Large scale queueing systems in the Quality/Efficiency driven regime and applications

Abstract

A parallel server queueing system in the so-called Halfin-Whitt (Quality- and Efficiency-Driven) heavy traffic regime has been recently widely recognized as a particularly succinct model of large scale call centers. While such queueing systems are well understood when the service times have exponential distribution, the case of non-exponential service times presents a major challenge and is of importance due to practical considerations. We will discuss the behavior of this system in steady-state when the service times have lattice-valued distribution with a finite support. We characterize the limiting queue length and waiting time distributions in terms of the stationary distribution of an explicitly constructed Markov chain. As a consequence, the "correct" scaling behavior of this system is established. Then we obtain an explicit expression for the critical exponent for the moment generating function of a limiting (scaled) steady-state queue length. This exponent has a compact representation in terms of three parameters: the amount of spare capacity and the coefficients of variation of interarrival and service times. Interestingly, it matches an analogous exponent corresponding to a single-server queue in the conventional heavy-traffic regime. The results are derived by constructing an appropriate Lyapunov function.

Joint work with P. Momcilovic.