M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 10, No. 3, Summer 2008, pp. 484–505
ISSN 1523-4614 | EISSN 1526-5498 | 08 | 1003 | 0484

informs

DOI 10.1287/msom.1070.0182

Facility Location with Stochastic Demand and Constraints on Waiting Time

Opher Baron, Oded Berman, Dmitry Krass
Joseph L. Rotman School of Management, University of Toronto, Toronto, Ontario, Canada M5S 3E6
{baron@rotman.utoronto.ca, berman@rotman.utoronto.ca, krass@rotman.utoronto.ca}

We analyze the problem of optimal location of a set of facilities in the presence of stochastic demand and congestion. Customers travel to the closest facility to obtain service; the problem is to determine the number, locations, and capacity of the facilities. Under rather general assumptions (spatially distributed continuous demand, general arrival and service processes, and nonlinear location and capacity costs) we show that the problem can be decomposed, and construct an efficient optimization algorithm. The analysis yields several insights, including the importance of equitable facility configurations (EFCs), the behavior of optimal and near-optimal capacities, and robust class of solutions that can be constructed for this problem.

Key words: facility location; stochastic demand; queueing; service level
History: Received: August 3, 2006; accepted: March 23, 2007. Published online in *Articles in Advance* January 4, 2008.

1. Introduction

Two key issues that must be addressed when locating service facilities are ensuring convenience and ensuring sufficient capacity. From the customer's point of view, convenience means not having to travel too far to obtain service, and sufficient capacity means being able to obtain service without an overly long wait at the facility. From the decision maker's point of view, these two aspects are related because the same budget could be used to locate a larger number of smaller-capacity facilities (focusing on convenience) or a smaller number of larger-capacity facilities (enjoying capacity-pooling effects and thus focusing on sufficient capacity). These issues arise in a variety of applications, ranging from location of public facilities such as hospitals, medical clinics, or government offices to location of private facilities such as stores, service centers, and warehouses.

This problem, which is rather fundamental in the facility location theory, has received a significant amount of attention in the literature; the interplay of locational and stochastic (queuing) aspects makes it particularly challenging. The problem belongs to the general class of location problems with stochastic demand and congestion with fixed servers, reviewed in Berman and Krass (2002). The study of models of this type originated with Marianov and Serra (1998). For further discussion of this class of problems, we refer the reader to Berman et al. (2006), Marianov and Rios (2001), Marianov and Serra (2002), Wang et al. (2002), as well as references in Berman and Krass (2002). Due to the complexity of the underlying problem, all papers listed above make very strong assumptions: The demand is assumed to be discrete. Either the number or the capacity of the facilities (or both) are assumed to be fixed. The set of potential facility locations is assumed to be discrete and finite. The demand arrival process is assumed to be Poisson, and the service process is usually assumed to be exponential.

We will consider the problem under a much more general setting: allowing for general spatial distribution of the demand, arrival, and service processes, and without fixing in advance either the number or the capacity of the facilities, or their potential locations. In fact, our model belongs to the class of continuous facility models with continuously distributed demand—to the best of our knowledge, no prior work on multifacility stochastic location problems has been attempted in this setting. Specifically, we assume that customer demand is distributed over a certain space (special cases include demand distributed over

a line, planar region, or a network) and that both the demand and service processes are stochastic. As a result, congestion may occur at the facilities. Two types of constraints are imposed to ensure adequate service at the facilities: (a) the maximum travel distance constraint that ensures convenience, and (b) the service-level constraint that limits the waiting times at the facilities. Costs are incurred for locating new facilities and for adding capacity to the facilities. Initially, we assume that all the facilities must have identical capacity and that customers always travel to the closest facility to obtain service; both of these assumptions are relaxed later. We seek to simultaneously optimize three types of decision variables: (a) the number of facilities to be located, (b) the location of the facilities, and (c) the service capacity of each facility. We will refer to this problem as the Stochastic Capacity and Facility Location Problem (SCFLP).

