1. Abstract

The thesis includes three parts, all of them dealing with a control of many–server queueing systems. A queueing model has $J \geq 2$ heterogeneous service stations, each consisting of many independent servers with identical capabilities. Customers of $I \geq 2$ classes can be served at these stations at different rates, that depend on both the class and the station. A system administrator dynamically controls scheduling and routing. We study this model in the Central Limit Theorem (or heavy traffic) regime proposed by Halfin and Whitt [18]. Both the model and the parametric regime have recently received much attention, especially in relation to large telephone call centers.

When studying queueing models in heavy traffic, one considers a sequence of models parameterized by $n \in \mathbb{N}$ that, under a Law of Large Numbers (LLN) limit, give rise to a static fluid model, which is critically loaded in a standard sense. The arrival rates and the number of servers are scaled up in such a way that the processes representing the number of class-i customers in the system, $i \in \mathcal{I}$, exhibit diffusive fluctuations about the fluid model.

In Part I we derive a diffusion model on $\mathbb{R}^I$ with a singular control term, that describes the scaling limit of the queueing model. The singular term may be used to constrain the diffusion to lie in certain subsets of $\mathbb{R}^I$ at all times $t > 0$. We say that the diffusion is *null-controllable* if it can be constrained to $\mathbb{X}_-$, the minimal closed subset of $\mathbb{R}^I$ containing all states of the prelimit queueing model for which all queues are empty. We give sufficient conditions for null controllability of the diffusion. Under these conditions we also show that an analogous, asymptotic result holds for the queueing model, by constructing control policies under which, for any given $0 < \varepsilon < T < \infty$, all queues in the system are kept empty on the time interval $[\varepsilon, T]$, with probability approaching one. This introduces a new, unusual heavy traffic 'behavior': On one hand the system is critically loaded, in the sense that an increase in any of the external arrival rates at the 'fluid level' results with an overloaded system. On the other hand, as far as queue lengths are concerned, the system behaves as if it is underloaded.

Part II introduces and analyzes the notion of throughput sub-optimality for queueing systems in heavy traffic. The underlying fluid model is assumed to be throughput sub-optimal. Roughly, this means that the servers can be allocated so as to achieve a total processing rate that is greater than the total arrival rate, while, for every $i \in \mathcal{I}$, the 'mass' of servers allocated to serve class i does not exceed the 'mass' of class-i 'material.' We show that there

exists a dynamic control policy for the queueing model that is efficient in a strong sense: Under this policy, for every finite T, the measure of the set of times prior to T, at which at least one customer is in the buffer, converges to zero in probability at the scaling limit. On the way to prove our main result, we also provide a characterization of throughput sub-optimality in terms of an algebraic condition.

In Part III we deal with the controlled diffusions. If a control problem is associated to the queueing model, then heavy–traffic approximations gives rise to a controlled diffusion model (CDM. It turns out that by letting service rates depend only on the station, or only on the customer class leads to significant simplifications of the CDM. We then indicate particular cases, when the exact solution is available, and describe how to construct control schemes for the queueing model that are asymptotically optimal.