Heavy-Traffic Approximations for Many-Server Service Systems

In the classical heavy-traffic scaling, a reflected Brownian motion (RBM) is used to approximate the queue-length process of a G/G/1 queue. Here we demonstrate the use of diffusion process in approximating queues with many servers under some scaling conditions on the system’s parameters. We will touch on some useful tools that allow us to move from a stochastic-process limit for X describing the number of customers in system to the limit of other quantities like queue-length process and the virtual waiting-time process.

Here we deliver a specific application of this technique. We apply heavy-traffic analysis to expose the impact of server-assignment rules on the idle-busy activity pattern of individual servers. This is important because how the available idleness is allocated has a direct impact on the performance of human-operated service systems like hospitals and call centers. In particular, we explore the possibility of creating work breaks from the naturally available idleness, assuming that the service system is staffed adequately to provide non-negligible idleness. We start by establishing many-server heavy-traffic limits to develop useful approximations for the distributions of server idle times with the customary longest-idle-server-first (LISF) rule and a random-routing (RR) alternative. We show that the pattern of idleness with these rules is totally different but neither produces effective work breaks. We then develop three dynamic priority (DP) rules and show through heavy-traffic analysis that these DP rules are asymptotically optimal in a sense that they maximize the number of work breaks among all work-conserving server-assignment rules.