We show that, under some mild assumptions, the SCFLP with a given number of facilities M can be decomposed into two subproblems: the deterministic Equitable Location Problem (ELP), which seeks to find facility locations that ensure that consumer demand is fairly distributed among the facilities, and the stochastic single-facility capacity determination problem, which computes the minimal required capacity at the busiest facility to ensure that the service-level constraint is met. By conducting a linear search with respect to the number of facilities M, the optimal number, locations, and capacities of the facilities can be determined. While the capacity determination problem cannot be solved in closed form for general service distributions, we develop efficient approximations of the required capacity using large-deviation results.

A key concept of this paper is an EFC—a location vector that ensures identical customer demand at all facilities. We show that if a feasible EFC location vector exists, it is optimal. We present conditions for the existence and feasibility of EFC for the case where demand is distributed over a line segment.

A series of computational experiments are conducted to assess the solvability of the SCFLP (and the ELP) and to develop managerial insights on the properties of the optimal solutions. We observe that, in the majority of cases, the optimization procedures select the minimal number of facilities for which a feasible EFC location vector exists. This leads to a simple and robust heuristic decision rule for SCFLP that is independent of the cost structure.

The paper is organized as follows: The SCFLP model and the required notation are formally defined in §2. In §3, we present key structural results that allow us to decompose the SCFLP, and the solution algorithm. The proofs of these and all other results that are not straightforward appear in the technical appendix available online.[1] Section 4 deals with the solution of the ELP and the existence and feasibility of EFC vectors. Section 5 presents computational experiments and discusses properties of optimal solutions and a heuristic decision rule. Section 6 describes several direct extensions and generalizations of the SCFLP model. In addition to a summary, §7 contains concluding remarks and directions for future research.

2. SCFLP Model

We formulate the SCFLP over a bounded space $P \subset R^N$ equipped with some norm $\|\cdot\|$. Note that since a graph can always be embedded in R^3, this includes problems on a network with the shortest distance norm.

On this space, the demand for service at any point $\mathbf{x} \in P$ (where small bold letters denote vectors) is assumed to follow a general renewal process. We assume that at each arrival epoch there is a single arrival with probability 1, and that the total demand over the space is Λ with $0 < \Lambda < \infty$. We denote the number of calls for service at $\mathbf{x}$ up to time t by $N_t(\mathbf{x})$ and the average demand arrival rate at $\mathbf{x}$ by $\lim_{t\to\infty} N_t(\mathbf{x})/t = \Lambda dF_\Lambda(\mathbf{x})$. Thus, the proportion of demand at any point $\mathbf{x} \in P$ is given by $dF_\Lambda(\mathbf{x})$ such that the Lebesgue integral is well defined and $\int_{\mathbf{x}\in P} dF_\Lambda(\mathbf{x}) = 1$. Finally, if $\lim_{\epsilon\to 0} \int_x^{x+\epsilon} dF_\Lambda(x) > 0$, we require that the demand at x follows a Poisson process, but we allow demand to follow a general renewal process whenever $\lim_{\epsilon\to 0} \int_x^{x+\epsilon} dF_\Lambda(\mathbf{x}) = 0$. Observe that $\lim_{\epsilon\to 0} \int_x^{x+\epsilon} dF_\Lambda(x) = 0$ implies that the general renewal process has an infinitesimal rate. Mathematically, the latter requirement is that the interrenewal time at $\mathbf{x}$

[1] An online appendix to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

with $\lim_{\epsilon\to 0}\int_x^{x+\epsilon} dF_\Lambda(x) = 0$ is distributed with a cumulative distribution function (CDF) $F_{\mathbf{x}}(t)$ such that for any $\varepsilon > 0$ and $t > 0$ we have $F_{\mathbf{x}}(t) \le \varepsilon$.

We consider the location of M identical service facilities, indexed $i = 1, \ldots, M$, where M is a decision variable. The cost for locating M facilities is $C_M > 0$ dollars. We assume $C_j \le C_{j+1}\ \forall j = 1, 2, \ldots$ and $\lim_{M\to\infty} C_M = \infty$.

We assume first come–first serve (FCFS) service discipline and an infinite buffer size. Service requirements are assumed to be independent and identically distributed (i.i.d.) with a CDF $F_S(s)$, a well-defined moment generating function (MGF) $G_S(\gamma)$ (service time), and a mean $E_S(S) = 1$. This assumption is made with no loss of generality since it simply rescales time.

