first Figure 4.68 both customer types have the same arrival rate function presented in Figure 4.52 on Page 132. In the second Figure 4.69 the arrival rate functions differ according to Figure 4.53 on Page 132.

[31] The aggregated waiting time is calculated by aggregating the time-dependent waiting time given in Equation (4.35) on Page 128 according to Equation (3.29) on Page 40.

[32] The aggregated utilisation is calculated by applying Equation (3.36) on Page 3.36 to Equations (4.39) and (4.40) on Page 130.

[33] See Pages 87 through 103.

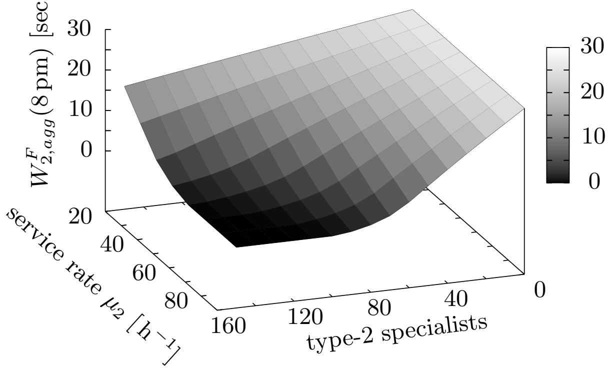

Fig. 4.68. Influence of the service rate μ_2 and the number of type-2 specialists $N_2(t)$ on the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ with equal arrival rate functions

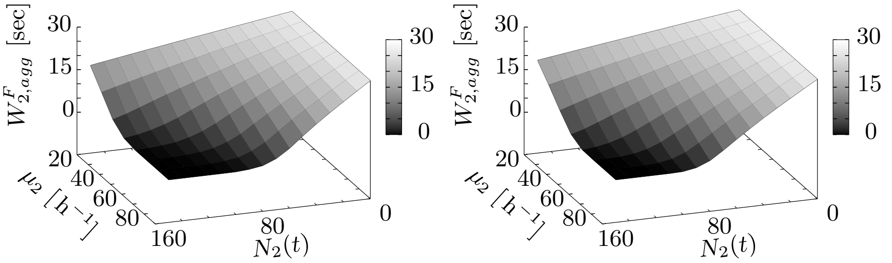

Fig. 4.69. Influence of the service rate μ_2 and the number of type-2 specialists $N_2(t)$ on the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ with different arrival rate functions

Comparing Figure 4.69 and the pictures in Figure 4.68 the curves for the aggregated waiting time for type-2 customers are almost identical. The waiting times have been aggregated by weighting the time-dependent waiting time by the departure rate as presented in Subsection 3.2.3 in Equation (3.29) on Page 40. The influence of these parameters is similar to the impact of the service rate μ_1 and the number of type-1 specialists $N_1(t)$ on the aggregated waiting times of type-1 customers presented in Figure 4.26 in the case of homogeneous customers and agents.

More interesting is the influence of the service rate μ_1 and the number of type-1 specialists $N_1(t)$ on the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ presented in Figures 4.70 and 4.71. In Figure 4.70 the arrival rate functions are identical, while in Figure 4.71 the arrival rate functions differ. Similar to the previous example the different arrival rate functions seem to have a minor influence on the aggregated waiting time of type-2 customers. If the maximum arrival rate of type-2 customers is reached before the maximum arrival rate of type-1 customers, as presented in the first picture of Figure 4.53 on Page 132, the aggregated waiting time of type-2 customers is

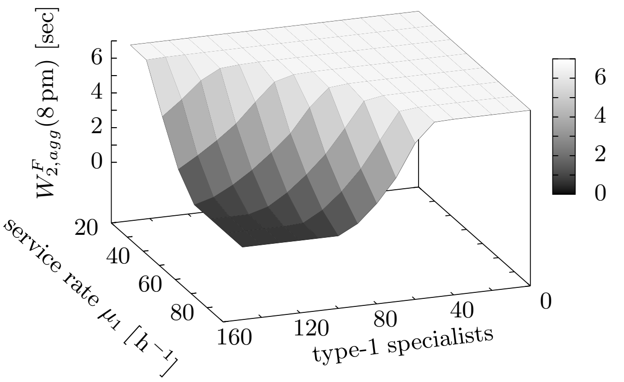

Fig. 4.70. Influence of the service rate μ_1 and the number of type-1 specialists $N_1(t)$ on the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ with equal arrival rate functions

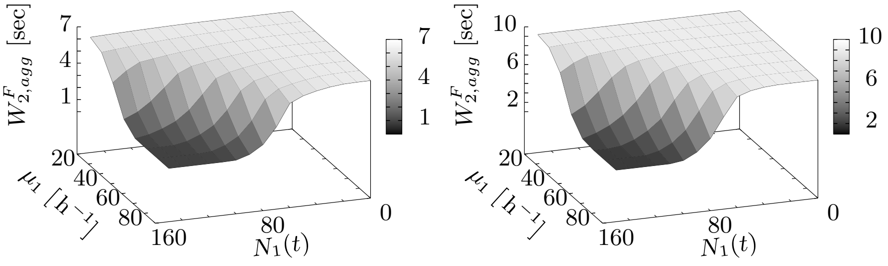

Fig. 4.71. Influence of the service rate μ_1 and the number of type-1 specialists $N_1(t)$ on the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ with different arrival rate functions

a little shorter than the aggregated waiting time if the maximum arrival rate of type-2 customers is reached afterwards.

The shorter aggregated waiting time in the case of few type-1 specialists on duty and small service rates result from the fact that a lot of type-2 customers arrive before most type-1 customers arrive. These type-2 customers are partly served by generalists and leave the system before the maximum arrival rate of type-1 customers is reached. These type-1 customers displace many type-2 customers from service by generalists. Therefore, the waiting time of type-2 customers increases. However, many type-2 customers already have been served in the first case presented in the first picture. If most type-2 customers arrive at a later time than most type-1 customers, the generalists are already occupied by type-1 customers. Therefore, most type-2 customers are attended to solely by their specialists and will have to wait for a longer time as depicted in the second picture of Figure 4.71. If more type-1 specialists are scheduled or they work much faster, generalists serve both kinds of customers. Therefore, the waiting time of type-2 customers decreases.

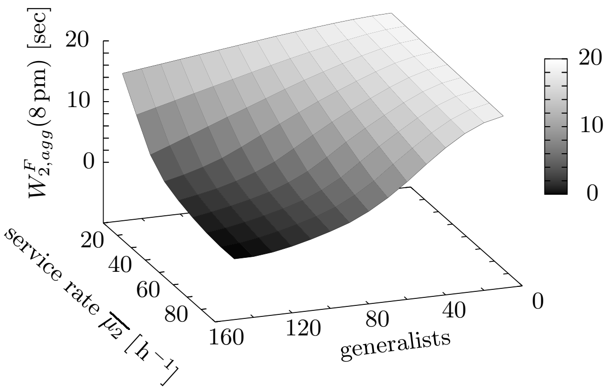

Fig. 4.72. Influence of the service rate for type-2 customers $\overline{\mu_2}$ and the number of generalists $N_G(t)$ on the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ with equal arrival rate functions

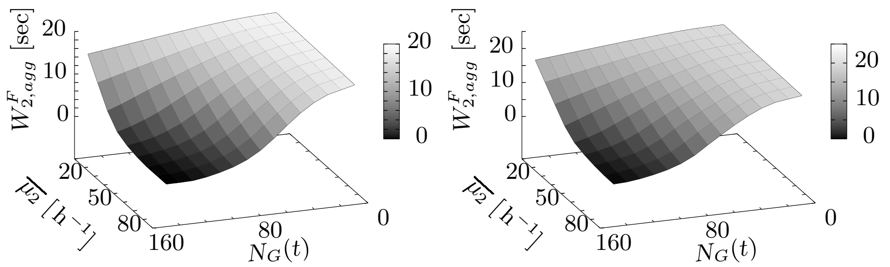

Fig. 4.73. Influence of the service rate $\overline{\mu_2}$ for type-2 customers and the number of generalists $N_G(t)$ on the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ with different arrival rate functions

The graph of the aggregated waiting time depending on the number of generalists $N_G(t)$ and the service rate $\overline{\mu_2}$ in Figures 4.72 and 4.73 of the generalists for type-2 customers does not differ much from the graph for the aggregated waiting time as the function of the number of specialists and their service rate in Figures 4.68 and 4.69 on Page 146. In these three examples the number of type-2 specialists was assumed to be 50 in order to visualise the influence of the number of generalists and their service rate more clearly.

In Figure 4.72 equal arrival rate functions for both kinds of customer classes were assumed. The graphs in Figure 4.73 refer to the arrival rate functions in Figure 4.53 on Page 132. Figure 4.73 underlines the observations of Figures 4.69 and 4.71 that the different arrival rate functions have a minor effect on the aggregated waiting time of type-2 customers.

An investigation of the other aggregated performance measures would lead to similar results. Different arrival rate functions have a big impact on the time-dependent performance measures but the influence on the aggregated

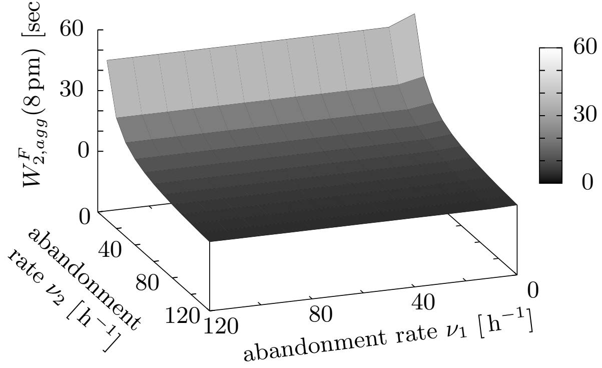

Fig. 4.74. Influence of the abandonment rates ν_1 and ν_2 on the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ with equal arrival rate functions

performance measures is very small. Therefore, we assume equal arrival rate functions as depicted in Figure 4.52 on Page 132 for the following examples and do not report the results of the aggregated performance measures with different arrival rate functions.

In Figure 4.74 the aggregated waiting time is a function of the abandonment rates of type-1 ν_1 and type-2 customers ν_2. If the type-2 customers become more impatient, i.e., the abandonment rate increases, the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ decreases until no customer has to wait at all. Contrary to the service rate μ_1 and the number of type-1 specialists $N_1(t)$, the abandonment rate ν_1 of type-1 customers seems to have no influence on the aggregated waiting time of type-2 customers although more patient customers of type 1 mean more customers in the system. Therefore, fewer customers of the second type can be served by generalists, so that they have to wait longer for an available agents. Only if type-1 customers are so patient that no type-1 customer abandons at all, the aggregated waiting time of type-2 customers increases slightly.

In Figure 4.75 the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ is presented as a function of the retrial parameters, γ_2 and p_2, of this customer class. The only difference between Figure 4.75 and Figure 4.28 on Page 101 is that the waiting time in Figure 4.28 is a little longer. This is due to the fact that in the homogeneous case 100 agents have been scheduled, while in Figure 4.75 100 type-2 specialists plus 50 generalists have been working. The aggregated waiting time increases if a higher percentage of customers retries and the retrial rate grows, i.e., the mean time to retrial shortens.

In order to intensify the effects of changing retrial rates γ_1 and probabilities p_1 of type-1 customers on the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$, the number of type-1 specialists $N(t)$ is assumed to be 80 in Figure 4.76. Although in this case many type-1 customers abandon, the influence of the retrial parameter on the aggregated waiting time is almost negligible.

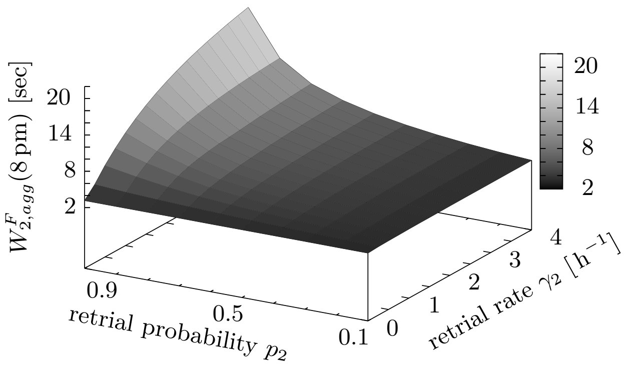

Fig. 4.75. Influence the retrial rate γ_2 and retrial probability p_2 of type-2 customers on the waiting time of type-2 customers $W^F_{2,agg}(T)$ with equal arrival rate functions

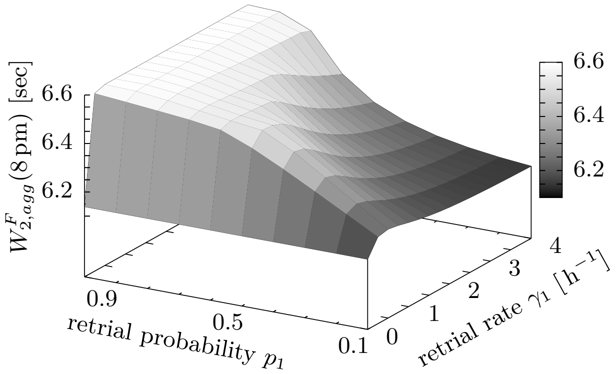

Fig. 4.76. Influence of the retrial rate γ_1 and retrial probability p_1 of type-1 customers on the aggregated waiting time of type-2 customers $W^F_{2,agg}(T)$ with equal arrival rate functions

If the retrial probability and the retrial rate increase, the aggregated waiting time increases by tenths of a second. Amazingly, the waiting time increases if the retrial rate increases from zero to $0.5\,\mathrm{h}^{-1}$. Thereafter the aggregated waiting time decreases again. The reason for this decrease might be that less work related to type-1 customers is shifted into periods of low load, because type-1 customers recall already when many primary arrivals occur.

As the influence of the different parameters on the number of type-2 customers and their aggregated waiting times is almost the same as the effects presented and explained in the previous Section 4.1, the graphs depicting the impact of these parameters on the probability of being served are not presented. Next the aggregated utilisation of the type-2 specialists as a function of the service rate and number of type-1 specialists and the number of gen-

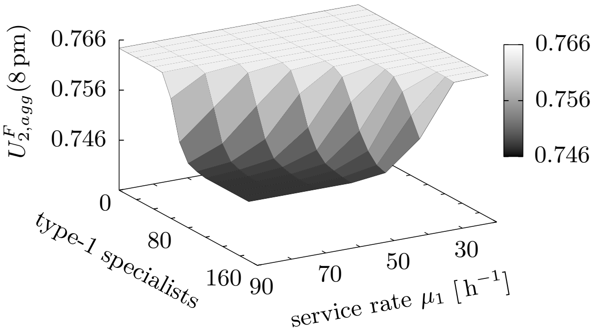

Fig. 4.77. Influence of the service rate μ_1 and the number of type-1 specialists $N_1(t)$ on the utilisation of type-2 specialists $U^F_{2,agg}(T)$

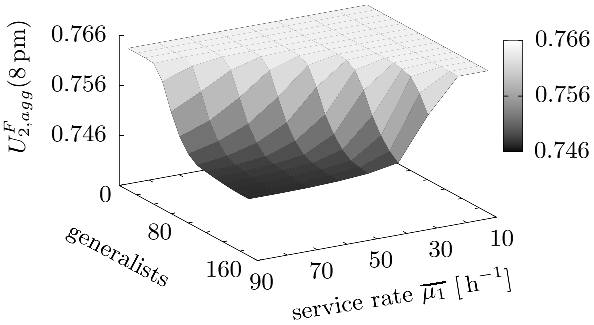

Fig. 4.78. Influence of the service rate for type-1 customers $\overline{\mu_1}$ and the number of generalists $N_G(t)$ on the aggregated utilisation of type-2 specialists $U^F_{2,agg}(T)$

eralists and their service rate for both kinds of customer classes is studied. Afterwards the aggregated utilisation of generalists is considered.

In Figure 4.77 the aggregated utilisation of type-2 specialists $U^F_{2,agg}(T)$ is a function of the number of type-1 specialists $N_1(t)$ on duty and their service rate μ_1. As argued on Page 149 we use for the calculation of the aggregated utilisation identical arrival rate functions for both customer classes as depicted in Figure 4.52 on Page 132. The parameters seem to have a minor influence on the aggregated utilisation of type-2 specialists. If few type-1 specialists are serving customers slowly, type-1 customers replace customers of the second type from the generalists. If enough type-1 specialists are scheduled or the service rate increases, the aggregated utilisation of type-2 specialists reduces, because more type-2 customers are served by generalists. However, the reduction of the utilisation is 2%, which is diminishingly small, because customers are served by a specialist first if a specialist is available.

A similar conclusion with respect to the influence of the other agents groups and their service rates can be drawn from Figure 4.78. This figure depicts the aggregated utilisation of type-2 specialists $U^F_{2,agg}(T)$ as a function

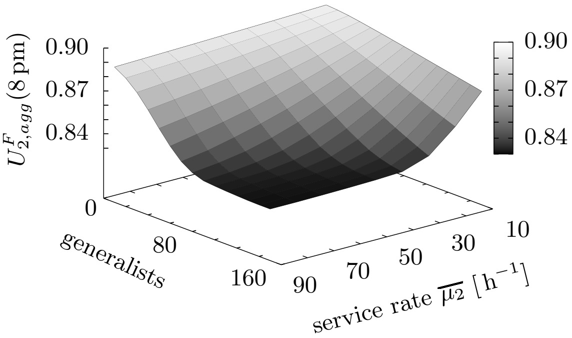

Fig. 4.79. Influence of the service rate $\overline{\mu_2}$ for type-2 customers and the number of generalists $N_G(t)$ on the utilisation of type-2 specialists $U^F_{2,agg}$

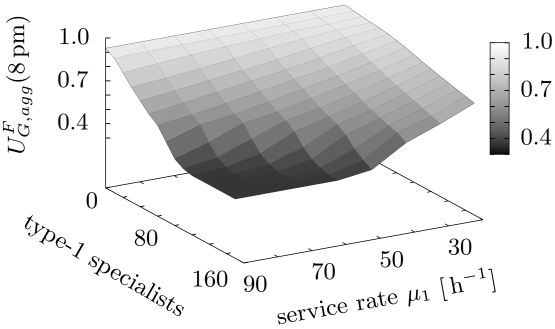

Fig. 4.80. Influence of the service rate μ_1 and the number of type-1 specialists $N_1(t)$ on the aggregated utilisation of generalists $U^F_{G,agg}$

of the number of generalists $N_G(t)$ and their service rate for type-1 customers $\overline{\mu_1}$. If the service rate and the number of generalists increase, the utilisation of type-2 specialists decreases slightly by about 2%, because more generalists can serve type-2 customers. Consequently, less type-2 customers have to wait and type-2 specialists have to serve fewer customers.

Finally, in Figure 4.79 the number of generalists $N_G(t)$ and the service rate for type-2 customers $\overline{\mu_2}$ have a higher impact on the utilisation of type-2 specialists. If more generalists are on duty, the utilisation decreases, because more type-2 customers are attended to by generalists and less customers have to wait. If the service rate of generalists for type-2 customers increases, the relief of type-2 specialists becomes even stronger.

As shown in Figure 4.25 on Page 98 the retrial rate does not influence the utilisation of agents. This is also true in the case of heterogeneous customers and agents. Therefore, we do not present figures showing the aggregated utilisation of type-2 specialist and generalists as functions of the retrial rate and the probability of retrial.

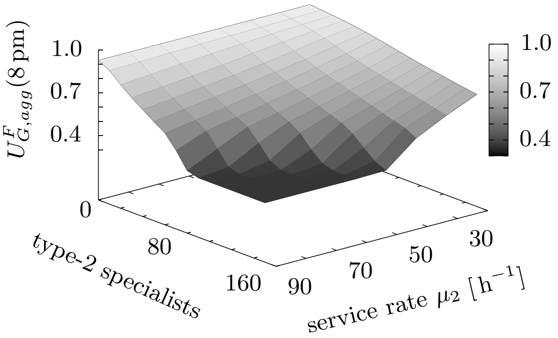

Fig. 4.81. Influence of the service rate μ_2 and the number of type-2 specialists $N_2(t)$ on the aggregated utilisation of generalists $U_{G,agg}^F$

We continue the investigation of the influence of the different parameters of the contact center model on the performance measures with the study of the aggregated utilisation of generalists $U_{G,agg}^F(T)$.

Figure 4.80 and Figure 4.81 are quite similar. They present the influence of the numbers of the different kinds of specialists $N_1(t)$ and $N_2(t)$ and their service rates μ_1 and μ_2 on the aggregated utilisation of generalists $U_{G,agg}^F(T)$. An increasing number of specialists and increasing service rates lead in both cases to an enormous reduction of the utilisation of generalists. The utilisation never falls below 0.3 because some generalists are needed to serve the customers of the other class. The strong impact of the number of specialists and their service rates is due to the fact that customers prefer to be served by the specialists. Therefore, the customers will be served by a specialist as soon as a specialist becomes available. In the fluid model this effect is intensified, because the customers are assumed to be a continuous fluid. In the fluid model every amount of customers that can be processed by specialists is served by specialists. The remainder is served by generalists, waits or abandons.

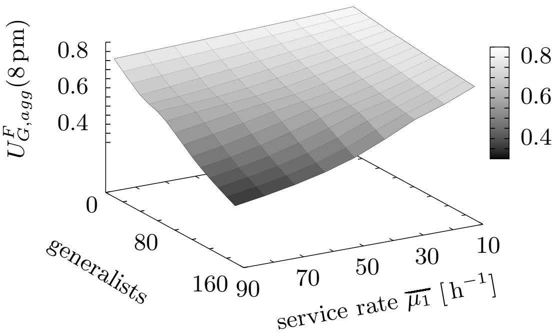

Fig. 4.82. Influence of the service rate for type-1 customers $\overline{\mu_1}$ and number of generalists $N_G(t)$ on the aggregated utilisation of generalists $U_{G,agg}^F$

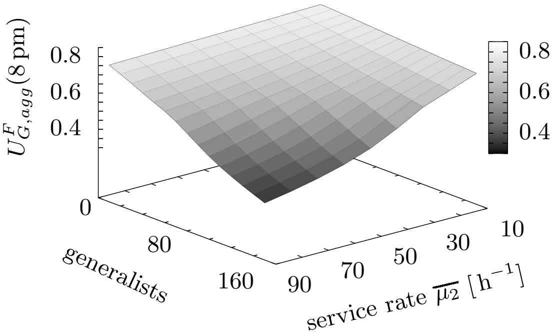

Fig. 4.83. Influence of the service rate for type-2 customers $\overline{\mu_2}$ and number of generalists $N_G(t)$ on the utilisation of generalists $U^F_{G,agg}$

Equivalently, the service rates of generalists for the different customer classes have similar effects on the aggregated utilisation of the generalists $U^F_{G,agg}(T)$ as shown in Figure 4.82 and 4.83. Contrary to the other investigations, we assume a constant number of 50 type-1 specialists in Figure 4.82 and a constant number of 50 type-2 specialists in Figure 4.83 to stress the effects of a changing number of generalists. A growing number of generalists leads to a decrease of the aggregated utilisation as well as growing service rates.

In this section we have observed that the influence of parameters referring to type-1 agents and customers on the aggregated waiting time of type-2 customers and the aggregated utilisation of type-2 specialists is very limited as long as most type-2 customers are served by their specialists. However, the influence of service rates and the number of specialists of both types on the utilisation of generalists is strong. This observation can be explained by the preference of customers for service by specialists.

4.2.5.5 Economical Performance Measures

In this section the profit function (4.43) on Page 131 is investigated with respect to the number of agents and their service rates. Therefore, we assume the revenue and cost parameters given in Table 4.5.

Revenue for served type-1 customers	$r_1 =$ 0.5 €
Revenue for served type-2 customers	$r_1 =$ 0.5 €
Hourly costs for occupied trunks	$\ell =$ 6.0 €
Hourly wages of type-1 specialists	$w_1 =$ 10.0 €
Hourly wages of type-2 specialists	$w_2 =$ 10.0 €
Hourly wages of generalists	$w_G =$ 11 €

Table 4.5. Cost and revenue parameters for the investigation of the profit function

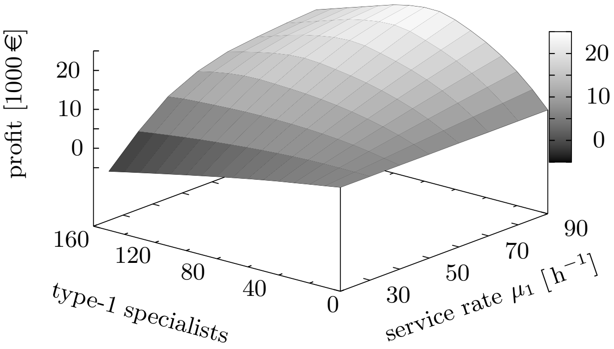

Fig. 4.84. Influence of the service rate μ_1 and number of type-1 specialists $N_1(t)$ on the daily profit with equal arrival rate functions

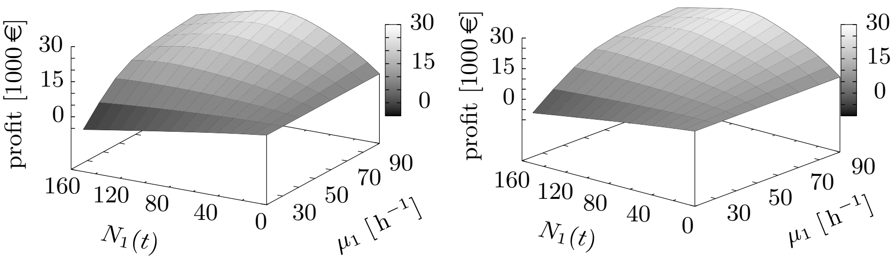

Fig. 4.85. Influence of the service rate μ_1 and number of type-1 specialists $N_1(t)$ on the daily profit with different arrival rate functions

We consider the profit for one working day starting at 7 am and ending at 8 pm and for the three arrival rate functions $\lambda_1(t)$ and $\lambda_2(t)$ presented in Figures 4.52 and 4.53 on Page 132. The parameters of the arrival rate functions are given in Equations (4.44) through (4.47). The other parameters of the examples are given in Table 4.4 on Page 133 with retrial rate $\gamma_1 = \gamma_2 = 0.5\,\text{h}^{-1}$ for both customer classes.

In Figures 4.84 and 4.85 the impact of the number of type-1 specialists $N_1(t)$ and their service rate μ_1 is depicted. The three pictures are almost identical with respect to the shape of the curves. Similar to the aggregated performance measures, the shape of the arrival rate functions and their location to each other seem to have a minor influence on the profit of one day. This effect may be due to the constant number of agents staffed for one day, as assumed here. If the number of agents can be adjusted to the demand, the influence of the different arrival rates may become visible. However, we show the profit function for all cases to have a closer look at the shape of the profit function.

The maximum values of the different profit functions in Figures 4.84 and 4.85 are similar. In the first Figure 4.84 the arrival rate functions of both customer classes are the same, while in Figure 4.85 the arrival rate functions differ. The first graph refers to the arrival rate function presented in the first picture of Figure 4.53 on Page 132 and the second graph to the second picture of Figure 4.53. If the arrival rate functions differ, the daily profit is slightly higher, because fewer agents are needed to serve customers. Furthermore, fewer customers have to wait. These two facts give rise to lower costs for paying agents and occupied lines.

The faster the agents work, i.e., the higher the service rate is, the fewer agents are needed to reach the maximum of the profit function associated with this service rate. If, e.g., the service rate is 90 h^{-1}, about 80 agents of type 1 suffice to reach the maximum. If more agents are scheduled, the profit decreases again. Contrarily, in the case of a small service rate more than 160 agents are needed to approach the optimum profit.

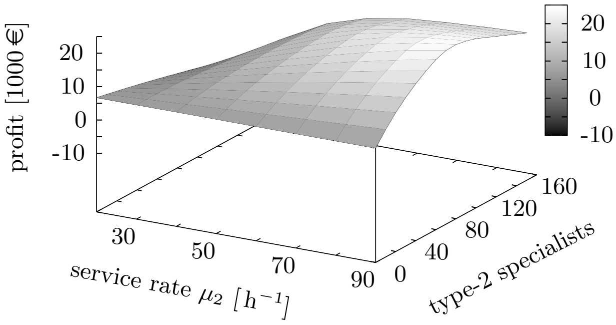

Fig. 4.86. Influence of the service rate μ_2 and number of type-2 specialists $N_2(t)$ on the daily profit with equal arrival rate functions

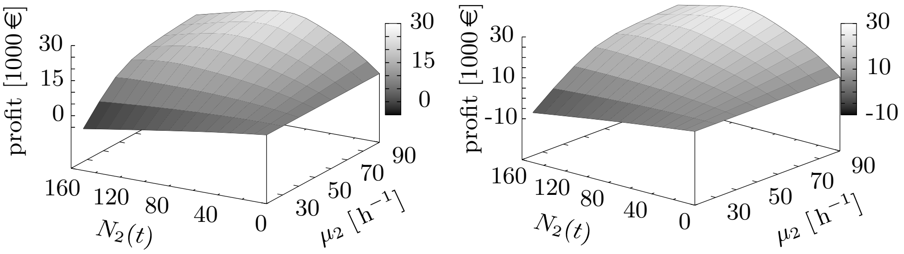

Fig. 4.87. Influence of the service rate μ_2 and number of type-2 specialists $N_2(t)$ on the daily profit with different arrival rate functions

Similar conclusions with respect to the impact of the number of type-2 specialists $N_2(t)$ and their service rate μ_2 can be drawn from Figures 4.86 and 4.87 as for the number of type-1 specialists and their service rate. For the first Figure 4.86, which represents the case of equal arrival rate functions for both kinds of customers, another perspective on the curve was chosen for a closer look at the profit curve around the maximum value. As in Figure 4.84 the graph of the profit function seems to be smooth and concave. The graphs in the second figure underline that the arrival rate functions have a minor influence on the profit. If the service rate increases fewer type-2 specialists are needed to reach the maximum of the profit function.

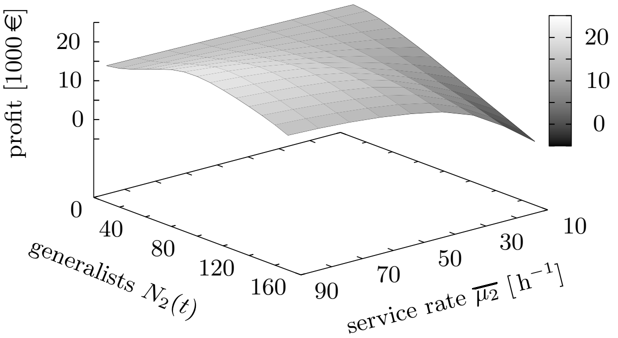

Fig. 4.88. Influence of the service rate $\overline{\mu_2}$ for type-2 customers and number of generalists $N_G(t)$ on the daily profit with equal arrival rate functions

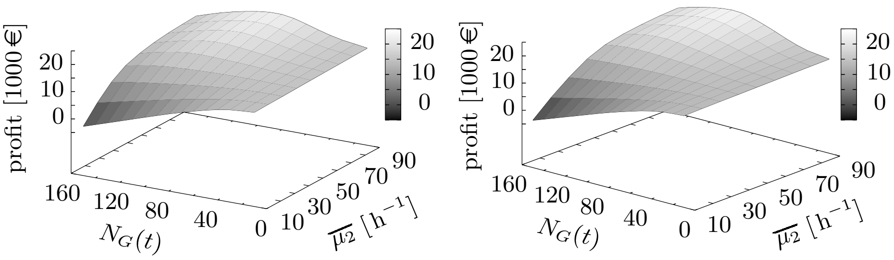

Fig. 4.89. Influence of the service rate $\overline{\mu_2}$ for type-2 customers and number of generalists $N_G(t)$ on the daily profit with different arrival rate functions

In Figures 4.88 and 4.89 the number of generalists $N_G(t)$ and their service rate $\overline{\mu_2}$ for type-2 customers is varied. In this case we assumed 50 type-1 specialists and 50 type-2 specialists to intensify the effects of a varying number of generalists. For low service rates an increase of the number of generalists leads to a smaller profit. In this case it is counterproductive to staff any additional generalist. If the agents work very quickly, the profit increases firstly in an almost convex manner for a growing number of generalists. If more than 50 generalists are staff the curve becomes concave.

4.3 Literature on Retrial Queues

Already in 1957 Cohen (1957) studied the influence of repeated calls in telephone traffic. Since 1957 a lot of work has been done on investigating single as well as multiserver retrial systems. A detailed overview on models investigated up to 1997 can be found in Falin (1990) and Falin and Templeton (1997).

