An Overview of Routing and Staffing Algorithms in Multi-Skill Customer Contact Centers

Ger Koole & Auke Pot

Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands

Submitted version

6th March 2006

Abstract

This paper gives an overview of routing and staffing algorithms in multi-skill contact centers. Related issues and problems are characterized, and models and mathematical tools from the literature for modeling and optimizing contact centers are described. In addition, a general model of a multi-skill contact center is described and the papers from the literature are put in that framework.

keywords: multi-skill call centers, literature overview, contact centers, skill-based routing, planning, workforce management

1 Introduction

During the last decennia call centers have become a well-known phenomenon. The total number of call centers increased substantially and call centers became bigger: the larger call centers employ 500–2000 agents, there are ten thousands of call centers across the globe, and between 1,500,000 and 2,000,000 agents. Outsourcing to for example India is becoming very popular. It is expected that this country will build up an industry that is worth 17 billion by 2008. In India 100,000s of engineers who can work in call centers, graduate each year (see NASSCOM (2006)). A reason for the growth of the industry is that customer relationships and customer services have become more important. There is also an explanation for the fact that call centers became larger; more representatives are assembled together in one building to improve service, reduce costs, and obtain a higher productivity. However, maximizing the economies of scale also requires improvement of the operational processes, especially concerning job routing and workforce management (WFM). Hence, forced by the pressure of improving these processes, new companies have arisen that support the call centers by delivering software, hardware and consulting services. The developments that concern the hardware and the corresponding software such as routing devices and databases are of central importance to the developments of the industry. The popularity triggered researchers to work on many different subjects. The work of mathematicians on planning and job routing is the subject of this paper.

Contact centers

In this paper we focus on a special type of contact center, namely multi-skill contact centers. The difference between contact centers and call centers is that in contact centers also other channels than telephony are used, for example fax and email. We will explain the meaning of the term 'call center' first. Mehrotra (1997) defines a call center as "any group whose principal business is talking on the telephone to customers or prospects." The employees talking on the telephone are commonly called agents or telephone service representatives. Similarly, we define a contact center as any group whose principal business is communicating to customers or prospects. The term 'multi skill' will be explained in Section 2.

Service level

Many problems concerning routing and employee planning are restricted by service-level (SL) constraints. Different factors determine the service quality of a customer, e.g., whether or not the customer has been served, whether or not the customer was rejected, the voice of the agent, etc. We note that some of these factors are difficult to measure. In practice, call centers do not take all of them into account when optimizing and scheduling the operational processes of workforce management. Both the quality of the service itself and the waiting time distribution are considered as very important quantities. An example of a service level measure often used in call centers is the percentage of all callers served within twenty seconds waiting. Eighty percent is considered as an acceptable level, in the sense that the costs of agents and the service level are well-balanced. In addition, companies also like to monitor abandonments, rejections and blockings, but these are less important and often not used for planning and optimization. The different service level measures are related to each other. High service levels often imply low abandonment and blocking percentages.

Economies of scale

Much work in the literature is related to the size of a call center. The size is an important quantity, because of the relation with costs and productivity. In general costs decrease relative to the size and productivity increases when call centers grow. These economic advantages are the so-called economies of scale. Much literature about routing and planning pays attention to economies of scale. The papers from the literature often try to characterize and quantify the relation between on the one hand routing policies and planning methods and on the other hand the economies of scale. Hence, the economies of scale are important to the subject of this paper and deserve to be discussed in more detail. This will be done in Section 3.

Content

This paper is structured as follows. Section 2 elaborates on the historical changes in the functionality of contact centers and treats the various decisions that have to be made on different levels. Section 3 explains the potential benefits of enlarging call centers. Section 4 summarizes relevant issues concerning routing and staffing. In Section 5 we present a framework for the basic problem of job routing and employee planning, which is used to characterize the models from the literature. Section 6 elaborates on routing policies. A number of approximation techniques to measure the performance in the multi-skill environment is listed. Staffing, i.e., the planning of employees, is discussed in Section 7, where we especially focus on the multi-server queue and refer to papers for important structural results. Theory about queueing systems with homogeneous servers is also relevant with regard to multi-skill systems because it has been shown that under certain conditions the behaviour of homogenous servers has similarities with contact centers having fully cross-trained agents. In Section 8 we describe the literature about routing and staffing in the most elementary queueing systems, the so-called canonical designs. Section 9 discusses the most relevant papers in more depth, treating different optimization methods. Finally, in Section 10 we discuss directions for future research.

The scientific literature pays increasingly attention to call centers. In Koole and Mandelbaum (2002) queueing models that are relevant to call centers are discussed. We refer to Gans, Koole, and Mandelbaum (2003) for a complete overview of mathematics in call centers. It contains a tutorial on how call centers function, a survey of academic research and an outline of important problems that had not been addressed earlier.

2 Trends and diversity

There are many different types of contact centers. We partition the types by discussing the different ways in which calls can be initiated and we describe recent trends. In addition, the differences in routing and planning are explained.

2.1 Traffic types

In this section we make a classification into the way that calls are initiated, the so-called traffic types. It describes the trend from outbound call centers to inbound call centers, as occurred in the eighties and nineties. For each type we address the routing and staffing problems and explain the differences.

Outbound

In outbound call centers agents used to dial manually. In general, each call is served by only one agent, i.e., the agent that initiated the call. Thus, the assignment of work to agents is trivial and there is always a match between the workload and the capacity of the agents. Hence, routing is not relevant in this type of outbound call centers.

The planning of agents, in particular shift scheduling and rostering, is easy. The agents generate their own work such that a productivity of one hundred percent can easily be obtained, assuming that sufficiently many telephone numbers of potential customers are available. The problem that remains is the find a match between the availability of employees and the number of employees required to meet economical objectives, while taking the preferences of the employees into account. Thus, there is hardly or no stochastics involved and in many realistic cases straightforward deterministic techniques for workforce management will suffice.

Predictive dialing

Within the class of outbound call centers a further division is possible. If calls are initiated by the system, instead of agents, we speak of predictive dialing. Otherwise we speak of the classical type of outbound calling.

For example, predictive dialing occurs in a call center that contacts potential customers via telephone. A potential customer hears a recorded message and can decide to have contact with an agent. Thus, the searching process for new customers often occurs completely automatically. In this way, agents talk to interested customers out of the pool of potential customers only. Usually, only a small percentage of the potential customers is interested. We say that there is a probability of success involved with each call. This type of outbound traffic is called predictive dialing. Just like in inbound call centers, the epochs at which interested customers are found is a stochastic process that depends on the number of lines in use for searching. But these processes are not exactly the same. An essential difference with inbound call centers is that the number of lines in use for searching is controllable (remind that the calls are initiated by the system), by which it is possible to anticipate successful calls that occur in the near future. We can ask ourselves the same questions concerning inbound call centers: Idleness, over-staffing and waiting times play also an important role in this type of call centers. Note that the effectiveness of predictive dialing depends on the hour of the day that one calls.

In Samuelson (1999) a heuristic method is described (without giving all technical details) to determine the optimal number of lines to initiate. The model exploits general service time distributions because estimations of the remaining service times are used. With that data it is possible to make better decisions about the number of new lines to initiate. They claim that it performs amazingly well.

In call centers that use predictive dialing, agents usually have a single skill. For example, agents do one campaign after the next, and they are trained in advance. With a single skill, numerical methods for routing and staffing are easy to derive by standard techniques.

Inbound

In inbound call centers, a call is initiated by the customer. The call is for that reason called an inbound call.

Nowadays, most call centers treat inbound traffic. Inbound call centers support (existing) customers, e.g., customers having questions about new products, people having products that need repair and executing transactions about financial products. This type of call center has many different functions, e.g., supplying information and selling.

The time epochs at which an inbound call center receives calls are random in time. This complicates the staffing of agents. Typically, by staffing exactly the number of agents required to handle all work a high productivity is easily obtained. However, the service level will be low because waiting times will be too long. Hence, to meet service level requirement it is necessary to staff additional agents, called safety staffing.

The previous sections explained that routing and staffing in inbound call centers are difficult issues and, hence, are the main subjects of the remainder of this paper.

2.2 Trends

This section illustrates the growing diversity by treating: the mixture of jobs from different traffic types, the difference between cost and profit centers, and the growing importance of multi-skill call centers.

Call blending

Currently telephone is not the only channel for the communication between companies and and its customers any more. Nowadays, fax, and increasingly email, represent a substantial part of all work. Also other types of services have gained populairy. Especially the services by which customers are helped without human interaction, for example by means of video and recorded messages offered via internet, often called self-internet services. Therefore, we no longer speak of call centers, the term contact center is more appropriate. Because jobs from the different channels have different properties, e.g., requiring different qualities of service, the variety of mathematical models increased significantly in recent years. This paper treats routing in the broad sense; it not only focuses on telephone services but also on the other types of communication between call centers and customers.

We speak of call blending when agents work on inbound calls as well as on other activities such as outbound calls, emails, or faxes. Typically, the outbound calls, emails and faxes have lower priority than inbound calls. For that reason agents only start these blending activities at moments that the workload of the call center is low and agents are idle. To determine the optimal number of agents working on the blending activities it is important to anticipate calls that will arrive in the near future. Consider for example call blending of inbound and outbound traffic with inbound traffic having a higher priority than outbound traffic. Then, it might be attractive at busy moments to interrupt an outbound job to handle a high priority job. This would increase the service level of the high priority jobs. In practice, call centers avoid that inbound or outbound calls are interrupted too frequently. If agents have to switch between the jobs too often, the productivity and service levels of both types of calls might decrease by the switching times, see for example Koole (2005), Chapter 7.

Cost and profit centers

Most call centers are part of a large organization and their primary role is supporting customers in using their products or services. An example is the helpdesk of an internet provider. Because these centers only incur costs and do not generate income directly, they are so-called cost centers.

If rewards can be attributed to calls and if these rewards compensate the costs, then we speak of a profit center. An example is a call center that belongs to a sales department, because selling products or services are commercial activities. A second example is a call center that handles calls for other call centers that outsource work.

Multiple skills

In single-skill call centers there is no distinction between the handling of different call types. Agents are supposed to handle all types of calls.

We speak of multi-skill call centers in case of different job types and agents having multiple skills. An example of a multi-skill call center is an insurance company that receives calls with damage reports about cars, houses, boats, and so forth. It is imaginable that agents are trained to handle calls concerning only one of these subjects. These agents are called specialists. Agents that are specialized in multiple skills are called cross-trained agents. Finally, agents that can handle all call types are called generalists or fully cross-trained agents.

Implementing single or multiple skills in a call center is a tactical decision. A single skill call center can be inefficient if agents need much knowledge in order to handle calls. Two reasons are given. In the first place, the training of new agents can be time consuming because they need to know a lot before they start working. In the second place, it has been shown that service times increase and productivity decreases when agents handle calls with a high diversity of subjects. Hence, in case of long handling times it might be attractive to introduce multiple skills. Then, the required knowledge and the capabilities of agents are split up in several groups, called skills. The benefit is that agents do not need all skills to become operational. This reduces education and training costs. Whitt (1999) investigates the tradeoff between economies of scale associated with larger systems and the benefit of assigning customers with different service times to different agent groups.

Introducing multiple skills has some drawbacks. The costs of the technological infrastructure are high and it brings increasing complexity into the routing of jobs and planning of workforce.