There are two common methods to model flexible capacity of a queuing system. One is to assume multiple parallel servers each with a given service rate $\hat{\mu} = 1$ (in accordance with our assumption that $ES = 1$ and without loss of generality). The decision variable is N, for the number of servers. For the second method, we assume a single server with a flexible service rate μ. The control variable is this server's capacity μ. In this case, the mean service time is $1/\mu$, and it is easy to show that the CDF and the MGF of the service time are $F_S(\mu s)$ and $G_S(\gamma/\mu)$ for $\gamma > 0$, respectively. If the system utilization is reasonably high, a system with $N = \mu$ parallel servers will perform similarly to a single server with capacity μ (assuming integer μ). Thus, the main difference between a single adjustable-capacity server and multiple unit servers is that the system capacity can be adjusted continuously in the former case, and in identical discrete steps in the latter.

While the discrete-capacity model may be more accurate for a simple facility (e.g., a car wash, where adding capacity means adding bays), we regard the continuous-capacity model as more suitable for more complex facilities: For example, it is not clear what a "server" represents in the case of a hospital, where capacity can be adjusted in a variety of ways (through better technology, more nursing support, more examination rooms, as well as more doctors). We consider both types of capacity models in this paper.

When we have M facilities, adding one unit of capacity to each facility costs $c_M > 0$ dollars. One could think of c_M/M as the discounted operation cost of a unit capacity when M facilities are operating; we assume $c_j \le c_{j+1}$ for all j. While all facilities are initially assumed to be identical, this assumption can be relaxed in some cases; see §6 for further discussion.

Let $\mathbf{x}_j \in P$ represent the location of facility j, for $j = 1, \ldots, M$. We assume that collocation is not allowed, i.e., there exists some minimal interfacility distance $\varepsilon > 0$, such that

$$\|\mathbf{x}_i, \mathbf{x}_j\| \ge \varepsilon \quad \forall i, j \in \{1, \ldots, M\},\ i \ne j. \tag{1}$$

To obtain service, customers are assumed to travel, at a fixed velocity v with $0 < v < \infty$, to a closest facility, with ties broken arbitrarily. Let $I_{\mathbf{x}}^j = 1$ if customers from $\mathbf{x}$ are served by the jth facility and $I_{\mathbf{x}}^j = 0$ otherwise. Note that every facility location vector induces a Voronoi partition of $V_1, \ldots, V_M$ of P, where $\mathbf{x} \in V_j$ implies that $\mathbf{x}_j$ is the closest facility to $\mathbf{x}$ for $j = 1, \ldots, M$. The indicator function $I_{\mathbf{x}}^j = 1$ if and only if $\mathbf{x} \in V_j$. We note that a Voronoi partition on a line is trivial and that efficient methods exist for obtaining Voronoi partitions on a plane and higher-dimensional spaces: see, e.g., Aurenhammer (1991) and Suzuki and Okabe (1995). While we believe that the closest assignment rule is the most realistic for many customer service facilities, and assume this rule throughout the paper, we discuss the relaxation of this assumption in §6.

With these definitions, the arrival rate to the jth facility is given by

$$\lambda^{\mathbf{x}_j} = \Lambda \int_{\mathbf{x}\in P} I_{\mathbf{x}}^j \, dF_\Lambda(\mathbf{x}) \quad \forall j = 1, \ldots, M. \tag{2}$$

We assume that the different stochastic elements described above are independent of each other. For example, the renewal processes that describe customer demand for service are independent of each other, of the location of the closest facility, and of the service processes.

We include two types of constraints to ensure adequate customer service: (a) the coverage constraint that requires a facility to be within a certain coverage radius of customer location, and (b) the service-level constraint to ensure that, on average, customers do not wait too long once they arrive at the facility.

