Performance Evaluation 80 (2014) 63–81

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Optimal capacity management and planning in services delivery centers

Aliza R. Heching, Mark S. Squillante *

Mathematical Sciences Department, IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA

ARTICLE INFO

Article history:
Available online 11 March 2014

Keywords:
Human server systems
Services delivery centers
Simulation optimization
Stochastic modeling and analysis
Stochastic optimization

ABSTRACT

This paper considers human server systems of queues that arise within the information technology services industry. We develop a two-phase stochastic optimization solution approach to effectively and efficiently address the capacity management and planning processes of information technology services delivery centers. A large collection of numerical experiments of real-world human server system environments investigates various issues of both theoretical and practical interest, quantifying the significant benefits of our approach as well as evaluating the financial-performance trade-offs often encountered in practice.

1. Introduction

System efficiency and process improvement have been studied extensively in computer systems, communication networks and other settings where non-human resources are used for service delivery. Significant research effort has been dedicated to the study and optimal design of these systems. Of growing business concern and research interest are innovative methods to evaluate system behavior and improve system performance in human capacity services systems; see, e.g., [1]. Such human capacity systems are characterized by the presence of human servers who manage the queues of customer requests and serve these requests. Customer requests are grouped into classes based on various attributes, which can include service performance guarantees and the skills required to respond to the request.

The human servers, or agents, often have different skills and different levels of experience and expertise within these skills that restrict the request classes they may serve or impact the rate at which they serve these request classes. Agents are typically grouped into teams based on various attributes such as customers supported, similarity of agent skills, and geographic location. Agent teams are associated with physical or virtual services delivery locations (SDLs) from which services are provided to customers. An SDL may support a subset of services, specializing in a subset of skills, or support a broad range of services. Services delivery providers (SDPs) offer service support from one or more potentially globally distributed SDLs. A services delivery center (SDC) represents a collection of constituent SDLs such that the SDLs comprising a single SDC may share resources and workloads, have common processes, operate as a single profit-and-loss center, or be related in other ways resulting in coordinated decision-making across the SDLs. Examples of such SDCs abound and cover a broad array of services areas including healthcare delivery, information technology delivery, and food services.

Similar to systems with non-human servers, of key operational and strategic interest for SDC environments is how to improve service delivery performance and reduce total delivery cost (driven largely by capacity staffing costs). In addition to skills matching, SDCs have some unique constraints introduced by the behavioral dynamics that are not present in

* Corresponding author. Tel.: +1 914 945 3360.
E-mail address: mss@us.ibm.com (M.S. Squillante).

http://dx.doi.org/10.1016/j.peva.2014.01.003
0166-5316/

non-human server systems. For example, restrictions such as shift schedules, legal limitations on agent utilization (often geography specific), training and learning effects, and fatigue effects must be considered. Moreover, contractual constraints may place restrictions on the minimum number of agents that must be available during various hours of the day as well as how agents may be shared across different customers (e.g., due to privacy concerns). SDCs where the end-customer directly observes agent behavior have additional constraints driven by expectations relating to customer experience.

Given the foregoing complexities (described in more detail below), simulation-based optimization is the primary solution approach for capacity management and planning of a wide range of SDC environments. The advantages of simulation-based optimization concern accuracy and solution quality, including the ability to use a detailed stochastic performance model that captures all of the characteristics and complexities of real-world SDCs and the ability to determine a high-quality optimal solution within the context of this high-fidelity stochastic model. The disadvantages of simulation-based optimization, however, concern the prohibitive costs in both time and resources required to obtain optimal solutions in practice. There are two general categories of simulation-based optimization techniques: (1) a broad spectrum of metaheuristics, such as tabu or scatter search, to control a sequence of simulation runs to find an optimal solution (e.g., see [2] and [3, Chapter 20]); (2) methods that directly solve the problem with a more rigorous mathematical foundation, such as stochastic approximation algorithms (e.g., see [4, Chapter 8] and [3, Chapter 19]). Although the latter category has a strong theoretical foundation in the case of continuous decision variables, the underlying theory in the presence of discrete decision variables is far less well understood [4]. Since the agent team capacities of SDCs are integer valued, the metaheuristics-based approach, employed in nearly all major simulation software products that support optimization (e.g., offerings from AnyLogic and Arena), is the dominant solution approach for capacity management and planning in many SDC environments.

The diverse costs in both time and resources of simulation-based optimization are particularly prohibitive in certain SDC environments (as demonstrated and quantified in Section 4). Our objective in this study is to develop a stochastic optimization solution approach for capacity management and planning that provides the advantages of simulation-based optimization while eliminating its disadvantages, within the context of a specific class of motivating SDC environments. We then apply our solution approach to support case studies of capacity management and planning decision-making in various real-world SDC environments.

1.1. Services delivery centers

While most SDCs have received far too little attention, the class of call center environments (CCEs) has been the focus of many different research studies. Aspects of CCEs include distinguishing attributes such as fully flexible agents, high-volume workloads, relatively short task processing times, relatively simple tasks, and tasks that are highly repetitive in nature; see, e.g., [5]. Several studies have considered diverse stochastic models of the capacity staffing problem in CCEs with skills-based routing. For example, Gurvich and Whitt [6] analyze a policy for assigning service requests to agent teams based on state-dependent thresholds for team idleness and service class queue length. Gurvich et al. [7] propose a formulation of the skills-based routing problem under stationary arrival rates by converting mean performance constraints to chance-based constraints such that a service delivery manager can select the risk of failing to meet the former constraints. Another body of work has attempted to more closely incorporate the complexities of various operational decisions in real-world CCEs using a wide range of approaches that combine simulation and optimization techniques. For example, Atlason et al. [8] solve a sample-mean approximation of the capacity staffing problem using a simulation-based analytic center cutting-plane method. Feldman and Mandelbaum [9] employ a stochastic approximation approach to determine optimal capacity staffing, where service levels (SLs) are modeled in the constraints or the objective and simulation is used to evaluate SL attainment.

Although CCEs have received a great deal of attention in the research literature, the class of human server systems arising in information technology (IT) SDC environments has received far less research attention. Moreover, there are many fundamental differences between CCEs and the comparatively understudied IT SDCs motivating our present study. This includes the degree of agent work-shift flexibility, where CCEs tend to employ greater flexibility in agent work-shift schedules than in IT SDCs to accommodate significantly higher degrees of nonstationary workload arrival patterns and significantly higher volumes of workload intensity. CCEs also tend to have a relatively lower fragmentation in skills required to respond to the different classes of requests, whereas IT SDCs tend to require greater expertise and deeper skills relative to the volume of requests for each of the different skills. This is one reason why many classes of requests require significantly more time to resolve by agents in an IT SDC than in a CCE.

We shall primarily focus on IT SDCs in this paper, though our general methodology can be applied to a wide variety of SDC environments. Arriving customer requests are tagged with attributes that include the request class, type of customer, breadth and depth of skills required to serve the request, geographic location from which the request was generated, and urgency of the request. The urgency of requests guides the order in which agent teams resolve the requests, where urgency and other factors (e.g., business needs and the criticality of supported systems) associated with each request class dictate how the request classes are prioritized for service. These attributes in combination further dictate the performance and quality guarantees associated with customer requests, which can involve system availability, sojourn time (total time in system), waiting time (total time until first touched by an agent), and residence time (difference between sojourn and waiting times). Such guarantees are provided in the form of service level agreements (SLAs), representing contractual agreements between the SDP and the customer that define the service performance the SDP must deliver to the customer. An SDP typically has multiple SLAs in place with a single customer, each of which specifies the scope, time frame, target and