2.3 Decision levels

We mention three important issues concerning job routing and workforce management: design, planning, and control. The differences between these three issues are explained and several examples are given. The examples are classified into operational, tactical, and strategic decisions, ordered from short term to long term. But not all combinations of issues and decision types are discussed. For example design involves mainly strategic decisions, instead of tactical and operational decisions.

Design

The design of call centers is concerned with structural long-term changes. It is by definition a long-term decision and it is mainly determined by strategic decisions. Examples of decisions about design are the choice or lay-out of the building, and the number of different skills that is distinguished among the agents. Long term decisions have impact on short-term types of decisions. These are being discussed next.

Planning

Planning is concerned with the scheduling of available resources in order to meet economical objectives. We distinguish four steps within the basic planning process of workforce management, namely: workload prediction, determining staffing levels, shift scheduling, and rostering. These are typically operational decisions. Workload prediction is concerned with the prediction of future workload. Staffing expresses the expected workload as numbers of required agents, which are the so-called staffing levels. Shift scheduling is the generation of shifts such that the staffing levels are met. We define a shift as a part of the day in which an agent can work. One shift can consist of multiple time intervals. Finally, rostering refers to the pairing of shifts into rosters and the assignment of the rosters to the employees. We define a roster as a set of shifts. When assigning rosters to employees their preferences and labor rules need to be taken into account.

In multi-skill call centers it is common that different agent groups are distinguished. The partitioning of agents into groups occurs usually in such a way that agents from the same group have (almost) the same sets of skills.

In multi-skill call centers the determination of staffing levels is more complicated and staffing levels require a more detailed description, as opposed to single-skill call centers. Because groups have different characteristics, the service level depends on the division of the agents over the groups. Hence, instead of a grand total it is beneficial to express the staffing levels per group, such that enough capacity is available for each call type.

Groups complicate the translation of predicted workload to staffing levels. If a skill occurs in the skill set of several groups, the activities of different groups become dependent.

Control

This section discusses the daily control of business processes. We define control as adjustments that are executed within short term and triggered by external factors. Control is related to WFM because it has the potential to improve service levels and reduce personnel costs. We remark that control involves operational, tactical, and strategic decisions.

On the operational level staffing deals with the re-scheduling of agents when the service level is low/high. A typical tactical decision involves acquiring new agents and determining their type of contracts. The training of the current agents for new skills lies in between and belongs therefore to the tactical or operational decisions. Decisions concerning shift scheduling and rostering are considered as operational decisions.

A subject that is closely related to staffing, is routing. An example of a control issue of routing is the optimization of policies during operations. Routing policies are described in Section 4.

3 Economies of scale

When call centers grow the productivity of manpower can increase without loss of service level. This is explained next. The arrival process of jobs is unpredictable. There is a continuous deviation between the expected and the actual number of arrivals during a short time interval. Let us assume that the length of this interval is of the same order of magnitude as the acceptable waiting time and consider such an interval. It is likely to happen that during this interval more work arrives than is expected on average, which can yield a low service level because customers have to wait. In small contact centers the error between predictions and realizations can be considerable because of the small numbers and the high variance of the total number of jobs that arrives during a short interval.

The variability in the number of arrivals is often expressed as the coefficient of variation (see Tijms (1986)). The number of arrivals X during the interval $(0, t)$ of a Poisson process with rate λ has a Poisson distribution with parameter λt. The coefficient of variation is given by $c_X = \frac{1}{\sqrt{\lambda t}}$, i.e., the standard deviation divided by the expectation. Note that c_X is large when λ is small, as is the case in small call centers. Whereas, in large call centers the relative deviation between actual realisations and predictions becomes neglectible, when considering intervals of the same length as in small call centers. Mathematically, c_X goes to zero while λ increases. Therefore, we conclude that the arrival process of the workload becomes better predictable because the randomness decreases relative to the expected number of arrivals. We remark that this can also be explained for a larger time-scale by using the law of large numbers.

Workload predictions are related to staffing levels. To meet service level conditions it is important that all jobs that arrive within a short time interval are served immediately. This requires that more agents are scheduled than the workload requires if all agents work continuously. The reason is that the exact arrival epochs of arrivals are unknown and the number of arrivals fluctuates over time even with a fixed arrival rate. Scheduling additional agents in order to meet service level constraints is called safety-staffing. Logically, by scheduling extra agents, agents are idle at time with fewer arrivals.

When call centers grow larger and larger, the randomness of X decreases relative to the expected number of arrivals and c_X goes to zero. Thus, the number of arrivals becomes better predictable. Eventually, to meet the service level conditions, the staffing levels can almost equal the expected workload, relative to the expected workload. The workload will eventually approach the maximum workload that the agents can handle, relatively, yielding a productivity of almost a hundred percent. This simplifies the determination of staffing levels.

In multi-skill call centers there is an additional factor that complicates staffing. Consider a small multi-skill call center with only specialists. Since there are no agents with multiple skills, jobs of different types are served by different groups of agents and there is no dependence between the number of busy agents in the different groups. The groups behave like separate call centers. Next, we schedule generalists instead of specialists. The benefit is that more agents are available to handle a certain job type. This minimizes the probability of queueing and maximizes the service level. (Assuming optimal routing of jobs and that all agents having the same skills scheduled, behave similarly, i.e., with the same service time distributions.) If the service level exceeds the lower bound some employees become redundant. By cancelling their shifts or rescheduling them on other tasks the service level remains acceptable and the productivity increases.

When the size of a multi-skill call center increases and a call center becomes large, it can be beneficial that agents specialize in specific tasks and do not combine different types of activities. The reason is that limiting the number of different tasks of an agent decreases the handling time of a job. In general, this is beneficial because shorter service times increase the productivity. Hence, in very large call centers it is almost optimal to use only specialists. Moreover, routing policies become less crucial because the gain of adding flexibility has become neglectibly small and therefore only specialists will be preferred.

However, in medium-sized call centers the use of only specialists or only generalists is inefficient. Consider a call center with three arrival streams and three skill types of calls: A, B, and C. There is one fully cross-trained agent and one specialist to serve type A calls. As we explained, by using specialists instead of generalists more waiting occurs in this example and the service level decreases. Hence, additional agents are required to meet the service level constraints. This decreases the overall productivity. By using also generalists, high productivity can be obtained without much loss in performance and without additional manpower, due to their flexibility. Hence, to maximize the success in medium-size contact centers it is important to find a balance between both: the number of specialized agents and the number of agents with multiple tasks.

Moreover, in medium-size contact centers with agents having different skills, the routing of jobs is very important. The reason is that the combination of the routing policy and the skills of agents can have a relatively large influence on the service level, which we show by means of an example:

Example: Consider a service center with 3 job types and 3 arrival processes, denoted by $M = 3$. Exactly one agent skill is associated to each job type. We assume that all service rates are exponentially distributed with rate 1, and the arrival rates are 3. Jobs that are served by an agent leave the system. There are 3 agent groups, group 1 with skill 3, group 2 with skills 2 and 3, and group 3 with skills 1 and 2. Calls are at arrival assigned to the agents according to an overflow policy, we refer to Section 6 for a description of this type of policy. The priorities of the agent groups are depicted graphically, see Figure 1. There are no waiting queues and calls that find all agents with the required skill busy are rejected. We define the service level as the percentage accepted customers. We compare two situations with different sizes of agent groups, s_i denotes the size of group i:

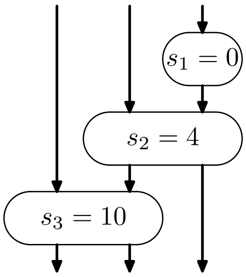

Figure 1: Example of a service center

1. $s_1 = 0$, $s_2 = 4$ and $s_3 = 10$ yields a service level of 84%, and
2. $s_1 = 3$, $s_2 = 1$ and $s_3 = 10$ yields 89%.

The only difference between situation 1 and 2 is that in situation 2 three agents are moved from group 2 to group 1, such that agents become dedicated to type 3. Thus, fewer jobs from class 3 are rejected and it appears that the service level increases substantially. The service level of jobs from class 1 and 2 will hardly decrease because sufficiently many agents are available in group 3.

Job routing in multi-skill contact centers is called skill-based routing (SBR). Moreover, salaries also play an important role in finding the right balance between the number of specialists and generalists. Adding skills to agents by training will increase the salaries in most cases.

4 Routing and staffing problems

This section treats the most relevant problems concerning routing and staffing. These are problems that much literature is written on. The relevant literature is discussed in later sections.

4.1 Routing

Routing policies specify the assignment of jobs to agents. With jobs we refer to calls, emails, and faxes. These policies are important because they have the potential to increase the efficiency of resource usage. Job routing policies usually consist of two parts: agent selection and job selection. Agent selection denotes the way that arriving jobs are assigned to the agents. This happens usually immediately after an arrival. Whereas, job selection denotes the selection of a job being assigned to an agent, either directed by the system or chosen by the agent. This often occurs immediately after an agent completes a job (assuming that there is a job present in the queues).

A special case of job routing is call routing. Call routing denotes the process between the arrival of a call and the assignment of the call to an agent. This is physically directed by dedicated hardware, called automatic call distributors (ACDs). Next to the hardware implementation, usually most call types also differ from faxes and emails because short waiting times are required.

The requirements on the waiting times makes routing (as is most frequently treated in the literature) important because of the high potential of waiting time reduction. Reconsider the example with skills A,B, and C. If a call of type A and B arrive it might be attractive to assign the type A-call to the specialist such that the other call is served by the generalist. Realize that if the type-A call would have been assigned to the generalist, it would not be possible to serve the B-type call immediately. This would decrease the service level. Besides, routing potentially has other benefits. In general, routing enables contact centers to reduce also the resolution times, which is the sum of the waiting time and the service time. This is because specialists often work faster than cross-trained agents. Then, maximizing the number of jobs handled by specialists decreases the average service time of a customer.

We remark that one should be careful when changing staffing levels because counter-intuitive effects are possible. For example see Section 3. It holds in general that when agents handle additional types of jobs the service level increases, assuming that the service time distributions are not affected by the additional job types. But for certain types of routing policies it is not difficult to compose examples such that the overall service level decreases by scheduling agents on additional job types. According to the literature routing issues are very complicated. Also the determination of staffing levels is difficult because it depends on many factors, including the routing policy.

4.2 Staffing

Routing in conjunction with staffing in multi-skill call centers is a difficult problem. The difficulty is easy to explain by means of an example. Reconsider the example with skills A,B, and C. There are two agent groups: group 1 with skills A and B and group 2 with skills B and C. The question is: What to do when a job of type B arrives? This depends on the number of available agents in each group and many other factors. This type of problem is computationally intractable in call centers with many agents. The same holds

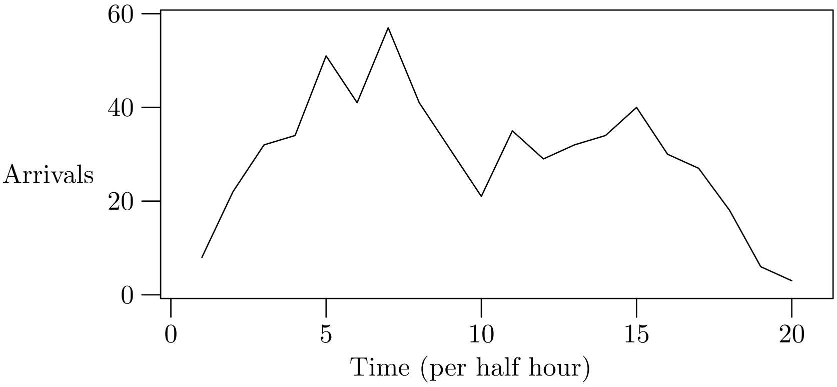

Figure 2: Daily pattern of work arrival

for more skill types, different service times and different priorities of calls. These factors make it even harder.