Let $r > 0$ be the exogenous coverage radius of a facility. The coverage constraint requires that

$$\min_{j=1,\ldots,M} \|\mathbf{x}_j, \mathbf{x}\| \le r, \quad \forall \mathbf{x} \in P.$$

Alternatively, denote by $R(\mathbf{x}_j)$ the maximum travel distance for customers patronizing facility j (note that $R(\mathbf{x}_j)$ is the radius of the jth Voronoi region V_j). Then the coverage constraints can be rewritten as follows:

$$R(\mathbf{x}_j) = \max_{\mathbf{x}\in P} \|\mathbf{x}, \mathbf{x}_j\| I_{\mathbf{x}}^j \le r, \quad j = 1, \ldots, M. \tag{3}$$

The service-level constraint requires that the probability of waiting in a queue for more than d time units does not exceed α for some finite $d > 0$ and $\alpha \in (0, 1)$. We discuss additional service-level measure in §6. Let W_i^j be the waiting time of the ith customer that arrives to the jth facility. Any control policy that satisfies the service-level constraint must ensure that the steady-state waiting time at the jth facility exists: that is, the arrival rate to each facility is lower than the service rate. We denote the steady-state waiting time at the jth facility by W^j.

Then the service-level constraint can be expressed as follows:

$$P(W^j > d) \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I\{W_i^j > d\} \le \alpha, \quad j = 1, \ldots, M, \tag{4}$$

where $I\{\cdot\}$ is the indicator function.

Let τ be an indicator variable that equals one if a single-server queue is considered and zero if a multiple-server queue is considered. Then, the stochastic capacity and facility location problem (SCFLP) is

$$\min_{M,\mathbf{x},\mu} \; C_M + \tau c_M \mu + (1-\tau) c_M N \tag{5}$$

$$\text{subject to} \quad (1, 3, 4) \quad \text{and}$$

$$\mathbf{x}_j \in P, \quad j = 1, \ldots, M, \; M - \text{positive integer},$$

$$\mu > 0, \quad N - \text{positive integer}.$$

There are several challenges in solving this problem. First, the service-level constraint (4) must be reformulated as a function of the decision variables. Another difficulty is that the decision variable M appears in all constraints as the upper limit of the index j. In fact, it is not at all clear how to solve SCFLP using standard mathematical programming techniques. In the following sections, we show how SCFLP can be decomposed into a series of subproblems, and discuss solution approaches for each subproblem.

3. Decomposition of SCFLP

In this section, we establish important structural results for SCFLP and use them to decompose SCFLP into three subproblems: (a) given the number of facilities M, find optimal facility locations; (b) given M and optimal locations, specify optimal processing capacity μ or N; and (c) find the optimal value of M.

3.1. Equitable Location Problem and Equitable Facility Configuration

We let $W(\lambda, \mu)$ be distributed as the steady-state waiting time in a single-server queue with service rate μ and arrival rate λ whenever such a steady state exists. Similarly, we let $W(\lambda, N)$ be distributed as the steady-state waiting time in an N-server queue whenever such a steady state exists. We define

$$\mu(\lambda) = \inf\{\mu > \lambda \mid P(W(\lambda, \mu) > d) \le \alpha\}, \tag{6}$$

$$N(\lambda) = \min\{N\text{—positive integer} \mid P(W(\lambda, N) > d) \le \alpha\}, \tag{7}$$

$$\lambda(\mu) = \sup\{\lambda > 0 \mid P(W(\lambda, \mu) > d) \le \alpha\}, \quad \text{and} \tag{8}$$

$$\lambda(N) = \sup\{\lambda > 0 \mid P(W(\lambda, N) > d) \le \alpha\}. \tag{9}$$

In §3.2, we show that our model results in a Poisson arrival process to facilities, and consequently that Assumption 1 below holds for our model.

Assumption 1. *In the single-server case, $\mu(\lambda)$ is strictly increasing with λ and $\lambda(\mu)$ is strictly increasing with μ for each facility and each pair of facilities.*

In the multiple-server case, $N(\lambda)$ is increasing with λ and $\lambda(N)$ is strictly increasing with N for each facility and each pair of facilities.

Note that, in general, Assumption 1 may not hold. Indeed, if all facilities operate as a GI/G/N queue, then the "for each facility" part of Assumption 1 might look intuitive from a queueing theory view point. However, because demand follows a general renewal process and the aggregation of that process is not necessarily a renewal process, facilities are not necessarily GI/G/N queues. Moreover, the "for each pair of facilities" part of Assumption 1 is fairly restrictive. Consider two facilities facing general arrival renewal processes with rates λ_1 and $\lambda_2 > \lambda_1$. Observe that if the variance of the arrival process to Facility 1

is higher than that to Facility 2 we might have $\mu(\lambda_1) > \mu(\lambda_2)$, and similarly $N(\lambda_1) > N(\lambda_2)$.

