Staffing and Routing in large-scale service systems

Mor Armony
Stern School of Business
New York University
44 West 4th Street
New York, NY 10012, USA

Abstract

Modern service systems thrive to provide customers with personalized service, which is customized to the customers needs. Recent trends include self selecting marker segmentation, multi-lingual customer support, customized cross-sales offerings, to name a few. With this growing level of service customization, the variety of services provided by any given organization is increasingly high. This requires service personnel with a large skill set. To avoid having to keep a very large group of customer service representatives that would cover the different skills, most organizations realize the importance of cross-training and server flexibility. However, to take full advantage of these high flexibility levels, one needs to make efficient customer-server assignments and sensible staffing and cross-training decisions. This latter challenge is now receiving increasing attention, and where this work's contribution lies.

Three main issues are typically addressed when dealing with the operations management of large-scale service systems. Given a forecast of the customers' arrival rates and their service requirements, these issues are: design, staffing and control. These three problems are all interrelated and should, therefore, be discussed in conjunction with one another. Yet, because of the complexity involved in addressing all these three combined, they are typically addressed hierarchically and unilaterally in the literature.

Even when one addresses the three issues separately, a general solution for all possible system configurations is yet to be achieved. Instead, we approach the problem by studying relatively simple models in order to gain insights to the more general model. The models we focus on in this work are the V-design and the $\wedge$-design. The V-design has multiple customer classes and a single server type. The $\wedge$-design, on the other hand, has homogeneous customers, but the servers are heterogeneous with multiple server types that have full overlap of their skills, but differ in the speed in which they serve the customers. With respect to these two designs we ask the following two questions:

1. Given a fixed set of servers, how to assign customers to servers so as to optimize system performance, and
2. How many servers (of each type) are required in order to minimize staffing costs while maintaining pre-specified performance goals.

We address these two questions by first characterizing simple scheduling schemes which are asymptotically optimal as the arrival rates and the number of servers increase to infinity. We then identify a simple form for an asymptotic feasible region. This region is the set of all staffing vectors that can obtain a pre-specified performance level, asymptotically as the arrival rate grows large. Finally, the asymptotic optimality of the staffing vector that minimizes the staffing costs within the asymptotically feasible region is established for a range of cost functions.

The asymptotic framework considered in this work is the many-server-traffic regime, first introduced by Halfin and Whitt. We refer to this regime as the QED (Quality and Efficiency Driven) regime. Systems that operate in QED regime enjoy an unusual combination of high efficiencies together with high quality of service.

Based on joint work with Avi Mandelbaum and Itay Gurvich.