We elaborate on several reasons that make staffing a difficult problem. First, staffing is a complicated task because there are often multiple objectives involved, for example multiple service-level conditions. Second, staffing is complicated by the fact that the arrival process is stochastic, i.e., the exact arrival epoch of a job is unpredictable. Third, the total daily workload is hard to predict, since there are many external factors involved. For example, an insurance company for cars observes a correlation between the weather and the number of accidents. Good service levels require accurate forecasts of the workload and thus are very important. Fourth, the productivity of an agent is not constant during the day, and is hard to predict. It is influenced by different factors. An example is shrinkage, denoting unproductivity due to tasks other than serving customers, e.g., meetings and breaks. Holidays can also be considered as a shrinkage factor, but differ from the other mentioned factors in the sense that holidays are well predictable (and known in advance).

5 Model

In this section we present a general model of a multi-skill contact center. We aim to establish a model that is not too complicated to analyze and one that approaches reality sufficiently close to be useful in practice.

5.1 Arrival process

Several studies show that the arrival intensity of work in inbound contact centers depends on the hour of the day. A typical pattern of the number of arrivals against time is plotted in Figure 2. The busiest hours are in the morning from 10:00 till 12:00 and in the afternoon from 13:30 till 15:30. From 19:00 in the evening till 7:00 in the morning hardly any work arrives. Of course, most of the work that arrives at night occurs via email and fax. Calls

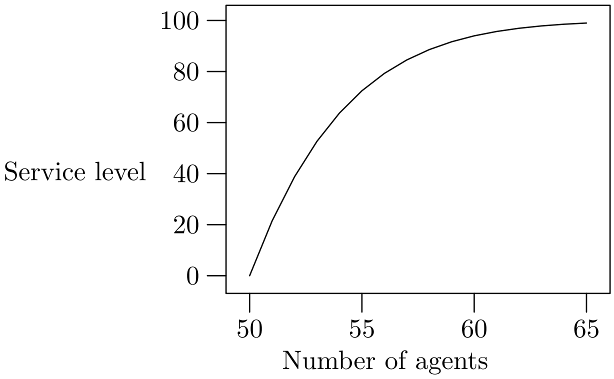

Figure 3: Concavity of a service-level measure

are rare at night because many call centers are closed during that hours, but differs per type of service and per country.

Call centers predict the workload in order to plan agents and meet service-level objectives. Case studies show that it is very difficult to do accurate predictions about future workload. Although the pattern with the two peaks from Figure 2 is roughly every day in a call center the same, there are still big differences between the processes during different days. Not only the total number of calls fluctuates strongly but the heights and the time epochs of the peaks also differ. As an illustation we note that contact centers are often clearly under- or over-staffed, yielding either very high or low service levels, respectively.

Predictions can be inaccurate in call centers because of randomness and unpredictable factors such as weather. Inaccurate predictions usually have much impact on service levels. For example, due to underestimations of workload service levels can be expected to be low during a day. To meet service level targets call centers have to compensate by providing service levels above the targets at other days. These compensations are undesired because of inefficiency and extra costs. We give two reasons to predict the future workload as accurate as possible, i.e., to minimize deviations between predictions and actual realisations of the workload. In the first place, long waiting times on days of under-estimations against short waiting times in case of over-estimations implies a high variance in waiting times over the different days. In the second place, on the long run costs increase because additional agents are required to meet service-level conditions. This is illustrated by means of an example. Assume that during 2 periods the workload requires 56 agents, in order to obtain a service level of 80 percent. We compare two scenario's by using the Erlang-C formula:

1. due to bad predictions someone schedules in period one and two 54 and 58 agents, respectively, yielding an average service level of 77%,and

2. 56 agents are scheduled in both periods by assuming that the predictions are accurate, which yields a service level of 80%.

The average service level is in scenario 1 on average below 80 percent, as opposed to system 2 with a service level of exactly 80 percent. Obtaining a service level of 80 percent in scenario 1 requires an additional agent during 1 of both periods, which increases costs. The reason is that the service level is a concave function of the staffing level, as depicted in Figure 3, see also Jagers and van Doorn (1986).

With regard to models, certain definitions have undesired mathematical properties. For example, an efficient staffing algorithm by local search exists if the service level is assumed to be concave with respect to the total number of agents, see Koole and Sluis (2003). However, with abandonments the concavity no longer holds. This kind of properties are undesired because it complicates the determination of staffing levels. The service level also needs to be easy calculated for planning purposes. Performance measures that are often used by the industry are calculated using simple queueing models. Especially, the standard Erlang-C model is frequently used to calculate measures on the waiting time distribution. In Koole (2003) the importance of the service level definition is discussed. It explains that an expectation of the waiting time distribution W_q and the acceptable waiting time (abbreviated by AWT):

$$E(W_q - \text{AWT})^+,$$

has several benefits, for example $E(W_q - \text{AWT})^+$ is convex in W_q.

For call centers that distinguish multiple job types predicting is even more difficult, as opposed to single-skill call centers. The reason is that multi-skill call centers not only require estimations over all job types, but also per job type. As explained in Section 3, smaller numbers have a relatively larger variance. This explains that predictions per job type are less accurate. Moreover, studies show that the daily pattern of different job types are sometimes different, which complicates the predictions even further.

The total number of job types is M and we define $\{1, 2, \ldots, M\} =: \mathcal{M}$. For the research that is described in this paper we define the arrival rate by

$$\lambda_m(t), \quad \text{the arrival rate of jobs of type } m \text{ at time } t,$$

in which we assume that jobs of type $m \in \mathcal{M}$ arrive according to a Poisson process with parameter $\lambda_m(t)$ at time t. This choice is often made in the literature because in practice many arrival processes are Poisson processes, and a Poisson process is memoryless, which the simplifies analysis.

A priority or weight is associated to the jobs from each class,

$$w_m, \quad \text{the weight of jobs of type } m.$$

We make no further assumptions about the usage of these priorities in models. Priorities are allowed to have different meanings. They can be implemented, for example, as holding costs of queues, or the reward of a service completion.

5.2 Agents and groups

The workload that is offered to the call center is handled by agents. We assume a one-to-one relation between job types and skills and an agent can have one or multiple skills.

In our model we consider agent groups. We define

$$N, \qquad \text{the total number of agent groups.}$$

and

$$S_g, \qquad \text{the skill set of group } g,$$

with $S_g \subset \mathcal{M}$ and $g \in \{1, \ldots, N\} =: \mathcal{G}$. The number of groups and the associated skills are fixed during the day. Agents work according to a non-preemptive service discipline, i.e., the services of jobs are not interrupted and the service proceeds until the job is completed. We define a collection of working hours $\mathcal{T} := \{1, 2, \ldots, T\}$, with T denoting the total number of consecutive periods, and

$$s_g^t, \qquad \text{the number of agents in group } g \text{ in period } t, t \in \mathcal{T}.$$

The service of a customer usually consists of two part: talk time and after-call work, also called wrap-up time. Talk time requires no further explanation. After-call work includes all activities related to the service of a call, like filling in a form, writing an email, and calling outbound.

The service time of an agent for a certain job is influenced by several factors. For the estimation of the service time, call centers can theoretically take as much information into account as is available, for example data about the historical service times of the customer, the subject of the call, characterists of the agent, etc. In our model service times are skill- and group-dependent, and exponentially distributed with

$$\mu_{mg}, \qquad \text{the service rate associated to skill } m \text{ and group } g.$$

We remark that for simplicity the working experience of the agents is not taken into account, but in reality agents often have a prefered skill.

An ambitious example of a call-center model is one that takes the properties and preferences of every agent into account, but in reality agents often have aprefered skill. Thus, every agent is considered to be unique. The call center may aim to optimize their economical objectives and meanwhile satisfy the personal requests of agents. Also this type of call center is a special case of our model; individual properties can be included by creating a different group for every agent.

5.3 Dynamics

Services of customers consist of several parts, for example dialing, waiting, and talking. We consider the service process of a call in a multi-skill call center as a dynamic process. It is in our opinion dynamic because some customers have to wait before they are served, while others are served immediately. The dynamics of the process is explained by means of a flow chart, see Figure 4 (we refer to Stolletz (2003) for a more extended version). The chart depicts the possible states of a call and the possible transitions between these states. For simplicity skills and job types are discarded in this figure. A description of

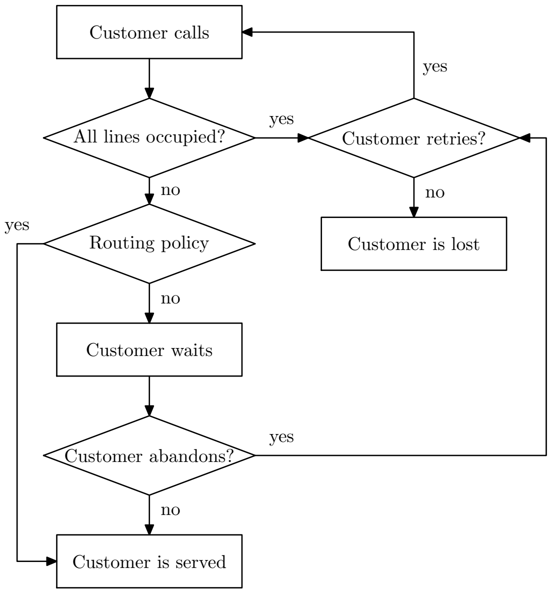

Figure 4: Flow chart of a customer

the process follows next. Customers that find all telephone lines busy are blocked. Calling customers that are not blocked enter the system and reveal the topic that they need to talk about. This is for example implemented by means of an automatic-voice-response system, a press-button-menu, or a department having agents dedicated for this purpose. When the topic of a call is determined, a new job is initiated. The topic determines the type of the job. In our model we assume a one-to-one relation between the topic of the job and the skill of an agent that is required to handle the job. Thus, from the moment that the job is initiated, it is known which agents are capable to serve the customer. When one of these agents is available it depends on the routing policy if the customer is immediately served by that agent. If the customer has to wait according to the routing policy, he/she ends up in a waiting queue. In our model we also assume that there exists a waiting room for every job type and

$$L_m, \quad \text{the number of waiting spots of queue } m.$$

Customers of a certain type are served according to a first-in-first-out service discipline, i.e., agents choose the job that arrived earliest. Waiting customers have a patience for waiting, which we model by a stochastic variable. If the waiting time exceeds the patience, the customer abandons the system by hanging up the phone. Abandoned customers may redial later when they expect shorter waiting times. We remark that due to redials the time epochs at which a customer calls are correlated to each other, which is in conjection with the assumption that calls arrive according to a Poisson process. Every customer who does not abandon eventually gets to talk to an agent and is served. The agent is determined by the routing policy. It can also occur that for some reason the customer cannot be helped and is directed to another agent. This can be an agent from the back-office, depending on the difficulty of the job.