A direct consequence of Assumption 1 and that the per-unit capacity cost c_M is positive is

Corollary 1. *Let $\lambda^{\max}$ denote the highest arrival rate to any facility in an optimal solution of SCFLP. Then, if Assumption 1 holds, in the single-server case the service-level constraint in any facility with arrival rate $\lambda^{\max}$ is active, i.e., $P(W(\lambda^{\max}, \mu) > d) = \alpha$, and in the multiple-server case for any facility with arrival rate $\lambda^{\max}$ there is a unique N such that $P(W(\lambda^{\max}, N) > d) \le \alpha$ and $P(W(\lambda^{\max}, N-1) > d) > \alpha$.*

For a fixed number of facilities M, we use Corollary 1 to replace the stochastic model SCFLP with a simpler deterministic problem. Suppose the total number of facilities M is fixed, and let SCFLP(M) represent the corresponding version of the stochastic location problem defined in §2.

Consider the following optimization problem, which we call the equitable location problem, ELP(M)

$$\begin{aligned} \min \quad & \lambda^{\max} \\ \text{subject to} \quad & (1,3) \quad \text{and} \\ & \lambda^{\max} \ge \lambda^{\mathbf{x}_j}, \quad j = 1, \dots, M \\ & \mathbf{x}_j \in P, \quad j = 1, \dots, M. \end{aligned} \tag{10}$$

The objective of the ELP(M) is to locate M distinct (i.e., separated by at least ε) facilities so that the coverage constraint is satisfied and the demand faced by the *busiest* facility (i.e., the facility with the largest arrival rate) is as small as possible. Minimizing the demand faced by the busiest facility reduces the differences in arrival rates between the busiest and the least busy facilities—hence the name of the problem.

Note that ELP(M) is a deterministic location problem since $\lambda^{\mathbf{x}_j}$ represents the total demand in the region assigned to facility j and does not depend on the behavior of the queuing system. The following result shows the equivalence of ELP(M) and SCFLP(M).

Theorem 1. *Consider $M > 0$.*

1. *ELP(M) is feasible if, and only if, SCFLP(M) is feasible.*

2. *Suppose ELP(M) is feasible and let $\mathbf{x}^*, \lambda^{\max}$ be the optimal location vector and objective function value in ELP(M), respectively. Let $\mu(\lambda^{\max})$ and $N(\lambda^{\max})$ be defined in (6) and (7), respectively. Then, if Assumption 1 is satisfied, $\mathbf{x}^*$ is an optimal location vector for SCFLP(M) leading to an optimal objective function value $z^* = C_M + \tau c_M \mu(\lambda^{\max}) + (1-\tau) c_M N(\lambda^{\max})$.*

3. *Suppose SCFLP(M) is feasible and let $\mathbf{x}^*$ be the optimal location vector and μ^* and N^* be the optimal capacity in the single- and multiple-server case, respectively. Then, if Assumption 1 is satisfied, $\mathbf{x}^*$ is also optimal in ELP(M). Moreover, for the single-server case the optimal objective function value in ELP(M) is $\lambda^*(\mu^*)$ and for the multiple-server case the corresponding value satisfies $\lambda^*(N^* - 1) < \lambda^{\max} \le \lambda^*(N^*)$.*

Thus, for a fixed number of facilities M, the original stochastic model can be replaced by the equivalent deterministic model ELP(M). While the latter can present computational challenges of its own (mostly due to the nonlinearity implicit in the definition of λ^{x_j}), it is relatively easy to solve in some cases, as discussed in §4.

We point out several interesting observations. First, the model ELP(M) is independent of the cost structure of the SCFLP. This indicates that the costs influence only the number of facilities to be located, but not the specific locations of the facilities. (This is, in part, due to our assumption that location-specific costs are similar for all locations or are dominated by the capacity-related costs, or both.) Second, while the equity considerations have to be enforced through separate constraints in most location models, they occur naturally in the SCFLP. We will return to this point in §4.