Most models were examined with respect to deriving ergodicity conditions[34] and stationary distributions for performance analysis. As the calculation of stationary distributions is often difficult and sometimes even impossible because of the multi-dimensionality of the state space, numerical algorithms as well as approximations of the stationary distribution have been developed. Ramalhoto and Gómez-Corral (1998) develop a decomposition formula for the steady state distribution of a multiserver system with impatient customers and an infinite number of sources. As the generator is a matrix of matrices, Hanschke (1999) deploys matrix continued fractions, Diamond and Alfa (1995, 1999) and Anisimov and Artalejo (2001) utilise the matrix-geometric method developed by Neuts (1981).

Other numerical exact and approximative algorithms where developed by Falin and Artalejo (1995), Artalejo (1995), Dudin and Klimenok (1999), Artalejo and Pozo (2002), Chakravarthy and Dudin (2002), and Almási et al. (2004).

Although some authors, e.g., Rodrigo et al. (1998) use the time-dependent modelling of the retrial system to develop steady state approximations, the behaviour of these time-dependent models is not studied. Only few authors considered time-dependent queueing systems with repeated attempts. Grier et al. (1997) investigate a multiserver retrial system without extra waiting places, a so-called Erlang-loss system. They approximate the time-dependent mean number of customers in the system and the times of peak blocking. Furthermore, they compare their approximation results to simulation results. The time-dependent blocking probabilities are used to predict the number of lines required to fulfil given blocking limits.

Mandelbaum et al. (1999a,b, 2002) investigate a multiserver-retrial system with homogeneous agents serving impatient customers as described in Section 4.1. By means of fluid and diffusion approximations they analyse the queue and the virtual waiting time process of a single customer class. An ad-

[34] see, e.g., Afanas (1994); Dudin and Klimenok (1999); Artalejo and Lopez-Herrero (2000); Falin and Gomez-Corral (2000); Breuer et al. (2002)

vantage of these methods is that the transient behaviour under high loads can be modeled as well[35].

Besides the papers analysing retrial queues, some effort has been put into estimating the retrial parameters from call center data by Hoffman and Harris (1986) and Aguir et al. (2004). Hoffman and Harris (1986) use a stationary approach to evaluate the percentage of blocked or abandoning customers who recall. In an empirical study about the usage of a taxpayer telephone information service they find that the average daily percentage of retrials is almost stable with a value of about 69%.

Aguir et al. (2004, 2005) deploy stationary analysis and the fluid approach to estimate the retrial rate. They show that the retrial rate of the fluid model is independent of the abandoning parameter. Furthermore, the retrial rate of the stochastic analysis converges to the retrial rate of the fluid model if the primary arrival rate increases. In the second paper Aguir et al. (2005) investigate the influence of retrials of blocked calls on the performance and the staffing decision.

To consider retrials, different assumptions with respect to the reasons for retrials of customers and the mean time to retrial are made. Some authors assume that a recall may occur if a customer is blocked. These authors mainly suppose that there is no extra waiting space, e.g. Grier et al. (1997) and Aguir et al. (2005). Other authors[36] assume that the customers are impatient and balk or renege. The third reason for retrial is a so-called return of a customer who has been served completely or partly as investigated by de Véricourt and Zhou (2005). All these papers have in common that a single class of calls or customer types is studied.

Contrarily, Kalyanaraman and Srinivasan (2003) investigate a single server system with two types of customers. Customers of the first type have priority over the other customers. If the server is busy customers of the first type wait in an infinite capacity queue, while customers of the second type are blocked and have to retry. The joint stationary distribution is obtained. Heterogeneous sources are considered by Almási et al. (2004) as well. They study a computer network with a finite number of sources by means of stationary analysis.

As retrials are a very complex problem in the contact center and telecommunication industry, a lot of work has been done on this topic in order to estimate the amplitude of the retrial parameters and to model the impact of these parameters on the performance. Most papers focus on a stationary analysis of the underlying queueing models. However, retrials often lead to strong interdependencies of subsequent periods, which are addressed in the papers using fluid and diffusion models. The analysis of the two queueing systems with retrials done in this chapter belongs to these papers.

[35] See also Anisimov and Atadzhanov (1994)

[36] See, e.g., Artalejo and Lopez-Herrero (2000); Fayolle and Brun (1988); Mandelbaum et al. (2002).

5

Personnel Staffing and Shift Scheduling based on Fluid Models

5.1 Formulation of a Basic Staffing and Shift Scheduling Problem

In the previous chapter we utilised the fluid approach and the diffusion refinement to analyse different kinds of contact centers with impatient customers and retrials. This type of analysis of a contact center is needed to determine the performance of a contact center and to detect room for improvement. Most parameters like the arrival rate, the service rate, the abandonment rate, or the retrial rate are hardly influenceable by the management. However, the number of agents staffed at each moment in time is a decision variable. The staffing of agents and scheduling of shifts associated with agents belongs to the operational planning process presented in Figure 2.6 on Page 19.

We will utilise the fluid approach to solve an integrated staffing and shift scheduling problem linked to the second and third phases of the operational planning process. In other words, we formulate a mathematical problem, which includes the initial value problem of the fluid approximation as an additional constraint in order to determine the optimal number of shifts needed. In this model the staffing of agents in isolated time intervals appears to be a special case of the more general scheduling of shifts which was defined by Ernst et al. (2004) as the problem of "selecting a set of the best shifts from a (large) pool of candidate shifts on a single day". In general, this problem itself is NP-complete[1].

We consider profit-orientated problems with a revenue r per served customer. The aggregated probability of being served is considered implicitly in the objective function as shown in Equation (3.41) on Page 43[2].

[1] Musliu (2001), Musliu et al. (2004) and Fukunaga et al. (2002).

[2] Other formulations of the staffing problems with respect to time-dependent revenues or costs can be found, e.g., in Helber and Stolletz (2004, pp. 46-55), Koole and Pot (2005), or Hampshire and Massey (2005) and references therein.

As in the previous chapter, we investigate the staffing and shift scheduling problem of the contact center with homogeneous agents and customers first. Then we analyse the problem for heterogeneous agents and customers. Customers are assumed to be impatient and some of them retry after waiting in the orbit.

For each occupied phone line the contact center has to pay an amount of ℓ per hour. Furthermore, the hourly salaries of agents are denoted by w. For the staffing and shift scheduling problem the time horizon is divided into a set of time intervals $j = 1, \ldots, \mathcal{J}$ of equal length δ. A typical length of the time interval in practice is thirty minutes[3]. During these time intervals the number of agents is fixed, i.e., the number of agents in subsequent intervals can be changed only at the beginning of each time interval. The j-th time interval is the time interval which starts at time t_{j-1} and ends at time t_j.

Each agent works according to a specific shift. A shift is a sequence of subsequent time intervals during which an agent is assumed to be present. It contains both working and resting time intervals. A shift of type k, $k = 1, \ldots, K$, is represented by a $\mathcal{J}$-dimensional, boolean vector

$$\boldsymbol{s}_k = (s_{k,1}, \ldots, s_{k,\mathcal{J}}). \tag{5.1}$$

Each entry $s_{k,j}$ in this shift vector equals 1, i.e. ***true***, if an agent of shift type k is on duty during the time interval j ($j = 1, \ldots, \mathcal{J}$). Equivalently, each shift is represented by an indicator function $s_k(t)$. The indicator function is one or ***true*** when an agent is on duty at time t, and zero, i.e. ***false***, otherwise. In Figure 5.1 and Equation (5.2) a long shift starting at 7 am and ending at 3 pm with a break from 10:30 am to 11 am and a short shift starting at 3:30 pm and ending at 7:30 pm without a break are shown.

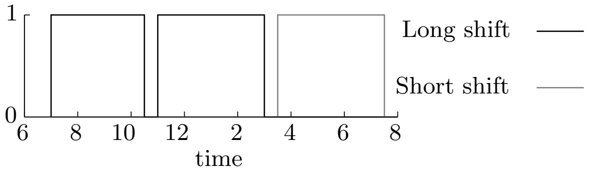

Fig. 5.1. Representation of a long shift including a rest and short shift without a rest

$$\begin{aligned} \boldsymbol{s}_{\text{long}} &= (1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0) \\ \boldsymbol{s}_{\text{short}} &= (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,0) \end{aligned} \tag{5.2}$$

The indicator function is related to the shift vector by

$$s_k(t) = s_{k,j} \quad \text{for } t_{j-1} \leq t \leq t_j \text{ and } k = 1, \ldots, K. \tag{5.3}$$

[3] See Gans et al. (2003) or Koole and Pot (2006).

We aim to determine a shift schedule, such that the profit of the contact center is maximised. A schedule is defined as a set of different shifts and the number of agents working according to each shift. We assume that for each chosen shift out of the set of possible shifts

$$\mathbb{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_K\} \tag{5.4}$$

an agent is available to work according to this shift, i.e., we do not consider the problem of assigning agents with difference preferences to the shifts. We denote a schedule by a vector

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_K) \in \mathbb{Z}^K \tag{5.5}$$

with x_k representing the number of agents working according to shift type k.

The number of agents on duty $N(\boldsymbol{x}, t)$ at time t is a function of the schedule, i.e, is the sum of all agents whose shifts cover the considered time. Formally, we have

$$N(\boldsymbol{x}, t) = \sum_{k=1}^{K} s_k(t) x_k \quad \text{for all } t \in \mathbb{R}_0^+. \tag{5.6}$$

As the number of agents is discrete and the shift function can only take values zero or one in each time interval $[t_{j-1}, t_j]$ for $j = 1, \ldots, \mathcal{J}$, it follows that $N(\boldsymbol{x}, t)$ is an integer valued step function. An example of the function of the number of agents at each time t is given in Figure 5.2.

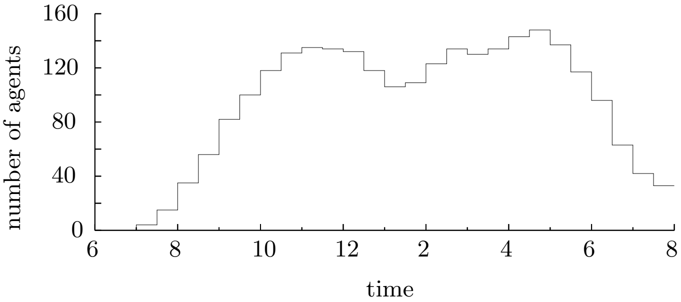

Fig. 5.2. Example of the step function $N(\boldsymbol{x}, t)$ representing the number of agents at each moment t during the working day

Each shift causes costs c_k for the wages of the agents working according to the shift type k. The hourly wage of an agent is w, which is assumed to identical for all shift types. However, we can easily extend the calculation to the case of different wages for each shift type. Therefore, the cost associated with shift type k is given by the time δ the agent is on duty in hours multiplied by the hourly wage, i.e,

$$c_k = \sum_{j=1}^{J} w \delta s_{k,j} \quad \text{for all } k = 1, \ldots, K. \tag{5.7}$$

Under these assumptions and the notation of Section 4.1, the profit function for a contact center with one customer class and a single group of homogeneous agents is the cumulative number of served customers minus the costs for occupied trunks and salaries[4], i.e.,

$$\text{profit}(\boldsymbol{x}, T) = \int_0^T r\mu \min\bigl\{Q^F(\boldsymbol{x}, t), N(\boldsymbol{x}, t)\bigr\} - \ell Q^F(\boldsymbol{x}, t)\, dt - \sum_{k=1}^{K} c_k x_k. \quad (5.8)$$

Other authors[5] also consider penalties for abandonment, blocking, or waiting of customers. From a management accounting perspective this is extremely problematic as these are non-observable quantities.

The number of customers in the system as well as the number of customers served is determined by means of the initial value problem given by the differential equations in Equation (4.1) on Page 75 of the fluid approach and some initial conditions

$$Q_S^F(\boldsymbol{x}, t_0) = Q_{S,0} \quad \text{and} \quad Q_{\mathcal{O}}^F(\boldsymbol{x}, t_0) = Q_{\mathcal{O},0}. \quad (5.9)$$

In general, we assume that the orbit and the system are empty at the beginning of the observation.

In real world call centers the total number of agents who can be staffed is usually limited. Therefore, we introduce another condition which limits the total number of agents by some fixed number M. As each entry in the schedule vector represents the number of agents scheduled for a special shift type, the total number of agents is the sum of all entries, i.e.

$$\sum_{k=1}^{K} x_k \leq M. \quad (5.10)$$

As mentioned before, the number of agents for shift type k must be nonnegative and integer valued, i.e.,

$$x_k \geq 0, \quad x_k \in \mathbb{N}_0 \quad \text{for all } k = 1, \ldots, K \quad (5.11)$$

By means of Equations (5.6) through (5.11) we can now formulate the integrative staffing and shift scheduling problem for the contact center with retrials of impatient customers and homogeneous customers and agents. In this thesis we aim to determine a shift schedule $\boldsymbol{x}$ which maximises the daily profit subject to the initial value problem and the constraints on the number of applicable shifts and staffable agents. If we assume, that the revenue gained from each served customer is zero, our formulation also encloses the cost minimisation problem. In this case additional service orientated restriction are needed. Our approach can be extended to the case of maximising weekly

[4] Compare Equation (4.10) on Page 78.

[5] See Hampshire and Massey (2005), Whitt (2006a) and references therein.

or monthly profit by extending the considered time period. The optimisation problem is given by

$$\max_{\boldsymbol{x}} \quad \text{profit}(\boldsymbol{x}, T) \tag{5.12a}$$

subject to

$$\begin{aligned} \frac{d}{dt} Q_S^F(\boldsymbol{x}, t) &= \lambda(t) - \mu(t) \min\{Q_S^F(\boldsymbol{x}, t), N(\boldsymbol{x}, t)\} \\ &\qquad - \nu(t)\{Q_S^F(\boldsymbol{x}, t) - N(\boldsymbol{x}, t)\}^+ + \gamma(t) Q_O^F(\boldsymbol{x}, t) \qquad (5.12\text{b}) \\ \frac{d}{dt} Q_O^F(\boldsymbol{x}, t) &= p\nu(t)\{Q_S^F(\boldsymbol{x}, t) - N(\boldsymbol{x}, t)\}^+ - \gamma(t) Q_O^F(\boldsymbol{x}, t) \\ Q_S^F(\boldsymbol{x}, t_0) &= 0 \\ Q_O^F(\boldsymbol{x}, t_0) &= 0 \qquad (5.12\text{c}) \\ N(\boldsymbol{x}, t) &= \sum_{k=1}^{K} s_k(t) x_k \qquad \text{for } t \in [t_{j-1}, t_j], j = 1, \dots \mathcal{J} \qquad (5.12\text{d}) \\ \sum_{k=1}^{K} x_k &\leq M \qquad (5.12\text{e}) \\ x_k &\in \mathbb{N}_0 \qquad \text{for all } k = 1, \dots, K \qquad (5.12\text{f}) \end{aligned}$$

If we assume that the shift lasts just one time interval of length δ, each shift type k is related to a specific time interval j, i.e., K and $\mathcal{J}$ coincide. Then the number of agents $N(t)$ in each time interval $t \in [t_j - 1, t_j)$ and the number of agents per shift x_j are identical. The costs per shift are given by the product of the length of the time interval and the hourly wage. In this case the optimisation program (5.12) becomes the more simple personnel staffing problem[6].

The optimisation problem in (5.12) is non-linear and dynamic. The decision variables are integer valued. We are not aware of an exact algorithm to solve this type of problem to optimality. Therefore, we use heuristics to solve the problem. The optimisation procedure iterates between calculating the profit function and solving the initial value problem until no improvement of the objective function can be found. In order to develop an effective optimisation procedure, we have to analyse the structure of the profit function with respect to the decision variable $\boldsymbol{x}$, the number of shifts scheduled. This is done in the next section.

[6] The personnel staffing problem is considered e.g. in Whitt (1999c), Helber and Stolletz (2003, 2004), Koole and Pot (2005), Feldman et al. (2005), Green et al. (2005), Hampshire and Massey (2005), Harrison and Zeevi (2005), and the references therein. See also Section 5.6 for further literature about staffing and shift scheduling problems.

5.2 Numerical Analysis of the Profit Function

In this section we investigate the profit function of the optimisation problem formulated in the previous section with respect to changes in the number of agents for each shift type scheduled. For many scheduling heuristics the concavity of the objective function is important to solve the problem[7]. A function $f : \mathbb{R}^{\mathcal{J}} \to \mathbb{R}$ is called concave, if for all vectors $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^{\mathcal{J}}$ and scalars $\alpha \in [0,1]$ the inequality

$$\alpha f(\boldsymbol{y}) + (1-\alpha) f(\boldsymbol{z}) \leq f(\alpha \boldsymbol{y} + (1-\alpha)\boldsymbol{z}) \tag{5.13}$$

holds. Koole and Pot (2005) are able to show that some profit functions used in contact center analysis and staffing are not concave in the number of agents. In the case of concavity, a simple heuristic like progression in direction of the steepest increase can be utilised. If the profit function is not concave, more sophisticated heuristics have to be used to leave local optima.

In order to get some first impressions of the profit function we use the parameters given in Table 5.1, the arrival rate function shown in Figure 4.2 on Page 79, and the shift types $\boldsymbol{s}_k$ given in Table 5.2 to calculate the profit for different schedules $\boldsymbol{x}$.

Service rate:	$\mu = \ 60\,\mathrm{h}^{-1}$
Abandonment rate:	$\nu = \ 120\,\mathrm{h}^{-1}$
Retrial rate:	$\gamma = \ 12\,\mathrm{h}^{-1}$
Probability of retry:	$p = \ 0.5$
Revenue per served customer:	$r = \ 0.5$ €
Hourly cost per occupied line:	$\ell = \ 6$ €/h
Hourly wage:	$w = \ 10$ €/h
Length of a time interval	$\delta = \ 0.5$ h

Table 5.1. Parameters of the optimisation problem for the investigation of the profit function

The contact center opens at 7 am and closes at 8 pm. Therefore we consider $\mathcal{J} = 26$ half-hour intervals. If agents work according to a long shift, they are present for 7.5 hours and have a rest break of half an hour after 3.5 hours. If agents work according to a short shift they are on duty for 4 hours. Therefore we get 31 different shift types as shown in Table 5.2. For the analysis we assume the following initial schedule

[7] See, e.g., Koole and Van der Sluis (2003), Atlason et al. (2004), and Koole and Pot (2006).

Type	Interval																									
$	1				5	6				10	11				15	16				20	21					26
	7:00 - 9:30					9:30-12:00					12:00-2:30					2:30 - 5:00					5:00 - 8:00					
$\boldsymbol{s}_1$	1	1	1	1	1	1	1		1	1	1	1	1	1	1											
$\boldsymbol{s}_2$		1	1	1	1	1	1	1		1	1	1	1	1	1	1										
⋮			⋱					⋱					⋱					⋱						⋱		
$\boldsymbol{s}_{11}$											1	1	1	1	1	1	1		1	1	1	1	1	1	1	
$\boldsymbol{s}_{12}$												1	1	1	1	1	1	1		1	1	1	1	1	1	1
$\boldsymbol{s}_{13}$	1	1	1	1	1	1	1	1																		
$\boldsymbol{s}_{14}$		1	1	1	1	1	1	1	1																	
⋮			⋱					⋱					⋱					⋱						⋱		
$\boldsymbol{s}_{30}$																		1	1	1	1	1	1	1	1	
$\boldsymbol{s}_{31}$																			1	1	1	1	1	1	1	1

Table 5.2. Schematic presentation of the basic shift types of the set $ for the investigation of the profit function and the scheduling problem

$$\begin{aligned} \boldsymbol{x}_0 = (&\overbrace{1, 8, 7, 10, 16, 23, 24, 25, 20, 9, 0, 0}^{\text{long shift types}}, \\ &\underbrace{3, 6, 9, 10, 7, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 8, 13}_{\text{short shift types}}) \end{aligned} \tag{5.14}$$

with an initial profit of 15 962.40€ and 203 agents staffed.

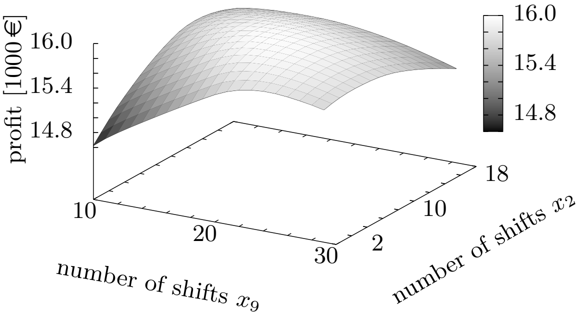

Fig. 5.3. Profit function with varying number of agents for shifts $\boldsymbol{s}_2$ and $\boldsymbol{s}_9$

In Figure 5.3 the number of agents working according to shift types $\boldsymbol{s}_2$ and $\boldsymbol{s}_9$ are varied. Both types of shifts are long shifts with one rest break in the middle. Agents with shift type 2 start their work at 7:30 am and rest between 11:00 and 11:30 am. At 3:00 pm they leave their seats. Agents with shift type 9 start at 11 am, rest from 2:30 to 3:00 pm and quit at 6:30 pm. If the initial schedule $\boldsymbol{x}_0$ is fixed in all other shift types, the maximum profit of 15 972.90€ is reached for $x_9 = 20$ and $x_2 = 7$. The profit function seems

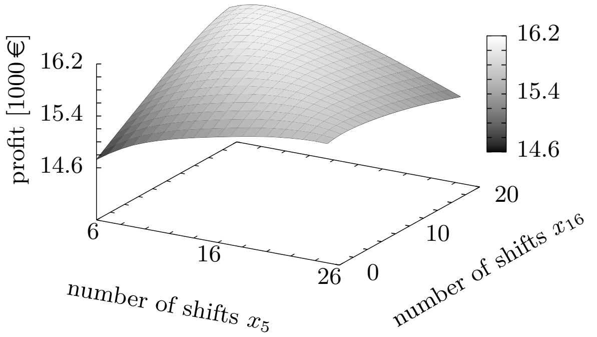

Fig. 5.4. Profit function with varying number of agents in the first and fourth time interval of the working day

to be monotonically increasing in the number of shifts until the optimum is reached. Furthermore, the shape appears to be quite regular and concave.

In Figure 5.4 the numbers of agents x_5 and x_{16} for a long shift $\boldsymbol{s}_5$ and a short shift $\boldsymbol{s}_{16}$ are varied. As in the previous figure the profit function seems to be concave and monotonically increasing to the optimum. The maximum profit of 16072.60€ is generated if 10 agents of shift type $\boldsymbol{s}_5$ and 16 agents of shift type $\boldsymbol{s}_{16}$ are working.

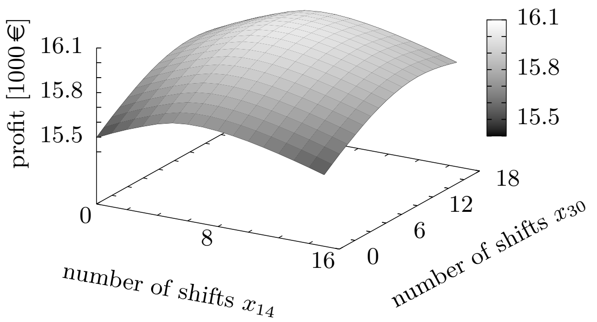

Fig. 5.5. Profit function with varying number of agents in the first and fourth time interval of the working day

Similar conclusions with respect to the profit function can be drawn from the final Figure 5.5. In this figure the number of shift type $\boldsymbol{s}_{14}$ and $\boldsymbol{s}_{30}$ are varied. These shifts are both short shifts. In the previous examples only few of these shifts were scheduled. For $x_{14} = 7$ and $x_{30} = 12$ the maximum profit of 16 010.70€ is reached.

From these three figures we are not able to make any conclusions with respect to the concavity of the profit function. As the differential equations for the number of customers in the system can only be solved numerically, the profit function can only be calculated numerically. The profit function consists of different parts for the cost and the revenue. The costs for the agents are linear in the number of agents. Therefore it depends on the integral term in Equation (5.8) on Page 164 whether the profit function is concave in the number of agents.

We tested the profit function numerically and found some rare examples for non-concavity of the profit function. However, the difference between the value of $\alpha\text{profit}(\boldsymbol{y}) + (1-\alpha)\text{profit}(\boldsymbol{z})$ and $\text{profit}(\alpha\boldsymbol{y} + (1-\alpha)\boldsymbol{z})$ for different values of α and schedules $\boldsymbol{y}$ and $\boldsymbol{z}$ were very small. They might be due to numerical instabilities. They cannot be presented graphically. We show a numerical example. In order to calculate Equation (5.13) in the example the numbers of agents have to be a continuous quantities as well. The parameter α was chosen to be 0.5. Furthermore, for all three examples the vector $\boldsymbol{y}$ is

$$\begin{aligned}(&\overbrace{2,10,5,9,16,22,23,22,28,3,1,1,}^{\text{long shift types}}\\ &\qquad\underbrace{1,1,12,12,9,6,1,0,0,0,0,0,0,0,0,0,0,8,22}_{\text{short shift types}}).\end{aligned} \tag{5.15}$$

This leads to a profit of 15 656.70 €. The vector $\boldsymbol{z}$ is chosen by changing two positions of the vector $\boldsymbol{y}$. We add in each of these positions one agent for the shift. Then we get the results of Table 5.3 which conflict with Equation (5.13) on Page 166.

Changes in $\boldsymbol{y}$		$\text{profit}(\boldsymbol{z})$	$\alpha\text{profit}(\boldsymbol{y})$	$\text{profit}(\alpha\boldsymbol{y}$
first	second		$+(1-\alpha)\text{profit}(\boldsymbol{z})$	$+(1-\alpha)\boldsymbol{z})$
position	position	[€]	[€]	[€]
12	26	15 639.90	15 648.30	15 647.90
14	20	15 630.00	15 643.35	15 643.00
14	26	15 637.40	15 647.05	15 646.60

Table 5.3. Counter example

Other investigations of objective functions can be found, e.g., in Atlason et al. (2004) and Cezik and L'Ecuyer (2005). Atlason et al. (2004) develop another algorithm to test the concavity of their objective function almost everywhere, as this property is needed in the cutting plane method. Cezik and L'Ecuyer (2005) assumed an S-shaped objective function. Koole and Van der Sluis (2003) constructed a concave objective function to apply a local search based on a mathematical concept named multimodularity for the scheduling and staffing problem.

5.3 Outline of the Heuristic Optimisation Procedure

5.3.1 Opening Procedure

As we can not be sure whether the profit function is concave everywhere, we should take this into consideration for our optimisation procedure. The heuristic procedure should on the one hand search in the direction of increasing profit, on the other hand it should have the ability to leave a local optimum in search of a better solution.

In order to accelerate the search we use an opening procedure, which generates a good schedule based on the suggestion made by Feldman et al. (2005). To solve the optimisation problem presented above, a simulated annealing algorithm is used.

A starting solution is determined by taking the arrival rate and the service rate as well as the revenue and cost into consideration. As argued by Feldman et al. (2005) the maximum profit is almost reached by staffing according to the workload of the contact center which is the arrival rate divided by the service rate.

The initial procedure presented in Algorithm 1 below starts with an empty schedule $\boldsymbol{x} = \mathbf{0}$ and determines the average arrival rates $\overline{\lambda_i}$ in each time interval $j = 1, \ldots, \mathcal{J}$, i.e.,

$$\overline{\lambda_i} = (t_j - t_{j-1})^{-1} \int_{t_{j-1}}^{t_j} \lambda(t)dt \quad \text{for all } j = 1, \ldots, \mathcal{J}. \tag{5.16}$$

This leads to a vector $\overline{\boldsymbol{\lambda}}$ of the average arrival rates which is used to estimate relative profit margins $\mathrm{marg}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_k)$ for each applicable shift $\boldsymbol{s}_k$, $k = 1, \ldots, K$. The relative profit margin of shift $\boldsymbol{s}_k$ is the relative increase in profit per time interval on duty if another shift of type k is staffed. These margins of the shifts are compared and the shift $\boldsymbol{s}_\kappa$ with the highest positive profit margin is scheduled.

If the highest relative profit margin is not unique, i.e., several shifts have the same highest relative profit margin, these shifts are compared with respect to the amount of work that has to be done during the intervals covered by each shift. Then the shift with the maximum amount of accumulated work $\mathrm{cum\_work}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_k)$ is staffed. In this case the algorithm prefers long shifts. Afterwards the vector of the average arrival rates is reduced by the number of customers who can be served by the additional shift. Then the relative profit margins are recalculated and the margins are compared again. This procedure is repeated until the margins of all shifts are negative, i.e.,

$$\mathrm{marg}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_k) < 0 \quad \text{for all } k = 1, \ldots, K \tag{5.17}$$

or the total number of shifts reached the maximum number of applicable shifts, i.e.

Algorithm 1 Determination of an Initial Solution

```
Require: $\boldsymbol{x} = \mathbf{0}$
for $j = 1, \dots, \mathcal{J}$ do                                   ▷ Estimation of the average
    $\overline{\lambda_j} = \frac{1}{t_j - t_{j-1}} \int_{t_{j-1}}^{t_j} \lambda(t)\,dt$          ▷ arrival rates per interval
end for
repeat
    for $k = 1, \dots, K$ do                          ▷ Calculation of the relative profit margins
        $\text{marg}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_k) = \frac{1}{|\boldsymbol{s}_k|} \sum_{j=0}^{\mathcal{J}} \left( r \min\{\overline{\lambda_j}, \mu\} - \ell \frac{\min\{\overline{\lambda_j}, \mu\}}{\mu} - w \right) \delta s_{k,j}$
    end for
    Determine: $\mathbb{K} = \{\kappa | \text{marg}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_\kappa) = \max\{\text{marg}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_k), k = 1, \dots, K\}\}$
    if $\kappa$ is not unique, i.e., $|\mathbb{K}| > 1$ then
        for $k \in \mathbb{K}$ do                          ▷ Calculation of the accumulated work
            $\text{cum\_work}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_k) = \sum_{j=1}^{\mathcal{J}} \frac{\overline{\lambda_j}}{\mu} s_{k,j}$
        end for
        Determine: $\kappa = k \in \mathbb{K} | \text{cum\_work}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_\kappa) = \max\{\text{cum\_work}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_k)\}$
    end if
    $x_\kappa \leftarrow x_\kappa + 1$
    for $j = 1, \dots, \mathcal{J}$ do                                      ▷ Calculation of the
        $\overline{\lambda_j} \leftarrow \max\{0, \overline{\lambda_j} - \mu s_{\kappa,j}\}$                        ▷ residual arrival rate
    end for
until $\text{marg}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_k) < 0$ for all $k = 1, \dots K$   or   $\sum_{k=1}^{K} x_k \geq M$
Calculate: $\text{profit}(\boldsymbol{x}_{initial})$
```

$$\sum_{k=1}^{K} x_k \geq M. \tag{5.18}$$

The relative profit margins are approximated by accumulating the maximum revenue achievable and the costs of each shift. In order to make the profit margins of short and long shifts comparable we divide the result by the length of the shift $|\boldsymbol{s}_k|$ in consideration, i.e.,

$$\text{marg}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_k) = \frac{1}{|\boldsymbol{s}_k|} \sum_{j=0}^{\mathcal{J}} \left(r \min\{\overline{\lambda_j}, \mu\} - \ell \frac{\min\{\overline{\lambda_j}, \mu\}}{\mu} - w \right) \delta s_{k,j} \tag{5.19}$$

for all $k = 1, \dots, K$. Here the length of the shift is defined as the sum of the entries of the shift vector, i.e.,

$$|\boldsymbol{s}_k| = \sum_{j=1}^{\mathcal{J}} s_{k,j} \quad \text{for all } k = 1, \dots, K. \tag{5.20}$$

This calculation is an approximation because we do not use the profit function defined in Equation (5.8) on Page 164 but the estimated or forecasted offered work of the contact center.

If the arrival rates $\overline{\lambda}_j$ $j = 1, \ldots, \mathcal{J}$ are greater than the service rates μ, a single agent can at most serve $\delta\mu$ customers in the time interval j. Therefore, the maximum revenue gained from a served customer by a single agent in the time interval is $\delta r\mu$. If the residual arrival rate $\overline{\lambda}_j$ is smaller than the service rate, an additional agent can at most serve $\delta\overline{\lambda}_j$ customers. Consequently, the maximum revenue is $\delta r\overline{\lambda}_j$. The same argument is used in the second term for the costs of occupied lines.

If more than one shift is associated with the highest profit margin, the accumulated work is calculated for the shifts with the highest margin by adding up the average offered load $\frac{\overline{\lambda}_j}{\mu}$, $j = 1, \ldots, \mathcal{J}$ during the working time of the shift, i.e.,

$$\text{cum\_work}(\overline{\boldsymbol{\lambda}}, \boldsymbol{s}_k) = \sum_{j=1}^{\mathcal{J}} \frac{\overline{\lambda}_j}{\mu} s_{k,j} \quad \text{for } k = 1, \ldots, K. \tag{5.21}$$

Finally, for this initial schedule $\boldsymbol{x}_{initial}$ the profit function in (5.8) on Page 164 is calculated, i.e., we solve the initial value problem for this schedule and use the results to compute the inital profit value.