The dynamics of the call center are modeled as follows. Jobs are served by exactly one agent. Thus, redirections of jobs, e.g., to the back-office, are not explicitly modeled. However, they can be modeled by adjusting the arrival rates. Moreover, routing policies consist of two parts in our model: call-to-agent and agent-to-call. Call-to-agent policies prescribe actions at the arrival epochs of jobs, and agent-to-call policies prescribe actions at service completions, see also Chapter 1.

In our model an email or a fax is only assigned to an agent if the agent is idle and no calls are available. The presence of emails and faxes improves the utilization of labor resources, because idle times are reduced. Idle times even disappear if sufficiently many emails and faxes are available. In our model we assume that faxes and emails have a lower priority than calls. Therefore, emails and faxes are served according to a preemptive service discipline, i.e., agents handling an email or fax are interrupted for calls. Due to the preemptive service discipline, the presence of emails and faxes has no influence on the service process of calls. This enables us to analyze calls without taking emails and faxes into consideration. However, we remark that preemption only approximates reality because sometimes agents complete the email or fax first, or start the service again after the interruption by a call.

Emails and faxes complicate the determination of staffing levels because also the workload of faxes and emails need to be taken into account. In case of preemption it is relatively simple because the staffing levels associated to calls can be calculated independent of faxes and emails, and the expected idle times can be allocated to faxes and emails.

A graphical representation of the topology of a multi-skill contact center is depicted in Figure 5, a similar (but less detailed) representation is given in Cezik & L'Ecuyer (Cezik and L'Ecuyer 2005). The arrows at the top represent the arrival processes of the different job types and each circle at the bottom denotes an agent group. The lines in the middle represents the waiting buffers and the policy that assigns the jobs to the agents. The arrows leaving the agent groups denote that a job is completed after the service by an agent and, hence, disappears from the system.

In our model we consider dynamic routing policies that take the queue length and the sizes of the agent groups into account.

5.4 Teams

In practice the way in which a call center is organized and agents are directed can influence decisions about the routing of jobs. This section describes a particular organisation form and the associated routing policies, as seen in practice. We discuss their impact on the service level, especially on the productivity of agents and the waiting times of customers.

A method that has the potential to increase productivity is introducing teams. In teams agents feel themselves more responsible for the offered work. It can avoid that agents slow down, since the whole team has to compensate for such behaviour. With respect to responsibility and competition between the agents, we can expect that teams

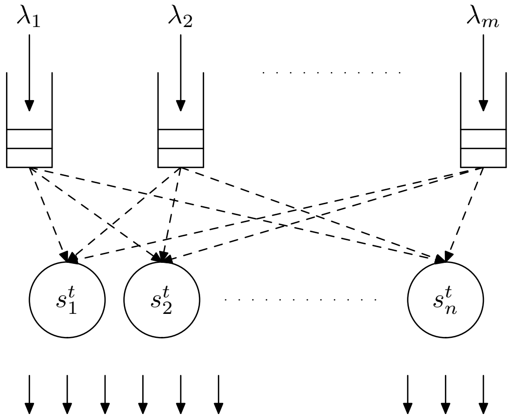

Figure 5: Model

are most effective if: (1) each team consists of a moderate number of agents, (2) each skill occurs in only one team, and (3) teams are composed of agents with almost the same set of skills. It can be beneficial that each team has a leader, which can strengthen the mentioned advantages. Productivity can also increase because sometimes a competitive atmosphere arises among the different teams. However, a big disadvantage of using teams is that economies of scale are (partly) lost. Because the dependencies between the teams is minimal, every team behaves as a separate call center. Hence, routing and planning become much easier. We remark that teams can easily be translated to our model; the team setting is incorporated by taking a separate agent group for every team.

5.5 Channels, service levels and flexibility

In practice there exist several definitions of service levels. For example, the percentage callers that is served within 20 seconds, the average waiting time, and the percentage that abandons. Service-level measures can differ per channel. Usually, the handling of emails and faxes is less urgent than calls. Waiting times of several hours or even one day are often allowed, while calls need to be answered immediately. Therefore, emails and faxes create flexibility in the number of agents available for telephone traffic and they offer an effective instrument to control the corresponding service level. Notice that we can handle an infinite backlog of emails and faxes in the model, by adding a queue with infinitely many jobs. The objective can be to maximize throughput while meeting service level constraints on calls. We will discuss in Section 8 a paper that considers this particular case. If agents are idle they have the possibility to handle these jobs. Of course, the arrival process of emails and faxes can also be modeled as, for example, a Poisson process, such that the queueing process behaves similarly to calls.

5.6 Planning of resources

Call centers aim to schedule the employees in such a way that objectives are met. A common objective is to satisfy a constraint on a service level measure and to minimize personnel costs. In this section we firstlist a number of properties that make planning a difficult problem. Next, we describe the features of the model with respect to planning.

The relation between the staffing levels and the total productivity of a call center fluctuates in practice. Often there is a deviation between the number of agents that is available according to the system and the number of scheduled agents. The reason is that agents are not fully productive during the whole day. Agents usually have several interruptions, for example, breaks and meetings. We learned from practice that the impact of these shrinkage factors can be huge; to meet service-level agreements, interruptions like breaks and meetings need definitely to be taken into account.

The shrinkage factors must be carefully analyzed when building and validating models. It is easy to see that the shrinkage percentage varies during the day. For example, if people take fewer breaks during busy hours, the shrinkage is lower. However, the impact of a change in shrinkage can be high during periods in which much work arrives; a small decrease in productivity can easily result in much lower service levels.

Obtaining good schedules is a difficult problem in multi-skill contact centers. A reason is that the skill set of each individual agent has influence on the performance and needs to be taken into account while measuring performance, as shown by means of an example in Section 3. Moreover, bad workload predictions easily result in a mismatch between on the one hand the total working capacity available for a job type and on the other hand the arriving workload of that type, yielding lower overall service levels. Thus, inaccurate predictions also contribute to the high complexity.

Also labor rules play a role in the planning of resources. It determines for example the maximum length of a shift and the minimum time between two consecutive shifts.

In our model, a shift is defined by a collection of working hours from $\mathcal{T}$ and a subset of skills from $\mathcal{M}$. We assume that for each shift there is a group of agents that has the skills to work that shift. Hence, for notational convenience we can denote the skill set of each shift with f_k, i.e., the index of the corresponding agent group for shift k, $k \in \{1, \ldots, K\}$, with K denoting the total number of shifts. In order to meet the service level constraints we suppose that for every agent group there is a set of workable shifts such that for some agent configuration the requirements are met. In this context a shift k is workable if there is a group g such that $f_k = g$.

A lower bound for the number of employees is the workload. The workload can easily be expressed in agent numbers if the service rates are class-dependent (and not group-dependent) $\rho_m = \frac{\lambda_m}{\mu_m}$ with μ_m the service rate of type m and λ_m the arrival rate of type m. But a drawback is that the workload of a type is not integral. For example, a workload of 0.75 for 4 skill types requires 4 specialists or 4 × 0.75 = 3 generalists.

In case of group- and skill-dependent rates a simple lower bound would be $\sum_{m=1}^{M} \rho_m = \frac{\lambda_m}{\max \mu_{mg}}$. However, these lower bounds give no guarantee with respect to stability of the queue lengths. The error due to integrality can be even larger than in case of a single skill.

To obtain an accurate lower bound one needs to take the staffing levels of the groups into account. Almost similar to Harrison and Zeevi (2004), we formulate

$$\begin{gathered}\min \sum y_g \\ \text{subject to} \\ Ax - Ey = 0 \\ (Rx)_m \geq \lambda_m, \quad \forall m \in \mathcal{M} \\ y \geq 0 \text{ and integer,}\end{gathered}$$

with E denoting the diagonal-identity matrix, y the vector with the staffing levels of each group, x_{mg} representing an activity, denoting the number of agents from group g that work on jobs from class $m \in S_g$. A column of matrix A takes the value 1 in row g if the activity associated with the column belongs to group g, otherwise it is 0. A column of matrix R takes the value μ_{mg} in row m if the activity associated with the column is to serve jobs of type m, otherwise it is 0. By adding

$$y_g \leq q_g^+ \text{ and } y_g \geq q_g^-, \quad \forall g \in \mathcal{G},$$

with q_g^- and q_g^+ the minimum and maximum size of group g, respectively, it is possible to control the sizes of the agent groups.

Moreover, it can be interesting to calculate minimal staffing levels that minimize costs. To achieve this, we adjust the objective function of the integer programming model by multiplying the staffing levels by costs, yielding $\sum c_g y_g$, with

c_g, the costs of staffing an agent in group g during one unit of time.

Armony and Bambos (2003) is also related because it treats the maximization of throughput subject to stability conditions.

5.7 Other models

We discuss several aspects that complicate the mathematical analysis of call centers, while they also show the diversity of call centers. We decided to exclude these aspects from the model for simplicity and because they are considered in only a few papers from the literature. The different aspects are treated by summarizing the papers from which the findings originate.

Brown, Gans, Mandelbaum, Sakov, Shen, and Zhao (2002) characterize the arrival process at an inbound call center as a nonstationary Poisson process. We remark that even this is slightly inaccurate and does not give a complete description because for example redials (possibly caused by abandonments) are not modeled. We refer to Aguir, Aksin, Karaesmen, and Dallery (2004) for a model with redials.

Inaccuracy between models and reality is often present in service time distributions. The literature most often assumes that they are exponentially distributed variables, while Bolotin (1994) and Mandelbaum, Sakov, and Zeltyn (2000) show lognormality.

Steckley, Henderson, and Mehrotra (2004) explain and analyze the stochastic relation between the prediction and the realization of the arrival process, for different time horizons. They relate the decisions about staffing levels to the real workload, by means of service level measures.

Whitt (2004a) and Sisselman and Whitt (2004) try to improve the routing such that the satisfaction of the customers and agents increases by paying attention to the working circumstances of the agents. They write: "skill-based routing is designed to ensure that calls are not only handled promptly but also resolved properly." Their solution is an extension to the priority routing from Wallace and Whitt (2004) because also the preferences of agents, with respect to their skills and schedule, are taken into account.

Gans and Zhou (2004) consider an outsourcing model with high and low priority customers. High priority customers should get a high service level, while the throughput of low priority customers is maximized. The decision is to accept or reject low priority customers. The rejected customers are outsourced.

5.8 Models from the literature

While studying the literature we observed that routing and staffing are most often analyzed separately. This section addresses the main characteristics of the corresponding models from the literature.

The literature treats at least three different settings of job routing, namely:

- call routing in multi-skill inbound call centers,
- call blending with faxes and emails in contact centers, and
- call routing in single-skill call centers with calls having different priorities.

We note that from a mathematical point of view the different routing problems have much in common. Many models are a special case of a multi-skill inbound call center with different priorities associated to each skill. This explains the importance of research on routing in multi-skill call centers and that results can be useful for solving different problems.

Models that include routing and staffing in multi-skill call centers have in common that agents are grouped. Agents from the same group are assumed to behave statistically identically. We note that, in conjunction with agent groups, the assignment of jobs to employees in multi-skill call centers often occurs according to priority routing policies, see Section 6.3.

Call center models are based on many assumption and simplifications. For example abandonments are often discarded. Also queues are sometimes omitted, either when the abandonment rate is assumed to be very high or when the number of lines is assumed to be very limited. The benefit of such simplifications is to make a tractable analysis possible. However, even with such simplifications, it remains very difficult to obtain results for multi-skill call centers, both analytically and numerically. Only small call centers are tractable.