We also point out that the feasibility of ELP(M) is not assured—it depends on the number of facilities M being large enough to ensure that a facility can be located within distance r of every point in P.

One case where the solution of ELP(M) is trivial is when an equitable facility configuration EFC exists. The latter is defined as follows:

Definition 1. Equitable Facility Configuration (EFC). We say that a location vector $\mathbf{x}$ represents an EFC if the arrival rates to all facilities are the same, i.e., $\lambda^{x_j} = \Lambda/M$ for all $j \in \{1, \dots, M\}$.

The following result now follows immediately since $\sum_j \lambda^{x_j} = \Lambda$ for any feasible location vector $\mathbf{x}$, and thus the value of the maximal summand is minimized by setting all terms to the same value.

Corollary 2. *Suppose a feasible EFC location vector exists in the ELP(M) model. Then this location vector is optimal.*

While it is easy to check whether a certain location vector represents EFC, it may be harder to find a feasible EFC vector (or to determine that one does not exist). These issues are further discussed in §4.

3.2. Arrival Process and Waiting Times

We show in Theorem 2 that the arrival process to each facility is Poisson, and thus each facility can be analyzed as an M/G/N queue. We then establish that Assumption 1 and therefore Theorem 1 hold for our model.

Theorem 2. *The arrival process to facilities in the SCFLP is Poisson.*

Therefore every facility operates as an M/G/N queue, and establishing the "for each" facility part of Assumption 1 will also prove the "for each pair" of facilities part. We now move to proving the strict changes in the waiting time assumed in Assumption 1. We denote the interarrival time of the ith customer to the jth facility by A_i^j for $i = 1, 2, \ldots$. By Theorem 2, A_i^j are i.i.d. and $A_1^j \sim \exp(\lambda^{x_j})$.

We start with the single-server case where Proposition 1 below is somewhat more general than we require. Consider a GI/G/1 queue with an FCFS service discipline, let $A_i \sim F_A(a)$ be the i.i.d. interarrival periods with expectation $E(A) < \infty$, and let $S_i \sim F_S(s)$ be i.i.d. service requirements with expectation $E(S) < \infty$. Suppose that the interarrival and service rates are set to λ and μ, respectively, (such that the interarrival and service times of the ith customer are a_i/λ and s_i/μ, respectively). If $E(A)/\lambda > E(S)/\mu$ a steady-state distribution for the waiting time in this GI/G/1 queue exists and, as before, we let $W(\lambda, \mu)$ be a random variable with this distribution. Then, extending the result from Weber (1983),

Proposition 1. *If either A or S are continuous random variables, then for every λ and μ that satisfy $E(A)/\lambda > E(S)/\mu$ and $P(A/\lambda < S/\mu) > 0$, we have that $W(\lambda, \mu)$ is well defined and for any $d > 0$, $P(W(\lambda, \mu) > d)$ is strictly decreasing with respect to μ and strictly increasing with respect to λ.*

For the multiple-server case, we establish the following result for GI/G/N queue.

Proposition 2. *If either A or S is a continuous random variable, then for every λ and N that satisfy $E(A)/\lambda > E(S)/N$, $P(A/\lambda < S/N) > 0$, and $P(A/\lambda > S) > 0$, we have $W(\lambda, n)$ is well defined for each $n \geq N$ and for any $d > 0$, $P(W(\lambda, N) > d)$ is strictly decreasing with respect to N and strictly increasing with respect to λ.*

Since, by Theorem 2, the arrival processes to queues at each facility are Poisson, and for the single-server case satisfying the service-level constraints (4) require $E(A)/\lambda > 1/\mu$, Proposition 1 is satisfied for any facility. Thus, the strict changes of $\mu(\lambda)$ and $\lambda(\mu)$ assumed in Assumption 1 follow from the strict changes in $P(W(\lambda, \mu) > d)$. For the multiple-server case, Assumption 1 follows from Proposition 2 in a similar manner.

3.3. Determining the Optimal Service Capacity in SCFLP(M)

In this section, we assume that the number of facilities M has been fixed and that the optimal solution to the ELP model defined in the previous subsection is available. We address the question of how to determine the optimal capacity.

