Asymptotic dimensioning of stochastic service systems

Stochastic service systems describe situations in which customers compete for service from scarce resources. Think of check-in lines at airports, waiting rooms in hospitals or queues in supermarkets, where the scarce resource is human manpower. Next to these traditional settings, resource sharing is also important in large-scale service systems such as the internet, wireless networks and cloud computing facilities. In these virtual environments, geographical conditions do not restrict the system size, paving the way for the emergence of large-scale resource sharing networks. This thesis investigates how to design large-scale systems in order to achieve the dual goal of operational efficiency and quality-of-service, by which we mean that the system is highly occupied and hence efficiently utilizes the expensive resources, while at the same time, the level of service, experienced by customers, remains high.

The intrinsic stochastic variability of arrival and service processes is the predominant cause of delays experienced by customers. Queueing theory and stochastics provide the tools to describe and evaluate congestion in these systems. An important insight obtained through queueing analysis is the effect of resource pooling for systems with many servers and corresponding economies-of-scale that can be achieved by increasing the scale of the system. Although classical queueing theory allows for exact evaluation of the performance of queueing systems of moderate size, exact analysis becomes intractable as demand R and capacity s become large. In those cases, one typically resorts to asymptotic approximation techniques, such as heavy-traffic diffusion approximations: the analysis of a sequence of queueing processes, scaled in space, in which the server utilization level approaches 100%. The resulting probabilistic limiting processes are easier to analyze. Moreover, the diffusion approximations have direct interpretations in terms of the original systems and lead to tractable characterizations of their performance.

The heavy-traffic regime that plays a central role in this thesis is the Halfin-Whitt regime, also known as the Quality-and-Efficiency Driven (QED) regime, which dictates that capacity should be equal to the nominal demand plus an additional variability hedge which is proportional to the square-root of the nominal load, i.e. $s = R + \beta\sqrt{R}$ for some $\beta > 0$. The driving force behind this scaling regime is the central limit theorem (CLT). The rule $s = R + \beta\sqrt{R}$, commonly known as the square-root staffing principle, has been proved to secure both efficiency (utilization approaches 100%) and quality-of-service, since the mean waiting time is negligible under this scaling as the system grows large. Since the QED regime allows coexistence of the two seemingly conflicting objectives in large-scale service systems, the paradigm has been implemented in a wide variety of operational settings. However, the standard QED regime fails to acknowledge features that play a dominant role in practice. This thesis contributes to the existing literature by identifying these distinctive traits and showing how to account for them in a modified QED framework.

In Chapters 2 & 3, we study how the limiting behavior of many-server queues is affected when one deviates from the standard square-root staffing principle. In Chapter 2 we investigate a novel family of scaling regimes, in which the amount of overcapacity $s - R$ is not necessarily of the order $\sqrt{R}$, which gives rise to a novel family of heavy-traffic regimes and corresponding scaling limits. Continuing our study of alternative scaling regimes, we investigate in Chapter 3 how to adapt the square-root staffing paradigm in case the system faces demand patterns that are stochastically more volatile than anticipated. This phenomenon is known as overdispersion and can be caused by e.g. the existence of correlation between the sources generating demand, or uncertainty about the arrival volume.

In Chapters 4 & 5, we review a family of queueing models in the QED regime in which the total number of customers that can reside in the system simultaneously is limited. As a result, customers may be denied access in case they find a full system on arrival. This fraction of arrivals may either reattempt later or leave the system directly. The impact of retrials on scaling rules in the QED regime is the focus of Chapter 4. Since the volume of initially blocked customers is proportional to $\sqrt{R}$, that is, the same order as the variability hedge in the staffing rule, retrials are prone to have a non-negligible effect on performance. We propose a heuristic method for the performance analysis of these types of queueing models with finite-size restrictions, which is based on a fixed-point equation. As a by-product this yields a two-fold square-root staffing principle, which prescribes a synchronous scaling for both the system capacity and waiting space. Chapter 5 describes how these ideas can be applied in the context of an emergency department.

Chapter 6 studies a cost minimization problem in a single-server queue with non-stationary input. The bulk of the queueing literature concerns performance analysis assuming that steady state is reached. However, the validity of this assumption in practice is questionable, for the simple fact that no service system runs infinitely long. Moreover, system parameters, such as the arrival volume, are likely to change over time. In this chapter, we characterize the error in performance metrics that follows from this transient nature of queues, and present a correction to the original staffing rule to account for the finite time horizon.

Finally, we analyze in Chapter 7 a specific stochastic service system: an inventory model of a blood bank with backlogs, perishable goods and consumer impatience. We obtain the stationary distribution of the inventory level, and deduce under appropriate scaling the stochastic process limit in terms of a diffusion process. This process limit allows for a more tractable approximate analysis of the model in case the number of blood deliveries and demand is large.