5.3.2 The Improvement Algorithm

To solve the maximisation problem and cope with the occasional non-concavity of the profit function, a simulated annealing algorithm[8] has been implemented. This algorithm reduces the possibility of getting trapped in a poor local optimum by allowing moves to a neighbouring but inferior solution $\boldsymbol{x}_{new} \in \mathbb{X}_{current}$ with

$$\mathbb{X}_{current} = \{\boldsymbol{x} | \boldsymbol{x} \text{ is neighbour of } \boldsymbol{x}_{current}\}, \tag{5.22}$$

where a neighbour has to be defined according to the problem. The algorithm was first published in 1953[9] and called simulated annealing because it simulates the cooling of material in a heat bath. An inferior solution is accepted if the simulated annealing condition

$$U(0,1) < \exp\left(\frac{\Delta\text{profit}}{\boldsymbol{temp}}\right)$$

is fulfilled, where $\boldsymbol{temp}$ is a control parameter, $U(0,1)$ is a uniformly distributed random number on $[0,1]$, and Δprofit is the change in the profit value, i.e.,

$$\Delta\text{profit} = \text{profit}(\boldsymbol{x}_{new}) - \text{profit}(\boldsymbol{x}_{current}). \tag{5.23}$$

Algorithm 2 provides a pseudocode of the simulated annealing algorithm adjusted to the optimisation problem in (5.12) on Page 165.

[8] See, e.g., Reeves (1996) and Thompson (1996).

[9] See Reeves (1996).

Algorithm 2 Adjusted Simulated Annealing Algorithm

```
$\boldsymbol{temp} \leftarrow \boldsymbol{temp}_0$
$\boldsymbol{x}_{current} \leftarrow \boldsymbol{x}_{initial}$
repeat
    Choose a feasible schedule $\boldsymbol{x}_{new} \in \mathbb{X}_{current}$ randomly
    Solve the initial value problem of Equations (5.12b) and (5.12c) on Page 165
    Calculate: profit($\boldsymbol{x}_{new}$)
    if profit($\boldsymbol{x}_{new}$) > profit($\boldsymbol{x}_{current}$) then
        $\boldsymbol{x}_{current} \leftarrow \boldsymbol{x}_{new}$
        profit($\boldsymbol{x}_{current}$) $\leftarrow$ profit($\boldsymbol{x}_{new}$)
    else if U(0,1)$\leq \exp\left(\frac{\triangle\text{profit}}{\boldsymbol{temp}}\right)$ then
        $\boldsymbol{x}_{current} \leftarrow \boldsymbol{x}_{new}$
        profit($\boldsymbol{x}_{current}$) $\leftarrow$ profit($\boldsymbol{x}_{new}$)
    end if
    $\boldsymbol{temp} \leftarrow f_{mul}\boldsymbol{temp}$
until $\boldsymbol{temp} \leq \boldsymbol{temp}_{final}$
```

The parameter $\boldsymbol{temp}$ can be interpreted as the temperature of the bath which is reduced by some cooling scheme. The initial temperature $\boldsymbol{temp}_0$ has to be chosen high in order to allow many inferior moves to be accepted in the early phases of the search. It is slowly reduced until almost all inferior solutions are rejected. The initial and final temperature as well as the cooling scheme depend on the problem itself and are important for the effectiveness and efficiency of the algorithm. For the cooling scheme a homogeneous and an inhomogeneous type are distinguished[10]. We use the inhomogeneous type, in which the temperature $\boldsymbol{temp}$ is reduced after every move by a very small amount.

For the shape of the cooling curves two methods are popular. In a geometric scheme the new temperature is generated by multiplying the temperature by a constant $f_{mul} \in (0, 1)$ close to 1. The second method reduces the temperature by dividing the temperature $\boldsymbol{temp}$ by the sum of 1 and the product of a positive constant $f_{div} \in (0, 1)$ close to zero and $\boldsymbol{temp}$, i.e.

$$\boldsymbol{temp} \leftarrow f_{mul}\boldsymbol{temp} \quad \text{or} \quad \boldsymbol{temp} \leftarrow \frac{\boldsymbol{temp}}{1 + f_{div}\boldsymbol{temp}}.$$

The final temperature $\boldsymbol{temp}_{final}$ has to be sufficiently small and depends on the parameters of the profit function (5.8) on Page 164. We use the first method, because the parameter $\boldsymbol{temp}$ stays high for more iterations such that more inferior moves are allowed. The initial and final temperatures as well as the factor f_{mul} determine the number of iterations of the algorithm by[11]

$$\text{no. of iterations} = \left\lceil \frac{\log(\boldsymbol{temp}_{final}) - \log(\boldsymbol{temp}_0)}{\log(f_{mul})} \right\rceil \tag{5.24}$$

[10] See Reeves (1996).

[11] See Reeves (1996).

The Algorithm 2 starts with the initial schedule generated by the opening procedure Algorithm 1 on Page 171 and tries to improve the profit repeatedly by choosing a feasible schedule from the neighbourhood $\mathbb{X}_{current}$ randomly. The neighbourhood of the schedule is defined as the set of all schedules which fulfil one of the following conditions:

- The schedule contains one additional shift.
- The schedule is reduced by one shift.
- A shift of the schedule is exchanged by another shift.

For this schedule the initial value problem in Equations (5.12b) and (5.12c) on Page 165 is solved and the profit is calculated. If the profit of the neighbouring schedule $\boldsymbol{x}_{new}$ is higher than the profit of the current schedule $\boldsymbol{x}_{current}$, the new schedule is accepted. Otherwise the simulated annealing condition is checked. If the simulated annealing condition is fulfilled the new schedule is also accepted. Then the temperature is reduced and the algorithm starts again until a final cooling temperature $\boldsymbol{temp}_{final}$ is reached.

We consider three possible operations to generate neighbouring schedules of the current schedule. Algorithm 3 presents how a feasible schedule from the neighbourhood $\mathbb{X}_{current}$ is chosen without generating the whole neighbourhood of the current schedule $\boldsymbol{x}_{current}$. Three possible operations are distinguished to change the current schedule which are:

- The first operation is adding a shift $\boldsymbol{s} \in \mathbb{S}$ if the total number of staffed agents remains below the maximum number of applicable agents M. This leads to K possible changes of the schedule, because we have K shift types.
- The second operation is removing a shift already scheduled. If the current schedule $\boldsymbol{x}_{current}$ contains at least one shift of each type this again leads to K possible changes or possible neighbours.
- Finally, the last operation is an exchange of a shift, i.e., an existing shift is removed and another shift $\boldsymbol{s} \in \mathbb{S}$ is scheduled, if such a change is feasible. This leads to at most K^2 applicable operations including exchanges of the same shift type.

Therefore, we get $(K+2)K$ possible operations to determine the new schedule from the current schedule.

In order to choose the next operation, the algorithm determines a uniformly distributed integer n on the interval $[0,(K+2)K)$, where K is the number of shift types in $\mathbb{S}$. From this random number n we determine two integers $i, j \geq 0$, to choose one of the three operations. The integer j determines the operation to generate the new schedule and is given by the greatest integer smaller than the fraction of the random number n and K, i.e.,

$$j = \left\lfloor \frac{n}{K} \right\rfloor . \tag{5.25}$$

If j has value *zero*, a shift is added. In this case the integer i indicates the number of the additional shift. For this purpose, the integer i is given by the remainder of integer division of n by K, i.e.,

$$i = n \mod K. \tag{5.26}$$

Therefore, i is an integer from the interval $[0, K-1]$. As the shifts are numbered from 1 to K, the integer i refers to shift $\boldsymbol{s}_{i+1}$. If j has value one a shift is removed. In this case the integer i again determines the shift $\boldsymbol{s}_{i+1} \in \$$ which is removed. Finally, if the value of j is any integer from the interval $[2, K+1]$ two shifts are exchanged. In this last case j is also needed to indicate the shift which should be removed. As j is an integer between 2 and $K+1$, j refers to shift $\boldsymbol{s}_{j-1} \in \$$. The integer i identifies the additional shift $\boldsymbol{s}_{i+1} \in \$$. This algorithm effects that mainly different shifts are exchanged during the simulated annealing algorithm.

Algorithm 3 Choose a Feasible Schedule $\boldsymbol{x}_{new}$

```
Require: $i, j, n \geq 0$ as integer
$K \leftarrow |\$|$
repeat                                                ▷ random selection of the
    $n \leftarrow \lfloor U(0, (K+2)K) \rfloor$                  ▷ neighbourhood operation
    $i \leftarrow n \mod K$                            ▷ remainder after division
    $j \leftarrow n \text{ div } K$                             ▷ $\lfloor \frac{n}{K} \rfloor$
until OPERATION_APPLICABLE($i, j, \boldsymbol{x}_{current}$)
if $j = 0$ then
    Add one agent of shift type $i + 1$
else if $j = 1$ then
    Remove one agent of shift type $i + 1$
else
    Add one agent of shift type $i + 1$
    Remove one agent of shift type $j - 1$
end if

function OPERATION_APPLICABLE($i$,$j$,$\boldsymbol{x}_{current}$)
    if $j = 0$ and $\sum_{k=1}^{K} x_k < M$ then         ▷ Condition (5.12e) on Page 165
        return true
    else if $j = 1$ and $x_{i+1} > 0$ then               ▷ Condition (5.12f) on Page 165
        return true
    else if $j \geq 2$ and $\sum_{k=1}^{K} x_k \leq M$ and $x_{j-1} > 0$ and $i \neq j - 1$ then
                                                  ▷ Conditions (5.12f) and (5.12e)
        return true
    else
        return false
    end if
end function
```

The boolean function OPERATION_APPLICABLE(i, j, $\boldsymbol{x}_{current}$) in Algorithm 3 determines whether the chosen operation can be executed. If the new schedule $\boldsymbol{x}_{new}$ violates at least one of the conditions presented in Equations (5.12e) or (5.12f) on Page 165 then the function returns the boolean value ***false***. If the parameter j is greater than one and i equals $j-2$, the function returns ***false*** as well, because in this case a shift would be substituted by itself. This would lead to $\boldsymbol{x}_{new} = \boldsymbol{x}_{current}$. If the operation is not applicable, another operation and shifts have to be chosen, i.e., n, i, and j are calculated afresh, until the new schedule is valid.

The advantage of this heuristic is that it can deal with profit functions which are not concave and might have multiple local optima, as argued by Jiménez and Koole (2004) and others. A simpler heuristic like hill climbing might end in a local maximum and progression in the direction of the steepest increase give rise to difficulties with respect to the non-concavity. Furthermore, this last algorithm is very time consuming because, contrary to the simulated annealing algorithm, the initial value problem has to be solved for all neighbours of the current solution, which needs a lot of time.

5.4 Numerical Results

5.4.1 An Unlimited Total Number of Agents

In this section we analyse the performance of the algorithm and the structure of the solution for a contact center with homogeneous agents and customers as described in Section 4.1. We assume that the arrival rate function is given by Equation (3.13) on Page 29 and Figure 4.2 on Page 79 with parameters according to (4.12) on page 79. The other parameters are given in Table 5.1 on Page 166.

We first investigate the influence of the cooling factor f_{mul} and the initial and final temperatures, if the number of agents who can be staffed is not limited, i.e., $M = \infty$. To solve the initial value problem we used the Runge-Kutta method, although the algorithms works on average four times as fast with the Euler method but less accurately.

The algorithm was tested on a notebook with a Pentium IVm processor with 2 GHz and 1024 MB RAM.

The initial schedule $\boldsymbol{x}_0$ found by the opening procedure in the case of an infinite total number of available shifts is given by

$$\begin{aligned}\boldsymbol{x}_0 = (&1, 8, 7, 10, 16, 23, 24, 23, 19, 5, 0, 1, 2, 3, 9, 13, 8,\\ &3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 9, 22)\end{aligned} \tag{5.27}$$

with an initial profit of 16 078.50 € and 207 agents staffed who work altogether 1239 hours.

As shown in Table 5.4 the cooling factor f_{mul} has a major influence on the performance with respect to the number of iterations and the running time of the heuristic solution algorithm. If the cooling factor reaches one, the number of iterations and the running time including the computing time of the initial schedule approaches infinity. However, the best solution found improves as well. Obviously, the initial solution found by the opening procedure is already very accurate. If the cooling factor is very small, the simulated annealing

schedule	f_{mul}	iterations	time	profit		agents	
$\boldsymbol{x}_i$			$[sec]$	[€]	rel. increase	number	hours
$\boldsymbol{x}_1$	0.9	51	1	16 054.50	-0.15%	207	1 239
$\boldsymbol{x}_2$	0.99	528	4	16 166.00	0.54%	211	1 243
$\boldsymbol{x}_3$	0.999	5 296	42	16 425.30	2.15%	263	1 232
$\boldsymbol{x}_4$	0.9999	52 981	426	16 466.20	2.41%	268	1 228
$\boldsymbol{x}_5$	0.99999	529 830	4 655	16 466.30	2.41%	277	1 228

Table 5.4. Influence of the cooling factor on the performance of the algorithm with $\boldsymbol{temp}_0 = 200$ and $\boldsymbol{temp}_{final} = 1$

algorithm is not able to find a better solution. The profit reduces by 0.15%. The major increase in profit occurs if the cooling factor rises from 0.99 to 0.999. In this case the profit increases by 1.61% from 16 166.00€ to 16 425.30€. In Table 5.4 the relative increases of the profit with respect to the initial solution are shown. The number of agents staffed grows from 207 to 277 but the total number of working hours reduces from 1 239 to 1 228. As we assumed no limit for the total number of agents M who can be staffed, the algorithm tends to schedule more agents with short shifts, because these shifts are more flexible.

Table 5.5 and Figure 5.6 present the effect that the algorithm substitutes long shifts by short shifts, if the number of iteration increases. A higher cooling factor reduces the temperature parameter more slowly, i.e., more iterations are needed. Consequently, the temperature stays high for more iterations such that more negative moves are allowed. Therefore, more long shifts are exchanged for short shifts.

shifts	schedules					
$\boldsymbol{s}_i \in \$$	$\boldsymbol{x}_0$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$
$\boldsymbol{s}_1$	1	2	2	0	1	1
$\boldsymbol{s}_2$	8	8	5	1	0	0
$\boldsymbol{s}_3$	7	7	11	3	2	2
$\boldsymbol{s}_4$	10	10	10	7	4	2
$\boldsymbol{s}_5$	16	16	12	7	3	1
$\boldsymbol{s}_6$	23	23	21	12	2	2
$\boldsymbol{s}_7$	24	24	21	13	5	2
$\boldsymbol{s}_8$	23	23	23	13	7	6
$\boldsymbol{s}_9$	19	19	17	15	8	7
$\boldsymbol{s}_{10}$	5	5	7	13	11	11
$\boldsymbol{s}_{11}$	0	0	1	10	8	5
$\boldsymbol{s}_{12}$	1	0	3	2	1	1
$\boldsymbol{s}_{13}$	2	2	1	2	1	0
$\boldsymbol{s}_{14}$	3	3	3	11	9	11
$\boldsymbol{s}_{15}$	9	9	7	13	15	15
$\boldsymbol{s}_{16}$	13	13	15	18	19	21
$\boldsymbol{s}_{17}$	8	8	12	13	26	26
$\boldsymbol{s}_{18}$	3	3	4	15	23	28
$\boldsymbol{s}_{19}$	0	0	1	15	20	18
$\boldsymbol{s}_{20}$	0	0	0	3	12	14
$\boldsymbol{s}_{21}$	0	0	0	0	2	1
$\boldsymbol{s}_{22}$	0	0	0	0	0	3
$\boldsymbol{s}_{23}$	0	0	0	0	0	1
$\boldsymbol{s}_{24}$	0	0	0	0	2	4
$\boldsymbol{s}_{25}$	0	0	0	3	12	13
$\boldsymbol{s}_{26}$	0	0	1	9	18	21
$\boldsymbol{s}_{27}$	0	0	1	14	21	21
$\boldsymbol{s}_{28}$	0	0	2	11	17	20
$\boldsymbol{s}_{29}$	1	1	1	5	13	10
$\boldsymbol{s}_{30}$	2	9	9	6	5	8
$\boldsymbol{s}_{31}$	22	22	21	2	1	2

Table 5.5. Comparison of schedules of Table 5.4

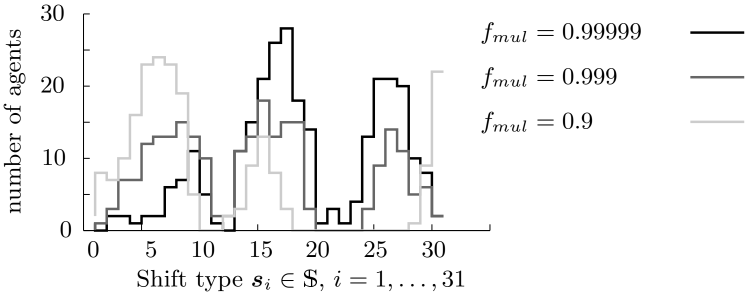

Fig. 5.6. The number of agents per shift type and the influence of the cooling factor f_{mul} on the performance of the heuristic algorithm

Substituting a long shift by a short shift means reducing the number of agents $N(t)$, $t_{j-1} \leq t < t_j$. In some intervals j, $j = 1, \ldots, \mathcal{J}$. On the one hand this might reduce the revenue, because some customers might not be served. On the other hand the costs for salaries of agents are reduced. Therefore an exchange of a long and short kind of shift might lead to either a lower or a higher profit.

In general, if too few agents are scheduled, such a substitution leads to lower profit, therefore inferior moves are needed to escape from the region of local optima with many long shifts scheduled.

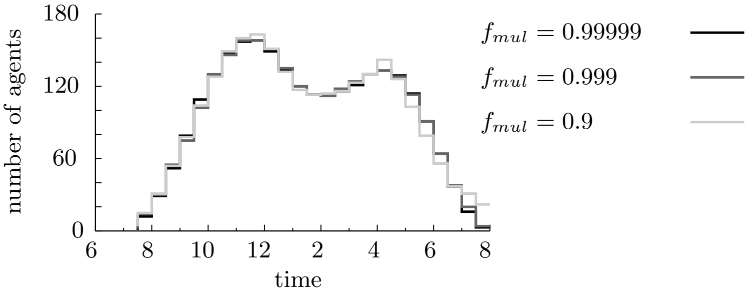

Fig. 5.7. The number of agents in each time interval and the influence of the cooling factor f_{mul} on the performance of the heuristic algorithm

In Figure 5.7 the number of agents on duty in each half-hour time interval is shown. Obviously, the different schedules do not lead to major differences with respect to the number of agents in each time interval.

Figure 5.8 compares the number of agents on duty and the number of customers in the system for the best schedule $\boldsymbol{x}_5$. The fit of both curves is remarkably good. At the beginning of each time interval the number of customers is smaller than the number of agents, and at the end the number of customers exceeds the number of agents if the arrival rate increases. Otherwise the number of customers is greater at the beginning of the time interval and smaller than the number of agents at the end. As the number of customers

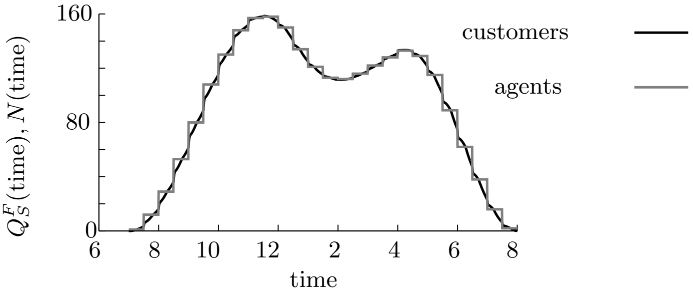

Fig. 5.8. Example for the number of agents $N(\text{time})$ and the number of customers in the system $Q_S^F(\text{time})$ in the contact center with schedule $\boldsymbol{x}_5$

f_{mul}	$W_{agg}^F(T)$[sec]	$P_{agg}^F(\text{served}, T)$	$U_{agg}^F(T)$
0.9	1.039	0.965	0.961
0.99	0.818	0.973	0.962
0.999	0.578	0.981	0.975
0.9999	0.576	0.981	0.978
0.99999	0.573	0.981	0.978

Table 5.6. Aggregated performance measures of the contact center configuration for different cooling factors

and the number of agents almost coincide, few customers have to wait and almost no customer abandons. This leads to a high quality of service.

Table 5.6 stresses the high quality of service for all schedules. It presents the aggregated performance measures for the contact center configuration given in Table 5.1 on Page 166 and the schedules calculated by the algorithm for different cooling factors. For the best schedule found, the aggregated waiting time is half as long as the aggregated waiting time for the worst schedule. However, in all cases the waiting time is very small. The difference in the aggregated probability of being served and the aggregated utilisation of agents is even smaller. The aggregated waiting time and the probability of being served improve if the profit increases, but the agents are even more utilised. Even in the model with the lowest profit the performance of the contact center is more than acceptable. The waiting times are extremely small and almost all customers are served. The aggregated utilisation of agents is very high. However, the agents might suffer from the heavy workload.

In order to show that not only the aggregated performance of the different systems but also the time-dependent performance is amazingly good, we compare the fluid results for the time-dependent waiting time $W_S^F(t)$, the probability of being served $P^F(\text{served}, t)$, and utilisation $U^F(t)$ of agents for the schedule $\boldsymbol{x}_5$ to simulation results. The simulation results are based on 500 simulation runs for the contact center staffed according to the same schedule.

At the beginning of each time interval, whenever the number of agents changes, the time-dependent performance measures change drastically. In the

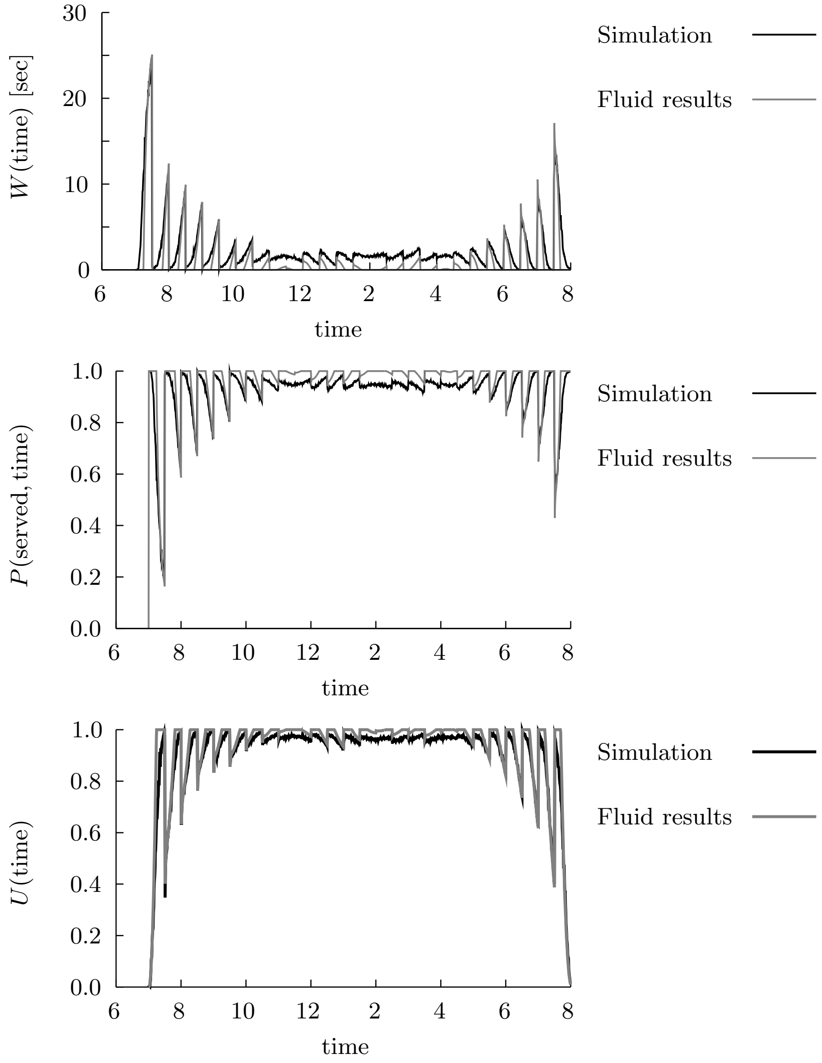

Fig. 5.9. Comparison of the time-dependent performance measures for the schedule $\boldsymbol{x}_5$

morning and after the lunch break[12] the waiting time and the utilisation decreases, because more agents are available than needed for attending to customers. The probability of being served is one. At the end of these intervals the waiting time and utilisation increases, while the probability of being served decreases, because more customers enter the system than can be served by the agents on duty.

[12] See Figure 3.10 on Page 44.

The fluid and simulation results almost coincide. Solely during the period of critically loading and the local minimum of the arrival rate function at about 2 pm does the fluid approach slightly underestimate the waiting time and overestimate the probability of being served and the utilisation.

Before the lunch break and in the evening the performance is poor at the beginning of each time interval and becomes better at the end, because the arrival rate function decreases during these periods. Consequently, the number of customers entering the system is higher at the beginning of each time interval than at the end[13].

At the beginning and the end of the working day the performance fluctuates much more strongly than during the periods of a high arrival rate. This effect can be explained in a similar way to the higher stochasticity of small system. If only few agents are on duty and the number of arrivals is small, an arriving customer who finds all agents busy will have to wait relatively longer until an agent finishes its service.

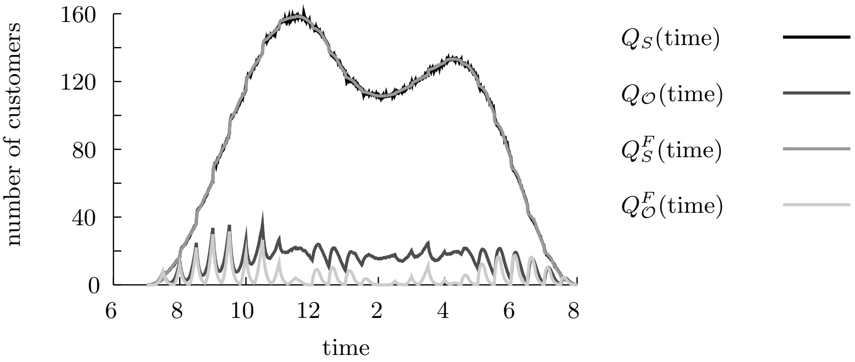

Fig. 5.10. The number of customers in the system and the orbit for the schedule $\boldsymbol{x}_5$

In addition to the performance measures we compare the results for the number of customers in the system and the orbit calculated by the simulation, i.e., $Q_S(t)$ and $Q_{\mathcal{O}}(t)$, and the fluid approximation, i.e., $Q_S^F(t)$ and $Q_{\mathcal{O}}^F(t)$, in Figure 5.10. While the curves for the number of customers in the system are almost indistinguishable, the number of customers in the orbit calculated by the simulation differs from the number calculated by the fluid approximation. In the simulation more customers are at any time in the orbit. However, the number of customers in the orbit is small. The different results for the number of customers in the orbit is caused by the neglected randomness in the fluid approach. This is in line with the worse time-dependent performance measures during these time periods.

[13] Similar results of the time dependent performance can be found in Green et al. (2001) and Ingolfsson et al. (2003).

schedule	$\boldsymbol{temp}_0$	iterations	time	profit		agents	
$\boldsymbol{x}_i$			$[sec]$	[€]	rel. increase	number	hours
$\boldsymbol{x}_6$	2	693	7	16 254.50	1.09%	213	1 239
$\boldsymbol{x}_7$	20	2 995	24	16 392.10	1.95%	226	1 228
$\boldsymbol{x}_8$	200	5 296	42	16 425.30	2.15%	263	1 232
$\boldsymbol{x}_9$	2 000	7 598	60	16 409.40	2.05%	237	1 230
$\boldsymbol{x}_{10}$	20 000	9 899	78	16 376.40	1.85%	235	1 234

Table 5.7. Influence of the cooling factor on the performance of the algorithm with $f_{mul} = 0.999$ and $\boldsymbol{temp}_{final} = 1$

Next we investigate the influence of the initial temperature on the heuristic scheduling algorithm in Table 5.7. In contrast to the cooling factor the influence of the initial temperature on the running time is smaller. If we increase the initial temperature by multiplying by a factor of 10, the number of iterations and the computing time increase by a fixed amount. A higher temperature does not imply a better solution of the heuristic. The best result was generated with an initial temperature of 200. Higher initial temperatures allow too large inferior moves. While the temperature decreases in the course of the calculation, the algorithm becomes unable to leave the inferior region again. In the best schedule the most agents are scheduled. The number of working hours is almost stable for all schedules. This implies that the best schedule contains more short shifts than the other schedules.

shifts	schedules					
$\boldsymbol{s}_i \in \$$	$\boldsymbol{x}_0$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$	$\boldsymbol{x}_9$	$\boldsymbol{x}_{10}$
$\boldsymbol{s}_1$	1	0	0	0	1	1
$\boldsymbol{s}_2$	8	8	3	1	5	8
$\boldsymbol{s}_3$	7	11	7	3	7	11
$\boldsymbol{s}_4$	10	12	9	7	8	5
$\boldsymbol{s}_5$	16	11	7	7	7	8
$\boldsymbol{s}_6$	23	17	15	12	10	12
$\boldsymbol{s}_7$	24	20	14	13	14	13
$\boldsymbol{s}_8$	19	21	17	13	13	15
$\boldsymbol{s}_9$	5	20	16	15	14	16
$\boldsymbol{s}_{10}$	0	7	12	13	10	7
$\boldsymbol{s}_{11}$	1	2	8	10	5	2
$\boldsymbol{s}_{12}$	2	0	0	2	0	0
$\boldsymbol{s}_{13}$	3	1	2	2	0	1
$\boldsymbol{s}_{14}$	9	2	7	11	7	3
$\boldsymbol{s}_{15}$	13	10	13	13	11	9
$\boldsymbol{s}_{16}$	8	15	16	18	14	20
$\boldsymbol{s}_{17}$	3	10	11	13	20	14
$\boldsymbol{s}_{18}$	0	6	14	15	12	12
$\boldsymbol{s}_{19}$	0	2	9	15	14	11
$\boldsymbol{s}_{20}$	0	0	2	3	5	4
$\boldsymbol{s}_{21}$	0	0	0	0	0	3
$\boldsymbol{s}_{22}$	0	0	0	0	0	0
$\boldsymbol{s}_{23}$	0	0	0	0	0	0
$\boldsymbol{s}_{24}$	0	0	0	0	0	0
$\boldsymbol{s}_{25}$	0	0	2	3	3	3
$\boldsymbol{s}_{26}$	0	2	5	9	11	7
$\boldsymbol{s}_{27}$	0	2	12	14	11	12
$\boldsymbol{s}_{28}$	0	3	5	11	11	9
$\boldsymbol{s}_{29}$	1	3	6	5	9	11
$\boldsymbol{s}_{30}$	9	12	10	6	11	13
$\boldsymbol{s}_{31}$	22	16	4	2	4	5

Table 5.8. Comparison of schedules of Table 5.7

Table 5.8 and Figure 5.11 present the associated schedules generated by the heuristic optimisation algorithm. The schedules as well as the number of agents staffed do not differ much. The maximum profit calculated for all initial

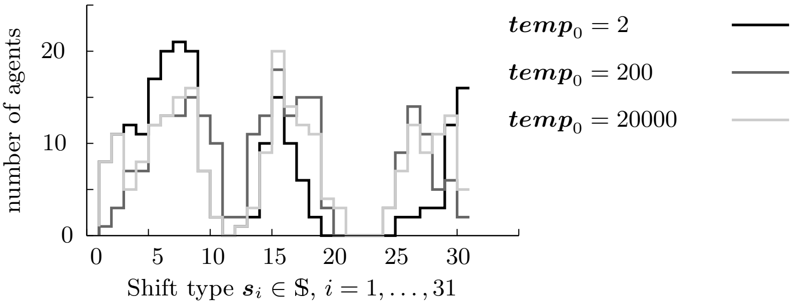

Fig. 5.11. The number of agents per shift type and the influence of the initial temperature $\boldsymbol{temp}_0$ on the performance of the heuristic algorithm

$\boldsymbol{temp}_0$	$W^F_{agg}(T)$[sec]	$P^F_{agg}(\text{served}, T)$	$U^F_{agg}(T)$
2	0.737	0.975	0.966
20	0.695	0.977	0.976
200	0.578	0.981	0.975
2 000	0.635	0.979	0.975
20 000	0.619	0.979	0.972

Table 5.9. Aggregated performance measures of the contact center configuration for different initial temperatures

temperatures are again almost the same. Therefore, an initial temperature of about 200 should be chosen.