3.3.1. Single-Server Case. Here capacity is μ. Let $\lambda^{\max}$ be the optimal solution to ELP for a specific value of M. It follows from Theorem 1 that the optimal service capacity is given by

$$\mu^* = \mu(\lambda^{\max}). \tag{11}$$

When the service time is exponential, the distribution of waiting times is available in closed form, allowing us to solve (11) directly. In this case, each facility operates an M/M/1 queue, and for the facility with the maximal arrival rate we have

$$P(W(\lambda^{\max}, \mu) > d) = (\lambda^{\max}/\mu)\exp(-(\mu - \lambda^{\max})d).$$

Thus, Equation (11) is equivalent to

$$(\lambda^{\max}/\mu)\exp(-(\mu - \lambda^{\max})d) = \alpha. \tag{12}$$

The capacity that satisfies (12) is

$$\mu^* = \frac{LW((d/\alpha)\lambda^{\max}\exp(d\lambda^{\max}))}{d}, \tag{13}$$

where the $LW(x)$ denotes the Lambert W(x) function that satisfies $LW(x)\exp(LW(x)) = x$, and is discussed

in Corless et al. (1996). This function is implemented in several common mathematical packages; there are also standard techniques for evaluating it numerically.[2]

How can Equation (11) be solved in general? Since each facility (in particular the one corresponding to the maximal arrival rate) operates as an $M/G/1$ queue, we can, in principle, determine the MGF for the waiting time $W(\lambda, \mu)$ and then invert it to find the probability distribution of the waiting times, substitute it into the service-level constraint (which we know is active at $\lambda^{\max}$), and then solve for μ to find the value of μ^*. However, this approach may be difficult to implement because, in most cases, the inversion of the MGF above can only be done numerically, and thus the substitution into the service-level constraint may be difficult to execute. For this reason, we present an alternative approach based on large-deviation bounds.

In the following discussion, we focus on the single facility with the arrival rate $\lambda^{\max}$ located at x; therefore, we drop the superscript j. As before, we let A_i, $i = 0, 1, 2, \ldots$ represent the interarrival periods and S_i, $i = 0, 1, 2, \ldots$ the service requirements. Let $S_0 = 0$ and given capacity μ we define $Y_i \equiv S_{i-1}/\mu - A_i$. We recall that any optimal policy to SCFLP(M) will guarantee the existence of a steady-state distribution for the waiting times in each facility, thus it will choose μ such that $\lambda^{\max} = 1/E(A) < \mu$ or $E(Y) = E(S)/\mu - E(A) < 0$. Moreover, if $S_{i-1}/\mu \le A_i$ for every i there is some $\varepsilon > 0$ such that for any $d > 0$ the service-level constraint (4) will also hold for $\mu_\varepsilon = \mu - \varepsilon$. Therefore, any optimal control will choose μ such that $F_Y(0) < 1$.

Assuming that the zeroth customer arrives to an empty server, the embedded waiting time in a G/G/1 queue can be described as a one-sided regulated random walk that is regulated at zero, known also as the Lindley recursion:

$$W_0 = 0 \quad \text{and} \quad W_i = \max\{W_{i-1} + Y_i, 0\} \quad \text{for } i = 1, 2, \ldots. \tag{14}$$

Using (14), we can interpret the event that a customer in a queue with infinite waiting room waits more than d time units as the event that a one-sided regulated random walk with a negative drift crosses a threshold; e.g., see Cohen (1982).

There is extensive literature developing bounds for the threshold-crossing probability of regulated random walks provided that $F_Y(0) < 1$ and that the conjugate point defined below exists. Let $G_Y(\gamma)$ represent the MGF of Y and suppose

$$\gamma^* = \arg\{\gamma > 0 \mid G_Y(\gamma) = 1\}, \tag{15}$$

then γ^* is called the conjugate point of Y. The conjugate point exists whenever there is some γ_0 such that $1 < G_Y(\gamma_0) < \infty$, which is true for most commonly used distributions, including exponential and normal (see, e.g., Chapter 7 of Gallager 1996). Under these assumptions, the service-level constraint (4) can be rewritten as

$$P(W(\lambda^{\max}, \mu) > d) \le e^{-\gamma^* d} = \alpha, \tag{16}$$

where we used the fact that the service-level constraint is tight at the facility corresponding to $\lambda^{\max}$. Using (16), we can write

$$\gamma^* = -\frac{\ln \alpha}{d}. \tag{17}$$

Because the MGF of Y_i at the conjugate point γ^* satisfies (15) (i.e., $G_Y(\gamma^*) = 1$), we have

$$G_S(\gamma^*/\mu)\frac{\lambda^{\max}}{\lambda^{\max} + \gamma^*} = 1. \tag{18}$$

