HOW DO YOU GET DELAYED FOR LONG IN MULTI-SERVER QUEUES?

DR. KARTHYEK R. A. MURTHY

Abstract

We consider two-server queues as an example of multi-server queues and attempt to understand how large delays happen in steady-state. In particular, we focus on the case where the incoming job sizes are heavy-tailed. We shall discuss the tail asymptotics in various regimes. Specifically, we shall discuss in detail our work on the half-loaded regime where the arrival rate is equal to the service rate, in which we make the following interesting observations: When the incoming jobs have finite variance, there are basically two types of effects that dominate the tail asymptotics. While the quantitative distinction between these two manifests itself only in the slowly varying components, the two effects arise from qualitatively very different phenomena (arrival of one extremely big job (or) two big jobs). Then there is a phase transition that occurs when the incoming jobs have infinite variance. In that case, only one of these effects dominate the tail asymptotics, the one involving arrival of one extremely big job.

The talk assumes only basic probability and not any prior knowledge on queuing systems.