Design and control of a large call center: Asymptotic analysis using two-scale fluid limits

Assaf Zeevi
Columbia Graduate School of Business
Uris Hall, Room 406
Columbia University
New York, NY 10027, USA

Abstract

We consider a stochastic model of a call center with m customer classes and r server pools. The model is one with doubly stochastic arrivals, which means that the m-vector of instantaneous arrival rates is allowed to vary both temporally and stochastically. Two levels of call center management are considered:

(i) A *static design problem* which determines staffing levels in the r server pools; and

(ii) a *dynamic control problem* whose solution determines the real-time assignment of incoming customer calls to servers.

The system manager's objective is to minimize the sum of personnel costs associated with staffing and abandonment penalties; the latter are incurred by customers abandoning their calls if forced to wait too long before commencement of service.

We consider a limiting parameter regime that is natural for large-scale call centers with rapid turnover. Mathematically this is expressed by means of a two-scale acceleration of arrival rates and service/abandonment rates, resulting in simplified limiting dynamics. For this parameter regime we prove an asymptotic lower bound on expected total cost, which uses a strikingly simple distillation of the original system data. We then propose a method for staffing and dynamic control based on linear programming (LP), and show that it

achieves the asymptotic lower bound on expected total cost; in that sense the proposed method is *asymptotically optimal*.

Joint work with Achal Bassamboo and Mike Harrison (both from Stanford GSB), and the material will be related to that covered in Mike Harrison's talk.