Table 5.9 reports the performance of the contact center with the schedule generated with the different initial temperatures. The waiting time in the contact center is negligible and the probability of being served is almost one. From the point of view of customers and managers the performance of the contact center is very good. The agents might suffer from the high utilisation. Although the profits of the different runs differ, the aggregated waiting time, the aggregated probability of being served, and the aggregated utilisation are almost identical.

schedule	$\boldsymbol{temp}_{final}$	iterations	time	profit		agents	
$\boldsymbol{x}_i$			[sec]	[€]	rel. increase	number	hours
$\boldsymbol{x}_{11}$	100	693	6	16 054.50	-0.15%	207	1 239
$\boldsymbol{x}_{12}$	10	2 995	23	16 208.40	0.96%	226	1 243
$\boldsymbol{x}_{13}$	1	5 296	42	16 425.30	1.34%	263	1 232
$\boldsymbol{x}_{14}$	0.1	7 590	61	16 435.90	0.06%	237	1 230
$\boldsymbol{x}_{15}$	0.01	9 899	78	16 440.40	0.03%	237	1 230

Table 5.10. Influence of the cooling factor on the performance of the algorithm with $f_{mul} = 0.999$ and $\boldsymbol{temp}_0 = 200$

Finally the influence of the final temperature is analysed for the contact center with an unlimited number of agents who can be staffed. Table 5.10 reports the results from the heuristic optimisation procedure. High final temperatures let the algorithm stop too early. If the final temperature is decreased by multiplying by 0.1, the number of iterations and the running time increase by a fixed amount. The influence of the final temperature on the number of iterations and the running time is similar to the influence of the initial temperature, which can also be derived from Equation (5.24) on Page 173 for the number of iterations of the simulated annealing algorithm. The best solution is found for a final temperature of 0.01. However, the increase of the profit from a final temperature of 0.1 to a final temperature of 0.01 is very small.

shifts	schedules					
$s_i \in \mathbb{S}$	x_0	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
s_1	1	2	2	0	0	0
s_2	8	8	1	1	1	1
s_3	7	7	4	3	3	3
s_4	10	10	4	7	8	8
s_5	16	16	15	7	7	7
s_6	23	23	16	12	10	11
s_7	19	24	18	13	13	12
s_8	5	23	16	13	13	13
s_9	0	19	15	15	15	15
s_{10}	1	5	11	13	14	14
s_{11}	2	0	11	10	9	9
s_{12}	3	0	0	2	1	1
s_{13}	9	2	4	2	2	2
s_{14}	13	3	4	11	11	12
s_{15}	8	9	23	13	13	12
s_{16}	3	13	13	18	17	17
s_{17}	0	8	10	13	16	15
s_{18}	0	3	5	15	14	16
s_{19}	0	0	12	15	15	14
s_{20}	0	0	1	3	3	3
s_{21}	0	0	0	0	0	0
s_{22}	0	0	0	0	0	0
s_{23}	0	0	0	0	0	0
s_{24}	0	0	0	0	0	0
s_{25}	0	0	0	3	4	4
s_{26}	0	0	6	9	10	10
s_{27}	0	0	6	14	14	14
s_{28}	0	0	6	11	8	9
s_{29}	1	1	6	5	8	7
s_{30}	9	9	6	6	5	5
s_{31}	22	22	9	2	3	3

Table 5.11. Comparison of schedules of Table 5.10

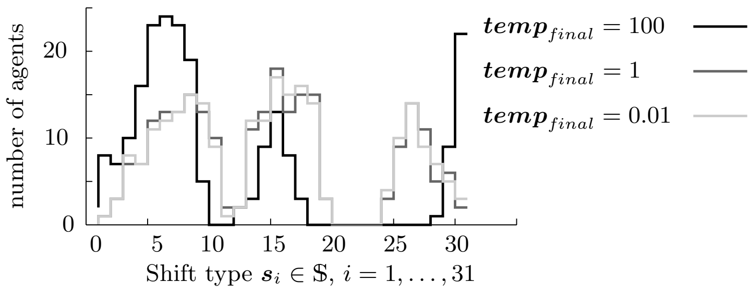

Fig. 5.12. The number of agents per shift type and the influence of the final temperature $\boldsymbol{temp}_{final}$ on the performance of the heuristic algorithm

In Table 5.11 and Figure 5.12 the schedules of the different solutions are shown. The schedule for $\boldsymbol{temp}_{final} = 0.1$ and $\boldsymbol{temp}_{final} = 0.01$ are almost identical. If the final temperature is high, almost no changes happen. Then a lot of long kinds of shifts are staffed. If the final temperature is smaller several long shifts are substituted by short shifts, which are more flexible. However, more agents are needed.

$\boldsymbol{temp}_{final}$	$W^F_{agg}(T)$[sec]	$P^F_{agg}(\text{served}, T)$	$U^F_{agg}(T)$
100	1.039	0.965	0.961
10	0.744	0.975	0.963
1	0.578	0.981	0.975
0.1	0.592	0.981	0.976
0.01	0.582	0.981	0.976

Table 5.12. Aggregated performance measures of the contact center configuration for different final temperatures

The aggregated performance measures in Table 5.12 stress the high quality of service of the contact center. As before, the waiting times are negligible and almost all customers are served. The agents are very highly utilised.

5.4.2 A Limited Total Number of Agents

Next we investigate the influence of the algorithm parameters on the performance of the solution and the algorithm, when the total number of agents M to be staffed is limited. We assume a maximum number M of 200 agents. As in the previous section we assume that the arrival rate function is given by Equation (3.13) on Page 29 and Figure 4.2 on Page 79 with parameters according to (4.12) on page 79. The other parameters are given in Table 5.1 on Page 166. Then the initial schedule $\boldsymbol{x}_0$ found by the initial procedure is given by

$$\begin{aligned}\boldsymbol{x}_0 = (&1, 8, 7, 10, 16, 23, 24, 23, 19, 5, 0, \mathbf{0}, \\ &2, 3, 9, \mathbf{11}, 8, \mathbf{0}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 9, \mathbf{19}).\end{aligned} \tag{5.28}$$

The bold numbers indicate the differences from the initial schedule (5.27) on Page 176. The profit gained by this schedule is 15 948.20 € and the 200 agents work altogether 1 208 hours, i.e., the profit reduces by 130.30 € and the number of working hours by 31.

In Table 5.13 the cooling factor is varied. The influence of the cooling factor on the performance of the algorithm and the solution is similar. The closer the cooling factor is to one the more iterations have to be done and the more slowly the temperature reduces. The best solution was found for the highest cooling factor. Except for schedule $\boldsymbol{x}_{17}$, all schedules lead to a total number of working hours of 1 208 hours. In schedule $\boldsymbol{x}_{17}$ the 200 agents work 1 211 hours in total. The increase of the profit function is even smaller than in the case of an unlimited total number of agents.

schedule $\boldsymbol{x}_i$	f_{mul}	iterations	time [sec]	profit [€]	agents
$\boldsymbol{x}_{16}$	0.9	73	< 1	15 956.10	200
$\boldsymbol{x}_{17}$	0.99	757	5	16 019.50	200
$\boldsymbol{x}_{18}$	0.999	7 598	44	16 131.50	200
$\boldsymbol{x}_{19}$	0.9999	76 006	427	16 143.50	200
$\boldsymbol{x}_{20}$	0.99999	760 087	4625	16 147.50	200

Table 5.13. Influence of the cooling factor on the performance of the algorithm with $\boldsymbol{temp}_0 = 200$ and $\boldsymbol{temp}_{final} = 0.1$

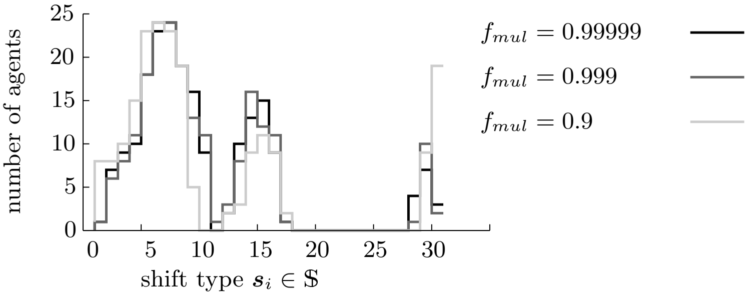

Fig. 5.13. The number of shifts of the schedules calculated by the heuristic algorithm with the different cooling factors

f_{mul}	$W^F_{agg}(T)$[sec]	$P^F_{agg}(\text{served}, T)$	$U^F_{agg}(T)$
0.9	1.582	0.947	0.974
0.99	1.492	0.950	0.978
0.999	1.400	0.953	0.980
0.9999	1.400	0.953	0.980
0.99999	1.398	0.953	0.980

Table 5.14. Aggregated performance measures of the contact center configuration for different cooling factors and a limited total number of agents

Figure 5.13 compares the different schedules calculated by the algorithm and Table 5.14 reports the aggregated performance measures. Obviously, the different schedules in Figure 5.13 do not differ much, if the total number of agents is limited. Furthermore, more shifts of the long kind are scheduled, because those shifts cover more time intervals. Although the number of agents is limited, the performance of the system is still very good as shown in Table 5.14. Almost all customers are served and the waiting time is diminishingly small. The only disadvantage is that the agents might suffer as the aggregated utilisation is almost one.

Therefore, it might be more useful to add an constraint to the scheduling problem, which limits the aggregated utilisation or fixes a minimum total number of agents, so that the agents do not suffer so much.

The influence of the initial and final temperature on the performance of the algorithm for the case of a limited total number of agents is equal to the

schedule $\boldsymbol{x}_i$	$\boldsymbol{temp}_{final}$	iterations	time $[sec]$	profit [€]	agents
$\boldsymbol{x}_{21}$	2	6905	6	16 102.10	200
$\boldsymbol{x}_{22}$	20	8512	24	16 138.70	200
$\boldsymbol{x}_{23}$	200	8003	44	16 131.50	200
$\boldsymbol{x}_{24}$	2 000	8290	64	16 140.50	200
$\boldsymbol{x}_{25}$	20 000	8513	78	16 119.80	200

Table 5.15. Influence of the cooling factor on the performance of the algorithm with $f_{mul} = 0.999$ and $\boldsymbol{temp}_{final} = 0.1$

case of an unlimited number. Therefore, solely the results of the algorithm and for the aggregated performance measures are presented.

In Table 5.15 the results for different initial temperatures are shown. Amazingly, the profit is higher if the initial temperature $\boldsymbol{temp}_0$ is 20 or 2 000. However, the difference between the schedules and the gained profit is very small. In the three cases $\boldsymbol{x}_{22}$, $\boldsymbol{x}_{23}$, and $\boldsymbol{x}_{24}$ with the highest profit, 200 agents work 1 208 hours in total whereas in the other two cases they work 1211 hours.

$\boldsymbol{temp}_0$	$W^F_{agg}(T)$[sec]	$P^F_{agg}(\text{served}, T)$	$U^F_{agg}(T)$
2	1.422	0.953	0.977
20	1.408	0.953	0.980
200	1.423	0.953	0.979
2 000	1.406	0.953	0.980
20 000	1.395	0.954	0.977

Table 5.16. Aggregated performance measures of the contact center configuration for different cooling factors and a limited total number of agents

Table 5.16 reports the aggregated waiting time, the aggregated probability of being served, and the aggregated utilisation of agents. The waiting time measured in seconds is negligible. The utilisation of agents is almost one and 95.3% of all entering customers are served. The number of customers entering the system consists of both retrials and primary calls, wherein the number of recalls is very low.

schedule $\boldsymbol{x}_i$	$\boldsymbol{temp}_{final}$	iterations	time $[sec]$	profit [€]	agents
$\boldsymbol{x}_{26}$	100	7598	6	15 956.10	200
$\boldsymbol{x}_{27}$	10	5989	24	15 959.30	200
$\boldsymbol{x}_{28}$	1	5296	44	16 131.50	200
$\boldsymbol{x}_{29}$	0.1	4891	60	16 141.10	200
$\boldsymbol{x}_{30}$	0.01	4603	78	16 141.10	200

Table 5.17. Influence of the cooling factor on the performance of the algorithm with $f_{mul} = 0.999$ and $\boldsymbol{temp}_0 = 200$

In Table 5.17 the results of the optimisation heuristic for different final temperatures are presented. Similar to Table 5.10 on Page 183 we chose final temperatures between 0.01 and 100. The results for final temperatures

$\boldsymbol{temp}_{final}$ of 0.01 and 0.1 are identical, i.e., a smaller final temperature does not lead to any improvement. If the final temperature is very low, only few inferior moves are allowed such that the algorithm may be trapped in a local optimum. The best solutions were found for the lowest and second to lowest final temperature. In these and the other cases except for schedule $\boldsymbol{x}_{27}$, the 200 agents again work 1 208 hours in total. In schedule $\boldsymbol{x}_{27}$ 1202 hours are distributed over the agents.

$\boldsymbol{temp}_{final}$	$W^F_{agg}(T)$[sec]	$P^F_{agg}(\text{served}, T)$	$U^F_{agg}(T)$
100	1.681	0.944	0.975
10	1.780	0.941	0.978
1	1.423	0.953	0.979
0.1	1.405	0.953	0.980
0.01	1.405	0.953	0.980

Table 5.18. Aggregated performance measures of the contact center configuration for different final temperatures and a limited total number of agents

The aggregated performance measures are reported in Table 5.18. As explained before, the performance of the contact centers is amazingly good. Nearly all customers are served and the waiting time is again negligibly small.

In this section we have shown that the heuristic optimisation algorithm works well. The initial schedule determined by the initial procedure is already a good solution to the optimisation problem. If the total number of agents is limited, the determined schedules are quite similar. The initial and final temperature as well as the cooling factor should be carefully adjusted. A high cooling factor leads to better solution but also to a very long running time of the algorithm. The influence of the initial and final temperature is less important. A too high initial temperature allows for too many inferior moves, so that the heuristic algorithm may become trapped in a local optimum. A too low final temperature increases the running time, but leads to no significant improvements of the solution. The time the algorithm needs for optimisation is independent of the dimension of the contact center. If the contact center is very small, the optimisation should be done more carefully, as the performance is much more sensitive to changes in the parameters.

5.5 Shift Scheduling for a Contact Center with Heterogeneous Customers and Agents

5.5.1 Formulation of a Generic Shift Scheduling Problem

In this section we aim to formulate the simultaneous staffing and shift scheduling problem for the contact center model depicted in Figure 4.51 on Page 120 and analysed in Section 4.2. In this case with three different types of agent groups, the row vector $\boldsymbol{x}$ representing the schedule consists of three row vectors $\boldsymbol{x}^{(1)} = (x_1^{(1)}, \ldots, x_{K_1}^{(1)})$, $\boldsymbol{x}^{(2)} = (x_1^{(2)}, \ldots, x_{K_2}^{(2)})$, and $\boldsymbol{x}^{(G)} = (x_1^{(G)}, \ldots, x_{K_G}^{(G)})$ with $K = K_1 + K_2 + K_G$, i.e.,

$$\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(G)}). \tag{5.29}$$

The first vector $\boldsymbol{x}^{(1)}$ describes the number of type-1 specialists working according to the shift types, the second vector $\boldsymbol{x}^{(2)}$ contains the numbers of type-2 specialists and the third $\boldsymbol{x}^{(G)}$ the number of generalists assigned to the respective shift type.

If different shifts for each agent type are considered, the set of all possible shifts $\$$ may consist of three subsets

$$\$_1 = \left\{ s_1^{(1)}, \ldots, s_{K_1}^{(1)} \right\}, \tag{5.30a}$$

$$\$_2 = \left\{ s_1^{(2)}, \ldots, s_{K_2}^{(2)} \right\}, \tag{5.30b}$$

and

$$\$_G = \left\{ s_1^{(G)}, \ldots, s_{K_G}^{(G)} \right\} \tag{5.30c}$$

for each type of agents. Otherwise K_1, K_2, and K_G are identical and $\$$ reduces to $\$_1$. In the optimisation procedure we order the shift types equivalently to the schedule vector, i.e., the shift vectors of type-1 specialists have number 1 to K_1, the shift vectors of type-2 specialists have number $K_1 + 1$ to $K_1 + K_2$ and the shift vectors of generalists have number $K_1 + K_2 + 1$ to K. Type-1-specialists earn an hourly wage $w^{(1)}$, type-2 specialists $w^{(2)}$, and generalists $w^{(G)}$. By means of the hourly wages of each agent group we are able to calculate the costs $c_k^{(i)}$, $i = 1, 2, G$, for each shift type k, $k = 1, \ldots, K_i$, according to Equation (5.7) on Page 163.

Two different customer classes are considered in this contact center model shown in Figure 4.51 on Page 120, which leads to different revenues for served customers. Served customers of the first type lead to a revenue of r_1 and served customers of type 2 yield a revenue of r_2. Both customer classes cause costs for occupied lines which are assumed to be equal without loss of generality for both customer classes, because the customers in the system cannot be distinguished from outside by the telephone provider. If different telephone numbers for the two customer classes are assumed, the costs for the occupied lines could differ as well. The other parameters of the contact center are as described on Page 120.

By means of these parameters and the notation used in Section 4.2 on Pages 119ff. we are able to formulate the profit function of the contact center with two customer classes and three kinds of agents. As explained on Page 164 for the contact center with homogeneous agents and customers the profit function is given by

$$\begin{aligned} \text{profit}(\boldsymbol{x}, T) = \int\limits_{t_0}^{T} & r_1 \mu_1 \min\{Q_1^F(\boldsymbol{x}, t), N_1(\boldsymbol{x}, t)\} + r_1 \overline{\mu_1} B_1^F(t) \\ & + r_2 \mu_2 \min\{Q_2^F(\boldsymbol{x}, t), N_2(\boldsymbol{x}, t)\} + r_2 \overline{\mu_2} B_2^F(t) \\ & - \ell \left(Q_1^F(\boldsymbol{x}, t) + Q_2^F(\boldsymbol{x}, t) \right) \, dt \\ & - \sum_{k=1}^{K_1} c_k^{(1)} x_k^{(1)} - \sum_{k=0}^{K_2} c_k^{(2)} x_k^{(2)} - \sum_{k=0}^{K_G} c_k^{(G)} x_k^{(G)}. \end{aligned} \tag{5.31}$$

The constraints of the optimisation problem are similar to those illustrated in Section 5.1. The only difference between this and the optimisation problem formulated in (5.12) on Page 165 is that here the different classes of customers and agents lead to additional constraints.

First of all, the initial value problem in Equation (5.12b) and (5.12c) is enlarged to the problem analysed in Subsection 4.2.5 on Pages 131ff, i.e., we have four instead of two differential equations and initial value conditions. For each agent group the number of available agents may be limited. We assume that the maximum number of agents of each type is limited by M_1, M_2, and M_G, respectively. The number of agents of each type in the different time intervals are calculated according to Equation (5.6) on Page 163. Furthermore, the number of staffed shifts in the schedule $\boldsymbol{x}$ must be non-negative and integer-valued.

In order to make notation shorter in the formulation of the staffing and shift scheduling problem we use the departure rate $d_1(t)$ and $d_2(t)$ given in Equations (4.18) and (4.21) on Pages 121f. as well as the number of waiting customers $L_1^F(t)$ and $L_2^F(t)$ given by Equations (4.17) and (4.22) on Page 123f. Consequently, the staffing and shift scheduling optimisation problem in the case of the contact center discussed in Section 4.1 on Pages 73ff. can be formulated as follows:

$$\max_{\boldsymbol{x}} \quad \text{profit}(\boldsymbol{x}, T) \tag{5.32a}$$

subject to

$$\begin{aligned} \frac{d}{dt} Q_1^F(\boldsymbol{x}, t) &= \lambda_1(t) + \gamma_1 Q_{\mathcal{O}1}^F(\boldsymbol{x}, t) - d_1(t) \\ \frac{d}{dt} Q_2^F(\boldsymbol{x}, t) &= \lambda_2(t) + \gamma_2 Q_{\mathcal{O}2}^F(\boldsymbol{x}, t) - d_2(t) \\ \frac{d}{dt} Q_{\mathcal{O}1}^F(\boldsymbol{x}, t) &= p_1 \nu_1 L_1^F(t) - \gamma_1 Q_{\mathcal{O}1}^F(\boldsymbol{x}, t) \\ \frac{d}{dt} Q_{\mathcal{O}2}^F(\boldsymbol{x}, t) &= p_2 \nu_2 L_2^F(t) - \gamma_2 Q_{\mathcal{O}2}^F(\boldsymbol{x}, t), \end{aligned} \tag{5.32b}$$

$$Q_1^F(t_0) = 0, \quad Q_2^F(t_0) = 0, \quad Q_{\mathcal{O}1}^F(t_0) = 0, \quad Q_{\mathcal{O}2}^F(t_0) = 0 \tag{5.32c}$$

$$N_i(\boldsymbol{x}, t) = \sum_{k=0}^{K_i} s_k^{(i)}(t) x_k^{(i)} \qquad \text{for all } t \in [t_0, T] \text{ and } i = 1, 2, G, \tag{5.32d}$$

$$\sum_{k=0}^{K_i} x_k^{(i)} \leq M_i \qquad \text{for } i = 1, 2, G, \tag{5.32e}$$

$$x_k^{(i)} \in \mathbb{N}_0 \qquad \text{for all } k = 1, \ldots, K \text{ and } i = 1, 2, G. \tag{5.32f}$$

This optimisation problem leads to the same problems with respect to solving as the simpler model discussed before. The occasional non-concavity of the profit function[14] becomes even more obvious as well as the non-linearity of the problem.

5.5.2 Modification of the Heuristic Optimisation Procedures

In this part we briefly describe how we adjusted the opening procedure of Subsection 5.3.1 and the parameters of the problem, so that we can easily apply the simulated annealing algorithm explained in Subsection 5.3.2.

Similar to the homogeneous case the algorithm determines the average arrival rates for type-1 $\overline{\lambda}_{1,j}$ and type-2 customers $\overline{\lambda}_{2,j}$ in each time interval $j = 1, \ldots, \mathcal{J}$ according to Equation (5.16) on Page 170 first. Additionally, three vectors $\overline{\boldsymbol{\lambda}}_G$, $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$ are introduced which have the same dimension as $\overline{\boldsymbol{\lambda}}_1$ and $\overline{\boldsymbol{\lambda}}_2$. The entries of the vector $\overline{\boldsymbol{\lambda}}_G$ are the sum of the average arrival rates of type-1 and type-2 customers, i.e.,

$$\overline{\lambda}_{G,j} = \overline{\lambda}_{1,j} + \overline{\lambda}_{2,j} \quad \text{for all } j = 1, \ldots, \mathcal{J}. \tag{5.33}$$

The other two vectors are needed to consider the priority rule. The entries of the first vector $\boldsymbol{\theta}_1$ have value one, if the generalists serve only type-1 customers. Otherwise the entries give the fraction of generalists who could serve type-1 customers, i.e.,

$$\theta_{1,j} = \begin{cases} 1, & \overline{\lambda}_{1,j} \geq \delta \overline{\mu_1} \\ \dfrac{\overline{\lambda}_{1,j}}{\overline{\lambda}_{G,j}}, & \text{otherwise} \end{cases} \qquad \text{for all } j = 1, \ldots, \mathcal{J}. \tag{5.34}$$

Contrarily, the vector $\boldsymbol{\theta_2}$ is defined to have entries zero if $\theta_{1,j}$ is one. Otherwise the entry denotes the fraction of generalists who could serve type-2 customers. These vectors are needed to calculated the relative profit margins of the shifts associated with generalists and to update the average residual arrival rate if a generalist is staffed in the opening procedure. Therefore $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ have to be recalculated in each iteration.

Next the relative profit margins $\operatorname{marg}\left(\overline{\boldsymbol{\lambda}}_i, \boldsymbol{s}_k^{(i)}\right)$ for $k = 1, \ldots, K_i$ and $i = 1, 2, G$ are determined. For the shifts associated with the specialists these

[14] Compare the Figures 4.84-4.88 on Pages 155-157.

margins are calculated according to Equation (5.19) on Page 171. In order to compute the profit margins of the shifts associated with generalists, the different arrival and service rates as well as the different revenues for type-1 and type-2 customers have to be considered. For this purpose we need the vectors $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$. By means of these vectors an average service rate

$$\overline{\mu_{G,j}} = \theta_{1,j}\overline{\mu_1} + \theta_{2,j}\overline{\mu_2} \quad \text{for all intervals } j = 1, \ldots, \mathcal{J} \tag{5.35}$$

is estimated, which is compared to the average total arrival rate $\overline{\lambda}_{G,j}$ in each interval. If the average service rate is smaller than or equal to the average total arrival rate, i.e., $\overline{\mu_{G,j}} \leq \overline{\lambda}_{G,j}$, the profit margin in interval j, $j = 1, \ldots \mathcal{J}$ is given by

$$\operatorname{marg}\left(\overline{\lambda}_{G,j}, s_{k,j}^{(G)}\right) = \left(r_1\theta_{1,j}\overline{\mu_1} + r_2\theta_{2,j}\overline{\mu_2} - \ell - w^{(G)}\right)\delta s_{k,j}^{(G)} \tag{5.36}$$

for all $k = 1, \ldots, K$. Otherwise the profit margin is computed according to

$$\operatorname{marg}\left(\overline{\lambda}_{G,j}, s_{k,j}^{(G)}\right) = \left(r_1\theta_{1,j}\overline{\lambda}_{1,j} + r_2\theta_{2,j}\overline{\lambda}_{2,j} - \ell\frac{\overline{\lambda}_{G,j}}{\overline{\mu_{G,j}}} - w^{(G)}\right)\delta s_{k,j}^{(G)} \tag{5.37}$$

for all $k = 1, \ldots, K$. The profit margin for a generalist working according to shift $\boldsymbol{s}_k^{(G)}$, $k = 1, \ldots, K_G$ is the sum of these partial margins divided by the length of the shift, i.e.

$$\operatorname{marg}\left(\overline{\boldsymbol{\lambda}}_G, \boldsymbol{s}_k^{(G)}\right) = \frac{1}{|\boldsymbol{s}_k^{(G)}|}\sum_{j=1}^{\mathcal{J}} \operatorname{marg}\left(\overline{\lambda}_{G,j}, s_{k,j}^{(G)}\right). \tag{5.38}$$

If the total numbers of agents M_1, M_2, and M_G are limited and the number of staffed agents of one class has reached the respective number, the margins are set to be -1. Consequently, no agents of this class are staffed any more.

Similar to the opening procedure for the homogeneous case, the profit margins are compared and the shift with the highest margin marg $\left(\overline{\boldsymbol{\lambda}}_i, \boldsymbol{s}_k^{(i)}\right)$ is staffed if the maximum is unique. If the maximum is not unique, then the set $\mathbb{K}$ of pairs (κ, ι) identifying the maximum profit margins contains more than one element where κ refers to the shift type and ι to the agent type. In this case the accumulated work for the shift and agent type pairs (k, i) from the set $\mathbb{K}$ with the highest margin is determined by means of Equation (5.21) on Page 172. In the case of generalists the average total arrival rates $\overline{\lambda}_{G,j}$ and the average service rates $\overline{\mu_{G,j}}$ are chosen for the calculation. Then the shift and agent type with the highest cumulative work is staffed. Finally, the residual average arrival rates have to be updated. Algorithm 4 shows the adjusted initial procedure.

For the Algorithm 2 on Page 173 in Subsection 5.3.2 only the function OPERATION_APPLICABLE described in Algorithm 3 on Page 175 in lines 15-26 has to be adjusted with respect to the maximum number of agents of each type M_1, M_2, and M_G. Therefore, we have to distinguish whether the indices i and j are between 1 and K_1 for type-1 specialists, between $K_1 + 1$ and $K_1 + K_2$ for type-2 specialists, or between $K_1 + K_2 + 1$ and K for generalists.

Algorithm 4 Modified Initial Procedure

```
Require: $\boldsymbol{x} = \mathbf{0}$
for $j = 1, \ldots, \mathcal{J}$ do                ▷ Estimation of the average arrival rates
    Determine $\overline{\lambda}_{1,j}$ and $\overline{\lambda}_{2,j}$ according to (5.16)
end for

repeat
    for $j = 1, \ldots, \mathcal{J}$ do
        $\overline{\lambda}_{G,j} = \overline{\lambda}_{1,j} + \overline{\lambda}_{2,j}$                ▷ Calculation of the total arrival rates
        if $\overline{\lambda}_{1,j} \geq \overline{\mu_1}$ then                ▷ Calculation of $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$
            $\theta_{1,j} = 1, \theta_{2,j} = 0$
        else
            $\theta_{1,j} = \overline{\lambda}_{1,j}/\overline{\lambda}_{G,j}$, $\theta_{2,j} = \overline{\lambda}_{2,j}/\overline{\lambda}_{G,j}$
        end if
        $\overline{\mu_{G,J}} = \theta_{1,j}\overline{\mu_1} + \theta_{2,j}\overline{\mu_2}$                ▷ Calculation of the average service rates
    end for

    for $k = 1, \ldots, K_1$ do                ▷ Calculation of the relative profit margins
        if $\sum_{k=1}^{K_1} x_k^{(1)} \leq M_1$ then
            Determine marg $\left(\overline{\boldsymbol{\lambda}}_1, \boldsymbol{s}_k^{(1)}\right)$ according to (5.19)
        else
            marg $\left(\overline{\boldsymbol{\lambda}}_1, \boldsymbol{s}_k^{(1)}\right) = -1$
        end if
    end for

    for $k = 1, \ldots, K_2$ do
        if $\sum_{k=1}^{K_2} x_k^{(2)} \leq M_2$ then
            Determine marg $\left(\overline{\boldsymbol{\lambda}}_2, \boldsymbol{s}_k^{(2)}\right)$ according to (5.19)
        else
            marg $\left(\overline{\boldsymbol{\lambda}}_2, \boldsymbol{s}_k^{(2)}\right) = -1$
        end if
    end for

    for $k = 1, \ldots, K_G$ do
        if $\sum_{k=1}^{K_G} x_k^{(G)} \leq M_G$ then
            Determine marg $\left(\overline{\boldsymbol{\lambda}}_G, \boldsymbol{s}_k^{(G)}\right)$ according to (5.36), (5.37), and (5.38)
        else
            marg $\left(\overline{\boldsymbol{\lambda}}_G, \boldsymbol{s}_k^{(G)}\right) = -1$
        end if
    end for

    $\mathbb{K} = \left\{ (\kappa, \iota) \,\middle|\, \text{marg}\left(\overline{\boldsymbol{\lambda}}_\iota, \boldsymbol{s}_\kappa^{(\iota)}\right) = \max\left\{ \text{marg}\left(\overline{\boldsymbol{\lambda}}_i, \boldsymbol{s}_k^{(i)}\right) \,\middle|\, \begin{array}{l} k = 1, \ldots, K_i, \\ i = 1, 2, G \end{array} \right\} \right\}$
```

Continued on the next page

Algorithm 4 Modified Initial Procedure – continued

```
    if $(\kappa, \iota)$ is not unique, i.e., $|\mathbb{K}| > 1$ then
        for $(k, i) \in \mathbb{K}$ do                ▷ Calculation of the accumulated work
            if $i = 1$ or $i = 2$ then
                $\text{cum\_work}\left(\overline{\boldsymbol{\lambda}}_i, \boldsymbol{s}_k^{(i)}\right) = \sum_{j=1}^{\mathcal{J}} \frac{\overline{\lambda}_{i,j}}{\mu_i} s_{k,j}^{(i)}$
            else
                $\text{cum\_work}\left(\overline{\boldsymbol{\lambda}}_G, \boldsymbol{s}_k^{(G)}\right) = \sum_{j=1}^{\mathcal{J}} \frac{\overline{\lambda}_{G,j}}{\overline{\mu_{G,j}}} s_{k,j}^{(G)}$
            end if
        end for
        $(\kappa, \iota) = (k, i) \in \mathbb{K} \,\Big|\, \text{cum\_work}\left(\overline{\boldsymbol{\lambda}}_\iota, \boldsymbol{s}_\kappa^{(\iota)}\right) = \max\left\{\text{cum\_work}\left(\overline{\boldsymbol{\lambda}}_i, \boldsymbol{s}_k^{(i)}\right)\right\}$
    end if

    $x_\kappa^{(\iota)} \leftarrow x_\kappa^{(\iota)} + 1$

    for $j = 1, \dots, \mathcal{J}$ do                ▷ Calculation of the residual arrival rate
        if $\iota = 1$ or $\iota = 2$ then
            $\overline{\lambda}_{\iota,j} \leftarrow \max\left\{0, \overline{\lambda}_{\iota,j} - \mu_\iota s_{\kappa,j}^{(\iota)}\right\}$
        else
            $\overline{\lambda}_{1,j} \leftarrow \max\left\{0, \overline{\lambda}_{1,j} - \overline{\mu_1}\theta_{1,j} s_{\kappa,j}^{(\iota)}\right\}$
            $\overline{\lambda}_{2,j} \leftarrow \max\left\{0, \overline{\lambda}_{2,j} - \overline{\mu_2}\theta_{2,j} s_{\kappa,j}^{(\iota)}\right\}$
        end if
    end for

until $\left\{\begin{array}{l} \text{marg}\left(\overline{\boldsymbol{\lambda}}_i, \boldsymbol{s}_k^{(i)}\right) < 0 \text{ for all } k = 1, \dots K_i, i = 1, 2, G \\ \textbf{or} \\ \sum_{k=1}^{K_i} x_k^{(i)} \geq M_i \text{ for all } i = 1, 2, G \end{array}\right\}$
Calculate: profit$(\boldsymbol{x}_{current})$
```

5.5.3 Numerical Results

In this subsection we present some results for the contact center with heterogeneous structures. We concentrate on the influence of the retrial and service rates linked to the wages, in order to show that the opening procedure as well as the main improvement algorithm work fine.

