Summary

Queueing models are typically used to analyze stochastic systems where congestion occurs. Prominent examples are grocery stores, amusement parks and road networks (visible queues), and call centers, communication networks, manufacturing and computer systems (at a more abstract level). In this thesis, we study multi-class queues, or more specifically, we consider a single queueing node that is used by multiple customer classes. Three common types of multi-class queues are: *priority* queues, *polling models*, and *Processor Sharing* queues. In priority queues, the customer classes are subject to a priority structure in which high priority classes have preferential treatment over lower priority classes. In polling models, customers of different classes arrive in different queues. There is a single server that can serve only one queue at a time and then switches to a different queue. Finally, Processor Sharing queues are characterized by the fact that all customers receive service simultaneously. Nonetheless, high priority customers may receive a larger share of the server. Following this hierarchy, priority queues are found in Chapters 6 and 7, polling models in Chapters 2–4, and Process Sharing queues in Chapter 5.

In Chapter 2 we analyze polling models with gated and globally gated service disciplines. We consider the following five local scheduling policies: FCFS, LCFS, ROS, PS and SJF (see Table 1.1 for a description). For each configuration, we derive the distribution of the waiting time in the heavy-traffic (HT) regime, i.e., when the load tends to 1. We show that the waiting-time distribution in HT is the product of two random variables. The first random variable captures the impact of the local scheduling policy, whereas the second random variable has a gamma distribution with known parameters and is the same for all local scheduling policies. These asymptotic results, combined with low-traffic results, are used to derive closed-form approximations for the waiting-time distributions in polling models with arbitrary load. The performance of these approximations is evaluated with simulations. The numerical results show that the approximations are accurate for all possible load values.

The model and type of results of Chapter 3 are similar to those in Chapter 2, but now the service discipline is exhaustive. This policy is more challenging to analyze than the gated policies, since we now have to deal with customers arriving during the service of the queue. We derive new closed-form expressions for the asymptotic

waiting-time distribution under exhaustive service. The waiting-time distribution in HT is the product of two random variables, where the first random variable captures the impact of the local scheduling policy. The second random variable has a gamma distribution and is the same for all local scheduling policies. The difference with the gated case is the fact that the first random variables are generally more complicated. The results lead again to closed-form approximations for the waiting-time distributions in polling models with arbitrary loads, which are evaluated using simulations. The approximations are accurate for all systems with reasonable loads.

In Chapter 4 we study polling systems with globally gated or gated service disciplines and FCFS as local scheduling policy. We are interested in the transient behavior of the cycle lengths. By deriving the joint LST of x cycles in terms of the first cycle, we are able to analyze the dependency structure between the different cycles. This is useful in, e.g., systems where breakdowns or other disruptions might occur, leading to long cycle lengths. The time to recover from such events is a primary performance measure. From the joint LST, we derive first and second moments and correlation coefficients between different cycles. Numerical results show the influence of cycle lengths on subsequent cycle lengths.

In Chapter 5 we analyze a Discriminatory Processor Sharing (DPS) queue. This is a queue, where all jobs that are present are served simultaneously. The different job types are assigned different weights and, depending on those weights, each job receives a share of the server's capacity. Jobs with higher weights receive more server capacity than jobs with lower weights and thus are served relatively fast. We assume that the service times are exponential and batches of jobs of various types arrive according to a Poisson process. We are interested in the joint queue-length distribution. We show that, in the HT regime, the scaled distribution is given by a vector of known constants multiplied by a single exponentially distributed random variable (with known parameter), also referred to as a state-space collapse. This simple result can be used to approximate the joint queue-length distribution in stable DPS systems. Numerical results show the usefulness of the asymptotic results for stable systems.

In Chapter 6 we study a specific single-server priority queue with two types of jobs. This chapter is motivated by a health-care application, more specifically, by access times for an appointment at a hospital's outpatient department. The type-1 jobs are patients arriving according to a Poisson process and the type-2 jobs are other tasks (e.g., administration tasks). We assume that there is an infinite number of type-2 jobs. If the queue length of type-1 jobs is above a certain threshold level, then more type-1 jobs are taken into service, by doing less type-2 jobs. This causes type-1 to be served faster. If the queue length of type-1 jobs drops below the threshold, more type-2 jobs will be taken into service again. We are interested in the waiting-time distribution of type-1 jobs and the fraction of time that less type-2 jobs can be done. To this end, we develop two different models, where the second model also allows for randomness in the number of type-2 jobs that can be done. Based on numerical experiments, we see that such systems may efficiently operate at high loads of type 1.

In Chapter 7 we study a specific multi-server priority queue with two types of jobs. This chapter is motivated by a call center application. Type-1 jobs (e.g., inbound calls) arrive according to a Poisson process and have non-preemptive priority over type-2 jobs (e.g., emails, outbound calls). We assume again an infinite number of type-2 jobs and the service-time distribution is exponential, with different means for the different job types. Type-1 jobs have a general patience distribution, they abandon the queue if their patience is smaller than their waiting time. If there is no queue of type-1 jobs, some of the servers will be kept idle, so that they are able to immediately handle arriving type-1 jobs. For the type-1 jobs, we derive the waiting-time distribution and the probability to abandon. The waiting-time distribution is given by the solution of second-order differential equations. When customers have infinite patience, the waiting-time distribution can be written as a mixture of exponentials. For the type-2 jobs, we determine the throughput.