6 Routing policies

This section treats some important subjects related to routing policies. In the remaining sections a number of papers from the literature is discussed. In case of differences with the model from Section 5, a short discussion is added.

6.1 Push and pull systems

There are several ways to implement call assignment policies. We make a distinction between push and pull systems.

- In a push system an arriving call is assigned to an available agent that is chosen by the computer system. That agent is responsible for the service of the customer.
- In pull systems different agents are simultaneously notified about calls in the queue. The agent that reacts fastest handles the call.

Pull systems give the agent more flexibility. But on the other hand, from the perspective of an agent, it might be attractive to ignore the call, and for example have a break instead. We note that in the literature it is often assumed that optimal control of the manager involves push systems.

6.2 Standard solution approach

A straightforward approach to analyze call centers is to apply the theory of Markov decision processes (see Puterman (1994) for details). When modeling a call center as a Markov decision process it is theoretically possible to take all relevant information (that is available at the decision epoch) into account, as long as the state is memoryless, for example the hour of the day, the queue lengths and the productivity of each individual agent. The information required to describe future events of the process is called the state. With respect to analysis it is important to keep the state as small as possible. The reason is that standard analysis techniques calculate a function for all possible states of the system, i.e., the state space. This can be a huge number of calculations. The fact that the number of states is exponential in the dimension of the problem is called the curse of dimensionality, see Bellman (1961).

In a Markov decision process the system moves from one state to another. This is called a transition. In most models from the literature the state changes when an arrival occurs or when the service of a job finishes. These are the two common epochs at which it is decided if an agent starts serving a call. At arrival, calls are either queued or immediately assigned to an agent, depending on the routing policy. If multiple agents are available the policy determines the agent that serves the job. Hence, the policies used at arrival epochs are also called agent-selection policies. At service completions agents become available again. A call-assignment policy prescribes if an agent becomes idle or starts to service another job. Relevant methods are discussed in Section 9, see the section about call selection.

6.3 Types of policies

Different types of routing policies exist. Dynamic routing is discussed first, hierarchial routing is explained next, and finally overflow routing or priority routing is treated.

Dynamic routing

Dynamic routing policies are policies that take the state of the system into account.

Optimal dynamic routing policies can be calculated by using methods from dynamic programming, which is part of the theory on Markov decision processes. We make a distinction between two types of applications of dynamic programming. In the first place, it is used in the literature for the derivation of structural results. Three examples are given:

- Hordijk and Koole (1993) study scheduling problems of multiclass customers on identical processors. Job arrive according to a controllable Markov process. As a special case they show the optimality of the μc-rule in the last node of a controlled tandem network.
- The two papers that we mentioned about call blending, Bhulai and Koole (2003) and Gans and Zhou (2003), are worth mentioning again, see Section 2.
- Carr and Duenyas (2000) analyze job admission to a single server queue with two job types, see Section 8.

In the second place, dynamic programming is useful to analyze models numerically. An example of an application is the calculation of the steady state distribution of the Erlang-C model. However, optimal routing policies are difficult to obtain for multi-skill call centers. The associated problems are hard to solve due to the huge number of different states. Several experiments show that only single-skill call centers and small multi-skill call centers are tractable. Notice that these models are already simplified as much as possible, such that the state contains a minimal amount of information. This is, for example, achieved by grouping agents together by assuming that agents with the same set of skills behave statistically the same. This reduces the size of the state space because no distinction between agents from the same group is required. In the literature a number of other simplifications is usually made, e.g., exponential service time distributions, a stationary arrival process, see for example Bhulai and Koole (2003) and Koole and Pot (2005).

Hierarchical routing

We characterize hierarchical routing as follows (see Franx, Koole, and Pot (2005)). Assume that jobs from different types arrive according to independent arrival processes and agent groups consist of agents with equal skill sets. An overflow routing policy, and in particular an agent selection policy, specifies for each call type an ordering of the agent groups. The ordering is chosen in such a way that jobs are assigned to the first available agent, while traversing the different groups according to the ordering. Thus, the group with the highest priority is considered first and if an agent is available the job is assigned to the agent, otherwise the group with the second-highest priority is considered etc. It is possible to visualize this in a network, with each node representing an agent group and the arcs denoting the priority ordering for each call type. A job is assigned to the first available agent in the network and is blocked if the agents along the path are all busy. Thus, they flow along the busy agents groups, explaining the term 'overflow routing'. Jobs leave the network immediately after service completion.

Next to agent-selection policies, call-selection policies can be specified in a similar way. In that case, priorities are used to denote for each group the type of job that is served at a service completion. If a job from the class with the highest priority is available, it is served, otherwise the queue that contains the jobs with the second-highest priority is considered, etc.

We discuss agent-selection policies in further detail. An hierarchical agent-selection policy is defined by an $N \times M$ matrix, which we denote by π. Each row specifies an agent group and each column a job type or skill. Column m specifies the priority of each group $g \in \mathcal{G}$, associated to skill or job type m. Value 1 indicates the highest priority and value k indicates the group with the k-highest priority. Value 0 indicates that the group does not have skill m. With hierarchical overflow routing there exists an ordering of the skill types such that $r_{rj} > r_{sj}$ and $r > s$. For example,

$$\pi = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 2 & 1 \\ 2 & 3 & 2 \end{bmatrix}. \tag{1}$$

Type 1 is handled by groups 1 and 3, and group 1 has a higher priority than group 3. Type 2 is served by all three groups and the priority increases with the group index. Finally, type 3 occurs in groups 2 and 3, and group 2 has a higher priority than group 3.

Routing policies and staffing levels are related to each other. With overflow routing it is possible that the service level increases when an agent with two skills is replaced by an agent with only one (changing the staffing levels of two groups), see Section 3 for an example, while with dynamic routing it is easy to show that the service level would definitely decrease.

With respect to storage space, overflow routing is one of the simplest kinds of routing policies to describe. Hence, the configuration of hardware is easy and the policy requires little memory. The required amount of stored data is moderate, because it is not necessary to store actions for each state of the system.

Section 6.4 gives a literature overview of approximation methods from the literature for blocking models with overflow routing.

Priority or overflow routing

Priority routing or overflow routing is an extension of hierarchical routing (see for example Wallace and Whitt (2004)). The priorities of the groups for each type can be described by a matrix, with a column representing a job type and the m-th column containing the priorities of the groups for the particular type, see for example (1). The difference with hierarchical routing is that no ordering of group priorities is required among the different job types.

Sisselman and Whitt (2004) consider priority-routing policies. They assume that there is a sufficient number of agents (with the right skills) available to satisfy the service level conditions. They develop an algorithm for setting up routing policies that support skill preferences of agents, for example, by giving sufficient idle time to agents. This is achieved by including different control parameters in the model. An advantage of the model is on one hand satisfaction of both customers and agents and on the other hand high flexibility by means of the control parameters. However, the control parameters make the implementation time-consuming and, additionally, the control parameters might also influence the computation times. A limitation is that the framework is not capable of identifying an optimal routing algorithm, as they mention in their paper. Service levels can only be measured by simulation.

6.4 Approximation techniques

Due to the high mathematical complexity of multi-skill call centers literature exists about approximation methods. The papers are discussed next.

Conditioning

In Shumsky (2003) the performance of a call center with two types of customers is approximated. The state space is partitioned into regions and the conditional system performance within each region is approximated. This yields a substantial reduction in computation times and in most cases errors smaller than 10 percent and on average less than 5 percent.

Overflow routing in blocking models

We refer to Section 6 for a description of overflow routing. Remember that with overflow routing, the agent groups are considered one after another according to the routing priorities. If all agents in a group are busy the job ‘flows over’ to the next group. In this section we consider blocking systems. If there is on arrival no agent available with the right skill, the call is blocked.

In blocking systems it is straightforward to express the service level as a function of the blocking probability. A great advantage of overflow routing in blocking systems is that fast approximation methods for calculating the blocking probabilities have been developed. These are discussed next.

The Equivalent Random Method is a well-known method in the area of computer networks and tele-communication. A description is given in Cooper (1981), pages 165–171. The method is derived from Kosten (1973), in which a formula is presented for the peakedness of the overflow process of an $M/M/s/s$ system. This formula has been developed by Wilkinson (1956) and Bretschneider (1956). Several other people contributed to the usefulness of the method by developing numerical approximation methods for the inverse of this formula. The method is generalized to call centers in Tabordon (2002).

Hayward and Fredericks together extended the work that was presented in Kosten (1973), see Fredericks (1980) and Wolff (1989) (pages 354–355), and developed the Hayward-Fredericks method. Their contribution was an approximation method for the overflow process of a $G/M/s/s$ system, where the arrival process is determined by the first moment and the peakedness factor. The resulting method decomposes a multi-skill blocking system with overflow routing into separate multi-server groups. The method is extended to call centers in Chevalier and Tabordon (2003).

The Interrupted-Poisson-Process Method is developed by Van Muylder. A description can be found in Tabordon (2002). The fundamental idea is that the superposition of different overflow streams is assumed to be a renewal process. Based on this assumption, an approximation is found for the Laplace transform of the inter-overflow times.

The Poisson method is developed by Koole and Talim, see Koole and Talim (2000). The overflow processes are approximated by Poisson processes. As a result, the superposition of overflow streams are approximated by the same type of process. This method provides an upper bound on the service level because the burstiness of the streams is underestimated. The impact of underestimating the burstiness is significantly present in systems with a low workload.

The Hyperexponential Decomposition method is described in Franx, Koole, and Pot (2005). It decomposes the routing network into separate agent groups. The overflow processes are approximated by processes with 2nd-order hyperexponentially distributed inter-arrival times. An exact analysis is applied to each group. The authors report that this method yields the most accurate results compared to the other methods, but it is also the most time consuming to implement.

We showed in this section that much literature exists about overflow routing in blocking systems. With regard to call centers, we remark that it is important to investigate if the resulting models and methods from blocking systems are useful in practice. The assumptions that are made implicitly, such as ignoring queues and abandonments, and assuming stationary arrival processes, can give inaccurate descriptions of reality. It is well known that (1) blocking and delay probabilities are different, (2) arrival processes fluctuate during the day and (3) most call centers operate in an overloaded situation, making abandonments very important to include in the models. Research about the impact of these characteristics on approximations are lacking. In our opinion, much work remains to be done about model validation in the future.

Approximate dynamic programming

Koole and Pot (2005) apply approximate dynamic programming, i.e., the value function is approximated by fitting a simple structure into the dynamic programming equations. In Bhulai (2005) a dynamic policy is derived by using the value function of a simpler queueing system. We refer to Section 9.3 for more details about the experiments of both papers.

7 Staffing and limiting regimes

This section discusses the literature about staffing and results obtained by considering limiting regimes. We note that most of the papers from the literature using limiting regimes are concerned with homogeneous single-station systems. The papers relevant to multi-skill and multi-station models are discussed in Section 9.

7.1 Classifications

Staffing is a broad subject and therefore several classifications are possible. We describe two classifications, related to: the planning horizon and the level of staffing.

Time horizon

With respect to the time horizon of decisions a proper classification of planning is: strategic, tactical and operational. These are discussed next. An example of a strategic problem is starting an additional call center in a different country to decrease personnel costs. An example of a tactical problem is the acceptance of new employees, the choice about their skills and also the career path of the current employees, including training, see Gans and Zhou (2002). An operational problem is, for example, the scheduling of employees for the next period (in terms of weeks or months). Examples of operational problems of very short-term are the determination of the kind of jobs that the employees work on, their skill set, the routing policy of the different activities to the employees and the use of agents with flexible contracts. Most of the references from this paper are related to this.

