Computers and Operations Research 94 (2018) 65–75

Contents lists available at ScienceDirect

Computers and Operations Research

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

Service system design with immobile servers, stochastic demand and concave-cost capacity selection

Samir Elhedhli[a,*], Yan Wang[a], Ahmed Saif[b]

[a] Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada
[b] Department of Industrial Engineering, Dalhousie University, 5269 Morris Street, Halifax, NS B3J 1B6, Canada

ARTICLE INFO

Article history:
Received 5 July 2016
Revised 5 January 2018
Accepted 27 January 2018
Available online 3 February 2018

Keywords:
Service system design
Economies-of-scale
M/M/1
Second-order cone programming
Special order sets of type 2 (SOS2)
Lagrangian relaxation

ABSTRACT

The service system design problem is a location-allocation problem with service quality considerations that is often modeled as a network of M/M/1 queues to minimize facility setup, customer access, and waiting costs. Traditionally, capacity decisions are either ignored or modeled as a selection among discrete capacity levels. In this work, we study the general continuous capacity case and account for economies-of-scale in its cost through an increasing concave function. We focus on the special square-root case that has been shown to model capacity in terms of the number of servers needed under Poisson arrivals and exponential service times. The problem is formulated as a mixed-integer nonlinear program with concave and convex terms in the objective function. Two novel resolution approaches are proposed: In the first, the problem is reformulated as a mixed-integer quadratic program with fourth-degree polynomial equality constraints. These constraints and the quadratic objective function are approximated using piecewise-linear segments. In the second, we use Lagrangian relaxation to decompose the problem and reformulate the subproblems as second-order cone programs that are solved at multiple utilization levels. The Lagrangian multipliers are updated using a cutting-plane method and a feasible solution is obtained by solving the corresponding set-covering formulation. The solution approaches are tested and compared. The linearization approach provides high quality solutions within short computational times for small instances and lower accuracy; whereas the Lagrangian approach scales well as size increases.

1. Introduction

Service time remains one of the primary criteria by which customers assess service quality, often measured by the degree to which the service meets customer requirements and expectations (Montgomery, 2013). Today's customers are increasingly impatient and have high expectations for service time. Therefore, organizations will often locate facilities close to demand sources and install sufficient service capacity to be able to serve customers promptly and reliably. Pressure to reduce service cost, however, is pushing towards capacity pooling and demand consolidation to take advantage of economies-of-scale, which may require few, large, and distant facilities. To minimize the total service cost, it is imperative to balance the trade-off between capacity, access and customer waiting costs.

Early work on the service system design problem includes the models and solution heuristics proposed by Amiri (1997, 1998). The same class of problems, also called "facility location problems with stochastic demand and congestion", was studied by Berman and Krass (2002). Wang et al. (2002) studied a facility location problem with stochastic demand and immobile servers that aims to minimize the customers' travelling and waiting costs, with limits on the number of facilities opened and the waiting time. Due to the complexity of the problem, heuristics and approximate solution methods were proposed, until Elhedhli (2006) developed an exact solution method based on outer approximation that can handle real-size instances of the problem efficiently. Syam (2008) considered multiple servers and priorities and modeled each service center as an independent M/M/c queue. The resulting nonlinear model was first linearized and a Lagrangian relaxation with a subgradient optimization was then used. Castillo et al. (2009) addressed a service system design problem with fixed servers, stochastic demand and congestion. They considered two scenarios: single-server centers in which the service rate is to be determined and multiple-server centers in which the number of servers in each center is unknown. In contrast to the inelastic arrival rate assumption made in the previous models, Aboolian et al. (2012) formulated a profit-maximizing service system design problem that explicitly accounts

* Corresponding author.
E-mail addresses: elhedhli@uwaterloo.ca (S. Elhedhli), y13wang@uwaterloo.ca (Y. Wang), asaif@dal.ca (A. Saif).

https://doi.org/10.1016/j.cor.2018.01.019
0305-0548/

for demand elasticity with respect to travel distance and congestion delays. Recently, Vidyarthi and Jayaswal (2014) modeled the service center design problem under Poisson demand arrivals and general service time distributions as a network of independent M/G/1 queues.

In all of the aforementioned models with capacity decisions, the capacity cost was assumed to increase linearly as a function of service rate or the number of servers. In reality, however, capacity costs often benefit from economies-of-scale. Concave cost functions resulting from economies-of-scale appeared frequently in facility location models, but not in service system design. This is probably due to the challenging structure of the resulting models. In the former, they typically have concave objectives, whereas in the latter they have both concave and convex objectives. Classical facility location problems with concave costs include the work of Feldman et al. (1966), Florian and Klein (1971) and Soland (1974), to name a few. More recently, Holmberg and Tuy (1999) studied a production-transportation problem with stochastic demand and concave production costs. They reduced the problem to a difference of convex functions (d.c.) optimization problem in a much smaller space, then solved it using branch-and-bound. Dasci and Verter (2001) formulated a facility location and technology acquisition problem as a mixed-integer concave minimization problem and developed a solution approach based on progressive piecewise linear underestimation. A Lagrangian heuristic was proposed in Saif and Elhedhli (2016b) to solve the same problem. Dupont (2008) considered a concave-cost facility location problem, studied the properties of its optimal solution and proposed a branch-and-bound algorithm. Baumgartner et al. (2012) developed a tri-echelon multi-product supply chain design model with economies-of-scale and transport frequencies, and proposed a successive linearization heuristic to solve it. Saif and Elhedhli (2016a), studied a cold supply chain design problem with economies-of-scale and proposed an exact solution approach combining Lagrangian decomposition, simulation-optimization, and branch-and-bound.