For this purpose, we assume the 31 shift vectors shown in Table 5.2 on Page 167 for each agent group, i.e., the sets of shifts $\mathbb{S}_1$. $\mathbb{S}_2$, and $\mathbb{S}_G$ are equal. As we assume that the contact center opens at 7 am and closes at 8 pm, the shift vectors are 26-dimensional. The schedule $\boldsymbol{x}$ is a 93-dimensional vector, because three types of agent are assumed who can be staffed according to 31 different shifts. The different time-dependent arrival rate functions of type-1 and type-2 customers are given by the three cases discussed in Section 4.2 on Pages 132f.[15].

[15] See Equations (4.44) through (4.47) and Figures 4.52 and 4.53 on Page 132.

Service rate					
	μ_1	μ_2	$\overline{\mu_1}$	$\overline{\mu_2}$	
	$60\,\mathrm{h}^{-1}$	$60\,\mathrm{h}^{-1}$	$50\,\mathrm{h}^{-1}$	$50\,\mathrm{h}^{-1}$	
			$60\,\mathrm{h}^{-1}$	$60\,\mathrm{h}^{-1}$	

Abandonment and retrial parameters					
ν_1	ν_2	γ_1	γ_2	p_1	p_2
$120\,\mathrm{h}^{-1}$	$120\,\mathrm{h}^{-1}$	$0.5\,\mathrm{h}^{-1}$	$0.5\,\mathrm{h}^{-1}$	0.5	0.5
		$12\,\mathrm{h}^{-1}$	$12\,\mathrm{h}^{-1}$		

Revenue and costs					
r_1	r_2	ℓ	$w^{(1)}$	$w^{(2)}$	$w^{(G)}$
0.5 €	0.5 €	6 €/h	10 €/h	10 €/h	12 €/h
			11 €/h	11 €/h	10 €/h

Table 5.19. Parameters of the contact centers in the shift scheduling problem

The parameters of the contact center model are given in Table 5.19. The values in the first line below the parameters are the default values of the contact center model, while the values in the second line are alternative values. These values were used in the investigation of the influence of the parameters on the performance of the contact center.

For the heuristic optimisation algorithm we use an initial temperature of $\boldsymbol{temp}_0 = 200$, final temperature of $\boldsymbol{temp}_{final} = 0.1$ and a cooling factor of $f_{mul} = 0.99999$, although this leads to long computing times. As shown in Section 5.4, good solutions can already be calculated with a cooling factor of $f_{mul} = 0.999$, which leads to a running time of less than 1 minute on the notebook computer used for calculation.

First of all the total numbers of agents M_1, M_2, and M_G are assumed to be big enough, such that they can be judged as unlimited. Afterwards we suppose that the maximum total numbers of type-1 and type-2 specialists are 200 and the maximum number of generalists is 50. For the initial solution the retrial rates γ_1 and γ_2 do not matter much because they are not considered in the initial procedure. However, the information is needed to calculate the profit of the initial schedule at the end of the opening procedure.

schedule $\boldsymbol{x}_{initial,1}$						
	γ_i	profit [€]	$W^F_{i,agg}(T)$	$P^F_{i,agg}(\mathrm{s},T)$	$\sum_{k=1}^{K_i} x_k^{(i)}$	$U^F_{i,agg}(T)$
1	0.5	30471.60	0.623 sec	0.979	198	0.958
2	0.5		0.623 sec	0.979	198	0.958
G					0	0.000
1	12.0	30321.50	0.841 sec	0.972	198	0.956
2	12.0		0.841 sec	0.972	198	0.956
G					0	0.000

Table 5.20. Solutions of the initial procedure for a contact center with equal arrival rate functions for both customer classes and different retrial rates

If both types of customers have identical arrival rate functions, the initial solution calculated by the initial procedure is already very good. The retrial rate has a significant influence on the quality of the initial solution as shown in Table 5.20. If the retrial rates are higher, i.e., the sojourn times in orbit are shorter, the profit, probability of being served, and the utilisation decrease, while the waiting time increases. In the case of the identical arrival rate functions considered here, fewer than 200 specialists of each type are needed. Therefore the initial solutions for limited and unlimited total numbers of agents are identical.

The performance of the contact center with the initial schedules is very high. Almost all customers are served and the aggregated waiting times are negligible. The utilisation of agents is high. In both cases no generalists are staffed, because the specialists work faster and are cheaper. Therefore, the estimated relative profit margins of the shifts associated with generalists are always smaller than the relative profit margins of the other shifts. After 198 shifts with specialists of each type are staffed, all the relative profit margins are negative and the opening procedure is stopped.

Schedule		γ_i	profit	$W^F_{i,agg}(T)$	$P^F_{i,agg}(\mathrm{s},T)$	$\sum_{k=1}^{K_i} x_k^{(i)}$	$U^F_{i,agg}(T)$
$\boldsymbol{x}_n$		$[\mathrm{h}^{-1}]$	[€]	[sec]			
$\boldsymbol{x}_1$	1	0.5	31 082.20	0.461	0.985	257	0.977
	2	0.5		0.466	0.985	251	0.977
	G					0	0.000
$\boldsymbol{x}_2$	1	12	31 070.00	0.578	0.981	252	0.978
	2	12		0.543	0.982	254	0.977
	G					0	0.000
$\boldsymbol{x}_3$	1	0.5	31 079.30	0.481	0.984	262	0.978
	2	12		0.540	0.982	251	0.977
	G					0	0.00

Table 5.21. Solutions of the optimisation heuristic for a contact center with equal arrival rate functions for both customer classes, unlimited total number of agents and different retrial rates

Compared to the initial solution all performance measures have improved and more shifts of specialists have been scheduled, because a lot of long shifts have been substituted by short shifts. Therefore more shifts are needed. The profit increases by 2% in the case of low retrial rates and by 2.5% in the case of high retrial rates. The optimum profit for all retrial rates are almost equal, i.e., the influence of the retrial rate decreases if the schedule is optimised. Furthermore, no generalists have been staffed, i.e., the contact center with equal arrival rates can be planned like two isolated contact centers.

Table 5.22 presents the resulting schedules referring to the contact center performance values in Table 5.21. As no shifts for generalists are staffed, the vector $\boldsymbol{x}^{(G)}$ is not shown. These schedules are similar to the schedules

Shift	schedule $\boldsymbol{x}_{initial,1}$		schedule $\boldsymbol{x}_1$		schedule $\boldsymbol{x}_2$		schedule $\boldsymbol{x}_3$	
$s_k \in \mathbb{S}$	$\boldsymbol{x}^{(1)}_{initial}$	$\boldsymbol{x}^{(2)}_{initial}$	$\boldsymbol{x}^{(1)}_1$	$\boldsymbol{x}^{(2)}_1$	$\boldsymbol{x}^{(1)}_2$	$\boldsymbol{x}^{(2)}_2$	$\boldsymbol{x}^{(1)}_3$	$\boldsymbol{x}^{(2)}_3$
s_1	1	1	0	0	0	0	1	1
s_2	8	8	0	3	0	0	1	1
s_3	5	5	1	5	3	1	3	2
s_4	9	9	3	4	3	5	1	5
s_5	16	16	1	3	3	2	1	2
s_6	21	21	3	4	3	3	2	3
s_7	23	23	4	5	4	4	3	5
s_8	22	22	5	7	7	5	4	7
s_9	18	18	9	9	9	9	7	8
s_{10}	4	4	10	8	11	11	8	11
s_{11}	1	1	7	4	6	8	5	7
s_{12}	1	1	1	0	1	0	1	0
s_{13}	2	2	1	1	1	1	0	0
s_{14}	4	4	10	6	9	10	7	9
s_{15}	10	10	15	13	14	14	16	14
s_{16}	11	11	21	18	18	19	20	19
s_{17}	8	8	23	21	24	23	25	23
s_{18}	4	4	23	24	23	23	24	22
s_{19}	0	0	21	17	18	20	21	18
s_{20}	0	0	9	10	10	10	12	11
s_{21}	0	0	2	4	1	1	3	0
s_{22}	0	0	1	0	2	0	1	0
s_{23}	0	0	1	0	0	2	1	1
s_{24}	0	0	4	1	2	2	4	2
s_{25}	0	0	12	10	11	11	13	11
s_{26}	0	0	17	18	17	17	18	17
s_{27}	0	0	21	18	19	20	21	18
s_{28}	0	0	16	15	16	17	19	17
s_{29}	0	0	9	12	9	9	11	10
s_{30}	8	8	5	7	6	4	7	4
s_{31}	22	22	2	4	2	3	2	3

Table 5.22. Schedules calculated by the optimisation heuristic for the contact center with equal arrival rate functions, unlimited total number of agents and different retrial rates

calculated in the Section 5.4 shown in Tables 5.5, 5.8, and 5.11 on Page 177f. for the contact center with a single class of customers and homogeneous agents. The improvement algorithm substitutes a lot of long shifts by short shifts so that more agents are staffed.

If we limit the total number of agents of each type such that at most 200 specialists of each type and 50 generalists can be scheduled, no generalists are staffed in the resulting schedules either. The difference of the profit calculated from the initial schedule and the final schedule of the optimisation heuristic has become smaller. Similar to the unlimited case the profit and the utilisation of the different agent groups have increased as well as the aggre-

Schedule		γ_i	profit	$W^F_{i,agg}(T)$	$P^F_{i,agg}(\mathrm{s},T)$	$\sum_{k=1}^{K_i} x_k^{(i)}$	$U^F_{i,agg}(T)$
$\boldsymbol{x}_n$		$[\mathrm{h}^{-1}]$	[€]	[sec]			
$\boldsymbol{x}_4$	1	0.5	30 919.30	0.692	0.977	200	0.978
	2	0.5		0.692	0.977	200	0.978
	G					0	0.000
$\boldsymbol{x}_5$	1	12	30 787.70	0.831	0.972	200	0.975
	2	12		0.829	0.972	200	0.975
	G					0	0.000

Table 5.23. Solutions of the optimisation heuristic for equal arrival rate functions for both customer classes, limited total numbers of agents and different retrial rates

gated waiting time. The probability of being served has slightly decreased. However, the performance of the solution remains remarkably good as shown in Table 5.23. The difference between the case of unlimited numbers of agents and limited numbers of agents is very small.

The influence of the retrial rate is very small, because almost all customers are served at their first attempt. The statements made on the influence of the retrial parameters in Sections[16] 4.1 and 4.2 are confirmed by the results shown here.

Next we compare the different initial solutions with respect to the profit for different arrival rate functions, limited and unlimited total numbers of agents, service rates of generalists, and wages of all agents in Tables 5.24 and 5.25. As the retrial rates do not influence the schedule of the initial solution, we assume identical retrial rates of both customer classes $\gamma_1 = \gamma_2 = 0.5\,\mathrm{h}^{-1}$. The other parameters are given in Table 4.4 on Page 133.

Schedule	Agents			Wages			service rates		Profit
$\boldsymbol{x}_{initial}$	1 (M_1)	2 (M_2)	G (M_G)	1	2	G	$\overline{\mu_1}$	$\overline{\mu_2}$	€
2	191 (∞)	206 (∞)	20 (∞)	10	10	12	50	50	30 633.40
3	190 (200)	200 (200)	21 (50)	10	10	12	50	50	30 623.30
4	0 (∞)	5 (∞)	397 (∞)	10	10	10	60	60	31 134.70
5	0 (∞)	0 (∞)	483 (∞)	11	11	10	50	50	23 613.50

Table 5.24. Comparison of the initial solutions with respect to the profit for different arrival rate functions presented in the first picture of Figure 4.53 on Page 132

If the arrival rate functions of the customer classes differ, in the opening procedure additional generalists are staffed, even if the number of agents is unlimited and the generalists are more expensive. The cheaper and the faster generalists are compared to the specialists, the more generalists are staffed.

[16] See Pages 83-86, 90-91, 94-96, 96-98, 100, 101, and 105 for the influence of the retrial parameters of type-1 customers and Pages 138-150 for the influence of the retrial parameters of type-2 customers.

Schedule	Agents			Wages			Service rates		Profit
$\boldsymbol{x}_{initial}$	1 (M_1)	2 (M_2)	G (M_G)	1	2	G	$\overline{\mu_1}$	$\overline{\mu_2}$	€
6	192 (∞)	200 (∞)	20 (∞)	10	10	12	50	50	32 162.80
6	192 (200)	200 (200)	20 (50)	10	10	12	50	50	32 162.80
7	0 (∞)	7 (∞)	401 (∞)	10	10	10	60	60	32 583.50
8	0 (∞)	0 (∞)	501 (∞)	11	11	10	50	50	24 532.30

Table 5.25. Comparison of the initial solutions with respect to the profit for different arrival rate functions presented in the second picture of Figure 4.53 on Page 132

The advantage of generalists is their ability to serve both customer classes. This advantage plays a major role if the arrival rate functions of the customers classes differ, i.e,, the minimum and maximum arrival rates are reached at different times of the day.

Based on these initial solutions the improvement algorithm searches for a better solution. The results of the optimisation heuristic for the initial schedules in Table 5.24 are given in Tables 5.26, 5.27, 5.28, and 5.29.

Schedule $\boldsymbol{x}_n$		M_i	profit [€]	$W^F_{i,agg}(T)$ [sec]	$P^F_{i,agg}(\mathrm{s},T)$	$\sum_{k=1}^{K_i} x_k^{(i)}$	$U^F_{i,agg}(T)$
$\boldsymbol{x}_6$	1	∞	31 666.40	0.439	0.985	274	0.981
	2	∞		0.808	0.971	244	0.964
	G	∞				4	0.808
$\boldsymbol{x}_7$	1	200	31 193.90	0.756	0.975	200	0.982
	2	200		0.765	0.939	200	0.972
	G	50				6	0.835

Table 5.26. Solutions of the optimisation heuristic for a contact center with different arrival rate functions presented in the first picture of Figure 4.53 on Page 132 and retrial rate $\gamma_1 = \gamma_2 = 0.5\,\mathrm{h}^{-1}$

In Table 5.26 the solution for limited and unlimited total numbers of agents are compared. Even in the case of unlimited numbers of agents some shifts for generalists are scheduled, although they work more slowly and are more expensive than the specialists. However, the major part of all customers is served by the specialists of their class. The performance of the contact center is high and the profit is only slightly higher in the case of unlimited total numbers of agents. If the number of agents is limited again, more long shifts are scheduled. The profit increases by 3.4% and by 1.9% compared to the initial solution, i.e., the initial solution is already very accurate.

In order to show that not only the aggregated performance measures but also the time-dependent performance measures in the case of a limited number of agents are very accurate, we show the time-dependent waiting time and probability of being served for type-2 customers for the final schedule $\boldsymbol{x}_7$ in Figure 5.14. The results for type-1 customers are similar to those presented

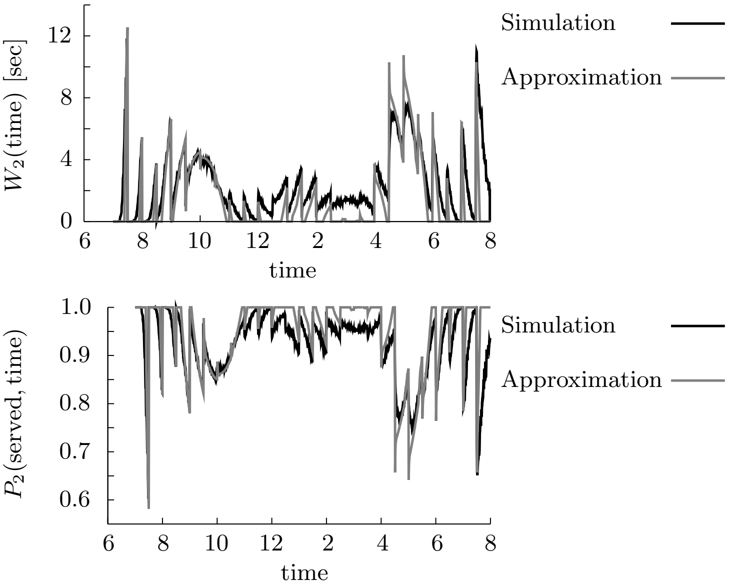

Fig. 5.14. Comparison of the approximation and simulation results for the time-dependence performance measures of type-2 customers for schedule $\boldsymbol{x}_7$

in Figure 5.9 on Page 180. If the contact center is staffed according to this schedule, it operates almost all the time at critical load. The waiting time and the probability of being served are approximated quite accurately. At the beginning and the end of the working day, when only a few customers call and few agents are on duty, the performance becomes worse. However, the customers have to wait only for few seconds and almost all customers are served during the day. The spikes in the curves are due to the fact that the numbers of agents can only be changed at the beginning of each time interval.

In addition to the time-dependent performance measures for type-2 customers we show the time-dependent utilisation of type-2 specialists and of generalists in Figure 5.15. While the utilisation of type-2 specialists is quite well approximated by the fluid approach, the results of the approximation and the simulation for the time-dependent utilisation of generalists differ extremely. This can be explained by the fluid assumption. In the simulation the customers are discrete so they remain with the generalists if their service has started. Therefore, the utilisation of the generalists is more constant. In the fluid approximation the customers are assumed to be continuous so that every part of a customer can be served. Consequently, one part might be served by a generalist while the other part is served by a specialist. As customers prefer to be attended to by a specialist, the generalists become available more often.

Comparing Tables 5.26 and 5.27 shows that the situation of type-2 customers improves with respect to the waiting time, if more calls of type-2 customers arrive later than the majority of the prioritised type-1 customers.

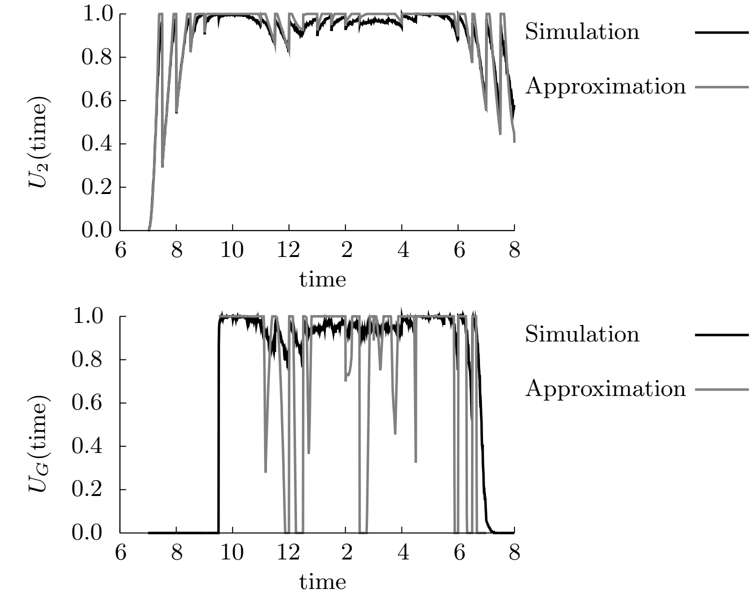

Fig. 5.15. Comparison of the approximation and simulation results for the time-dependence utilisation of type-2 specialists and generalists

Schedule $\boldsymbol{x}_n$		M_i	profit [€]	$W^F_{i,agg}(T)$ [sec]	$P^F_{i,agg}(\mathrm{s},T)$	$\sum_{k=1}^{K_i} x_k^{(i)}$	$U^F_{i,agg}(T)$
$\boldsymbol{x}_8$	1	∞	33 175.40	0.449	0.985	280	0.979
	2	∞		0.625	0.978	248	0.970
	G	∞				2	0.835
$\boldsymbol{x}_9$	1	200	32 730.50	0.877	0.971	200	0.982
	2	200		1.177	0.960	200	0.969
	G	50				3	0.784

Table 5.27. Solutions of the optimisation heuristic for a contact center with different arrival rate functions presented in the second picture of Figure 4.53 on Page 132 and retrial rate $\gamma_1 = \gamma_2 = 0.5\,\mathrm{h}^{-1}$

The reason for this result is that type-1 customers displace type-2 customers from the generalists. Then the value of the profit function increases as well, while the probability of being served and the utilisation stays almost the same.

Finally, we vary the service rate of generalists and the wages of agents to compare the influence of these parameters on the performance when the arrival rates of the two customer classes differ. For this purpose, we assume that the maximum number of agents is practically unlimited. Table 5.28 and Table 5.29 report the performance of the contact center with the resulting schedules.

Schedule		w_i	$\mu_i, \overline{\mu_i}$	profit	$W^F_{i,agg}(T)$	$P^F_{i,agg}(\mathrm{s},T)$	$\sum_{k=1}^{K_i} x_k^{(i)}$	$U^F_{i,agg}(T)$
$\boldsymbol{x}_n$		[€]	$[\mathrm{h}^{-1}]$	[€]	[sec]			
$\boldsymbol{x}_{10}$	1	10	60	31 972.80	0.006	0.999	179	0.997
	2	10	60		0.980	0.736	143	0.998
	G	10	60				200	0.946
			60					
$\boldsymbol{x}_{11}$	1	11	60	29 395.00	0.318	0.989	272	0.987
	2	11	60		0.818	0.956	237	0.973
	G	10	50				19	0.853
			50					

Table 5.28. Solutions of the optimisation heuristic for a contact center with different arrival rate functions presented in the first picture of Figure 4.53 on Page 132 and retrial rate $\gamma_1 = \gamma_2 = 0.5\,\mathrm{h}^{-1}$

If the generalists work as fast as the specialists and are as expensive, all types of agents are staffed. The total number of agents of each type are almost the same. Especially for type-1 customers the situation with respect to the waiting time and the probability of being served improves, while the situation of type-2 customers worsens, because some type-2 specialists are substituted by generalists. In the first case, with equal wages and service rates, the profit calculated based on the final schedule is by 2.7% better than the profit of the initial schedule. However, the number of agents staffed differs much. While in the initial solution almost no shifts for specialists are scheduled, the number of agents of each type are nearly the same in the final solution.The difference between the initial and final solution is enormous if the wages for generalists are lower than the wages for specialists but the generalists work a little more slowly. In this case the opening procedure seems to be less accurate. The profit increases by about 24.5%. In the initial schedule no shifts for specialists are scheduled, while in the final schedule the number of shifts for generalists is very small and the number of specialists has increased.

Similar conclusions can be drawn from Table 5.29 which presents the performance measures and profit gained by the final schedule in the case of the arrival rate function in the second picture of Figure 4.53 on Page 132. The profit and the utilisation increase if all agents serve the customers with equal mean time and equal hourly wage. The utilisation of specialists is almost one, because all customers prefer to be served by their specialists. Therefore, in the fluid model only the fraction of customers that exceeds the number of specialists in each moment, is served by generalists.

If the generalists serve customers more slowly than the specialists and the hourly wages of generalists are lower than the wages of specialists, only few generalists are staffed. However, more generalists and fewer specialists are staffed than in the cases of higher wages of generalists as shown in Tables 5.26

Schedule		w_i	$\mu_i, \overline{\mu_i}$	profit	$W^F_{i,agg}(T)$	$P^F_{i,agg}(\mathrm{s}, T)$	$\sum_{k=1}^{K_i} x_k^{(i)}$	$U^F_{i,agg}(T)$
$\boldsymbol{x}_n$		[€]	[h^{-1}]	[€]	[sec]			
$\boldsymbol{x}_{12}$	1	10	60	33 389.90	0.057	0.998	170	0.997
	2	10	60		0.886	0.733	180	0.996
	G	10	60				174	0.948
			60					
$\boldsymbol{x}_{13}$	1	11	60	30 773.00	0.326	0.989	267	0.987
	2	11	60		0.718	0.964	238	0.981
	G	10	50				13	0.804
			50					

Table 5.29. Solutions of the optimisation heuristic for a contact center with different arrival rate functions according to the second picture of Figure 4.53 on Page 4.53 and retrial rate $\gamma_1 = \gamma_2 = 0.5\,\mathrm{h}^{-1}$

and 5.27. Therefore, the utilisation of specialists and the waiting time of type-2 customers increase.

The examples of this section show that the heuristic optimisation determines good schedules. Even the schedule created by the opening procedure leads to the high performance of the contact center. The utilisation of agents is often very high, such that agents might suffer from stress. Almost all customers are served and the waiting times are short.

5.6 Overview of Current Literature on Contact Center Staffing and Scheduling

During the last ten years the amount of literature on staffing and shift scheduling in call and contact center has increased, caused by the growing importance of call and contact centers in the service sector and the fact that 60%-70% of the costs in contact centers are personnel related. Koole and Pot (2006) give an overview of algorithms and problems related to routing and staffing in contact centers. In most staffing and shift scheduling tools for contact centers the simple Erlang-C queueing model is implemented[17], although a lot of work has already been done to develop better approaches and rules of thumb[18] based on more realistic models.

The literature can be roughly subdivided into three categories according to the phases of the operational planning process depicted in Figure 2.6 on Page 19. The first category is dedicated to the long-term and mid-term planning of call and contact centers. This category includes papers on hiring

[17] See Gans et al. (2003).

[18] For example, the well known square-root staffing rule for the Erlang-A model analysed by Garnett et al. (2002).

and firing employees, as described by Gans and Zhou (2002). Furthermore, the so-called call center outsourcing analysed in Gans and Zhou (2004) and Ren and Zhou (2004) falls into this category.

The second and largest category contains the staffing of agents in isolated or interrelated time intervals, where in this category the staffing goals give rise to a further differentiation. On the one hand economic staffing objectives are considered, e.g., in Helber and Stolletz (2003, 2004), Koole and Pot (2005), and Hampshire and Massey (2005). On the other hand most authors[19] aim to meet certain service level constraints. However, the border between these groups sometime blurs.

Based on the stationary approach for a call center with impatient customers and different cost and revenue structures found in german call centers, Helber and Stolletz aim to maximise the profit or to minimise the cost of the call center subject to aggregated performance measures. Koole and Pot investigate the profit function of this staffing optimisation further. Contrarily, Hampshire and Massey (2005) use a fluid approach for the profit function with a penalty for abandonment and blocking. In their optimisation approach the number of agents N_j and the number of telephone lines K_j in each period are determined. For this model they are able to show that the number of agents N_j and the number of additional telephone lines $K_j - N_j$ are complementary, i.e., either $N_j = 0$ or $K_j - N_j = 0$.

The staffing according to certain performance goals in telephone traffic has a longer tradition. In 1984 Sze described a queueing model with abandonment and time-dependent arrival rates for telephone operator staffing by means of simulation and approximation. A widely used rule of thumb is the so-called square-root staffing principle based on the Erlang-C model, which dates back to a study by A. K. Erlang[20]. If $\rho = \frac{\lambda}{\mu}$ is the offered load of the systems measured in *Erlangs* the number of agents needed is given by

$$N^* = \rho + \beta\sqrt{\rho}. \tag{5.39}$$

Borst et al. (2004) determine the parameter β such that the costs of staffing and queueing is slightly balanced. Garnett et al. (2002) extended the principle to the Erlang-A model. Based on these results Aguir et al. (2005) showed that disregarding retrials can lead to under- and overstaffing and incorporated the retrial parameters into the square-root staffing rule. Jennings et al. (1996) adjusted this principle to a model with time-dependent arrival and service rates via an infinite server model as presented in Sections 3.2 and 3.3. Feldman et al. (2005) are able to extended the approach by Jennings et al. further, such that not only a target for the probability of delay is met but also time-stable performance is achieved.

[19] Whitt (1999c), Chen and Henderson (2001), and Feldman et al. (2005)

[20] See Gans et al. (2003).

Besides these papers focussing on a homogeneous group of agents and statistically identical customer requests, the square-root staffing rule was also involved in staffing of multiclass[21] contact centers and skill-based routing[22] as a basic guess for the number of agents needed, where the staffing-rule proposed in Gurvich et al. (2004) is closely related to the control of a contact center [23], i.e., the assignment of customers to agents.

Other approaches for staffing of call centers which incorporated time-dependent and/or stochastic demand are based on fluid approximations[24], the so-called SIPP approach[25], or a numerical approximation by the Erlang-A model[26].

The last category mainly deals with shift scheduling for a given demand. The literature related to shift scheduling in call and contact centers dates back to the operator scheduling at telephone traffic switchboards. Segal (1974) determines the number of agents assigned to shifts at optimal costs based on a network-flow formulation. Henderson and Berry (1976) consider the problem of determining shifts as well as constructing a schedule. They use heuristic methods to solve the cost minimisation problem under the assumption of a predefined demand. More recently, Fukunaga et al. (2002) describe a staffing and shift scheduling software system which uses artificial intelligence search methods to solve a tour scheduling problem for a given demand. A tour scheduling problem integrates day off planning and shift scheduling[27]. This software is based on an Erlang-C model and aims to fulfil predefined technical performance targets.

Ingolfsson et al. (2002) use a genetic algorithm to search for good schedules of predefined shifts. They incorporate time-dependent and stochastic demand by solving the Chapman-Kolmogorov equations for a Erlang-C model and compare their results to the traditional SIPP[28] and an integer programming approach. In another paper Ingolfsson et al. (2003) improve this approach by involving the determination of employee requirements. The method iterates between the evaluation of a schedule and the generation of a new schedule by an integer programming approach until a feasible schedule is found. Similarly, Bhulai et al. (2006) use linear programming to solve the shift scheduling problem after having determined the staffing levels for the considered sequence of time intervals so that the service level is reached in each time interval. They extend their method to a multi-skill environment.

[21] See, e.g., Gurvich et al. (2004) and Harrison and Zeevi (2004).

[22] See Wallace and Whitt (2004)

[23] See, for example, Atar (2005a,b) and references therein

[24] See, Harrison and Zeevi (2005) and Whitt (2006a).

[25] See, Green et al. (2001) and Green et al. (2005).

[26] See Whitt (2006a).

[27] See Ernst et al. (2004) for a differentiation of these problem and further literature.

[28] See Section 3.1 and Green et al. (2001).

More theoretically orientated is the paper by Koole and Van der Sluis (2003), who introduce a local search algorithm which relies on the multimodularity of the objective function. In their approach the demand is given as in most other papers to shift scheduling[29]. Another more technical approach by Atlason et al. (2004) related to call center staffing is based on the concavity of the service level function[30]. They use simulation and cutting plane methods and show how the concavity of the objective function can be verified. Furthermore, they present some convergence results. Likewise, Cezik and L'Ecuyer (2005) use simulation and the cutting plane method on an integer programming approach to determine a schedule for a multi-skill call center.

Besides those call and contact center related papers, shift and tour scheduling problems are strongly linked to health care[31] and airline crew scheduling[32].

This review shows that the staffing and shift scheduling in contact centers is a widely discussed and ongoing research problem. Different methods have been used to determine the number of needed agents, i.e., the staff requirements, and to solve the shift scheduling problem for a given demand and service level. However, despite the papers by Ingolfsson et al. (2003) the determination of the staffing requirements and the shift scheduling problem were solved separately, whereby most authors stop when a feasible schedule was found. Furthermore, the staffing and shift scheduling approaches mainly rely on the either the stationary or the time-dependent Erlang-C model without abandonment and retrials.

[29] See, e.g., Bechtold and Jacobs (1996), Thompson (1996), Aykin (1998, 2000), Musliu (2001), Eveborn and Rönnqvist (2004), Eitzen et al. (2004), and the references therein.

[30] See also Koole and Pot (2006) for a brief discussion of this problem.

[31] See, e.g., Dowsland and Thompson (2000), Isken (2004), Moz and Vaz Pato (2004), Aickelin and White (2004), and especially the overview given in Burke et al. (2004).

[32] See, Kohl and Karisch (2004), Ernst et al. (2004) and the references therein.