Level of staffing

A second classification is the organization level of expressing staffing levels:

- agent groups (with each agent having the same set of skills),
- teams that contain agents with different skill sets, and
- the call center as a whole.

The same classification is possible with respect to leadership levels. The names of the corresponding managers are usually called group, team and call-center manager, respectively.

In software packages the procedures for determining the total staffing level are often based on elementary queueing models, see for example Cooper (1981). These ignore skill-based routing because all agents are assumed to be equal. All agents have the same skills and the same service time distributions. Our experience is that for more detailed studies simulation is frequently used in industry.

Optimal staffing is discussed in Akşin and Harker (2003). They consider a model with parallel independent multi-server queues and one shared information system. All servers that handle calls need to communicate with this system. This occurs simultaneously and is modeled as a processor-sharing service discipline. They obtain product form solutions and expressions for performance measures from their analysis. The model does not fit in our framework from Section 5, because of the shared information system.

7.2 Limiting regimes

An important finding with respect to staffing is the square-root-safety-staffing rule, see Borst, Mandelbaum, and Reiman (2000). This rule relates changes in the offered workload to the required number of agents s such that the service level remains equal. This rule is given by

$$s = a + \beta\sqrt{a},$$

where $a = \frac{\lambda}{\mu}$ is the offered load (λ=arrival rate,μ=service rate) and β represents the service grade. The square-root safety staffing rule has a wide range of applications. Illustrations are given in the papers mentioned below.

The Halfin-Whitt regime or Quality and Efficiency Driven (QED) regime, introduced in Halfin and Whitt (1981), is an example of a limiting regime. Under this regime the number of servers goes to infinity and the arrival rate is appropriately scaled such that the service-level remains constant (in the limit). This regime is based on the square-root safety staffing rule. In Whitt (1992) a multi-server queue is analyzed. On one hand the relation between workload and server utilization is considered and on the other hand the square-root safety staffing principle is discussed. The second one is extended to general arrival and service time distributions. Moreover, several implications are listed. Borst, Mandelbaum, and Reiman (2000) present an overview of the different limiting regimes and relate them to each other. Also the square-root safety staffing principle is revisited and extended with costs for staffing and delay.

A single skill

In Whitt (2004b) a multi-server station is considered and approximated by analyzing the corresponding fluid model. The model and the analysis support general distributions for the arrival process, the service times and the abandonment times. The main objective is to determine the optimal number of agents that maximizes a profit function built up out of revenue for throughput and costs concerning the number of servers in use, abandonments and waiting time. Under certain conditions, nondecreasing and convex cost functions and nondecreasing concave revenue functions, a solution is found. The effectiveness is extensively validated for different inter-arrival time distributions.

Multiple skills

Reiman (2000) is relevant to the multi-skill setting, in which diffusion limits are considered.

In Harrison and Zeevi (2004) and Bassamboo, Harrison, and Zeevi (2004) the arrival is scaled up by a super-linear function and the service and abandonment rates grow linearly. These papers are discussed in Section 9.

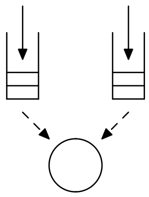

Figure 6: V-design

The conventional heavy-traffic limit theory is applied to a call center in Harrison and Lopez (1999). It is the first paper in which a dynamic control problem is explicitly solved in the multi-pool multi-class setting using conventional heavy-traffic limit theory. The routing policies are modeled by assuming that the system is in a heavy traffic situation. In this regime the arrival rates and services rate are accelerated linearly such that the system utilization approaches one. This requires only the first moments of the arrival and service times. The time is rescaled such that they obtain the formal heavy traffic limit. Using the central limit theorem a Brownian motion is recognized and a Brownian control problem is formulated. From this asymptotically optimal policies are derived. It is explained that these policies can be useful in controlling the original system. The policies suggest to hold backlog in one particular job class, and to utilize the servers fully unless the total backlog is small.

8 Canonical designs

Insight is often obtained from simple examples. In the context of contact centers these are called the canonical designs and will be discussed next. They are useful to get a better understanding of routing.

V-design

With V-design one refers to a model with 2 job classes and one group of homogeneous servers, as drawn in Figure 6. Perry and Nilsson (1992) consider a system with two classes of calls that are served by a single pool of agents. They determine the required number of agents and an assignment policy to satisfy a target on the expected waiting times.

We discuss two papers on call blending. The models are the same and treat the case with two type of jobs: inbound calls and outbound jobs. The inbound calls arrive according to a Poisson process. The number of outbound jobs is unlimited and they represent the backlog. (Unserved calls wait in a buffer until they go into service.) The objective is to maximize throughput of outbound calls. There is a constraint on the average waiting time of inbound jobs. The question is how to schedule an available agent. The choices are: keep the agent idle, assign an inbound job, or assign an outbound job.

In Bhulai and Koole (2003) the case with equal service rates for the inbound and blending work is solved to optimality, and the structure of the optimal policy is characterized. The starting point of their analysis is the value function, well-known from dynamic programming. In contrast, Gans and Zhou (2003) analyze the model using a linear programming formulation. In addition to the authors mentioned first, they obtain results for the case with unequal service rates. The optimal policy is a threshold-reservation policy.

Armony and Maglaras (2004b) and Armony and Maglaras (2004a) analyze a multi-server call center with two types of customers. At arrival, customers get information about their expected waiting time and can choose between hanging up, waiting until a server is available (the first type) and a call-back service (the second type). Customers get a guarantee on the maximum delay before receiving a reply. They propose a nearly optimal policy that serves customers from the second type of the queue length exceeds a threshold. The paper also gives approximation of the performance for different measures. It shows that simple routing schemes can be very effective and can lead to models that are easy to solve. The call-back service is not included in the model from Section 5 because, as far as we know, call-back options are not a standard feature of multi-skill contact centers and analyzing the model is already complicated enough.

Brandt and Brandt (1999) is to some extent related to call blending. It assumes two job types with equal service rates. High priority jobs are the live customers. If they decide to leave the queue, due to impatience, a low priority job is created that represents the call-back message. This paper develops performance measures for a fixed threshold policy. (Call backs do not exist in the model from Section 5.)

Van Mieghem (1995) proves asymptotic optimality of a simple generalized $c\mu$ rule with waiting costs that are convex increasing.

The literature on telecommunication contributes substantially to our knowledge of contact centers. Blanc, Waal, Nain, and Towsley (1992) consider a multi-server queue with two job types, different rewards per type and a first-in-first-out service discipline. The system is controlled by the admission of the job type with the lowest reward. The objective is to maximize the total discounted rewards. The optimal policy admits the jobs with the lowest reward only if the total number of jobs is below a fixed limit, the threshold. Admission is not considered in Section 5. Call centers usually serve all customers.

Guérin (1998) presents a model without waiting queues. It contains a multi-server station, which receives low- and high-priority jobs. He develops an admission policy for the low-priority jobs such that the fraction of blocked high-priority jobs is bounded and they analyze the system under that policy.

In the context of retail Berman and Larson (2000) consider a two type system. High-priority customers arrive to cash registers according to a stochastic process, and generate workload. Meanwile there is a quantity of low priority work to be handled by the servers, such as restocking and maintenance activities. A switching cost is included for each time that a server changes from high priority jobs to the low priority work. The paper describes two heuristics to control the servers. We remark that switching costs and times are also realistic in call centers because productivity decreases when cross-trained agents switch between jobs of different types. But it is not included in the model from Section 5 to keep it simple. Moreover, switching costs can be modeled implicitly by decreasing the service

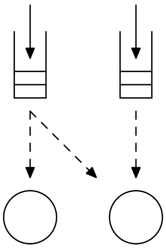

Figure 7: N-design

rates.

Schaack and Larson (1986) use generating functions to characterize the performance of threshold reservation policies for systems in which $M \geq 2$ classes of customers, all with the same exponential service time distribution. The paper concentrates on performance analysis and does not include a notion of optimality as it analyzes policies. It does not explicitly address service-level constraints.

Carr and Duenyas (2000) consider a single server queue with two job types and different service standards. They discuss a job admission and sequencing problem.

Örmeci, Burnetas, and Emmons (2002) consider dynamic admission control in a loss queueing system with two classes of jobs with different service rates and random revenues.

Peköz (2002) considers a multi-server nonpreemptive queue with high and low priority customers. The decision maker decides when waiting customers enter service. The goal is to minimize the mean waiting time for high-priority customers while keeping the queue stable. An asymptotically optimal policy is derived using a linear programming approach.

N-design

Stanford and Grassman (1998) consider a model with two skill types and two agent groups: one group of specialists and one group of generalists, as depicted in Figure 7. Fixed priority policies are used. Shumsky (2003) proposes an approximate analysis, see Section 6.4. Bell and Williams (2001) prove the asymptotic optimality of threshold controls in the conventional heavy-traffic limit.

M-design

The M-design, see Figure 8, is a model with two groups of specialists for both call types and one group of generalists that has both skills. There are no waiting rooms ($L_m = 0$). This model is analyzed in Örmeci (2003), under the assumption that the specialists work faster than the generalist. The paper shows that it is optimal to give the specialists a higher priority than the generalists. Concerning the generalists a sufficient condition is derived under which it is optimal to accept jobs if no specialists are available and a generalist is.

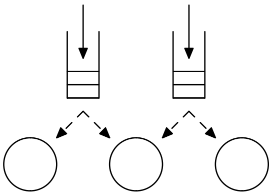

Figure 8: M-design

Others

The remaining designs are not relevant or are discussed elsewhere: The I-design represents a single-skill call center and is for that reason relevant to staffing, see Section 7 for a literature overview. The Λ-design classifies single-skill multi-server queues. If the servers are not homogeneous this design is concerned with agent selection policies, see Section 9. We consider the W-design in combination with the M-design as a multi-skill multi-server system, as described in Section 5. It is the general setting of multi-skill call centers. In Section 9 relevant literature is discussed.

9 Optimization

This section is devoted to optimization algorithms for multi-skill contact centers. We elaborate on a number of papers that contribute substantially to our understanding of multi-skill contact centers and especially of routing. The papers help to answer the following questions: How to improve routing policies and plan agents, and, given a policy, how to determine the service level?

9.1 Shift scheduling

In this section we address the most important papers about shift scheduling in single-skill call centers and we describe two papers about shift scheduling in multi-skill call centers. The different models and methods have in common that a period of one working day is considered.

Single-skill

With respect to single-skill call centers we mention three papers that are in our opinion most interesting. The most elementary model originates from Dantzig (1954)

$$\begin{gathered}\min \sum_{k\in\mathcal{K}} c_k x_k \\ \text{subject to} \\ \sum_{k\in\mathcal{K}} a_{k,t} x_k \geq s_t, \quad t \in \mathcal{T} \\ x_k \geq 0 \text{ and integer}, \quad k \in \mathcal{K}\end{gathered}$$

with c_k the price of shift k, s_t the staffing level in period t, and $a_{k,t}$ denotes 1 if shift k is active during interval t and otherwise 0. The decision variable x_k indicates the number of times that shift type k occurs.

