Chapter 2
Modeling and Optimization of Complex Service Delivery Systems

Yixin Diao
IBM T. J. Watson Research Center, USA

Aliza Heching
IBM T. J. Watson Research Center, USA

ABSTRACT
The services delivery business is highly dynamic and highly competitive, with thin profit margins. Strict service quality targets coupled with highly variable service request arrival patterns and ever increasing cost containment targets make it challenging for a service delivery provider to deliver consistent quality and remain profitable. In this chapter we present a modeling and optimization framework that provides a solution to this challenge. A discrete event simulation model is used as a proxy for the real world service delivery environment, containing the interaction among dynamic customer workload and strict service level constraints. Afterwards, an optimization model provides recommended staffing levels while considering the skills required to respond to different types of service requests, and the shift schedules that the service agents must follow. We demonstrate the applicability of the modeling and optimization framework in a large IT services delivery environment.

INTRODUCTION
The services sector represents a growing fraction of the U.S. and world economy. In fact, recent data from the U.S. Bureau of Labor Statistics reveals that services represents more than 70% of the total industry output (as measured in U.S. dollars) and more than 75% of the U.S. labor force is employed in the services sector. While employment in manufacturing has declined steadily since 1990, the number of employees and jobs in the services sector has grown and is anticipated to continue to increase in the upcoming years. This growth in the services sector has resulted in significant challenges to the viability and profitability of incumbent service providers, due to the growing number of competitors in the mar-

DOI: 10.4018/978-1-4666-8496-6.ch002
ket. The new entrants challenge incumbent providers, often providing richer and more diverse offerings at lower cost. Customers have thus come to expect price reductions from service providers at the same high quality of service. Service providers must therefore seek ways to improve efficiency and reduce the cost of delivery while maintaining or increasing the quality of service. The cost of human capital is the most significant cost faced by a service provider. Thus it is critical for service organizations to invest in methods that support accurate determination of the number of resources with different skills that are required at different times and in different locations from which customers are serviced.

Resource requirement planning is a well-studied problem in the Operations Research literature. However, this problem has primarily been studied in the contexts where the resource being managed is parts or objects whose attributes do not change dramatically over the planning horizon. Even perishable items, whose attributes or whose capacity to service demand change over the planning horizon, differ from resource planning with human capacity. For example, people learn and acquire new skills, lose skills that are not used, and have service rates that may be impacted by the composition of their workload or by the sequence in which they process different types of requests. Thus, new methods are being developed to study and optimize service systems, beyond the methods that have been developed for traditional manufacturing settings.

The methods that we describe falls in the area of optimal staffing with skills based routing (“SBR”), where customer requests arrive with specific skill requirements and are subsequently serviced by agents with corresponding skills. The SBR problem is known to be analytically complex with limited theoretical results. Gans et al. (2003) and Aksin et al. (2007) provide detailed surveys of the analytical approaches (in the form of analytical queueing methods) that have been undertaken. The most common approaches are to either simplify the topology of the delivery network or simplify the routing schemes. However, none of them are desirable in a service delivery environment where both the network and the routing schemes are complex and the service providers are seeking practical solutions but not conceptual guidance.

An alternative solution to the SBR problem is the simulation-based approach. Simulation allows capturing all of the intricacies of the system environment including the details of how requests for service arrive, the details of the service delivery process, and any rules for how requests are serviced. A common model in simulation-based approach is to adopt a two stage method wherein optimization is used to generate a “starting solution” and simulation is used to evaluate real system feasibility (e.g., service level attainment) of this analytical model suggested solution. A number of papers have begun to address the question of optimal staffing for service systems. Atlason et al. (2008) considers a multi-period problem of determining optimal staffing levels while meeting service level requirements. They solve a sample average approximation of the problem using a simulation based analytic center cutting plane method and assuming that the service level functions are pseudoconcave. Cezik et al. (2008) extends this approach by applying it to large problem instances and developing heuristic methods to handle the numerical challenges that arise. Feldman et al. (2010) uses stochastic approximation to determine optimal staffing levels, assuming that the service level functions are convex in the staffing levels. They consider two model formulations, one in which the service levels are strict constraints and the second in which the service levels enter as costs in the objective function, and use simulation to evaluate service level attainment. Robbins et al. (2008) considers a two stage approach for determining optimal staffing levels in a call center environment. In the first stage they solve for the staffing levels by using the per period attainment as an approximation for the true service level attainment. In the second stage, the simulation is used to evaluate true system performance and service level attainment. Bouzada (2009) describes the use of simulation to determine optimal staffing levels in a call center environment. The