6

Conclusions and Suggestions for Future Research

In this thesis dynamic inbound contact centers with heterogeneous agents and retrials of impatient customers were analysed. The term dynamic characterises the processes of inbound contact centers in two ways. On the one hand the processes of a real-world contact center are time-dependent and on the other they are random. The interarrival times, service durations and abandonment behaviour depend on the time of the day and the day of the week. Furthermore, customers contact the center at random times and the service durations and waiting time limits are random as well.

In order to model and analyse both aspects of dynamics, we used the so-called strong approximations which contain a fluid approximation and a diffusion refinement. By means of the fluid approach we showed how retrials and time-dependent parameters influence technical and economical performance measures. Furthermore, we used a diffusion refinement of the fluid approach to investigate the impact of the parameters on the variability of the queueing processes.

Finally, we developed an algorithm for integratively solving a staff requirement planning and shift scheduling problem which emerges in contact centers with time-dependent processes and retrials. This algorithm relies on the fluid approximation.

Besides heterogeneous and impatient customers, differently skilled agent groups, time dependencies and randomness, in many real-world contact centers the arrivals often consist of primary attempts and retrials of impatient customers. These retrials influence the performance of the contact center and can lead to under- or overestimation of the demand. In order to investigate the various impacts on the performance we use models with different kinds of customer requests and differently skilled agents. The customers are assumed to be impatient and a certain percentage of customers retries after having abandoned. Furthermore, the arrival rates strongly depend on time of the day.

In the traditional stationary queueing approach mainly used in practise just one aspect of dynamics can be considered. In this approach mainly the randomness of the arrivals, service completions and abandonments is addressed. This aspect has an enormous influence in very small contact centers, e.g., for emergency calls, while in big contact centers analysed in this thesis its influence diminishes. In the traditional approach the modelling of heterogeneous customers and agents as well as retrials is another serious problem because the stochastic processes become processes of higher dimensions. Consequently, the generator becomes a matrix of matrices and the stationary distibution becomes nearly uncomputable as argued in Section 3.1.4.

Therefore, we use the fluid approach. By means of this approach the influence of the time-dependencies can be shown and is stressed. Furthermore, the distribution of the random events becomes negligible, i.e., it does not matter whether the service times are exponentially, lognormally or normally distributed. We show, that the retrials can easily be modelled and even different customer classes and various groups of agents can be imbedded in this approach because the fluid approach leads to a simple first order und numerically stable initial value problem. This initial value problem can be solved by simple methods like the Euler method. Each customer class leads to two differential equations of the initial value problem: one to determine the number of customers in the system and one to determine the number of customers in the orbit waiting to recall. Additional agent groups lead to additional terms in the differential equations which describe the supplemental non-preemptive priority rule. The modelling by the fluid approach gives rise to a preemptive priority rule. Numerical examples show that the error due to the change is very small. Unfortunately, not as many performance measures are calculable as for the stationary approach, e.g., the well-known X/Y service level cannot be calculated. For this performance measure we need the distribution of the waiting time. However, we showed that only few performance measures are needed to analyse and benchmark the performance of a contact center. We derive the average waiting time, the probability of being served and the utilisation as technical performance measures. These measures are used to calculate the profit as an economical performance measure.

We investigate the influence of the service rate, the abandonment rate, number of agents, the mean time to retrial and the probability of retrial on the performance measures. Thereby we focus on the influence of the retrial parameter. Additionally, we compare the results to simulation results to show that the approximation is very accurate. Higher service rates and a greater number of agents lead to smaller waiting times, a higher probability of being served and less utilisation. The abandonment rates solely influence the waiting time such that higher rates leads to smaller waiting times. We show that the probability of being served and the utilisation are not effected. The influence of the mean time to retrial is shown to be smaller than the influence of the probability of retrial. The shorter mean times to retrial lead to a little longer waiting times in the case of high load. The influence on the probability of beeing served is similar. Higher probabilities of retrial give rise to higher waiting times and smaller probabilites of being served. Furthermore, we show that the retrial parameter do not influence the utilisation because retrial are solely caused, if the contact center is temporarily overloaded.

The diffusion refinement is based on the fluid approach. It considers the randomness of the processes. All random processes are approximated by normally distributed processes with a mean given by the fluid approach. That is why, some shortcomings of the fluid approach especially for the times of critical loading apply to the diffusion approximation as well. The diffusion refinement is needed to extend the initial value problem of the fluid approach for the variances and covariances of the processes describing the number of customers.

These differential equation for the variances and covariances are difficult to derive. In order to solve the extended initial value problem we had to use a Runge-Kutta-Method because the differential equations of the variances and covariances are quite sensitiv to the service and abandonment rates. If the service rate and the abandonment rate differ a lot, the variances and covariances change drastically. It was shown that the calculated variances and covariances are useful for the analysis of interdependencies of different customers processes. They might be of value to get more robust schedules of agents in a future optimisation approach. However, for a staff requirement planning and shift scheduling approach we do not need the diffusion refinement. This can be done solely by means of the fluid approximation.

We have modelled contact centers with retrials and time-dependent processes by means of fluid approximation and its diffusion refinement. Thereby we included the heterogeneity of customer requests as well as the different groups of agents. We restrict the model to the case of two customer classes and three agent groups. However, we showed that this model can easily be derived from the simple model and more complex models can be deduced as easily. We show that the number of customers in the system and the orbit are positive correlated. The correlation reduces if the contact center becomes more and more overloaded, i.e., the processes of the number of customers in the system and the number of customers waiting to retry become more and more independent.

The simulation we used is based on the traditional contact center model which assumes exponentially distributed interarrival and service times as well as waiting time limits and times to retrial. The advantages and disadvantages of this assumptions were discussed in Section 2.2. Therefore, other distributions for the service times, waiting time limits, and times to retrial should be simulated and the results should be compared to the approximation. Furthermore, in the simulation each priority rule can easily be simulated while the fluid approximations leads to a preemptive priority rule. We showed that the error due to this change in the priority rule is negligible small.

In addition the results of the approximation should be compared to the data of a real-world contact center. In order to collect and evaluate statistical significant retrial data it is necessary to mark a big number of calling and abandoning customer such that one can identify recalls. However, many customers may call several times from different telephones or locations. Therefore, it is difficult to identify a recall such that it is hard to get good data.

For the short-term economic optimisation and operational planning, the allocation of shifts and the agents associated with these shifts is a major problem. We have shown how the fluid approach can be incorporated in an optimisation problem to solve the staff requirement planning and a shift scheduling problem simultaneously. For this purpose, we have derived a general profit function for the contact center model which is maximised.

We showed that the structure of the profit function is quite regular and concave almost everywhere. Only few regions of the profit function seem to be non-concave. However, this discrepency might be due to numerical instabilities in the computation.

We used an initial procedure and simulated annealling to solve the optimisation problem. This simulated annealling heuristic is very simple and questionable because it does not converge to the optimal solution for sure. The parameters of the simulated annealling algorithm must be chosen very carefully and depend on the problem itself. Therefore, the implementation of other algorithms should be a fruitful task for future research.

The initial procedure is based on the relative profit margin of an additional shift. The profit margins are calculated from the arrival rates and the service rates of the different agents weighted by the revenues and costs. We have shown that this procedure of the staffing algorithm already leads to a very accurate solution, i.e. high performance and profit. As the simulated annealling heuristic starts with this solution we get almost optimal results.

We have shown that the optimisation problem can be easily extended. Even more general problems with various customer classes and more agent groups as discussed in Section 5.5 can be solved. A further customer class leads to at least two additional differential equations in the fluid approach and additional terms in the profit function which all have the same structure. One differential equation is needed to describe the number of customers in the system and one for the number of customers in the orbit. All differential equations have the same structure as well. With the fluid approach we are also able to model customers, who change their customer class. An example of such a change is a customer who first tries to phone and afterwards writes an e-mail.

If another agent group is added, we have to decide which customer class is served first and which second. Furthermore, the order by which customers are referred to the agent groups has to be determined. This leads to another term or a change in the terms of the differential equations. The number of differential equations remains constant and the new terms have a similar structure as the terms, which describe the service by agents and the priority of customers.

However, it will be necessary to improve the scheduling algorithm. Other optimisation algorithms may lead to even better results or find them more quickly. Furthermore, we used a fixed set of predefined shifts with fixed rest breaks where in practise more flexibility may be needed. Therefore, one should include more flexible rest breaks and shifts, e.g., one could consider time windows for rest breaks. In our model we assumed that agents are available to work according to each scheduled shift. However, in real-world contact centers the contracts of employment and the preferrences of the agents lead to additional constraints. Therefore, it would be worthwhile to enlarge the shift scheduling to a tour scheduling problem which includes day-off scheduling.

Finally, another extension could involve the results of the diffusion refinement in order to make the staffing and scheduling decision more robust with respect to the variability of the processes. However, this seems to be not as worthwhile in practise, because the diffusion approximation is difficult to derive and the benefits are small in a big contact center. The reason is that the effects of randomness become more and more negligible in big contact centers. Therefore the major focus should lie on improving the staffing and scheduling algorithm. By means of the methods presented in this thesis and improved staffing and shift scheduling algorithms the calculation of shift scheduling becomes much quicker and easier. That is why an integration of the methods into a software tool will be worthwhile for the management of call and contact centers.

A

Derivation of the Differential Equations for the Variances and Covariances in Contact Centers with Retrials

A.1 Contact Centers with Homogeneous Customers and Agents

In this section we derive the stochastic functional equation for the diffusion processes of the contact center model analysed in Section 4.1[1]. Based on the diffusion processes we determine the differential equations for the variances and covariances of these processes. These diffusion processes result from an application of the functional central limit theorem and its extension proven by Rao (1973). Furthermore, we use the approximation of the scaled stochastic processes describing the number of customers in the system and the orbit derived in Komlós et al. (1975).

The stochastic processes for the number of customers in the system and the orbit are given by the sum of Poisson Processes describing the arrival and departure of customers, i.e.,

$$\begin{aligned} Q_S(t) = Q_S(0) &+ A_1\left(\int_0^t \lambda(s)\,ds\right) + A_2\left(\int_0^t \gamma(s) Q_{\mathcal{O}}(s)\,ds\right) \\ &- A_3\left(\int_0^t \mu(s)\min\{Q_S(s), N(s)\}\ ds\right) \qquad \text{(A.1a)} \\ &- A_4\left(\int_0^t \nu(s)\{Q_S(s) - N(s)\}^+\,ds\right) \end{aligned}$$

[1] These equations were already derived by Mandelbaum et al. (1998). They derive the equations also in a more general way.

$$\begin{aligned} Q_{\mathcal{O}}(t) = Q_{\mathcal{O}}(0) &+ A_{\mathcal{O}}\left(\int_0^t p\nu(s)\{Q_S(s) - N(s)\}^+ \, ds\right) \\ &- A_2\left(\int_0^t \gamma(s) Q_{\mathcal{O}}(s)\, ds\right) . \end{aligned} \tag{A.1b}$$

As explained in Chapter 3 on Page 32, $A_i()$, $(i = 1, 2, 3, 4)$ are standard Poisson Processes with mean 1. $Q_S(0)$ and $Q_{\mathcal{O}}(0)$ represent the number of customers in the system and the orbit at the beginning of the observation.

We scale the arrival rate and the number of agents according to Halfin and Whitt (1981). Furthermore, we divide the abandonment process into two independent processes, which describe the number of customers moving into the orbit $A_{4\mathcal{O}}()$ and the number of lost customers $A_{4L}()$. If we approximate the scaled processes by means of Brownian motions, we get

$$\begin{aligned} Q_S^n(t) = Q_S^n(0) &+ \int_0^t n\lambda(s)\, ds + \int_0^t n\gamma(s) \frac{Q_{\mathcal{O}}^n(s)}{n}\, ds \\ &- \int_0^t n\mu(s) \min\left\{\frac{Q_S^n(s)}{n}, N(s)\right\} ds - \int_0^t p\nu(s) \left\{\frac{Q_S^n(s)}{n} - N(s)\right\}^+ ds \\ &+ B_1\left(\int_0^t n\lambda(s)\, ds\right) + B_2\left(\int_0^t n\gamma(s)\frac{Q_{\mathcal{O}}^n(s)}{n}\, ds\right) \\ &- B_3\left(\int_0^t n\mu(s) \min\left\{\frac{Q_S^n(s)}{n}, N(s)\right\} ds\right) \\ &- B_{4\mathcal{O}}\left(\int_0^t np\nu(s)\left\{\frac{Q_S^n(s)}{n} - N(s)\right\}^+ ds\right) \\ &- B_{4L}\left(\int_0^t (1-p)n\nu(s)\left\{\frac{Q_S^n(s)}{n} - N(s)\right\}^+ ds\right) + o(\log n) \end{aligned} \tag{A.2a}$$

$$\begin{aligned} Q_{\mathcal{O}}^n(t) = Q_{\mathcal{O}}^n(0) &+ \int_0^t np\nu(s)\left\{\frac{Q_S^n(s)}{n} - N(s)\right\}^+ ds - \int_0^t n\gamma(s)\frac{Q_{\mathcal{O}}^n(s)}{n}\, ds \\ &+ B_{4\mathcal{O}}\left(\int_0^t np\nu(s)\left\{\frac{Q_S^n(s)}{n} - N(s)\right\}^+ ds\right) - B_2\left(\int_0^t n\gamma(s)\frac{Q_{\mathcal{O}}^n(s)}{n}\, ds\right) \\ &+ o(\log n) , \end{aligned} \tag{A.2b}$$

where $B_i()$ $(i = 1, 2, 3, 4\mathcal{O}, 4L)$ are standard Brownian Motions with mean 0 and variance t at time t.

We assume that the set $\boldsymbol{S} = \{t|Q(t) = N(t)\}$ has measure zero to circumvent the difficulties discussed on Page 60[2]. Then we can derive the functional equations for the diffusion processes by applying the modified functional central limit theorem by Rao from

$$\lim_{n\to\infty} \frac{Q_S^n(t) - nQ_S^F(t)}{\sqrt{n}} = Q_S^D(t) \tag{A.3}$$

$$\lim_{n\to\infty} \frac{Q_\mathcal{O}^n(t) - nQ_\mathcal{O}^F(t)}{\sqrt{n}} = Q_\mathcal{O}^D(t) \tag{A.4}$$

with $Q_S^F(t)$ and $Q_\mathcal{O}^F(t)$ given in (4.1) on Page 75. Passing n to infinity gives rise to the stochastic functional equations for the diffusion processes, which are

$$\begin{aligned}
Q_S^D(t) = {} & Q_S^D(0) + \int_0^t \lambda(s)\,ds + \int_0^t \gamma(s)Q_\mathcal{O}^D(s)\,ds \qquad \text{(A.5a)}\\
& - \int_0^t \mu(s)Q_S^D(t)\mathbb{1}_{\{Q_S^F(s)\le N(s)\}}\,ds - \int_0^t \nu(s)Q_S^D(t)\mathbb{1}_{\{Q_S^F(s)>N(s)\}}\,ds\\
& + B_1\left(\int_0^t \lambda(s)\,ds\right) - B_3\left(\int_0^t \mu(s)\min\{Q_S^F(s), N(s)\}\;ds\right)\\
& + B_2\left(\int_0^t \gamma(s)Q_\mathcal{O}^F(s)\,ds\right) - B_{4\mathcal{O}}\left(\int_0^t p\nu(s)\{Q_S^F(s) - N(s)\}^+\,ds\right)\\
& - B_{4L}\left(\int_0^t (1-p)\nu(s)\{Q_S^F(s) - N(s)\}^+\,ds\right)
\end{aligned}$$

$$\begin{aligned}
Q_\mathcal{O}^n(t) = {} & Q_\mathcal{O}^D(0) + \int_0^t p\nu(s)Q_S^D(t)\mathbb{1}_{\{Q_S^F(s)>N(s)\}}\,ds - \int_0^t \gamma(s)Q_\mathcal{O}^D(s)\,ds \qquad \text{(A.5b)}\\
& + B_{4\mathcal{O}}\left(\int_0^t p\nu(s)\{Q_S^F(s) - N(s)\}^+\,ds\right) - B_2\left(\int_0^t \gamma(s)Q_\mathcal{O}^F(s)\,ds\right).
\end{aligned}$$

In order to derive the differential equations for the variances and covariances we use some properties of the Brownian Motions[3] and the chain rule

[2] Mandelbaum et al. (1998) and Mandelbaum et al. (1999a, 2002) consider both cases, i.e., with and without the assumption of measure zero.

[3] See Page 64.

of stochastic calculus[4]. As the variances and covariances of the processes are given by

$$\mathbf{VAR}\big[Q_S^D(t)\big] = \mathbf{E}\Big[\big(Q_S^D(t)\big)^2\Big] - \mathbf{E}\big[Q_S^D(t)\big]^2 \tag{A.6a}$$

$$\mathbf{VAR}\big[Q_{\mathcal{O}}^D(t)\big] = \mathbf{E}\Big[\big(Q_{\mathcal{O}}^D(t)\big)^2\Big] - \mathbf{E}\big[Q_{\mathcal{O}}^D(t)\big]^2 \tag{A.6b}$$

$$\mathbf{COV}\big[Q_S^D(t), Q_{\mathcal{O}}^D(t)\big] = \mathbf{E}\big[Q_S^D(t)Q_{\mathcal{O}}^D(t)\big] - \mathbf{E}\big[Q_S^D(t)\big]\,\mathbf{E}\big[Q_{\mathcal{O}}^D(t)\big] \tag{A.6c}$$

we determine

$$\frac{d}{dt}\mathbf{E}\Big[\big(Q_S^D(t)\big)^2\Big]\,, \frac{d}{dt}\mathbf{E}\big[Q_S^D(t)\big]\,, \frac{d}{dt}\mathbf{E}\Big[\big(Q_{\mathcal{O}}^D(t)\big)^2\Big]\,, \frac{d}{dt}\mathbf{E}\big[Q_{\mathcal{O}}^D(t)\big]\,,$$

and

$$\frac{d}{dt}\mathbf{E}\big[Q_S^D(t)Q_{\mathcal{O}}^D(t)\big]\,.$$

We start with differential equations for the mean values and get

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\big[Q_S^D(t)\big] &= \gamma(t)\mathbf{E}\big[Q_{\mathcal{O}}^D(t)\big] - \mu(t)\mathbf{E}\big[Q_S^D(t)\big]\,\mathbb{1}_{\{Q_S^F(t)\le N(t)\}} \\ &\quad - \nu(t)\mathbf{E}\big[Q_S^D(t)\big]\,\mathbb{1}_{\{Q_S^F(t)>N(t)\}}\end{aligned} \tag{A.7a}$$

$$\frac{d}{dt}\mathbf{E}\big[Q_{\mathcal{O}}^D(t)\big] = \nu(t)\mathbf{E}\big[Q_S^D(t)\big]\,\mathbb{1}_{\{Q_S^F(t)>N(t)\}} - \gamma(t)\mathbf{E}\big[Q_{\mathcal{O}}^D(t)\big] \tag{A.7b}$$

because the Brownian Motions have mean zero. By means of the chain rule of stochastic calculus we determine

$$\begin{aligned}&\frac{d}{dt}\mathbf{E}\Big[\big(Q_S^D(t)\big)^2\Big] \\ &\quad = -2\left(\mu(t)\mathbb{1}_{\{Q_S^F(t)\le N(t)\}} + \nu(t)\mathbb{1}_{\{Q_S^F(t)>N(t)\}}\right)\mathbf{E}\Big[\big(Q_S^D(t)\big)^2\Big] \\ &\qquad + 2\gamma(t)\mathbf{E}\big[Q_S^D(t)Q_{\mathcal{O}}^D(t)\big] \\ &\qquad + \lambda(t) + \gamma(t)Q_{\mathcal{O}}^F(t) + \mu(t)\min\big\{Q_S^F(t), N(t)\big\} + \nu(t)\big\{Q_S^F(t) - N(t)\big\}^+\end{aligned} \tag{A.7c}$$

$$\begin{aligned}&\frac{d}{dt}\mathbf{E}\Big[\big(Q_{\mathcal{O}}^D(t)\big)^2\Big] \\ &\quad = 2p\nu(t)\mathbb{1}_{\{Q_S^F(t)>N(t)\}}\mathbf{E}\big[Q_S^D(t)Q_{\mathcal{O}}^D(t)\big] - 2\gamma(t)\mathbf{E}\Big[\big(Q_{\mathcal{O}}^D(t)\big)^2\Big] \\ &\qquad + \gamma(t)Q_{\mathcal{O}}^F(t) + p\nu(t)\big\{Q_S^F(t) - N(t)\big\}^+\end{aligned} \tag{A.7d}$$

$$\begin{aligned}&\frac{d}{dt}\mathbf{E}\big[Q_S^D(t)Q_{\mathcal{O}}^D(t)\big] \\ &\quad = -\left(\mu(t)\mathbb{1}_{\{Q_S^F(t)\le N(t)\}} + \nu(t)\mathbb{1}_{\{Q_S^F(t)>N(t)\}}\right)\mathbf{E}\big[Q_S^D(t)Q_{\mathcal{O}}^D(t)\big]\end{aligned} \tag{A.7e}$$

[4] See Karatzas and Shreve (1991), pp. 149-159.

$$
\begin{aligned}
&+\gamma(t)\mathbf{E}\big[Q_S^D(t)Q_{\mathcal{O}}^D(t)\big] \\
&+p\nu(t)\mathbb{1}_{\left\{Q_S^F(t)>N(t)\right\}}\mathbf{E}\big[Q_S^D(t)Q_{\mathcal{O}}^D(t)\big]-\gamma(t)\mathbf{E}\Big[\big(Q_{\mathcal{O}}^D(t)\big)^2\Big] \\
&-\gamma(t)Q_{\mathcal{O}}^F(t)-p\nu(t)\big\{Q_S^F(t)-N(t)\big\}^+
\end{aligned}
$$

As

$$\frac{d}{dt}\left(\mathbf{E}\big[Q_i^D(t)\big]\,\mathbf{E}\big[Q_j^D(t)\big]\right)=\frac{d}{dt}\mathbf{E}\big[Q_i^D(t)\big]\,\mathbf{E}\big[Q_j^D(t)\big]+\mathbf{E}\big[Q_i^D(t)\big]\,\frac{d}{dt}\mathbf{E}\big[Q_j^D(t)\big]$$

for $i,j=S,\mathcal{O}$ the differential equations for the variances and covariance are given by Equations (4.2) on Page 76.

A.2 Contact Centers with Heterogeneous Customers and Agents

In this section we derive the functional equations for the diffusion process and the differential equations for the covariances of the contact center model analysed in Section 4.2. To the best of our knowledge such a model has not been analysed. However, special cases of this model are the contact center model in the previous section and the priority queue analysed by Mandelbaum et al. (1998). Equivalently to the previous Section A.1 we start from the stochastic processes for the number of customers in the system and the orbits as sums of standard Poisson Processes with rate 1. Thereby we use the following notation to arrange the equation more clearly und shortly. The random number of busy generalists serving customers of type $i=1,2$ at time t is denoted $B_I^G(t)$, i.e,

$$B_1^G(t)=\min\Big\{N_G(t),\{Q_1(t)-N_1(t)\}^+\Big\} \tag{A.8}$$

$$B_2^G(t)=\min\Big\{\{Q_2(t)-N_2(t)\}^+,\Big\{N_G(t)-\{Q_1(t)-N_1(t)\}^+\Big\}^+\Big\}. \tag{A.9}$$

This notation is in line with the number of busy generalists in the fluid model defined on Page 121. Furthermore we denote by $L_i(t)$ the time-dependent and stochastic number of customers of type $i=1,2$ waiting to be served, i.e.,

$$L_1(t)=\{Q_1(t)-N_1(t)-N_G(t)\}^+ \tag{A.10}$$

$$L_2(t)=\Big\{Q_2(t)-N_2(t)-\Big\{N_G(t)-\{Q_1(t)-N_1(t)\}^+\Big\}^+\Big\}^+. \tag{A.11}$$

With this notation the four stochastic processes of the vector $\boldsymbol{(Q(t))_{t\in\mathbb{R}_0^+}}$ introduced in (4.15) on Page 120 are given by

$$\begin{aligned}Q_1(t) = Q_1(0) &+ A_1^{\lambda}\left(\int_0^t \lambda_1(s)\,ds\right) + A_{\mathcal{O}1}\left(\int_0^t \gamma_1 Q_{\mathcal{O}1}(s)\,ds\right) \\ &- A_1^{\mu}\left(\int_0^t \mu_1(s)\min\{Q_1(s),N_1(s)\}\,ds\right) - A_1^{\overline{\mu}}\left(\int_0^t \overline{\mu_1}(s)B_1^G(s)\,ds\right) \\ &- A_{1\mathcal{O}}\left(\int_0^t p_1\nu_1(s)L_1(s)\,ds\right) - A_1^{\nu}\left(\int_0^t (1-p_1)\nu_1(s)L_1(s)\,ds\right)\end{aligned} \tag{A.12a}$$

$$\begin{aligned}Q_2(t) = Q_2(0) &+ A_2^{\lambda}\left(\int_0^t \lambda_2(s)\,ds\right) + A_{\mathcal{O}2}\left(\int_0^t \gamma_2 Q_{\mathcal{O}2}(s)\,ds\right) \\ &- A_2^{\mu}\left(\int_0^t \mu_1(s)\min\{Q_2(s),N_2(s)\}\,ds\right) - A_2^{\overline{\mu}}\left(\int_0^t \overline{\mu_2}(s)B_2^G(s)\,ds\right) \\ &- A_{2\mathcal{O}}\left(\int_0^t p_2\nu_2(s)L_2(s)\,ds\right) - A_2^{\nu}\left(\int_0^t (1-p_2)\nu_2(s)L_2(t)\,ds\right)\end{aligned} \tag{A.12b}$$

$$Q_{\mathcal{O}1}(t) = Q_{\mathcal{O}1}(0) - A_{1\mathcal{O}}\left(\int_0^t \gamma_1(s)Q_{\mathcal{O}1}(s)\,ds\right) + A_{\mathcal{O}1}\left(\int_0^t p_1\nu_1(s)L_1(s)\,ds\right) \tag{A.12c}$$

$$Q_{\mathcal{O}2}(t) = Q_{\mathcal{O}2}(0) - A_{2\mathcal{O}}\left(\int_0^t \gamma_2(s)Q_{\mathcal{O}2}(s)\,ds\right) + A_{\mathcal{O}2}\left(\int_0^t p_2\nu_2(s)L_2(s)\,ds\right). \tag{A.12d}$$

Next we scale these stochastic processes according to Halfin and Whitt and approximate the scaled stochastic processes by means of standard Brownian Motion with mean zero and variance t at time t as in the previous Section A.1. For the number of busy generalists and the number of customers waiting this leads to

$$B_1^{Gn}(t) = \min\left\{N_G(t), \left\{\frac{Q_1^n(t)}{n} - N_1(t)\right\}^+\right\} \tag{A.13}$$

$$B_2^{Gn}(t) = \min\left\{\left\{\frac{Q_2^n(t)}{n} - N_2(t)\right\}^+, \left\{N_G(t) - \left\{\frac{Q_1^n(t)}{n} - N_1(t)\right\}^+\right\}^+\right\} \tag{A.14}$$

$$L_1^n(t) = \left\{\frac{Q_1^n(t)}{n} - N_1(t) - N_G(t)\right\}^+ \tag{A.15}$$

and

$$L_2^n(t) = \left\{\frac{Q_2^n(t)}{n} - N_2(t) - \left\{N_G(t) - \left\{\frac{Q_1^n(t)}{n} - N_1(t)\right\}^+\right\}^+\right\}^+. \tag{A.16}$$

By means of these terms we can denote the scaled processes by

$$
\begin{aligned}
Q_1^n(t) = Q_1^n(0) &+ \int\limits_0^t n\lambda_1(s)\,ds + \int\limits_0^t n\gamma_1 \frac{Q_{\mathcal{O}1}^n(s)}{n}\,ds \qquad\qquad \text{(A.17a)}\\
&- \int\limits_0^t n\mu_1 \min\left\{\frac{Q_1^n(s)}{n}, N_1(s)\right\}\,ds - \int\limits_0^t n\overline{\mu_1} B_1^{Gn}(s)\,ds - \int\limits_0^t n\nu_1 L_1^n(s)\,ds\\
&+ B_1^{\lambda}\left(\int\limits_0^t n\lambda_1(s)\,ds\right) + B_{\mathcal{O}1}\left(\int\limits_0^t n\gamma_1 \frac{Q_{\mathcal{O}1}^n(s)}{n}\,ds\right)\\
&- B_1^{\mu}\left(\int\limits_0^t n\mu_1 \min\left\{\frac{Q_1^n(s)}{n}, N_1(s)\right\}\,ds\right) - B_1^{\overline{\mu}}\left(\int\limits_0^t n\overline{\mu_1} B_1^{Gn}(s)\,ds\right)\\
&- B_{1\mathcal{O}}\left(\int\limits_0^t np_1\nu_1 L_1^n(s)\,ds\right) - B_1^{\nu}\left(\int\limits_0^t n(1-p_1)\nu_1 L_1^n(s)\,ds\right) + o(\log n)
\end{aligned}
$$

$$
\begin{aligned}
Q_2^n(t) = Q_2^n(0) &+ \int\limits_0^t n\lambda_2(s)\,ds + \int\limits_0^t n\gamma_2 \frac{Q_{\mathcal{O}2}^n(s)}{n}\,ds \qquad\qquad \text{(A.17b)}\\
&- \int\limits_0^t n\mu_2 \min\left\{\frac{Q_2^n(s)}{n}, N_2(s)\right\}\,ds - \int\limits_0^t n\overline{\mu_2} B_2^{Gn}\,ds - \int\limits_0^t p_2\nu_2(s) L_2^n(s)\,ds\\
&+ B_2^{\lambda}\left(\int\limits_0^t n\lambda_2(s)\,ds\right) + B_{\mathcal{O}2}\left(\int\limits_0^t n\gamma_2 \frac{Q_{\mathcal{O}2}^n(s)}{n}\,ds\right)\\
&- B_2^{\mu}\left(\int\limits_0^t n\mu_1 \min\left\{\frac{Q_2^n(s)}{n}, N_2(s)\right\}\,ds\right) - B_2^{\overline{\mu}}\left(\int\limits_0^t n\overline{\mu_2} B_2^{Gn}(s)\,ds\right)\\
&- B_{2\mathcal{O}}\left(\int\limits_0^t np_2\nu_2 L_2^n(s)\,ds\right) - B_2^{\nu}\left(\int\limits_0^t n(1-p_2)\nu_2 L_2^n(s)\,ds\right) + o(\log n)
\end{aligned}
$$

$$
\begin{aligned}
Q_{\mathcal{O}1}^n(t) = Q_{\mathcal{O}1}^n(0) &+ \int\limits_0^t np_1\nu_1 L_1^n(s)\,ds - \int\limits_0^t n\gamma_1 \frac{Q_{\mathcal{O}1}^n(s)}{n}\,ds \qquad\qquad \text{(A.17c)}\\
&+ B_{\mathcal{O}1}\left(\int\limits_0^t np_1\nu_1 L_1^n(s)\,ds\right) - B_{1\mathcal{O}}\left(\int\limits_0^t n\gamma_1 \frac{Q_{\mathcal{O}1}^n(s)}{n}\,ds\right) + o(\log n)
\end{aligned}
$$

$$
\begin{aligned}
Q_{\mathcal{O}2}^n(t) = Q_{\mathcal{O}2}^n(0) &- \int\limits_0^t n\gamma_2 \frac{Q_{\mathcal{O}2}^n(s)}{n}\,ds + \int\limits_0^t np_2\nu_2 L_2^n(s)\,ds \qquad\qquad \text{(A.17d)}\\
&+ B_{\mathcal{O}2}\left(\int\limits_0^t np_2\nu_2 L_2^n(s)\,ds\right) - B_{2\mathcal{O}}\left(\int\limits_0^t n\gamma_2 \frac{Q_{\mathcal{O}2}^n(s)}{n}\,ds\right) + o(\log n)\,.
\end{aligned}
$$