Thompson (1997) considers two types of service levels: aggregate threshold service levels and minimum acceptable service levels. A method is introduced that takes both types of service-level constraints into account. The method integrates the determination of staffing levels and shift scheduling, and the problem is solved by using linear programming.

Atlason, Epelman, and Henderson (2002) and Atlason, Epelman, and Henderson (2004) consider optimal shift scheduling in single-skill call centers under the same types of service level constraints. They propose a general methodology based on the cutting plane method of Kelly Jr. (1960) that integrates the determination of on one hand staffing levels and the other hand shifts. They assume that the service level cannot be easily computed and instead, is evaluated using simulation. The problem is formulated as an integer programming problem in conjunction with the service level constraints, which are included via non-linear constraints. They relax the problem from the non-linear constraints and use standard methods to solve the remaining set-partitioning problem. In case service-level constraints are violated the cutting-plane technique is applied. First the violated constraints are detected and next cuts are added such that

- the current solution is removed from the feasible region and
- no optimal solution is lost.

This continues iteratively until a feasible solution is found. The algorithm produces good results, but the computation times are considerable due to the simulations.

Multi-skill

Cezik and L'Ecuyer (2005) extended the method of Atlason *et al.* to the equivalent problem in multi-skill call centers. Due to the higher complexity as opposed to single-skill call centers, the computation times are longer.

Bhulai, Koole, and Pot (2006) propose a two-step method to generate shifts in multi-skill call centers. In advance the day is split in different consecutive time intervals, typically half an hour or one hour. In the first step, the optimal staffing levels for each group and each interval are determined. In the second step, shifts are composed such that the staffing level in each interval is met. This approach reduces the required number of simulations. An example is discussed that is solved close to optimality.

9.2 Staffing

The papers that mainly contribute to our understanding of staffing in multi-skill call centers, are treated next.

Loss model

Chevalier, Shumsky, and Tabordon (2004) search for a compromise between service level and staffing budget. Fast optimization is achieved by looking at 'equivalent' blocking systems and using an efficient approximation algorithm, discussed in Section 6.4. Their model consists of an overflow routing scheme with several groups of specialists and one group of fully cross-trained agents. The latter mentioned group is added for flexibility. Their conclusion is to spend 80 precent of the staffing budget on specialists and 20 percent on cross-trained agents. The ratio between specialists and flexible agents in a call center with two skills has been analyzed before in Shumsky (2003).

Credit schemes and characterization of feasible agent configurations

Borst and Seri (2000) consider the model for inbound multi-skill call centers as we described in Section 5: Poisson arrivals, different job types, exponentially distributed service times, etc. The service level is measured by the average waiting time. To ensure that a desired level of service is reached, a set of sufficient conditions and a set of necessary conditions on the number of agents are derived. In order to derive these conditions, two situations are considered. Firstly, a call center with only specialists is analyzed. Secondly, they look at the case that all agents operate as generalists. The results from these two situations are applied to the case in which agents have different sets of skills. In addition, two routing schemes are introduced that prescribe the call selection decision. They ensure that the service level is similar to or better than one would have experienced in the first situation.

LP with recourse

In Harrison and Zeevi (2004) the total personnel cost and the expected abandonment cost are minimized, taking into account the number of agents per skill group and the routing strategy. The number of server pools and customer classes are fixed. The arrival rates are allowed to be time-dependent and to vary stochastically. The procedure is as follows. In the first stage, the system is considered as a regime that scales up the arrival rate super-linearly and the service and abandonment rates linearly; it is the so-called local fluid version of the dynamic scheduling problem. The model is formulated for a fixed capacity vector and the associated optimization procedure yields the optimal staffing vector. During the optimization, exactly one skill is assigned to each agent. The expected abandonment costs are approximated using fluid limits. In the second stage, the capacity vector is optimized by minimizing the associated personnel cost and a cost that represents the system performance. The model from the first stage is used for a fixed capacity vector. This is a so-called two-stage LP with recourse according to the literature about stochastic programming.

Bassamboo, Harrison, and Zeevi (2004) consider the same model. In addition, under a limiting parameter regime they give a lower bound on the expected total costs. They propose a method for staffing and routing that achieves this lower bound.

The question of how the algorithm behaves in realistic situations remains open.

Pooling

We treat pooling as a synonym for merging. In Tekin, Hopp, and Oyen (2004) methods are presented to measure the effect of pooling different departments. It is assumed that the departments are mutually independent. Each department is considered as a multi-server queue. The impact of pooling is analyzed by using approximations for the multi-server queue. They consider general service time distributions and arrival rates, which differ among the departments, and different department sizes. The underlying pooling models are described in Kleinrock (1975). In comparison to the model from Section 5, general service times are not modeled in this paper. We note that departments do fit in the model because a department is techniqually similar to an agent group. Smith and Whitt (1981) and Benjaafar (1995) showed that a pooled system is better than a dedicated one if arrivals and service times have the same distribution. In van Dijk (2002) is shown numerically that pooling is not necessarily an improvement, in particular if the service distribution among the different classes are different. Much other literature about pooling is dedicated to serial service environments.

9.3 Routing

We mention the following papers because they treat the optimization of routing policies in multi-skill call centers.

Simulation procedure

Wallace and Whitt (2004) consider priority routing. A procedure for optimizing staffing and call routing is proposed, which is achieved by exploiting limited cross-training. We will give a brief description of the algorithm: The initial number of agents is determined by assuming that all agents have all skills. Then, all agents are replaced by specialists. Next, additional skills are assigned to the agents. This yields the initial feasible solution. It is optimized by means of simulation. We remark that several types of performance measures are included. In addition, they numerically show (by means of simulation) that when each agent has at most two skills, the performance can be almost as good as when each agent has all skills, which pleads for the use of the classical Erlang models. Thus by adding a little flexibility for routing, it is possible to obtain the almost lowest possible staffing level. This holds even in conjunction with simple static routing policies.

However, the result is only shown for call centers with equal mean service times for each type of job and without holding costs. The question is if these assumptions are realistic and if these results also hold without them.

Overflow routing

For a description of overflow routing we refer to Section 6. This section presents a number of methods that are helpful to improve and optimize service levels. The methods use approximations of the blocking probability, see Section 6.4. One could expect that blocking models are not useful to call centers because call centers have waiting queues. However, they appear to be useful with respect to optimization. The methods compare the blocking probabilities of call centers that undergo small adjustments. For example, when the staffing level of a certain agent group is increased by one. It appears that the differences in blocking probability give a good indication of performance change for similar systems that include waiting queues. Applications of this kind of method are given in Koole, Pot, and Talim (2003), Chevalier, Shumsky, and Tabordon (2004) and Chevalier and Van den Schrieck (2005). The studies show that the optimized instances are nearly optimal, also in the corresponding delay systems. The second reference has been discussed in Section 9.2. The other two papers are described below.

In Koole, Pot, and Talim (2003) a local search method is proposed. The objective is twofold: low staffing levels and good routing policies under service level constraints. As already explained, both issues are related to each other. They demonstrate that even with simple optimization procedures good routing policies can be found. The calculation times are short. The optimization method consists of an evaluation- and an improvement-step. Section 6.4 presents several good alternatives for the evaluation step. The Poisson method is used in the article.

In Chevalier and Van den Schrieck (2005) an optimization method is applied to determine the optimal staffing level of each agent group. They combine a Branch and Bound algorithm with the Hayward method. The method is effective and performs well.

Approximate dynamic routing

In Koole and Pot (2005) approximate dynamic programming is applied to a call center with several groups of specialists and one group of fully cross-trained generalists. A different holding cost is associated to each job type, representing the priority of the job type. The objective is to minimize the average holding costs of calls in the waiting queues. Using Little's law, this is equivalent to minimizing the weighted average waiting time. It is shown that ADP works well for instances with two or three skills. Examples with up to 50 agents are provided and the performance of the policies is about 5 percent worse than optimality.

In Bhulai (2005) a heuristic method is proposed to assign calls to agents in blocking systems, which can be generalized to delay systems and behaves well in large call centers. The policies are nearly optimal.

Call selection

If an agent completes a job, the system assigns a new job to the agent. We are interested in good rules to determine the next job. Relevant literature are the papers discussing queueing systems with priority routing. However, most papers treat the single-server queue. But this is not really a restriction because these rules are technically also applicable to multi-server systems, such as call centers.

Fixed priority (FP) routing is the easiest policy to implement and analyze. It has been analyzed for general service time distributions, and non-preemptive and preemptive service disciplines, see Kleinrock (1975). A disadvantage is the limited possibility to control the first two moments of the waiting time distribution.

The shortest remaining processing time (SRPT) policy minimizes the average waiting time, see Schrage and Miller (1966). According to them the preemptive case is hard to analyze. They do not analyze the higher moments of the waiting time distributions.

The class of time function scheduling (TFS) offers a more general framework for routing policies (which includes FP as a special case). These policies schedule customers according to general functions of the time customers spend in the system. Most of the literature concerns TFS with linear functions of the time in system. Kleinrock (1975) discusses TFS with exponential service time distributions. They present a formula for the average waiting time in case of preemptive and non-preemptive scheduling. Lu and Squillante (2004) consider the TFS with general service time distributions under a non-preemptive service discipline. TFS policies can be fair in many situations and it can also be a mixture between first-in-first-out and smallest jobs first. They give closed-form expressions for the first and second moment of the waiting time distribution. They show that it is possible to control the first two moments very precisely.

For the multi-class single-server case (V-design) it is shown in Federgruen and Groenevelt (1988) and discussed in Walrand (1988) that the μc-rule is optimal among the work-conserving policies. The higher the value of $\mu_m c_m$ the higher the priority of class m.

Yahalom and Mandelbaum (2005) consider the multi-class multi-server queue with Poisson arrivals (V-design) and equal service rates. The objective is to minimize the discounted holding costs in the long run. Each job type has a fixed holding cost c_m, $m = 1, 2, ..., M$. It is assumed (without loss of loss of generality) that $c_1 \geq c_2 \geq \cdots \geq c_m$. They show that the optimal policy is a combination of the μc-rule with a threshold policy for each type m on the number of reserved agents $K_m(x)$, depending on state x. At service completion a server chooses to serve a type m customer if there are no customers in all higher-priority queues and there are more than $K_m(x)$ idle servers. The problem of choosing the appropriate threshold levels is not considered. Note that the policy is not work-conserving.

10 Future directions

During the last decades there is trend in the direction of ‘virtual contact centers’. The rapid developments in telecommunications (e.g., fast internet connections and protocols) and in computer technology (e.g., powerful automatic call distributors) make it possible to bring employees from different locations together, explaining the word ‘virtual’. Telephone is the traditional way of communication in call centers. Currently other media such as email and fax are also very popular. We expect that the movement to virtual contact centers will continue. The future expectations are that agents log on to the system from their home, resulting in higher flexibility of both parties. Also more types of media will be supported by the contact systems, for example chat and self-service media such as video. An appropriate synonym for virtual contact centers is “single-point-of-contact help desk”, mentioned in Wallace and Whitt (2004). We also refer to Whitt (2002) for a vision about the future developments.

An interesting subject for future research is the integration of routing and planning. In this way, by considering problems of a larger scope, it could be that results become even more useful for the industry.

Models from the literature have often many assumptions and require estimations, concerning for example the arrival-time and service-time distributions. The impact of assumptations and the estimation of model parameters are complicated to analyse. It is in our opinion important to validate the models in realistic situations. For example, how effective are call routing policies if the workload deviates from the predictions. In other words, research concerning model validation and robustness is often lacking, but important for the industry.

