SURVEY and REVIEW

Johan S. H. van Leeuwaarden, Britt W. J. Mathijsen, and Bert Zwart are the authors of the Survey and Review paper in this issue: "Economies-of-Scale in Many-Server Queueing Systems: Tutorial and Partial Review of QED Halfin–Whitt Heavy-Traffic Regime." We have all gone many, many times through the experience of waiting in a queue until eventually getting attention from one of the servers. In communication networks, data packets wait in line until a communication channel becomes available. Queueing theory provides mathematical tools to analyze and improve the performance of systems of this kind, where *jobs* (i.e., customers, patients, data packets, etc.) have to wait for one of the multiple *servers*. Queues behave in an inherently random way, and, if you are like me, you will often think you were unlucky and joined the line at what happened to be the busiest moment. Therefore, from a mathematical point of view, queueing theory deals with stochastic processes.

What is the optimal number of servers s for a given queue? Rescale the time variable in such a way that, on average, jobs arrive at a rate of one per unit time. If s is not larger than the average processing time λ of the jobs, then the queue will surely grow longer and longer. Thus $s > \lambda$ is a minimal requirement for the stability of the system. However, even when this requirement is fulfilled, due to the randomness of the arrivals and of the processing times, some (or even most) jobs will not be dealt with as they turn up. If the gap $s - \lambda$ is small, the quality of the queue will be small: jobs will very likely have to wait and the waiting time will be long on average (left column in Figure 2). On the contrary, if s is too big, the system will be inefficient, as several servers will probably remain idle for long periods, as depicted in the right column of the figure. The QED acronym in the title of the paper refers to the quality- and efficiency-driven regime, where one tries to choose s large enough to avoid low quality, but not so large that it leads to gross inefficiency (central column in Figure 2). The words "heavy-traffic regime" indicate that the authors are interested in cases where λ is large, so that much attention is paid to the queueing stochastic process in the limit $\lambda \uparrow \infty$.

The paper, which contains many enlightening figures, will convey to a general audience a flavor of queueing theory and, more generally, of the techniques used to study applied stochastic processes. For the expert, the authors provide an extensive survey of the literature and suggest many research opportunities.

J. M. Sanz-Serna
Section Editor
jmsanzserna@gmail.com