We apply the functional central limit theorem and its extension by Rao (1973) to the scaled and approximated processes $\boldsymbol{(Q^n(t))}_{\boldsymbol{t}\in\mathbb{R}_0^+}$. If the sets of critical times $\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3$ and $\boldsymbol{S}_4$ given in Equation (4.25) on Page 124 have measure zero, we get the stochastic functional equations of the diffusion processes $\big(\boldsymbol{Q^D(t)}\big)_{\boldsymbol{t}\in\mathbb{R}_0^+}$. These processes are given by[5]

$$\begin{aligned}
Q_1^D(t) = Q_1^D(0) &+ \int\limits_0^t \gamma_1 Q_{\mathcal{O}1}^D(s)\, ds - \int\limits_0^t \mu_1 Q_1^D(s) \mathbb{1}_{\left\{Q_1^F \le N_1\right\}}\, ds \\
&- \int\limits_0^t \overline{\mu_1} Q_1^D(s) \mathbb{1}_{\left\{N_1 < Q_1^F \le N_1 + N_G\right\}}\, ds - \int\limits_0^t \nu_1 Q_1^D(s) \mathbb{1}_{\left\{Q_1^F > N_1 + N_G\right\}}\, ds \\
&+ B_1^\lambda\Bigg(\int\limits_0^t \lambda_1(s)\, ds\Bigg) + B_{\mathcal{O}1}\Bigg(\int\limits_0^t \gamma_1 Q_{\mathcal{O}1}^F(s)\, ds\Bigg) \\
&- B_1^\mu\Bigg(\int\limits_0^t \mu_1 \min\Big\{Q_1^F(s), N_1(s)\Big\}\, ds\Bigg) - B_1^{\overline{\mu}}\Bigg(\int\limits_0^t \overline{\mu_1} B_1^F(t)\, ds\Bigg) \\
&- B_{1\mathcal{O}}\Bigg(\int\limits_0^t p_1 \nu_1 L_1^F(s)\, ds\Bigg) - B_1^\nu\Bigg(\int\limits_0^t (1-p_1)\nu_1 L_1^F(s)\, ds\Bigg)
\end{aligned} \tag{A.18a}$$

$$\begin{aligned}
Q_2^D(t) = Q_2^D(0) &+ \int\limits_0^t \gamma_2 Q_{\mathcal{O}2}^D(s)\, ds - \int\limits_0^t \mu_1 Q_2^D(s) \mathbb{1}_{\left\{Q_2^F \le N_2\right\}}\, ds \\
&- \int\limits_0^t \overline{\mu_2} Q_2^D(s) \mathbb{1}_{\left\{N_2 < Q_2^F \le N_2 + \left\{N_G - \left\{Q_1^F - N_1\right\}^+\right\}^+\right\}}\, ds \\
&+ \int\limits_0^t \overline{\mu_2} Q_1^D(s) \mathbb{1}_{\left\{N_1 < Q_1^F \le N_1 + N_G,\, N_2 < Q_2^F \le N_2 + N_G + N_1 - Q_1^F\right\}}\, ds \\
&- \int\limits_0^t \nu_2 Q_2^D(s) \mathbb{1}_{\left\{Q_2^F > N_2 + \left\{N_G - \left\{Q_1^F - N_1\right\}^+\right\}^+\right\}}\, ds \\
&- \int\limits_0^t \nu_2 Q_1^D(s) \mathbb{1}_{\left\{N_1 < Q_1^F \le N_G + N_1,\, Q_2^F > N_2 + N_G + N_1 - Q_1^F\right\}}\, ds \\
&+ B_2^\lambda\Bigg(\int\limits_0^t \lambda_2(s)\, ds\Bigg) + B_{\mathcal{O}2}\Bigg(\int\limits_0^t \gamma_2 Q_{\mathcal{O}2}^F(s)\, ds\Bigg)
\end{aligned} \tag{A.18b}$$

[5] For reasons of space we omit the time argument s in the indicator function $\mathbb{1}_{\{\}}$,i.e., we write Q_i^F and N_j instead of $Q_i^F(s)$ and $N_j(s)$ for $i = 1, 2, \mathcal{O}1, \mathcal{O}2$ and $j = 1, 2, G$.

$$- B_2^{\mu}\left(\int_0^t \mu_1 \min\left\{Q_2^F(s), N_2(s)\right\}\, ds\right) - B_2^{\overline{\mu}}\left(\int_0^t \overline{\mu_2} B_2^F(s)\, ds\right)$$

$$- B_{2\mathcal{O}}\left(\int_0^t p_2\nu_2 L_2^F(s)\, ds\right) - B_2^{\nu}\left(\int_0^t (1-p_2)\nu_2 L_2^F(s)\, ds\right)$$

$$Q_{\mathcal{O}1}^D(t) = Q_{\mathcal{O}1}^D(0) + \int_0^t p_1\nu_1 Q_{\mathcal{O}1}^D(s)\mathbb{1}_{\left\{Q_1^F > N_1 + N_G\right\}}\, ds - \int_0^t \gamma_1 Q_{\mathcal{O}1}^D(s)\, ds \tag{A.18c}$$

$$+ B_{\mathcal{O}1}\left(\int_0^t p_1\nu_1 L_1^F(s)\, ds\right) - B_{1\mathcal{O}}\left(\int_0^t \gamma_1 Q_{\mathcal{O}1}^F(s)\, ds\right)$$

$$Q_{\mathcal{O}2}^D(t) = Q_{\mathcal{O}2}^D(0) - \int_0^t \gamma_2 Q_{\mathcal{O}2}^D(s)\, ds \tag{A.18d}$$

$$+ \int_0^t p_2\nu_2 Q_2^D(s)\mathbb{1}_{\left\{Q_2^F > N_2 + \left\{N_G - \left\{Q_1^F - N_1\right\}^+\right\}^+\right\}}\, ds$$

$$+ \int_0^t p_2\nu_2 Q_1^D(s)\mathbb{1}_{\left\{N_1 < Q_1^F \leq N_1 + N_G,\, Q_2^F > N_2 + N_G + N_1 - Q_1^F\right\}}\, ds$$

$$+ B_{\mathcal{O}2}\left(\int_0^t p_2\nu_2 L_2^F(s)\, ds\right) - B_{2\mathcal{O}}\left(\int_0^t \gamma_2 Q_{\mathcal{O}2}^F(s)\, ds\right).$$

Thereby $B_i^F(t)$ and $L_i^F(t)$ $(i = 1, 2)$ are defined according to Equations 4.16, 4.20, 4.17, and 4.22 on Pages 121ff.

From the function equations for the diffusion processes we determine the differential equations for the mean values of the diffusion processes and the differential equations of the products of the processes, i.e., we have to derive

$$\frac{d}{dt}\mathbf{E}\left[Q_1^D(t)\right], \frac{d}{dt}\mathbf{E}\left[Q_2^D(t)\right], \frac{d}{dt}\mathbf{E}\left[Q_{\mathcal{O}1}^D(t)\right], \frac{d}{dt}\mathbf{E}\left[Q_{\mathcal{O}2}^D(t)\right]$$

and

$$\frac{d}{dt}\mathbf{E}\left[Q_i^D(t)Q_j^D(t)\right] \quad \text{for } i, j = 1, 2, \mathcal{O}1, \mathcal{O}2.$$

For this purpose we use the following notation for the conditions of the indicator function $\mathbb{1}_{\{\}}$ which were already introduced in Equation (4.27) on Page 125:

$$c_1 : Q_1^F(t) \leq N_1(t)$$
$$c_2 : N_1(t) < Q_1^F(t) \leq N_1(t) + N_G(t)$$
$$c_3 : Q_1^F(t) > N_1(t) + N_G(t)$$
$$c_4 : N_1(t) < Q_1^F(t) \leq N_1(t) + N_G(t),\ Q_2^F(t) > N_1(t) + N_2(t) + N_G(t) - Q_1^F(t)$$
$$c_5 : N_1(t) < Q_1^F(t) \leq N_1(t) + N_G(t),\ N_2(t) < Q_2^F(t) \leq N_1(t) + N_2(t) + N_G(t) - Q_1^F(t)$$
$$c_6 : Q_2^F(t) \leq N_2(t)$$
$$c_7 : Q_2^F(t) > N_2(t) + \Big\{ N_G(t) - \Big\{ Q_1^F(t) - N_1(t) \Big\}^+ \Big\}^+$$
$$c_8 : N_2(t) < Q_2^F(t) \leq N_2(t) + \Big\{ N_G(t) - \Big\{ Q_1^F(t) - N_1(t) \Big\}^+ \Big\}^+ .$$

As the means of the Brownian Motions are zero, we get for the mean values of the diffusion processes

$$\frac{d}{dt}\mathbf{E}\Big[Q_1^D(t)\Big] = \gamma_1 \mathbf{E}\Big[Q_{\mathcal{O}1}^D(t)\Big] - \mu_1 \mathbf{E}\Big[Q_1^D(t)\Big]\,\mathbb{1}_{\{c_1\}} - \overline{\mu_1}\mathbf{E}\Big[Q_1^D(t)\Big]\,\mathbb{1}_{\{c_2\}} - \nu_1 \mathbf{E}\Big[Q_1^D(t)\Big]\,\mathbb{1}_{\{c_3\}} \tag{A.19a}$$

$$\frac{d}{dt}\mathbf{E}\Big[Q_2^D(t)\Big] = \gamma_2 \mathbf{E}\Big[Q_{\mathcal{O}2}^D(t)\Big] - \mu_1 \mathbf{E}\Big[Q_2^D(t)\Big]\,\mathbb{1}_{\{c_6\}} - \overline{\mu_2}\mathbf{E}\Big[Q_2^D(t)\Big]\,\mathbb{1}_{\{c_8\}} + \overline{\mu_2}\mathbf{E}\Big[Q_1^D(t)\Big]\,\mathbb{1}_{\{c_5\}} - \nu_2 \mathbf{E}\Big[Q_2^D(t)\Big]\,\mathbb{1}_{\{c_7\}} - \nu_2 \mathbf{E}\Big[Q_1^D(t)\Big]\,\mathbb{1}_{\{c_4\}} \tag{A.19b}$$

$$\frac{d}{dt}\mathbf{E}\Big[Q_{\mathcal{O}1}^D(t)\Big] = +p_1\nu_1 \mathbf{E}\Big[Q_{\mathcal{O}1}^D(t)\Big]\,\mathbb{1}_{\{c_3\}} - \gamma_1 \mathbf{E}\Big[Q_{\mathcal{O}1}^D(t)\Big] \tag{A.19c}$$

$$\frac{d}{dt}\mathbf{E}\Big[Q_{\mathcal{O}2}^D(t)\Big] = -\gamma_2 \mathbf{E}\Big[Q_{\mathcal{O}2}^D(t)\Big] + p_2\nu_2 \mathbf{E}\Big[Q_2^D(t)\Big]\,\mathbb{1}_{\{c_7\}} + p_2\nu_2 \mathbf{E}\Big[Q_1^D(t)\Big]\,\mathbb{1}_{\{c_4\}}. \tag{A.19d}$$

To derive the means of the products we again need the chain rule of stochastic calculus and some properties of the Brownian Motion described on Page 64. Furthermore we use the the functions $b_1(t)$ through $b_4(t)$ introduced in Equation 4.30 on Page 126 which are

$$b_1(t) = \lambda_1(t) + \gamma_1(t) Q_{\mathcal{O}1}^F(t) + \mu_1 \min\big\{Q_1^F(t), N_1(t)\big\} + \nu_1 L_1^F(t) + \overline{\mu_1} B_1^F(t)$$
$$b_2(t) = \lambda_2(t) + \gamma_2(t) Q_{\mathcal{O}2}^F(t) + \mu_2 \min\big\{Q_2^F(t), N_2(t)\big\} + \overline{\mu_2} B_2^F(t) + \nu_2 L_2^F(t)$$
$$b_3(t) = \gamma_1(t) Q_{\mathcal{O}1}^F(t) + p\nu_1 L_1^F(t)$$
$$b_4(t) = \gamma_2(t) Q_{\mathcal{O}2}^F(t) + p\nu_2 L_2^F(t).$$

This gives rise to

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\Big[\Big(Q_1^D(t)\Big)^2\Big] &= 2\gamma_1\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}1}^D(t)\Big] - 2\mu_1\mathbf{E}\Big[\Big(Q_1^D(t)\Big)^2\Big]\mathbb{1}_{\{c_1\}} \\ &\quad - 2\left(\overline{\mu_1}\mathbb{1}_{\{c_2\}} + \nu_1\mathbb{1}_{\{c_3\}}\right)\mathbf{E}\Big[\Big(Q_1^D(t)\Big)^2\Big] + b_1(t)\end{aligned} \tag{A.19e}$$

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\Big[\Big(Q_2^D(t)\Big)^2\Big] &= 2\gamma_2\mathbf{E}\Big[Q_2^D(t)Q_{\mathcal{O}2}^D(t)\Big] - 2\mu_2\mathbf{E}\Big[\Big(Q_2^D(t)\Big)^2\Big]\mathbb{1}_{\{c_6\}} \\ &\quad - 2\overline{\mu_2}\mathbf{E}\Big[\Big(Q_2^D(t)\Big)^2\Big]\mathbb{1}_{\{c_8\}} + 2\overline{\mu_2}\mathbf{E}\Big[Q_1^D(t)Q_2^D(t)\Big]\mathbb{1}_{\{c_5\}} \\ &\quad - 2\nu_2\mathbf{E}\Big[\Big(Q_2^D(t)\Big)^2\Big]\mathbb{1}_{\{c_7\}} - 2\nu_2\mathbf{E}\Big[Q_1^D(t)Q_2^D(t)\Big]\mathbb{1}_{\{c_4\}} + b_2(t)\end{aligned} \tag{A.19f}$$

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\Big[\Big(Q_{\mathcal{O}1}^D(t)\Big)^2\Big] &= 2p_1\nu_1\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}1}^D(t)\Big]\mathbb{1}_{\{c_3\}} - 2\gamma_1\mathbf{E}\Big[\Big(Q_{\mathcal{O}1}^D(t)\Big)^2\Big] \\ &\quad + b_3(t)\end{aligned} \tag{A.19g}$$

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\Big[\Big(Q_{\mathcal{O}2}^D(t)\Big)^2\Big] &= 2p_2\nu_2\mathbf{E}\Big[Q_2^D(t)Q_{\mathcal{O}2}^D(t)\Big]\mathbb{1}_{\{c_7\}} - 2\gamma_2\mathbf{E}\Big[\Big(Q_{\mathcal{O}2}^D(t)\Big)^2\Big] \\ &\quad + 2p_2\nu_2\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}2}^D(t)\Big]\mathbb{1}_{\{c_4\}} + b_4(t).\end{aligned} \tag{A.19h}$$

The means of the products of the diffusion processes are given by

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\Big[Q_1^D(t)Q_2^D(t)\Big] &= \gamma_1\mathbf{E}\Big[Q_2^D(t)Q_{\mathcal{O}1}^D(t)\Big] + \gamma_2\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}2}^D(t)\Big] \\ &\quad - \left(\mu_1\mathbb{1}_{\{c_1\}} + \overline{\mu_1}\mathbb{1}_{\{c_2\}} + \nu_1\mathbb{1}_{\{c_3\}}\right)\mathbf{E}\Big[Q_1^D(t)Q_2^D(t)\Big] \\ &\quad - \left(\mu_2\mathbb{1}_{\{c_6\}} + \overline{\mu_2}\mathbb{1}_{\{c_8\}} + \nu_2\mathbb{1}_{\{c_7\}}\right)\mathbf{E}\Big[Q_1^D(t)Q_2^D(t)\Big] \\ &\quad + \overline{\mu_2}\mathbf{E}\Big[\Big(Q_1^D(t)\Big)^2\Big]\mathbb{1}_{\{c_5\}} - \nu_2\mathbf{E}\Big[\Big(Q_1^D(t)\Big)^2\Big]\mathbb{1}_{\{c_4\}}\end{aligned} \tag{A.19i}$$

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}1}^D\Big] &= \gamma_1\mathbf{E}\Big[\Big(Q_{\mathcal{O}1}^D(t)\Big)^2\Big] + p_1\nu_1\mathbf{E}\Big[\Big(Q_1^D(t)\Big)^2\Big]\mathbb{1}_{\{c_3\}} \\ &\quad - \left(\mu_1\mathbb{1}_{\{c_1\}} + \overline{\mu_1}\mathbb{1}_{\{c_2\}} + \nu_1\mathbb{1}_{\{c_3\}} + \gamma_1\right)\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}1}^D(t)\Big] - b_3(t)\end{aligned} \tag{A.19j}$$

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}2}^D(t)\Big] &= \gamma_1\mathbf{E}\Big[Q_{\mathcal{O}1}^D(t)Q_{\mathcal{O}2}^D(t)\Big] - \gamma_2\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}2}^D(t)\Big] \\ &\quad - \left(\mu_1\mathbb{1}_{\{c_1\}} + \overline{\mu_1}\mathbb{1}_{\{c_2\}} + \nu_1\mathbb{1}_{\{c_3\}}\right)\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}2}^D(t)\Big] \\ &\quad + p_2\nu_2\mathbf{E}\Big[Q_1^D(t)Q_2^D(t)\Big]\mathbb{1}_{\{c_7\}} + p_2\nu_2\mathbf{E}\Big[\Big(Q_1^D(t)\Big)^2\Big]\mathbb{1}_{\{c_4\}}\end{aligned} \tag{A.19k}$$

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\Big[Q_2^D(t)Q_{\mathcal{O}1}^D(t)\Big] &= \gamma_2\mathbf{E}\Big[Q_{\mathcal{O}1}^D(t)Q_{\mathcal{O}2}^D(t)\Big] - \gamma_1\mathbf{E}\Big[Q_2^D(t)Q_{\mathcal{O}1}^D(t)\Big] \\ &- \left(\mu_2\mathbb{1}_{\{c_6\}} + \overline{\mu_2}\mathbb{1}_{\{c_8\}} + \nu_2\mathbb{1}_{\{c_7\}}\right)\mathbf{E}\Big[Q_2^D(t)Q_{\mathcal{O}1}^D(t)\Big] \\ &+ \left(\overline{\mu_2}\mathbb{1}_{\{c_5\}} - \nu_2\mathbb{1}_{\{c_4\}}\right)\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}1}^D(t)\Big] + p_1\nu_1\mathbf{E}\Big[Q_1^D(t)Q_2^D(t)\Big]\mathbb{1}_{\{c_3\}}\end{aligned} \tag{A.19l}$$

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\Big[Q_2^D(t)Q_{\mathcal{O}2}^D(t)\Big] &= \gamma_2\mathbf{E}\Big[\Big(Q_{\mathcal{O}2}^D(t)\Big)^2\Big] - \gamma_2\mathbf{E}\Big[Q_2^D(t)Q_{\mathcal{O}2}^D(t)\Big] \\ &- \left(\mu_2\mathbb{1}_{\{c_6\}} + \overline{\mu_2}\mathbb{1}_{\{c_8\}} + \nu_2\mathbb{1}_{\{c_7\}}\right)\mathbf{E}\Big[Q_2^D(t)Q_{\mathcal{O}2}^D(t)\Big] \\ &+ \left(\overline{\mu_2}\mathbb{1}_{\{c_5\}} - \nu_2\mathbb{1}_{\{c_4\}}\right)\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}2}^D(t)\Big] \\ &+ 2p_2\nu_2\mathbf{E}\Big[\Big(Q_2^D(t)\Big)^2\Big]\mathbb{1}_{\{c_7\}} + 2p_2\nu_2\mathbf{E}\Big[Q_1^F(t)Q_2^D(t)\Big]\mathbb{1}_{\{c_4\}} - b_4(t).\end{aligned} \tag{A.19m}$$

$$\begin{aligned}\frac{d}{dt}\mathbf{E}\Big[Q_{\mathcal{O}1}^D(t)Q_{\mathcal{O}2}^D(t)\Big] &= -(\gamma_2+\gamma_1)\mathbf{E}\Big[Q_{\mathcal{O}1}^D(t)Q_{\mathcal{O}2}^D(t)\Big] \\ &+ p_1\nu_1\mathbb{1}_{\{c_3\}}\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}2}^D(t)\Big] \\ &+ p_2\nu_2\mathbf{E}\Big[Q_2^D(t)Q_{\mathcal{O}1}^D(t)\Big]\mathbb{1}_{\{c_7\}} + p_2\nu_2\mathbf{E}\Big[Q_1^D(t)Q_{\mathcal{O}1}^D(t)\Big]\mathbb{1}_{\{c_4\}}.\end{aligned} \tag{A.19n}$$

Based on Equations (A.19) we are able to generate the differential equations of the variances and covariances in Equation (4.29) on Page 126. In order to make notations shorter, we introduce the four-dimensional quadratic and symmetric matrix $\boldsymbol{E}^{\mathbf{2}}(t) := \mathbf{E}\Big[\big(\boldsymbol{Q}^{\boldsymbol{D}}(t)\big)^2\Big]$ which[6] consists of the means of the product of the processes, i.e.,

$$\boldsymbol{E}^{\mathbf{2}}(t) = \begin{pmatrix} \mathbf{E}\Big[\big(Q_1^D\big)^2\Big] & \mathbf{E}\big[Q_1^DQ_2^D\big] & \mathbf{E}\big[Q_1^DQ_{\mathcal{O}1}^D\big] & \mathbf{E}\big[Q_1^DQ_{\mathcal{O}2}^D\big] \\ \mathbf{E}\big[Q_1^DQ_2^D\big] & \mathbf{E}\Big[\big(Q_2^D\big)^2\Big] & \mathbf{E}\big[Q_2^DQ_{\mathcal{O}1}^D\big] & \mathbf{E}\big[Q_2^DQ_{\mathcal{O}2}^D\big] \\ \mathbf{E}\big[Q_1^DQ_{\mathcal{O}1}^D\big] & \mathbf{E}\big[Q_2^DQ_{\mathcal{O}1}^D\big] & \mathbf{E}\Big[\big(Q_{\mathcal{O}1}^D\big)^2\Big] & \mathbf{E}\big[Q_{\mathcal{O}1}^D, Q_{\mathcal{O}2}^D\big] \\ \mathbf{E}\big[Q_1^DQ_{\mathcal{O}2}^D\big] & \mathbf{E}\big[Q_2^DQ_{\mathcal{O}2}^D\big] & \mathbf{E}\big[Q_{\mathcal{O}1}^DQ_{\mathcal{O}2}^D\big] & \mathbf{E}\Big[\big(Q_{\mathcal{O}2}^D\big)^2\Big] \end{pmatrix}, \tag{A.20}$$

and the four-dimensional row vector

$$\boldsymbol{e}(t) := \mathbf{E}\Big[\boldsymbol{Q}^{\boldsymbol{D}}(t)\Big] = \Big(\mathbf{E}\Big[Q_1^D\Big], \mathbf{E}\Big[Q_2^D\Big], \mathbf{E}\Big[Q_{\mathcal{O}1}^D\Big], \mathbf{E}\Big[Q_{\mathcal{O}2}^D\Big]\Big). \tag{A.21}$$

By means of this matrix and the matrices $\boldsymbol{A}(t)$ and $\boldsymbol{B}(t)$ introduced in Equations (4.26) and (4.31) on Pages 125f, respectively, we can rewrite the differential equations given in Equations (A.19e) through (A.19n) by the matrix equation

$$\frac{d}{dt}\boldsymbol{E}^{\mathbf{2}}(t) = \big(\boldsymbol{E}^{\mathbf{2}}(t)\boldsymbol{A}(t)\big)^T + \boldsymbol{E}^{\mathbf{2}}(t)\boldsymbol{A}(t) + \boldsymbol{B}(t). \tag{A.22}$$

[6] For reasons of space we omit the time argument t in the matrix and the following vectors, i.e., $Q_i^D(t)$ is denoted by Q_i^D for $i = 1, 2, \mathcal{O}1, \mathcal{O}2$.

Furthermore the Equations (A.19a) through (A.19d) are given by

$$\frac{d}{dt}\boldsymbol{e}(t) = \boldsymbol{e}(t)\boldsymbol{A}(t). \tag{A.23}$$

Then the derivative of the four-dimensional quadratic and symmetric matrix of the covariances[7]

$$\mathbf{COV}\Big[\boldsymbol{Q^D}(t)\Big] =$$

$$\begin{pmatrix} \mathbf{VAR}[Q_1^D] & \mathbf{COV}[Q_1^D, Q_2^D] & \mathbf{COV}[Q_1^D, Q_{\mathcal{O}1}^D] & \mathbf{COV}[Q_1^D, Q_{\mathcal{O}2}^D] \\ \mathbf{COV}[Q_1^D, Q_2^D] & \mathbf{VAR}[Q_2^D] & \mathbf{COV}[Q_2^D, Q_{\mathcal{O}1}^D] & \mathbf{COV}[Q_2^D, Q_{\mathcal{O}2}^D] \\ \mathbf{COV}[Q_1^D, Q_{\mathcal{O}1}^D] & \mathbf{COV}[Q_2^D, Q_{\mathcal{O}1}^D] & \mathbf{VAR}[Q_{\mathcal{O}1}^D] & \mathbf{COV}[Q_{\mathcal{O}1}^D, Q_{\mathcal{O}2}^D] \\ \mathbf{COV}[Q_1^D, Q_{\mathcal{O}2}^D] & \mathbf{COV}[Q_2^D, Q_{\mathcal{O}2}^D] & \mathbf{COV}[Q_{\mathcal{O}1}^D, Q_{\mathcal{O}2}^D] & \mathbf{VAR}[Q_{\mathcal{O}2}^D] \end{pmatrix}.$$

is given by

$$\begin{aligned} \frac{d}{dt}\mathbf{COV}\big[\boldsymbol{Q^D}(t)\big] &= \frac{d}{dt}\boldsymbol{E^2}(t) - \left(\frac{d}{dt}\boldsymbol{e}(t)\right)^T \boldsymbol{e}(t) - (\boldsymbol{e}(t))^T \frac{d}{dt}\boldsymbol{e} \qquad \text{(A.24)} \\ &= \big(\boldsymbol{E^2}(t)\boldsymbol{A}(t)\big)^T + \boldsymbol{E^2}(t)\boldsymbol{A}(t) + \boldsymbol{B}(t) - (\boldsymbol{e}(t)\boldsymbol{A}(t))^T\boldsymbol{e}(t) - (\boldsymbol{e}(t))^T\boldsymbol{e}(t)\boldsymbol{A}(t) \end{aligned}$$

This leads to Equations (4.29) on Page 126.

[7] As in the previous matrix we omit the time argument t.

Glossary of Notation

Sets

$\mathbb{R}$	set of all real numbers
$\mathbb{R}_0^+$	set of all positive real numbers and zero
$\mathbb{Z}$	set of all integers
$\mathbb{N}_0$	set of all natural numbers and zero
$\mathbb{N}_0^m$	set of all integer valued vectors of dimension m with entries equal to or greater than zero
$\boldsymbol{S}$	set of times with critical loading in the diffusion approximation (See Pages 62 and 124)
$\mathbb{S}$	set of all shift types which can be scheduled (See Page 163)
$\mathbb{K}$	set of shifts with maximum relative profit margin in opening procedure of the shift scheduling algorithm (See Pages 171 and 193)

Parameters of the contact center models

$\lambda_i(t)$	time-dependent arrival rates of customer class $i = 1, 2$ (See Page 29)
$m_j^{(i)}, t_j^{(i)}$	parameters of the sinusoidal arrival rate function of customer class $i = 1, 2$ (See Page 29)
$\mu_i(t), \overline{\mu_i}$	time-dependent service rate of specialists and generalists for customer class $i = 1, 2$
$\nu_i(t)$	abandonment rate of customer class $i = 1, 2$
$\gamma_i(t)$	time-dependent retrial rate of customer class $i = 1, 2$
p_i	probability of retrial of customer class $i = 1, 2$
$N_i(t)$	time-dependent number of homogeneous servers of type $i = 1, 2, G$

Functions

$\{X\}^+$	maximum of 0 and X
$\mathbb{1}_{\{c\}}$	indicator function equals 1 if the condition c is fulfilled, otherwise it is 0

Stochastic and deterministic processes

$\boldsymbol{Q(t)}$	row vector describing the stochastic process of the numbers of customers in the treated model
$Q_i(t), Q_{\mathcal{O}i}(t)$	stochastic processes in vector $\boldsymbol{Q(t)}$ describing the number of customers of class $i = 1, 2$ in the system and the orbit
$A\Big(\int\limits_0^t \cdot\, ds\Big)$	Standard Poisson process with mean 1 (See Pages 32f.)
$B\Big(\int\limits_0^t \cdot\, ds\Big)$	Standard Brownian motion with mean 0 and standard deviation 1 (See Pages 60ff.)
$\boldsymbol{Q^F(t)}$	$= \big(Q_i^F(t), \ldots, Q_{\mathcal{O}i}^F, \ldots\big)$ row vector describing the fluid processes of the numbers of customers in the treated model
$\boldsymbol{Q^D(t)}$	$= \big(Q_i^D(t), \ldots, Q_{\mathcal{O}i}^D, \ldots\big)$ row vector describing the diffusion processes of the number of customers in the treated model
$\mathbf{E}\big[\boldsymbol{Q^D(t)}\big]$	mean row vector of the diffusion processes
$\mathbf{COV}\big[\boldsymbol{Q^D(t)}\big]$	covariance matrix of the diffusion processes
$\mathbf{COV}\big[Q_i^D(t), Q_j^D(t)\big]$	covariance of the diffusion processes modelling the number customers in the system or the orbit $i, j \in \{S, \mathcal{O}, 1, 2, \mathcal{O}1, \mathcal{O}2\}$
$\mathbf{VAR}\big[Q_i^D(t)\big]$	variance of the diffusion process describing the number of customers in the system or the orbit

Performance measures in the fluid model

$W_i^F(t)$	time-dependent waiting time of customers of type $i = 1, 2$ (See Pages 39, 76 and 127f.)
$W_{agg,i}^F(T)$	aggregated waiting time over the interval $[0, T]$ of customers of type $i = 1, 2$ (See Page 40)
$P_i^F(\text{served}, t)$	time-dependent probability of being served for customers of type $i = 1, 2$ (See Pages 40, 77, and 129)
$P_{agg,i}^F(\text{served}, T)$	aggregated probability of being served over the interval $[0, T]$ for customers of type $i = 1, 2$ (See Page 40)
$U_i^F(t)$	time-dependent utilisation of agents of type $i = 1, 2, G$ (See Pages 41, 78, and 130)
$U_{agg,i}^F(t)$	aggregated utilisation over the interval $[0, T]$ of agents of type $i = 1, 2, G$ (See Page 42)
$L_i^F(t)$	time-dependent queue length of customers of type $i = 1, 2$ (See Page 121f.)

$B_i^F(t)$	time-dependent number of busy generalists serving customers of type $i = 1, 2$ (See Page 121f.)
$d_i(t)$	time-dependent departure rate of customers of type $i = 1, 2$ (See Pages 38 and 121f.)

