Performance analysis and optimization in customer contact centers

Ger Koole
Vrije Universiteit, Department of Mathematics
De Boelelaan 1081a, 1081 HV Amsterdam
The Netherlands
koole@few.vu.nl

Abstract

We discuss performance models for telephone call center, or more generally, customer contact centers. These performance models are used for workforce planning. We discuss current practice, and discuss its weak points. Finally we discuss ways to plan better and put this also in the context of contact centers with multiple skills and communication channels.

1. Introduction

A *call center* is a collection of resources (typically agents and ICT equipment) capable of handling customer contacts by telephone. If the call center handles not only telephone contacts but also contacts by fax, email, and so forth, then it is usually called a *customer contact center*.

The objective of a call center is usually to obtain at least a certain minimal service level (SL) for minimal costs. The SL will in general consists of multiple criteria, some of which are related to the quality of the actual service that is delivered; some are related to the time before a *customer service representative* or *agent* is reached: the waiting time. We will concentrate on SL metrics that have to do with waiting times, although there is a relation between the effectiveness of the service delivered and the waiting time (e.g., when the wrong answer is given, the customer calls back later, thereby increasing traffic). The most usual way to define service level is by taking a percentage α and a number a for which must hold: α% of the customers must have a waiting time shorter than a seconds. Often we see values such as $\alpha = 80$ and $a = 20$ (the 'industry standard', for what it's worth). Whether this service level must hold for every time interval, or on average over the whole day, is not always clear.

Another way to define service times is by the number of *abandoned calls*. Customers abandon because they do not have the patience to wait for a agent to be available. Customer patience typically averages one minute. Of course the fraction of abandoned calls is strongly related to the waiting time; often call center managers will choose the service level parameters α and a such that abandonments occur rarely. Percentages up to 3 or 5% are often considered acceptable.

Let us now consider the costs. Highest are personnel costs, they usually account for 60 to 70% of the total costs. Other important types of costs are the costs of ICT equipment and telephone costs; the latter depends strongly on who pays for the incoming calls.

In the next sections we discuss quantitative methods useful to call center management. We start with performance models. Then we move on to a question that is very relevant to call center managers: how to schedule agents such that SL and cost constraints are satisfied? In the final section we discuss the shortcomings of the scheduling method and show in which directions solutions have to be sought.

Let us say a word on the available literature before moving to performance measures. Call centers already exist for many decades, but recently the number of call centers and their size increased dramatically. The number of papers on call centers in the scientific literature reflects this situation: although there always have been publications in Operations Research journals on this subject, this number has increased significantly over the last few years. In Gans et al. [4] a recent overview of the call center literature can be found.

2. Performance models

The basic model for a call center is the Erlang C or $M|M|s$ model. Consider a certain interval with on average λ arriving calls per time unit, an average call holding time of β, and s agents. If $\lambda\beta < s$, then there is a probability (usually denoted by C) that an arbitrary arriving call gets blocked. And if a call gets blocked, then the waiting time has an exponential distribution with parameter $s\mu - \lambda$ (with $\mu = \beta^{-1}$). Using this the service level can be approx-

imated. There is no need for the users to implement the formulas themselves: there are many so-called *call center calculators* freely available on the internet.

There are a number of reasons why the Erlang C formula might give bad approximations for the actual SL: the arrival rates and the number of agents varies, the service times are non-exponential, the number of lines is finite, and waiting calls abandon. We discuss them one by one.

Stationarity Arrivals in call center occur according to a Poisson process with a varying rate, and also the number of scheduled agents varies over time. The Erlang formula is valid under stationarity which assumes non-varying parameters, in contradiction with the above. The standard solution is to assume that the parameters are all piece-wise constant, and to apply the Erlang formula to each interval using the parameter values belonging to that interval. Often 15-minute intervals are taken. Thus we do not find a stationary situation; however, experience shows that this type of stochastic processes converges fast to its equilibrium state, meaning that in general we can take the stationary situation as approximation.

