Summary

Service-Level Variability and Impatience in Call Centers

A call center is defined as a group of telephone agents whose principal business is serving callers over the telephone. This thesis only considers the type of call center in which outside callers initiate the call to the call center. (The usual notation in the field of queueing theory is used where callers are called customers and agents are called servers.) The task of a call center manager is to keep costs as low as possible, which means keeping the number of servers as few as possible. On the other hand, a call center often has to obey a contract that prescribes that the service level has to be above a certain level. An example of this is that at least 80% of the customers should wait no longer than 20 seconds before receiving service.

A call center operates in a volatile environment with a high level of uncertainty. For example, it is impossible to know beforehand how many customers will call on a given day, when they will call, and how long a conversation will last. Under the Markovian assumptions, i.e., customers arrive according to a Poisson process and service times are assumed to be exponentially distributed, a call center can be modeled by a queueing system that is easy to analyze. For a given number of servers, it is possible to determine the service level, and therefore it is possible to find the optimal number of servers such that the service-level target will be met. However, there is a considerable problem that has not been recognized in the literature yet: the variability in the service level.

Most models only consider long-term performance, i.e., the service-level estimate will be met only in the long run. In practice however, service levels will be reported as an average over short intervals that typically range from several hours up to a day at most. In such short intervals, the service level is a random variable with a significant variability. The estimate provided by a queueing system is then only the expectation. By considering the whole distribution of the service level, improved decisions can be made regarding the optimal number of servers.
Chapters 2 and 4 concern the characterization of the distribution of the service level. In Chapter 2 a closed-form approximation for the variance is constructed. Based on this approximation, the normal probability distribution is used to characterize the form of the service-level distribution. This approximation is justified in the limit. The approximations turn out to be very accurate for shorter intervals starting from two or three hours. Furthermore, using the service-level distribution it is shown how better decisions can be made. Chapter 4 analyzes the distribution of the service level in an exact way by utilizing a double Laplace-Stieltjes transform. Instead of the customer-average service level, the time-average proportion of the time that the virtual waiting-time process is below the acceptable waiting time is considered. The time average converges to the customer-average service level due to a property of the Poisson process. The distribution of the service level is used in Chapter 3 to ensure that the service level at the end of the day is above the required target. In a call center a long-term planning has to be made concerning the number of scheduled servers, of which the level of uncertainty is high. At the start of the day and during the day more information becomes available. Using the idea of flexible servers, that can be added or removed, the number of servers is variable over the day and becomes a decision variable. The problem is modeled as a Markov decision process, that allows to find the optimal policy to meet the service-level target at minimal costs.

The second theme of this thesis is the impatience of customers. Customers that cannot directly get contact with a server have to wait in a virtual queue until a server becomes available. Customers have the property to be impatient: when the waiting time becomes too large, they hang up. This abandonment process influences the performance measures and therefore has to be taken into consideration in the models. An additional problem is that the service level is no longer unambiguously defined. Multiple definitions are used in practice.

Chapter 5 considers the problem of impatient customers. This chapter studies the different service-level definitions, including all those used in practice. Moreover, two new models are introduced based on data from call centers. Both models are shown to fit reality very well. The different service-level definitions are compared through numerical analysis. It is shown what the effect is on the required number of servers. Chapter 6 deals with a call center with impatience and multiple priorities. Customers are grouped by their priority and high-priority customers get priority over customers with a lower priority. It is shown how performance measures can be obtained for both types of customers under two kinds of service disciplines.