Profit and cost parameters

r_i	revenue gained from serving customers of type $i = 1, 2$
w_i	hourly wages of agents of type $i = 1, 2, G$
ℓ	hourly costs for occupied telephone lines
$\mathrm{cost}(T)$	cost function of the fluid model (See Page 130)
$\mathrm{profit}(T)$	profit function of the fluid model (See Pages 78 and 131)

Parameters of the staffing and shift scheduling problem

δ	length of the considered time intervals
$\mathbf{0}$	vector of 0s
$\boldsymbol{x}$	shift schedule, vector of the number of shifts scheduled
$x_k^{(i)}$	number of shifts of type $k = 1, \ldots, K_i$ for agents of type $i = 1, 2, G$
$\mathcal{J}$	number of time intervals j for the shift scheduling problem
$\boldsymbol{s}_k$	shift of type $k = 1, \ldots, K$, i.e., a boolean vector of length $\mathcal{J}$
$c_k^{(i)}$	costs for agents of type $i = 1, 2, G$ working according to shift type $k = 1, \ldots, K_i$
$\mathrm{profit}(\boldsymbol{x}, T)$	profit gained from staffing according to schedule $\boldsymbol{x}$ (See Pages 164 and 190)
M_i	maximum total number of agents of type $i = 1, 2G$ who can be staffed
$N_i(\boldsymbol{x}, t)$	function of the time-dependent number of agents of type $i = 1, 2, G$ staffed according to shift $\boldsymbol{x}$

Parameters and functions for the optimisation heuristic

$\overline{\boldsymbol{\lambda}}_i$	$\mathcal{J}$-dimensional vector of the average arrival rates to agents of type $i = 1, 2, G$ in the time intervals $j = 1, \ldots, \mathcal{J}$ (See Page 170)
$\mathrm{marg}(\overline{\boldsymbol{\lambda}}_i, \boldsymbol{s}_k)$	relative profit margin of shift type $i = 1, \ldots, K$ for agents of type $i = 1, 2, G$ (See Page 171)
$\mathrm{cum\_work}(\overline{\boldsymbol{\lambda}}_i, \boldsymbol{s}_k)$	cumulated work of shift type $i = 1, \ldots, K$ for agents of type $i = 1, 2, G$ (See Page 172)

List of Figures

List of Tables

References

M. Abramowitz and I.A. Stegun, editors. *Handbook of Mathematical Functions.* Dover Publications, Inc., 10 edition, 1974. ISBN 0-486-61272-4. 43

L.G. Afanas. Ergodicity conditions for queueing systems with repeated calls. *Journal of Mathematical Science*, pages 2835–2838, 1994. 34

S. Aguir, F. Karaesmen, O.Z. Akşin, and F. Chauvet. The impact of retrials on call center performance. *OR Spectrum*, 26:353–376, 2004. 2.2, 44, 40, 3.2.4.2, 3.4, 7, 4.1.5.1, 4.3

S. Aguir, O.Z. Akşin, F. Karaesmen, and Y. Dallery. On the interaction between retrials and sizing of call centers. Working paper, URL: http://home.ku.edu.tr/˜zaksin/Rappels_z.pdf, December 2005. 7, 4.3, 5.6

U. Aickelin and P. White. Building better nurse scheduling algorithms. *Annals of Operations Research*, 128:159–177, 2004. 31

B. Almási, G. Bolch, and J. Sztrik. Heterogeneous finite-source retrial queues. *Journal of mathematical sciences*, 121(5):2590–2596, 2004. 4.3

E. Altman, T. Jiménez, and G. Koole. On the comparison of queueing systems with their fluid limits. *Probability in the Engineering and Informational Sciences*, 15:165–178, 2001. 3.2.2, 3.2.4.1, 3.4

V.V. Anisimov and J.R. Artalejo. Analysis of Markov multiserver retrial queues with negative arrivals. *Queueing Systems*, 39:157–182, 2001. 4.3

V.V. Anisimov and K.L. Atadzhanov. Diffusion approximation of systems with repeated calls and an unreliable server. *Journal of Mathematical Science*, 72(2):3032–3034, 1994. 3.4, 35

A. Antipov and N. Meade. Forecasting call frequency at a financial services call centre. *Journal of the Operational Research Society*, 53(9):953–960, 2002. 46, 59

M. Armony and C. Maglaras. On customer contact centers with call-back option: Customers' decisions routing rules, and system design. *Operation Research*, 2004. to appear, URL: http://www2.gsb.columbia.edu/faculty/cmaglaras/papers/cc-static.html. 33

M. Armony and A. Mandelbaum. Design, staffing and control of large service systems: The case of a single customer class and multiple server types. Preprint, URL: http://iew3.technion.ac.il/serveng/References/InvertedV.pdf, 2004. 3.4

J.R. Artalejo. A queueing system with returning customers and waiting line. *Operations Research Letters*, 17:191–199, 1995. 31, 4.3

J.R. Artalejo and M. J. Lopez-Herrero. On the single server retrial queue with balking. *INFOR*, 38(1):33–50, 2000. 34, 36

J.R. Artalejo and M. Pozo. Numerical calculation of the stationary distribution of the main multiserver retrial queue. *Annals of Operations Research*, 116:41–56, 2002. 4.3

S. Asmussen. *Applied Probability and Queues*, volume 51 of *Stochastic Modelling and Applied Probability.* Springer, New York et al., second edition, 2003. 32, 4, 5, 33

R. Atar. A diffusion model of scheduling control in queueing systems with many servers. *Annals of Applied Probability*, 15:820–852, 2005a. 3.4, 23

R. Atar. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. Preprint URL: http://www.ee.technion.ac.il/people/atar/asymp2.ps, 2005b. 3.4, 23

R. Atar, A. Mandelbaum, and M.I. Reiman. A Brownian control problem for a simple queueing system in the Halfin-Whitt regime. *System and Control Letters*, 51(3-4):269–275, 2004a. 3.4

R. Atar, A. Mandelbaum, and M.I. Reiman. Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy-traffic. *Annals of Applied probability*, 14(3):1084–1134, August 2004b. 3.4

J. Atlason, M.A. Epelman, and S.G. Henderson. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*, 127:333–358, 2004. 7, 5.2, 5.6

T Aykin. A comparative evaluation of modeling approaches to the labor shift scheduling problem. *European Journal of Operational Research*, 125(2): 381–397, September 2000. 29

T. Aykin. A composite branch and cut algorithm for optimal shift scheduling with multiple break windows. *Journal of the Operations Research Society*, 49:603–615, 1998. 29

A. Bassamboo, J. M. Harrison, and A. Zeevi. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Operations Research*, 2005. to appear. 34

H. Bauer. *Measure and Integration Theory.* Studies in Mathematics 26. deGruyter, first edition, 2001a. ISBN 3-110-16719-0. 62

H. Bauer. *Wahrscheinlichkeitstheorie.* deGruyter, fifth edition, 2001b. ISBN 3-110-17236-4. 55

S.E. Bechtold and L.W. Jacobs. The equivalence of general set–covering and implicit integer programming formulations for shift scheduling. *Naval Research Logistics*, 43:233–249, 1996. 29

S. Bhulai, G. Koole, and A. Pot. Simple methods for shift scheduling in multi-skill call centers. submitted, April 2006. 53, 5.6

P. Billingsley. *Convergence of Probability Measures*. Probability and Statistics. Wiley-Interscience, second edition, 1999. ISBN 0-471-19745-9. 62

V.A. Bolotin. Modeling call holding time distributions for ccs network design and performance analysis. *IEEE Journal on Selected Areas in Communications*, 12(3):433–438, 1994. 2.2

F. Borchardt, M. Krafft, W. M., W. Schwetz, and P. Winkelmann. Das Jahresgutachten des CRM-Expertenrats für 2005, 2005. 6, 57

S. Borst, A. Mandelbaum, and M. Reiman. Dimensioning large call centers. *CWI Report, Probability, Networks, and Algorithms*, 2002. 50, 3, 47

S. Borst, A. Mandelbaum, and M. I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004. 50, 5.6

L. Breuer, A. Dudin, and V. Klimenok. A retrial $BMAP/PH/N$ system. *Queueing Systems*, 40:433–457, 2002. 34

L. Brown. The analogy between statistical equivalence and stochastic strong limit theorems. Technical report, Wharton School, University of Pennsylvania, 2002. 57

L. Brown. Empirical analysis of call center traffic. In *Call Center Forum*, 2003. Wharton Financial Institutions Center, May 8, 2003. 58

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of telephone call center: A queueing-science perspective. Working paper, URL: http://fic.wharton.upenn.edu/fic/papers/03/0312.pdf, November 2002. 2.2, 2.2, 29, 30, 2.2, 45, 58, 13, 14

E.K. Burke, P. de Causmaecker, G. vanden Berghe, and H. van Landeghem. The state of the art of nurse rostering. *Journal of Scheduling*, 7:441–499, 2004. 31

M. T. Cezik and P. L'Ecuyer. Staffing multiskill call centers via linear programming and simulation. URL: http://www.iro.umontreal.ca/˜lecuyer/papers.html, 2005. 23, 53, 5.2, 5.6

S.R. Chakravarthy and A.N. Dudin. A multi-server retrial queue with BMAP arrivals and group services. *Queueing Systems*, 42:5–31, 2002. 4.3

T. Chamberlain. Emerging contact center solutions: Outbound predictive dialing, June 2001. Summer Conference, URL: http://www.trmanet.org/members/mopdfs-ppt/su01_outbound_predictive_dialing.pdf. 17

J. Chang, H. Ayhan, J.G. Dai, and C.H. Xia. Dynamic scheduling of a multiclass fluid model with transient overload. URL: http://www.isye.gatech.edu/faculty/dai/Preprint/acdx0515.pdf, April 2004. 3.4

B.P.K. Chen and S.G. Henderson. Two issues in setting call centre staffing levels. *Annals of Operations Research*, 108:175–192, 2001. 19

J.W. Cohen. Basic problems of telephone traffic theory and the influence of repeated calls. *Philips Telecommunication Review*, 18(2):49–100, 1957. 4.3

D.J. Daley and L.D. Servi. Estimating customer loss rates from transactional data. *Statistics Research Report*, 12:1–17, 1997. 59

F. de Véricourt and Y.-P. Zhou. Managing response time in a call routing problem with service failure. *Operations Research*, 53(6):968–981, November–December 2005. 4.3

J.E. Diamond and A.S. Alfa. Matrix analytical methods for M/PH/1 retrial queues. *Commun. Statist.-Stochastic Models*, 11(3):447–470, 1995. 4.3

J.E. Diamond and A.S. Alfa. Matrix analytical methods for a multi-server retrial queue with buffer. *Sociedad de Estadistica e Investigación Operativa*, 7(2):249–266, 1999. 4.3

K.A. Dowsland and J.M. Thompson. Solving a nurse scheduling problem with knapsacks, networks, and tabu search. *Journal of the Operational Research Society*, pages 825–833, 2000. 31

A. Dudin and V. Klimenok. Queueing system $BMAP/G/1$ with repeated calls. *Mathematical and Computer Modeling*, 30:115–128, 1999. 4.3, 34

G. Eitzen, D. Panton, and G. Mills. Multi-skilled workforce optimisation. *Annals of Operations Research*, 127:359–372, 2004. 29

A.T. Ernst, H. Jiang, M. Krishnamoorthy, B. Owens, and D. Sier. An annotated bibliography of personnel scheduling and rostering. *Annals of Operations Research*, 12:21–144, 2004. 52, 5.1, 27, 32

P. Eveborn and M. Rönnqvist. Scheduler - a system for staff planning. *Annals of Operations Research*, 128:21–45, 2004. 29

A. Evenson, P.T. Harker, and F.X. Frei. Effective call center management: Evidence from financial services. *Working paper series / Financial Institutions Center, Wharton School, University of Pennsylvania*, 98(25B), 1999. 6, 56

G. Falin. A survey of retrial queues. *Queueing Systems*, 7:127–167, 1990. 4.3

G. Falin and J.R. Artalejo. Approximations for multi-server queues with balking/retrial discipline. *OR Spektrum*, 17:239–244, 1995. 4.3

G. Falin and A. Gomez-Corral. On a bivariate Markov process arising in the theory of single–server retrial queues. *Statistica Neerlandica*, 54(1):67–78, 2000. 34

G. Falin and J.G.C. Templeton. *Retrial Queues*. Chapman und Hall, 1997. 4.3

G. Fayolle and M.A. Brun. On a system with impatience and repeated calls. In *Queueing theory and its applications*, volume 7, pages 283–305. North-Holland, Amsterdam, 1988. 36

Z. Feldman. Staffing of time-varying queues to achieve time-stable performance. Master's thesis, Technion - Israel Institute of Technology, Haifa, Israel, 2004. 17, 3.2.2, 3.2.2, 3.2.4.1, 4.1.5.1

Z. Feldman, A. Mandelbaum, W.A. Massey, and W. Whitt. Staffing of time-varying queues to achieve time-stable performance. Preprint , URL: http://iew3.technion.ac.il/serveng/references, January 2005. 26, 50, 51, 3.3.2, 3.3.4, 3.4, 6, 5.3.1, 5.6, 19

A. Fukunaga, E. Hamilton, J. Fama, D. Andre, O. Matan, and I. Nourbakhsh. Staff scheduling for inbound call centers and customers contact centers. Technical report, American Association for Artificial Intelligence, Blue Pumpkin Software 884Hermosa Court, Suite 100 Sunnyvale, CA 94086, 2002. 55, 1, 5.6

N. Gans and Y.-P. Zhou. A call-routing problem with service-level constraints. *Operations Research*, 51(2):255–271, 2003. 33

N. Gans and Y.-P. Zhou. Overflow routing for call-center outsourcing. *Manufacturing and Service Operations Management*, 2004. to appear, URL: http://opim.wharton.upenn.edu/ gans/pubs/Gans-Zhou-Outsourcing.pdf. 5.6

N. Gans and Y.-P. Zhou. Managing learning and turnover in employee staffing. *Operations Research*, 50:991–1006, 2002. 5.6

N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review and research prospects. *Manufacturing and Service Operations Management*, 5:79–141, 2003. 11, 40, 2.5, 2, 11, 3, 17, 20

O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing and Service Operations Management*, 4(3):208–227, 2002. 18, 51, 2, 3, 3.2.2, 3.2.5, 47, 3.3.2, 3.3.2, 4.1.1, 18, 5.6

E. Gelenbe and G. Pujolle. *Introduction to Queueing Networks.* Jon Wiley & Sons, New York, second edition, April 1999. Reprint. 32

P.W. Glynn and W. Whitt. Heavy-traffic extreme-value limits for queues. *Operations Research Letters*, 18:107–111, 1995. 3.4

I. Grama and M. Nussbaum. A nonstandard Hungarian construction for partial sums, June 1997. URL: http://www.wias-berlin.de/ publications/Preprints/324/ wias_Preprints_324. pdf. 57

L. V. Green, P. J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. URL: http://www.columbia.edu/˜ww2040/, 2005. 6, 25

L.V. Green, P.J. Kolesar, and J. Soares. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, 49 (4):549–564, July–August 2001. 16, 13, 25, 28

N. Grier, W.A. Massey, T. McKoy, and W. Whitt. The time-dependent erlang loss model with retrials. *Telecommunication Systems*, 7:253–265, 1997. 4.3

D. Gross and C.M. Harris. *Fundamentals of Queueing Theory.* John Wiley & Sons, New York et al., third edition, 1998. 4

P. Gross and B. Badura. Sozialpolitik und soziale Dienste: Entwurf einer Theorie personenbezogener Dienstleistungen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 19:361–385, 1977. Sonderheft: Soziologie und Sozialpolitik. 3

I. Gurvich, M. Armony, and A. Mandelbaum. Staffing and control of large-scale service system with multiple customer classes and fully flexible servers.

URL: http://iew3.technion.ac.il/References/Vdesign.pdf, November 2004. 3.4, 5.6, 21

B. Haas Margolius. A sample path analysis of the $M_t/M_t/c$ queue. *Queueing Systems*, 31:59–93, 1999. 25, 27

S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–589, 1981. 3.2.2, 3.2.2, 3.3.1, 54, 3.4, 4.1.2, A.1

R.C. Hampshire and W.A. Massey. Call center staffing for profit optimality. submitted to Management Science, special issue on call center management, 2005. 3.4, 2, 5, 6, 5.6

T. Hanschke. Markov chains and generalized continued fractions. *Journal of applied Probability*, 29:838–849, 1992. 20

T. Hanschke. A matrix continued fraction algorithm for the multiserver repeated order queue. *Mathematical and Computer Modelling*, 30:159–170, 1999. 20, 4.3

J. M. Harrison and A. Zeevi. Dynamic scheduling of a multi-class queue in the Halfin-Whitt heavy traffic regime. *Operations Research*, 52(2):243–257, 2004. 34, 31, 3.4, 21

J. M. Harrison and A. Zeevi. A method for staffing large call centers based on stochastic fluid models. *Manufacturing and Service Operations Management*, 7:20–36, 2005. 34, 3.4, 6, 24

L. Hawkins, T. Meier, W.S. Nainis, and H.M. James. The evolution of the call center to 'customer contact center'. URL: http//www.itsc.state.md.us/PDF/Call_Center_Evolution_C4_Final.pdf, February 2001. ITSC White Paper. 2

S. Helber and R. Stolletz. *Call Center Management*. Springer–Verlag, 2003. 16, 2.2, 36, 2.5, 15, 6, 5.6, 6

S. Helber and R. Stolletz. Grundlagen und Probleme der Personalbedarfsermittlung. *Zeitschrift für Betriebswirtschaft: Produktion von Dienstleistungen*, 1:67–88, 2004. 39, 47, 59, 39, 2, 6, 5.6

W.B. Henderson and W.L. Berry. Heuristic methods for telephone operator shift scheduling: An experimental analysis. *Management Science*, 22(12): 1372–1380, August 1976. 5.6

K. Hoffman and C.M. Harris. Estimation of a caller retrial rate for a telephone information system. *European Journal of Operational Research*, 27:207–214, 1986. 4.3

D.L. Iglehart. Limit diffusion approximations for the many-server queue and the repairman problem. *Journal of Applied Probability*, 1965. 3.3.1

A. Ingolfsson, M.A. Haque, and A. Umnikov. Accounting for time-varying queueing effects in tour scheduling. *European Journal of Operational Research*, 139:585–597, 2002. 5.6

A. Ingolfsson, E. Cabral, and X. Wu. Combining integer programming and the randomization method to schedule employees. Technical Report 02-1, Department of Finance and Management Science, Faculty of Business, University of Alberta, 2003. 53, 13, 5.6

International Engineering Consortium. Speech-enabled interactive voice response systems. URL: http://www.iec.org/online/tutorials/speech_enabled/index.html, 2005. 15

M.W. Isken. An implicit tour scheduling model with applications in healthcare. *Annals of Operations Research*, 128:91–109, 2004. 31

K. Jackson. Thinking beyond the old 80/20 rule. *Call Center Magazine*, July 2002. URL: http://www.callcentermagazine.com/ shared/ article/ showArticle.jhtml?articleId=8701899. 36

P. Jelenkovic, A. Mandelbaum, and P. Momcilovic. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems*, 47:53–69, 2004. 3.3.2

O.B. Jennings, A. Mandelbaum, W. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, 42:1383–1394, 1996. 26, 50, 3.3.2, 70, 5.6

T. Jiménez and G. Koole. Scaling and comparison of fluid limits of queues applied to call centers with time-varying parameters. *OR Spectrum*, 26: 413–422, 2004. 22, 23, 29, 38, 3.2.4.1, 3.2.5, 72, 3.4, 5.3.2

G. Jongbloed and G. Koole. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318, 2001. 21

R. Kalyanaraman and B. Srinivasan. A single server retrial queue with two types of calls and recurrent repeated calls. *International Journal of Information and Management Sciences*, 14(4):49–62, 2003. 4.3

W. Kang, F.P. Kelly, N.H. Lee, and R. J. Williams. Fluid and Brownian approximation for an internet congestion control model. In *Proceedings of the 43rd IEEE Conference on Decision and Control*, pages 3938–3943, December 2004. 14

I. Karatzas and S.E. Shreve. *Brownian Motion and Stochastic Calculus.* Graduate Texts in Mathematics. Springer, 2 edition, 1991. 66, 4

S. Karlin and H.M. Taylor. *A First Course in Stochastic Processes.* Academic Press, 1975. 22, 1, 24

F.P. Kelly and R.J. Williams. Fluid model for a network operating under a fair bandwidth-sharing policy. *Annals of Applied Probability*, 14:1055–1083, 2004. 14

J.F.C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society*, 24:383 – 392, 1962. 3.3.1, 53

J.F.C. Kingman. The heavy traffic approximation in the theory of queues. In W. Smith and W. Wilkinson, editors, *Proceedings of the Symposium on Congestion Theory*, pages 137–159. University of North Carolina Press, 1965. 53

L. Kleinrock. *Queueing Systems, Volume I: Theory.* Wiley–Interscience, 1975. ISBN 0-471-49110-1. 1, 6, 9

N. Kohl and S.E. Karisch. Airline crew rostering: Problem types, modeling, and optimization. *Annals of Operations Research*, 127:223–257, 2004. 32

J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent RV'-s, and the sample DF.i. *Zeitschrift für Wahrscheinlichkeitstheorie verw. Gebiete*, 32:111–131, 1975. 3.3.2, 3.3.2, A.1

G. Koole. Optimization of business processes - applications and theory of stochastic modeling, December 2002. lecture notes, URL: http://www.math.vu.nl/˜koole/obp/obp.pdf. 1, 49, 2.5

G. Koole. Redefining the service level in call centers. Technical report, Department of Stochastics, Vrije Universiteit Amsterdam, 2003. Working paper, URL: http://www.math.vu.nl/˜koole/articles/report03b/. 35, 36, 48

G. Koole. Performance analysis and optimization in customer contact centers. In *Proceedings of QEST 04, IEEE, Enschede*, pages 2–5, 2004. 42

G. Koole. Call Center Mathematics, a scientific method for understanding and improving your contact center, June 2005. URL: http://www.math.vu.nl/˜koole/ccmath/book.pdf. 12, 41, 45, 2.5, 39

G. Koole and A. Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operation Research*, 113:41–59, 2002. 1, 7, 2.5, 2

G. Koole and A. Pot. A note on profit maximization and monotonicity results for inbound call centers. submitted, 2005. 2, 6, 5.2, 5.6

G. Koole and A. Pot. An overview of routing and staffing algorithms in multi-skill customer contact centers. Working paper, URL: http://www.math.vu.nl/˜koole/articles/report06a/art.pdf, March 2006. 33, 52, 60, 3, 7, 5.6, 30

G. Koole and J. Talim. Exponential approximation of multi-skill call centers architecture. In *Proceedings of QNETs 2000, Ilkley (UK)*, pages 23/1–10, 2000. 24

G. Koole and E. Van der Sluis. Optimal shift scheduling with a global service level constraint. *IIE Transactions*, 35(11), 2003. 49, 7, 5.2, 5.6

T.G. Kurtz. Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and their Applications*, 6:223–240, 1978. 57

M.A. Lariviere and J.A. Van Mieghem. Strategically seeking service: How competition can generate Poisson arrivals. *Manufacturing and Service Operations Management*, 6(1):23–40, 2004. 24, 12

R.V. Lüde and M.R. Nerlich. Call Center in Arbeits- und Betriebssoziologischer Perspektive: Chancen und Risiken im Betrieblichen Innovationsprozess. In *Call Center Evolution*, pages 1–19. Verlag Vahlen. München, 2002. 10, 56, 11

P. Maas and A. Graf. Leadership by customers? New roles of service companies' customers. *German Journal for Human Resource Research*, 18(3), 2004. 6

P. Maas and A. Graf. New roles and functions of customers in service companies - new challenges for HRM. In *EURAM 2005 - Annual Conference of the European Academy of Management*, Munich, May 2005. Paper 237. 3

A. Mandelbaum. Call centers: Research bibliography with abstracts, January 2004. version 4, URL: http://iew3.technion.ac.il/serveng/References/references.html. 2.5

A. Mandelbaum and W.A. Massey. Strong approximations for time-dependent queues. *Mathematics of Operations Research*, 20:33–64, 1995. 22, 3.4

A. Mandelbaum and N. Shimkim. A model for rational abandonments from invisible queues. *Queueing Systems: Theory and Applications (QUESTA)*, 36:141–173, December 2000. 7

A. Mandelbaum, W.A. Massey, and M. I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30:149–201, 1998. 21, 3.2.1, 3.2.2, 25, 3.2.2, 30, 34, 35, 3.2.2, 37, 3.3.1, 52, 3.3.2, 3.3.2, 58, 59, 3.3.2, 60, 61, 3.3.2, 65, 67, 68, 69, 3.4, 71, 4.1.1, 2, 4, 6, 4.2.1, 23, 26, 4.2.3, 1, 2, A.2

A. Mandelbaum, W.A. Massey, M.I. Reiman, and B. Rider. Time varying multiserver queues with abandonment and retrials. *ITC, Teletraffic Engineering in a Competitive World*, 16:355–364, 1999a. 25, 3.3.4, 3.4, 1, 3, 4, 5, 28, 4.3, 2

A. Mandelbaum, W.A. Massey, M.I. Reiman, and A.L. Stolyar. Waiting time asymptotics for time varying multiserver queues with abandonment and retrials. In *Allerton Conference Proceedings*, 1999b. 38, 25, 3.2.3.1, 3.3.4, 3.4, 1, 3, 28, 4.3

A. Mandelbaum, A. Sakov, and S. Zeltyn. Empirical analysis of a call center. Technical report, Technion, Israel Institute of Technology, 2000. URL: http://iew3.technion.ac.il/serveng/References/ccdata.pdf. 7, 27, 29, 2.2, 58, 13

A. Mandelbaum, W.A. Massey, M.I. Reiman, A.L. Stolyar, and B. Rider. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems: Modeling, Analysis, Design and Management*, 21(2):149–172, 2002. 3.2.3.1, 3.3.4, 3.4, 1, 3, 5, 28, 4.3, 36, 2

P. Marchal. Constructing a sequence of random walks strongly converging to Brownian motion. *Discrete Mathematics and Theoretical Computer Science*, pages 181–190, 2003. 51

W.A. Massey. The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems*, 21(2-4):173–204, 2002. 25

V. Mehrotra and J. Fama. Call center simulation modeling: Methods, challenges, and opportunities. In S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, editors, *Proceedings of the 2003 Winter Simulation Conference*, pages 135–143, 2003. 23

U. Mitzlaff. *Diffusionsapproximation von Warteschlangensystemen.* PhD thesis, Technical University of Clausthal, July 1997. 3.4

M. Moz and M. Vaz Pato. Solving the problem of rerostering nurse schedules with hard constraints: New multicommodity flow models. *Annals of Operations Research*, 128:179–197, 2004. 31

N. Musliu. *Intelligent Search Methods for Workforce Scheduling: New Ideas and Practical Applications.* PhD thesis, Vienna, University of Technology, September 2001. 1, 29

N. Musliu, A. Schaerf, and W. Slany. Local search for shift design. *European Journal of Operational Research*, 153(1):51–64, 2004. 1

M. F. Neuts. *Matrix-geometric solutions in stochastic models - An Algorithmic Approach.* Johns Hopkins series in the mathematical sciences. Johns Hopkins University Press, 1981. 20, 4.3

G.F. Newell. *Applications of Queueing Theory.* Chapman and Hall, London, New York, second edition, 1982. 29, 3.2.2, 36, 3.4, 72

S. Oh. Modeling ACD data to improve computer simulation of call center. Master's thesis, University of Calgary, Falculty of Management, 1999. 23

K. Rahko. Measurements for control and modeling of teletraffic. In *Proceedings of the 13th international Teletraffic Congress*, 1991. Copenhagen. 2.2

M.F. Ramalhoto and A. Gómez-Corral. Some decomposition formulae for $M/M/r/r+d$ queues with constant retrial rate. *Commun. Statist.-Stochastic Models*, 14(1-2):123–145, 1998. 4.3

C.R. Rao. *Linear Statistical Inference and its Applications.* Wiley, New York, second edition, 1973. 3.3.2, A.1, A.2

C.R. Reeves. *Modern Heuristic Search Methods*, chapter 1, pages 1–25. John Wiley & Sons Ltd., 1996. 8, 9, 10, 11

Z.J. Ren and Y.-P. Zhou. Call center outsourcing: Coordinating staffing level and service quality. under review, URL: http://smgpublish.bu.edu/ren/, October 2004. 5.6

A.D. Ridley, W. Massey, and M. Fu. Fluid approximation of a priority call center with time-varying arrivals. *The Telecommunications Review*, pages 69–77, 2004. 23

A. Rodrigo, M. Vázquez, and G. Falin. A new Markovian description of the $M/G/1$ retrial queue. *European Journal of Operational Research*, 104:231–240, 1998. 4.3

A. Schmidt. *Matrix-Kettenbrüche und Markovsche Prozesse.* PhD thesis, Clausthal, Technical University, 1997. 20

B.H. Schmitt. *Customer Experience Management: A Revolutionary Approach to Connecting with Your Customers.* John Wiley & Sons, Ltd., 2003. 13, 57

F. Schoenberg. On rescaled Poisson processes and the Brownian bridge. *Annals of the Institute of Statistical Mathematics*, 54(2):445–457, 2002. 51

M. Segal. The operator-scheduling problem: A network-flow approach. *Operations Research*, 24:808–823, 1974. 5.6

E. Seneta. Finite approximations to infinite non-negatve matrices. *Proc. Camb. Phil. Soc.*, 63:893–912, 1967. 19

E. Seneta. Finite approximations to infinite non-negative matrices II. *Proc. Camb. Phil. Soc.*, 64:465–470, 1968. 19

E. Seneta. Computing the stationary distribution for infinite Markov chains. *Linear Algebra and its Applications*, 34:259–267, 1980. 19, 20

M.E. Sisselman and W. Whitt. Empowering contact-center agents through preference-based routing. *Production and Operations Management*, December 2004. SeatLink Inc. 1, 33

W.J. Stewart. *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, 1994. 19

M. Stockmann. Aktuelle Kennzahlen der Call Center Branche. CallCenter-World 2005, Pressekonferenz, February 2005. 5, 7

R. Stolletz. *Performance Analysis and Optimization of Inbound Call Centers.* Number 528 in Lecture notes in Economics and Mathematical Systems. Springer, 2003. 9, 33, 40, 42, 2, 8, 3.1.3, 3.1.3

D.Y. Sze. A queueing model for telephone operator staffing. *Operations Research*, 32(2):229–249, 1984. 5.6

P. Taylor, G. Mulvey, J. Hyman, and P. Bain. Work organisation, control, and the experience of work in call centers. *Work, Employment and Society*, 16(1):133–150, 2002. 11

G.M. Thompson. A simulated-annealing heuristic for shift scheduling using non-continuously available employees. *Computers Ops Res*, 23(3):275–288, 1996. 8, 29

R.L. Tweedie. The calculation of limit probabilities for denumarable Markov processes from infinitesimal properties. *Journal of Applied Probability*, 10: 84–99, 1973. 19

P. Van der Cruyssen. Linear difference equations and generalized continued fractions. *Journal on Computing*, 22:269–278, 1979. 19

R.B. Wallace and W. Whitt. A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management*, 2004. submitted. 22

A.R. Ward and P. W. Glynn. Properties of the reflected Ornstein–Uhlenbeck process. *Queueing systems: theory and applications*, 44(2):109–124, 2003a. ISSN 0257-0130. 25

A.R. Ward and P.W. Glynn. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems: Theory and Applications*, 43(1):103–128, 2003b. ISSN 0257-0130. 51, 3.4, 25

A.R. Ward and P.W. Glynn. Diffusion approximation of $GI/GI/1$ queues with balking or reneging. *Queueing Systems*, 50(4), August 2005. 3.4

Wharton Business School. Web-based call centers transform customer service, February 2000. Internet publication, URL: http://www1.skillsoft.com/elearning/wharton/archives/oper_mgmnt.html. 10, 13

Wharton Business School. Telephone call centers: The factory floors of the 21st century, April 2002. Internet publication, URL: http://knowledge.wharton.upenn.edu/articles.cfm?catid=14&article id=540. 7

Wharton Business School. Call centers: How to reduce burnout, increase efficiency, May 2004. internet publication, URL: http://knowledge.wharton.upenn.edu/article/997.cfm. 37, 41

W. Whitt. An overview of Brownian and non-Brownian FCLTs for the single-server queue. *Queueing Systems*, 36:39–70, 2000. 64, 25

W. Whitt. *Stochastic-Process Limits.* Springer, 2002a. 22, 26, 27, 32, 33, 56, 60, 63, 71

W. Whitt. Stochastic models for the design and management of customer contact centers: Some research directions. Working paper, URL: http://www.columbia.edu/˜ww2040/directions.pdf, 2002b. 2.5

W. Whitt. How multiserver queues scale with growing congestion-dependent demand. *Operations Research*, 51:531–542, 2003. 28, 34, 50, 3.3.2

W. Whitt. Engineering solution of a basic call-center model. *Management Science*, 51(2), February 2005a. URL: http://www.columbia.edu/˜ww2040/submissionREV.pdf. 28, 3, 6

W. Whitt. Engineering solution of a basic call-center model: Supplementary material. Extra material related to Whitt (2005a), URL: http://www.columbia.edu/˜ww2040/supplementREV.pdf, February 2005b. 34

W. Whitt. Two fluid approximations for multi-server queues with abandonments. *Operations Research Letters*, 33:363–372, Aug 2005c. submitted to Operations Research, URL: http://columbia.edu/˜ww2040/CallFluid3.pdf. 47, 49

W. Whitt. Staffing a call center with uncertain arrival rates and absenteeism. *Production and Operations Management*, 2006a. to appear, URL: http://www.columbia.edu/˜ww2040/FluidStaff.pdf. 3.4, 5, 24, 26

W. Whitt. Sensitivity of performance in the erlang-a model to changes in the model parameters. *Operations Research*, (2):247–260, 2006b. submitted. 48

W. Whitt. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50(10):1449–1461, October 2004. 48

W. Whitt. Improving service by informing customers about anticipated delays. *Management Science*, 45(2):193–207, 1999a. 8

W. Whitt. Predicting queueing delays. *Management science*, 45(6):870–888, June 1999b. 29

W. Whitt. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24:205–212, June 1999c. 6, 19

R.S. Winer. Customer relationship management: A framework, research directions, and the future. URL: groups.haas.berkeley.edu/fcsuit/PDF-papers/CRMApril 2001. Haas School of Business, University of California at Berkeley. 4

Witness Systems. Die fünf Mythen des Qualitätsmanagements im Contact Center. URL: http://www.witness-systems.de/international/germany/Produkte%20%26%20Service/whitepapers/docs, July 2002a. Executive whitepaper. 10, 55

Witness Systems. Kosten senken – Kunden halten: Return on Investment im Call Center. URL: http://www.witness-systems.de/international/germany/Produkte%20%26%20Service/whitepapers/docs, November 2002b. Executive whitepaper. 13

Witness Systems. Momentaufnahme: Wie steht es um die Motivation, Training und Entwicklung in deutschen Call Centern. URL: http://www.witness-systems.de/international/germany/Produkte%20%26%20Service/whitepapers/docs, August 2004. executive whitepaper. 56, 11

R.W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2), 1982. 31, 38

E. Zohar, A. Mandelbaum, and N. Shimkin. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science*, 48(4):566–583, 2002. 58