Scaling limits for critical queues with diminishing populations

Gianmarco Bet

Queues are exceedingly common in the modern society. In fact, queueing phenomena occur whenever there is competition for a scarce resource. Often, this resource is time and the competition is over the attention of a server. Queues need not be physical queues: one can think of call centers where customers compete for an operator and hospital wards, where customers are competing for the time of a specialist. On an even more abstract level, queues play a prominent role in data handling and dissemination. For example, wireless communication protocols spread information in packets, which are then buffered before being received. These phenomena present a serious challenge, and as such they have stimulated the birth and development of a rich branch of mathematics dedicated to studying them. Our broad goal is therefore to develop techniques to study queueing models that are inspired by real-world systems. In particular, our research is motivated by the current lack of understanding of time-dependent queueing models. These are processes whose dynamics and structural properties change markedly over time. More specifically, we study queueing systems that serve only a finite number of arriving users, also known as transitory queueing models. This feature is shared by a large number of practical scenarios. Queues at post offices, at concerts, at voting booths, or at a hospital wards are all a consequence of a large, but finite pool of customers requesting for service independently from each other. In our stylized model, we assume an initial population of n customers. Each customer will, at some random time, require service and thus join a single queue leading to a server processing jobs in a First-Come-First-Served manner. If the customers are homogeneous, their arrival times are identically distributed. By further assuming that the arrival times are independent, we obtain the so-called Delta/G/1 queue. The characteristic feature of the Delta/G/1 queue is that as more customers join and leave the system, fewer customers can potentially join. To the queue we associate a graph, as follows: we regard customers as vertices and whenever customer i joins during the service of customer j we draw a directed edge from i to j. This scheme yields a directed random forest, where each connected component corresponds to a busy period in the original queue. The connection between the queue and the random tree is given by the (embedded) queue length process, that describes the length of the queue at the end of each service, and that also corresponds to the exploration process of the random graph. Our main goal is the study of the Delta/G/1 queue and its associated random graph as a means to understand the complex dynamics characterizing transitory queues. In order to obtain tractable expressions, we make some simplifying assumptions. First, we focus on the “large customers pool” regime, that is on the regime $n \to \infty$. Indeed, when n is very small, the resulting queueing model can be analyzed directly by numerical simulation. Moreover, our results yield approximations that are accurate even for relatively small n ($n \geq 1000$). Second, we consider the Delta/G/1 queue in the so-called heavy traffic, or critical, regime. Considering queueing systems in their heavy-traffic regime typically leads to a reduction in complexity because unnecessary details become negligible and the most relevant characteristics for performance emerge. In the case of the Delta/G/1 queue, the heavy-traffic condition amounts to assuming that, during rush hour, the influx of customers is approximately equal to the processing speed of the server. Under this assumptions, we prove several limit theorems for the queue length process of the Delta/G/1 queue (resp. embedded Delta/G/1 queue). The convergence of the embedded queue leads to various results describing the size of the connected components of the corresponding random graph. We consider three different variations of the Delta/G/1 queue. First, we

assume that the arrival times are independent and identically distributed (i.i.d.), and the service times are i.i.d. with finite second moment (the expectation of the second moment squared is finite). In this setting, the queue length process converges to a Brownian motion with negative polynomial drift. In fact, the drift represents the effect of the customers leaving the system and not being able to join the queue again. Different arrival patterns give polynomials of different degree. Second, we study how heavy-tailed services (i.e. infinite second moment) affect the performance of the system. We find that, in this setting, the queue length process converges to an α-stable motion with a negative drift. The α-stable motion is a process that is driven by many small jumps and occasionally one large jump. It is therefore appropriate for modeling phenomena that are characterized by a large volatility and sudden transitions, such as stock markets and insurance claims. Third, we introduce a new class of models regulated by a parameter α, which we refer to as the Deltaα/G/1 queue, where each customer samples its arrival time according to a distribution that depends on its service time. As a consequence of this, depending on the value of α, customers that require more attention from the server tend to join earlier (resp. $\alpha \geq 0$), or later (resp. $\alpha \leq 0$). Besides being of independent interest, this model interpolates between two known models in the literature. For $\alpha=0$, we retrieve the Delta/G/1 queue, while for $\alpha=1$ we obtain the well-known inhomogeneous random graph. Our results then generalize both models, and help explain the universality phenomenon (identical asymptotic behavior) that characterizes them.