Service times In reality service times are not exponential. However, in all commercial software tools service times are assumed to be exponentially distributed, and practice shows that this is in general a good approximation.

Number of lines There is a limited maximal number of connections between the call center and the public network. Every customer that is connected to the call center is using a connection, even while waiting. It can be the case that the number of lines is chosen such that even during peak traffic blocking rarely occurs; in that case an infinite number of lines (as in the Erlang C) is a very good approximation. In other situations the number of lines might have a significant influence on performance.

Abandonments As callers wait they tend to abandon; an average patience of 1 minute is no exception. Modeling abandonments is crucial, unless the service level is so high that abandonments rarely occur. Modeling abandonments also makes the model more robust: even in overload situations (i.e., $\lambda\beta > s$) the SL is well defined. Usually the patience is assumed to be exponential, the resulting model is sometimes called the Erlang A model. There exist a number of proprietary implementations of the Erlang A (including the possibility of having a finite number of lines, known under the names of Merlang or Hills-B). A tool in which this model is implemented can also be found on `http://www.math.vu.nl/~koole/ccmath`. Whitt [5] recently showed that the patience distribution can have a considerable impact on the performance. The phenomenon of *redials* remains until now almost unexplored.

3. Optimization

The Erlang formula, or one of its generalizations, is the starting point of the management of the workforce. Managing this workforce in such a way that service level targets are met and that costs are minimized is the subject of this section. This operational task consists mainly of scheduling agents such that there are enough agents at each time interval to meet the desired service levels. Decision Support Systems exist to support this task. We discuss the typical operation of this type of tool.

The first step of workforce management (WFM) is predicting the load of the system. Predicting call volume on the basis of historical data is quite difficult due to the diversity of events that have to be taken into account, ranging from special holidays to the introduction of new products that influence the number of calls. The forecasting step results in predictions of the arrival rate for each interval that the call center is operational.

Based on these predictions the minimum number of agents needed to reach the service level for each interval can be calculated, using (a generalization of) the Erlang formula. From this daily schedules can be made using a set of standard shifts. This is done using some optimization method. There can be numerous different shifts, depending on starting times, lengths and the moments of the breaks.

After the minimum occupancy levels have been translated into shifts, these shifts have to be assigned to agents. Often there are different groups of agents with for example different shift lengths. By assigning upper and lower bounds to certain categories of shifts this can be accounted for. Often agents can choose from the shifts based on their personal preferences.

By now we have a complete picture of the standard approach in which the problem is decomposed into four steps: call volume estimation, calculation of minimum number of agents, determining shifts, assigning agents to shifts. We find this approach in many software packages that are especially designed for this task.

4. Planning in practice

Planning does not stop when all shifts are assigned to agents. Sometimes updates are made, and during the execution the SL is monitored and measures are taken if necessary. This occurs rather frequently, because often the offered load deviates considerably from the forecast, or because the number of agents is lower than planned, due for example to illness. Of course the latter is accounted for in the planning: the number of agents needed according to the Erlang formula is raised by a percentage called the *shrinkage*. However, this is a fixed number, and the actual absence because of illness and other reasons varies from day to day.

Because the actual absence cannot be predicted at the moment the schedule is made the average is taken. The fluctuations make ad hoc decisions at the latest moment necessary.

Similar arguments apply to the offered load. The load is influenced by external factors that cannot be predicted in advance. Sometimes a higher offered load can only be predicted from the load in the previous intervals: the difference between actual and average load shows a positive correlation between successive intervals (Avramidis et al. [1]). A worst case approach at the time the schedule is made leads to structural overstaffing and thus to high costs.

The way to deal with these, at the time the schedule is made, unpredictable fluctuations is adding enough flexibility to the whole process such that changes can be made to the schedule as soon as better load and agent availability estimates are available. We discuss a number of ways to add flexibility to the process.