Our impression of the directions for future work is that there are still great challanges for mathematicians.

Acknowledgments
We would like to thank Maarten Soomer for improving the readability and the structure, and Sandjai Bhulai for his good ideas and for improving the readability.

References

Aguir, M., O. Aksin, F. Karaesmen, and Y. Dallery. 2004. On the interaction between retrials and sizing of call centers. Working paper.

Akşin, O., and P. Harker. 2003. Capacity sizing in the presence of a common shared resource: Dimensioning an inbound call center. *European Journal of Operational Research* 147 (3): 464–483.

Armony, M., and N. Bambos. 2003. Queueing dynamics and maximal throughput scheduling in switched processing systems. *Queueing Systems* 44 (3): 209–252.

Armony, M., and C. Maglaras. 2004a. Contact centers with a call-back option and real-time delay information. *Operations Research* 52 (4): 527–545.

Armony, M., and C. Maglaras. 2004b. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research* 52 (2): 271–292.

Atlason, J., M. Epelman, and S. Henderson. 2002. Combining simulation and cutting plane methods in service systems. In *Proceedings of the 2002 National Science Foundation Design, Service and Manufacturing Grantees Conference.*

Atlason, J., M. Epelman, and S. Henderson. 2004. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*:333–358.

Bassamboo, A., J. Harrison, and A. Zeevi. 2004. Design and control of a large call center: Asymptotic analysis of an LP-based method. Working paper.

Bell, S., and R. Williams. 2001. Dynamic scheduling of a system with two parallel servers in heavy traffic with complete resource pooling: Asymptotic optimality of a continuous review treshold policy. *Annals of Applied Probability* 11:608–649.

Bellman, R. 1961. *Adaptive control processes: A guided tour.* Princeton University Press.

Benjaafar, S. 1995. Performance bounds for the effectiveness of pooling in multi-processing systems. *European Journal of Operational Research* 87:375–388.

Berman, O., and R. Larson. 2000. A queuing control model for retail services having back-room operations and cross-trained workers. Working paper, Massachusetts Institute of Technology, Cambridge, MA.

Bhulai, S. 2005. Dynamic routing policies for multi-skill call centers. Technical report, Vrije Universiteit Amsterdam.

Bhulai, S., and G. Koole. 2003. A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control* 48:1434–1438.

Bhulai, S., G. Koole, and S. Pot. 2006. Simple methods for shift scheduling in multi-skill call centers. Submitted.

Blanc, J., P. d. Waal, P. Nain, and D. Towsley. 1992. Optimal control of admission to a multiserver queue with two arrival streams. *IEEE Transactions on Automatic Control* 37 (6): 785–797.

Bolotin, V. 1994. Telephone circuit holding time distributions. In *Proceedings of the 14th International Teletraffic Conference*, ed. J. Labetoulle and J. Roberts, 125–134.

Borst, S., A. Mandelbaum, and M. Reiman. 2000. Dimensioning large call centers. Working paper.

Borst, S., and P. Seri. 2000. Robust algorithms for sharing agents with multiple skills. Working paper.

Brandt, A., and M. Brandt. 1999. On a two-queue priority system with impatience and its application to a call center. *Methodology Computational Applied Probability* 1:191–210.

Bretschneider, G. 1956. Die Berechnung von Leitungsgruppen für überfliessenden Verkehr in Fernsprechwahlanlagen. *Nachrichtentechnische Zeitschrift* 9:533–540.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, and L. Zhao. 2002. Statistical analysis of a telephone call center: A queueing-science perspective. The Wharton School, University of Pennsylvania, Philadelphia, PA.

Carr, S., and I. Duenyas. 2000. Optimal admission control and sequencing in a make-to-stock/make-to-order production system. *Operations Research* 48 (5): 991–1006.

Cezik, M., and P. L'Ecuyer. 2005. Staffing multiskill call centers via linear programming and simulation. Working paper.

Chevalier, P., R. Shumsky, and N. Tabordon. 2004. Routing and staffing in large call centers with specialized and fully flexible servers. Submitted to Manufacturing and Service Operations Management.

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

Chevalier, P., and J. Van den Schrieck. 2005. Optimizing the staffing and routing of small size hierarchical call-centers. Working paper.

Cooper, R. 1981. *Introduction to Queueing Theory*. 2nd ed. North Holland.

Dantzig, G. 1954. A comment on Edie's 'traffic delays at toll booths'. *Operations Research* 2 (3): 339–341.

Federgruen, A., and H. Groenevelt. 1988. $M/G/c$ queueing systems with multiple customer classes: Characterization and control of achievable performance under nonpreemptive priority rules. *Management Science* 34:1121–1138.

Franx, G., G. Koole, and S. Pot. 2005. Approximating multi-skill blocking systems by hyperexponential decomposition. To appear in Performance Evaluation.

Fredericks, A. 1980. Congestion in blocking systems-a simple approximation technique. *The Bell System Technical Journal* 59 (6): 805–827.

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

Gans, N., and Y. Zhou. 2002. Managing learning and turnover in employee staffing. *Operations Research* 50 (6): 991–1006.

Gans, N., and Y. Zhou. 2003. A call-routing problem with service-level constraints. *Operations Research* 51 (2): 255–271.

Gans, N., and Y. Zhou. 2004. Overflow routing for call-center outsourcing. Working paper.

Guérin 1998. Queueing-blocking system with two arrival streams and guard channels. *IEEE Transactions on Communications* 36:153–163.

Halfin, S., and W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29 (3): 567–588.

Harrison, J., and M. Lopez. 1999. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems* 33:339–368.

Harrison, J., and A. Zeevi. 2004. A method for staffing large call centers based on stochastic fluid methods. Working paper.

Hordijk, A., and G. Koole. 1993. On the optimality of LEPT and μc rules for parallel processors and dependent arrival processes. *Advances in Applied Probability* 25:979–996.

Jagers, A., and E. van Doorn. 1986. On the continued Erlang loss function. *Operations Research Letters* 5:43–46.

Kelly Jr., J. 1960. The cutting-plane method for solving convex programs. *Journal of the SIAM* 8 (4): 703–712.

Kleinrock, L. 1975. *Queueing systems, volume II: Computer applications.* Wiley.

Koole, G. 2003. Redefining the service level in call centers. Working paper.

Koole, G. 2005. Call center mathematics: A scientific method for understanding and improving contact centers.

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

Koole, G., and S. Pot. 2005. Approximate dynamic programming in multi-skill call centers. In *Proceedings of the Winter Simulation Conference.*

Koole, G., S. Pot, and J. Talim. 2003. Routing heuristics for multi-skill call centers. In *Proceedings of the Winter Simulation Conference*, 1813–1816.

Koole, G., and H. v. d. Sluis. 2003. Optimal shift scheduling with a global service level constraint. *IIE Transactions* 35:1049–1055.

Koole, G., and J. Talim. 2000. Exponential approximation of multi-skill call centers architecture. In *Proceedings of QNETs 2000*, 23/1–10.

Kosten, L. 1973. *Stochastic theory of service systems.* Pergamon, Oxford, England.

Lu, Y., and M. Squillante. 2004. Scheduling to minimize general functions of the mean and variance of sojourn times in queueing systems. IBM Research Report.

Mandelbaum, A., A. Sakov, and S. Zeltyn. 2000. Empirical analysis of a call center. Working paper.

Mehrotra, V. August 1997. Ringing up big business. *OR/MS Today*:18–24.

NASSCOM 2006. Call-center statistics. http://www.outsource2india.com/services/callcenters.asp, data originates from the NASSCOM McKinsey Report.

Örmeci, E. 2003. Dynamic admission control in a call center with one shared and two dedicated service facilities. unpublished.

Örmeci, E., A. Burnetas, and H. Emmons. 2002. Dynamic policies of admission to a two-class system based on customer offers. *IIE Transactions* 34:813–822.

Peköz, E. 2002. Optimal policies for multi-server non-preemptive priority queues. *Queueing Systems: Theory and Applications* 42:91–101.

Perry, M., and A. Nilsson. 1992. Performance modeling of automatic call distributors: Assignable grade of service staffing. In *XIV International Switching Symposium*, 294–298.

Puterman, M. 1994. *Markov decision processes.* Wiley.

Reiman, M. 2000. Diffusion limits for multiskill call centers with many agents. Talk at Applied Probability Society of INFORMS 2000, San Antonio, Nov. 5-8.

Samuelson, D. 1999. Predictive dialing for outbound telephone call centers. *Interfaces* 29(5):66–81.

Schaack, C., and R. Larson. 1986. An N-server cutoff priority queue. *Operations Research* 34 (2): 257–266.

Schrage, L., and L. Miller. 1966. The queue $M/G/1$ with the shortest remaining processing time discipline. *Operations Research* 14 (4): 670–684.

Shumsky, R. 2003. Approximation and analysis of a queueing system with flexible and specialized servers. *OR Spektrum* 26:307–330.

Sisselman, M., and W. Whitt. 2004. Preference-based routing. Submitted.

Smith, D., and W. Whitt. 1981. Resource sharing for efficiency in traffic systems. *Bell System Technical Journal* 60 (13): 39–55.

Stanford, D., and W. Grassman. 1998. Bilingual server call centers. In *Analysis of Communication Networks: Call centers, traffic and performance*, 31–47.

Steckley, S., S. Henderson, and V. Mehrotra. 2004. Service system planning in the presence of a random arrival rate. Submitted.

Stolletz, R. 2003. *Performance analysis and optimization of inbound call centers.* Springer.

Tabordon, N. 2002. *Modeling and optimizating the management of operator training in a call center.* Ph. D. thesis, Université catholique de Louvain.

Tekin, E., W. Hopp, and M. V. Oyen. 2004. Pooling strategies for call center agent cross-training. Working paper.

Thompson, G. 1997. Labor staffing and scheduling models for controlling service levels. *Naval Research Logistics* 44 (8): 719–740.

Tijms, H. 1986. *Stochastic models. an algorithmic approach.* Wiley.

van Dijk, N. 2002. To pool or not to pool? The benefits of combining queueing and simulation. In *Proceedings of the Winter Simulation Conference.*

Van Mieghem, J. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Annals of Applied Probability* 5:1249–1267.

Wallace, R., and W. Whitt. 2004. A staffing algorithm for call centers with skill-based routing. Working paper.

Walrand, J. 1988. *An introduction to queueing networks.* Prentice-Hall, New Jersey.

Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Science* 38:708–723.

Whitt, W. 1999. Partitioning customers into service groups. *Management Science* 45 (11): 1579–1592.

Whitt, W. 2002. Stochastic models for the design and management of customer contact centers: some research directions. Working paper.

Whitt, W. 2004a. The impact of increased employee retention upon performance in a customer contact center. Submitted.

Whitt, W. 2004b. Staffing a call center with uncertain arrival rate and absenteeism. Working paper.

Wilkinson, R. 1956. Theories for toll traffic engineering in the U.S.A. *Bell System Technical Journal* 35 (2): 421–514.

Wolff, R. 1989. *Stochastic Modeling and the Theory of Queues.* Prentice Hall, Inc., New Jersey.

Yahalom, T., and A. Mandelbaum. 2005. Optimal scheduling of a multi-server multi-class non-preemptive queueing system. Preprint.