Empirical studies provide strong evidence for the existence of economies-of-scale in service systems. For instance, Doukas and Switzer (1991) and Zardkoohi and Kolari (1994) found that economies-of-scale are present in bank branches in Canada and Finland, respectively. Filippini and Zola (2005) reported a similar finding for postal services in Switzerland. In healthcare, the effect of economies-of-scale is well-studied and documented (see, for example, Pope and Burge (1996) and Preyra and Pink (2006)). On the other hand, queuing models are widely used in practice to study service centers in many areas such as healthcare (Gupta, 2013), call centers (Koole and Mandelbaum, 2002) and logistics (Benjaafar et al., 2008). M/M/1 queues, in particular, are widely used and are believed to provide a reasonable approximation for several sophisticated models, especially in large networks.

In this paper, we consider a service system design problem with immobile servers, stochastic demand and capacity selection under concave costs. Unlike previous work in which capacity and service rates are selected from a finite set of levels, we explicitly incorporate server capacity as a continuous decision variable with a concave cost structure to account for economies-of-scale. This approach provides a more realistic modelling framework in many practical settings. Although we limit this work to a square root function, the analysis extends to any concave increasing function. The square-root function has been validated for call center staffing through the famous square-root staffing formula (Whitt, 2007). The service system we consider falls under the general class of location problems with stochastic demand and congestion for which service is standard across facilities, customers are assigned centrally, and utility is primarily influenced by access costs and waiting time (Berman and Krass, 2002).

The problem is formulated as a nonlinear mixed-integer program with linear constraints and an objective function that has both convex and concave terms. Two solution approaches are proposed; The first starts by reformulating the model as a fourth-degree polynomial program, for which a piecewise linearization based on SOS2 constraints is used; The second uses Lagrangian relaxation to decompose the problem by service centers. The resulting subproblems are reformulated and solved as mixed-integer second-order cone programs (MISOCP) when some variables are fixed. A cutting plane method is used to find the Lagrangian bound and feasible solutions are constructed from the subproblem solutions by solving the corresponding set-covering formulation. The proposed solution approaches are first demonstrated on a small example, followed by extensive numerical testing on realistic instances.

Although the use of a square-root function may seem restrictive, we offer a motivation and a justification for this selection. First, the well-known and widely-implemented square-root staffing rule (Whitt, 2007) stipulates that the optimal service capacity is proportional to the square root of total demand for service centers that operate under the Quality-and-Efficiency-Driven regime (QED) characterized by remarkably high levels of both quality and efficiency (Halfin and Whitt, 1981). Second, we hope that the modelling framework and the proposed solution methodologies for the special square-root case will open the door for an approach that can handle the general case. In particular, the reformulation to a polynomial program, which is one of the main contributions of this work, is generalizable (see Appendix A). The SOCP approach, however, is not generalizable but can be used as an approximation.

The rest of this paper is organized as follows: Section 2 discusses the problem formulation. Section 3 presents a new model and studies the properties of the objective function. Sections 4 and 5 describe the SOS2 piecewise linearization and the SOCP-based Lagrangian solution approaches, respectively. An illustrative example is provided in Section 6. Numerical testing and results are discussed in Section 7. Conclusions are provided in Section 8.

2. Problem formulation

In this section, we introduce the service system design problem with immobile servers, stochastic demand, and concave-cost capacity selection and formulate it as a mixed-integer nonlinear programming (MINLP) problem. Consider a set of customers indexed by $i \in I$ and a set of potential facility locations indexed by $j \in J$. Customer demand can be modeled as a Poisson process with rate λ_i. Each customer is assigned to a single facility. The cost of serving a unit demand of customer i from facility j is c_{ij}. The service rate (*i.e.*, capacity) of facilities is assumed finite with a mean of μ_j, which is a decision variable. There is a linear response time cost of t per unit time per customer that penalizes the total time customers spend in the system. The setup cost of a facility is an increasing concave function $f_j(\mu_j)$ of its capacity to capture economies-of-scale. For example, it could take the form $a\mu^b$, where $a > 0$ and $0 < b < 1$. The service system design problem aims to determine the location and capacity of facilities and the assignment of customers to facilities in order to satisfy the demand at the minimum possible cost. The total cost is composed of the assignment cost, the response time cost and, in this work, a concave capacity cost. To formulate the problem, we introduce the following decision variables:

$$x_{ij} = \begin{cases} 1 & \text{if customer } i \text{ is assigned to facility } j, \ i \in I, \ j \in J. \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_j = \text{service rate/capacity of facility } j, \ j \in J.$$