Flexibility in agent contracts With flexibility in agent contracts we mean that for certain agents we can decide on a very short notice (e.g., at the beginning of the day) whether we require them to work or not (overtime is discussed later). Of course they get paid for being available, and often they are guaranteed a minimum number of working hours per week. This is an excellent solution to deal with variability in arrival rate and absence, but it is also costly: flexible agents are more expensive than agents with a fixed contract. It is therefore necessary to quantify the need for flexible agents. This can be done by giving bounds to the arrival rate and the absenteeism and to determine the minimal and maximal required workforce from this. The difference between the two should be filled in using the flexibility in contracts, or by other means that are discussed below.

Flexibility in task assignment Introducing flexible contracts gives us the possibility to handle days with a higher than usual traffic load. If the peaks are shorter, in the order of an hour, then we cannot require agents to come just for this short period of time. Sometimes it is possible to mobilize extra workforce by having personnel from outside to work in the call center. Although this seems a simple solution for emergency cases, one should realize that the extra agents should be trained and that the telephony and IT equipment should be in place to accommodate all agents. Again, the costs are considerable.

Another way of introducing flexibility in tasks assignment is as follows. Nowadays customers use increasingly internet-based communication *channels* such as email and web forms. The SL requirements for these contacts are different from telephone contacts: they usually have to be answered within one day. This means that agents who are scheduled to deal with contacts with less tight SL requirements are shifted to inbound traffic if the load requires this. If this is done dynamically and on a per-call basis then this is called (call) *blending*. This gives theoretically the best performance, but the efficiency of agents is reduced and it demands adaptations in the software that manages the call center. For these reasons a change in task assignment is usually planned for several intervals at once.

Note that in case of a shift in task assignment to inbound calls the "low" SL channels should be planned later, to avoid answering late and to avoid additional traffic from customer waiting for an answer to their email. This can typically be done in overtime. Note that the above is a good reason to stimulate customers to use email instead of the telephone for their contacts. This can be done by refering to the website in the welcome message, by charging for telephone contacts, and so forth.

Multiple skills When the number of contacts grow in a call center the moment occurs that certain agents specialize in certain skills. This increases the effictiveness and efficiency of agents, and reduces training costs. On the other hand, the call center is split up into smaller call centers, each with their own skill. Thus there are less economies of scale and there is less flexibility, as each agent can only handle a subset of the customer contacts. For this reason certain agents are cross-trained. These cross-trained agents can be scheduled in two ways, similarly to the way multi-channel agents are scheduled: they receive dynamically calls requiring different skills, or they are scheduled during one or more interval for the same skill. Again, dynamically assigning calls requiring different skills is best from a theoretical point of view. However, this decreases the efficiency and agents often have a preferred skill. Thus there is a trade-off between economies of scale and single-skill efficiency. (And when the economies of scale play a big role, then perhaps outsourcing is cheaper as the outsourcer probably can achieve economies of scale.) Note that finding optimal dynamic assignment rules is a challenging high-dimensional optimization problem, for which the current methods are not suitable. Only for restricted classes of routing policies approximations exist (Chevalier & Tabordon [3]). In practice it is often best to give an agent always a single type of calls, but to vary this type over the day depending on load and occupancy. Having cross-trained agents also gives the possibility to react to changes in load on one of the skills. According to Chevalier et al. [2] 20% cross-trained agents suffices in standard situations.

References

[1] A. Avramidis, A. Deslauriers, and P. l'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Science*, 2004. To appear.

[2] P. Chevalier, R. Shumsky, and N. Tabordon. Routing and staffing in large call centers with specialized and fully flexible servers. Submitted, 2004.

[3] P. Chevalier and N. Tabordon. Overflow analysis and cross-trained servers. *International Journal of Production Economics*, 85:47–60, 2004.

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

[5] W. Whitt. Engineering solution of a basic call-center model. *Management Science*, 2004. To appear.