Therefore, the (approximate) optimal service capacity is given by

$$\mu^* = \arg\left\{G_S\left(-\frac{\ln \alpha}{d\mu}\right)\frac{\lambda^{\max}}{\lambda^{\max} - \ln(\alpha)/d} = 1\right\}, \tag{19}$$

which can be solved numerically.

We note that further improvements can be made by tightening the bound in (16) using the techniques of Ross (1974). However, as shown in §5, the bounds above already provide very accurate estimates of the required capacity of the facilities, thus further tightening is unlikely to result in major improvements. Thus, in view of the lack of generality of the tighter bounds and the minor improvements they could bring, we chose to use the bounds above in the remainder of the paper.

[2] The zero branch of the Lambert W(x) is needed because it is the only branch with positive real values for positive arguments of the function.

It is interesting to analyze the optimal capacity obtained under the large-deviation approximation (19) when the service is exponential. In this case, $G_S(\delta) = 1/(1-\delta)$. Thus, from (19) we obtain

$$\mu^* = \lambda^{\max} + \gamma^*, \tag{20}$$

implying that the optimal capacity is equal to the arrival rate to the busiest facility plus a safety factor that depends on the required service level.

3.3.2. Multiple-Server Case. Here capacity is N (recalling, $\hat{\mu} = 1$) and we denote by $W(\lambda^{\max}, N)$ a random variable that is distributed as the steady-state waiting time in an M/G/N queue with arrival rate $\lambda^{\max}$. We observe that by (4) and (5) of Abate et al. (1995) for an M/G/N queue we can approximate for $N > \lambda^{\max}$ (under similar conditions on the service distributions as in the single-server case)

$$P(W(\lambda^{\max}, N) > d) \le e^{-\gamma^*(N)d}, \tag{21}$$

where $\gamma^*(N)$ in (21) is given similarly to (15) by

$$\gamma^*(N) = \arg\left\{\gamma > 0 \;\middle|\; G_S\left(\frac{\gamma}{N}\right) G_A(-\gamma) = 1\right\}. \tag{22}$$

Because the arrival rate is Poisson with rate $\lambda^{\max}$, we have, similar to (18), that $\gamma^*(N)$ is the unique positive solution of

$$G_S\left(\frac{\gamma}{N}\right)\frac{\lambda^{\max}}{\lambda^{\max}+\gamma} = 1.$$

Moreover, due to the discrete nature of capacity it is likely that the service-level constraint would not be active at either of the facilities. Because it is economical to reduce the number of servers to the minimal number of facilities that will result in $\gamma^*(N)$ larger than the one in (17), we have

$$N^* = \min\left\{N > \lambda^{\max}, N \text{ integer} \;\middle|\; \gamma^*(N) \ge \frac{-\ln\alpha}{d}\right\}. \tag{23}$$

For example, when service is exponential, using (22) we have

$$\frac{1}{1-\gamma/N} = \frac{\lambda^{\max}+\gamma}{\lambda^{\max}}$$
$$\gamma^*(N) = N - \lambda^{\max}.$$

Thus, we look for the smallest N that holds

$$N - \lambda^{\max} \ge \frac{-\ln\alpha}{d}$$
$$N \ge \lambda^{\max} - \frac{\ln\alpha}{d} \text{ and } N^* = \left\lceil \lambda^{\max} - \frac{\ln\alpha}{d} \right\rceil, \tag{24}$$

where $\lceil x \rceil$ denotes the smallest integer larger than x. Comparing (20) with (24) for exponential service, we observe that the service-level constraint dictates a utilization level of $\rho = \lambda^{\max}/(\lambda^{\max} - (\ln\alpha)/d)$. However, when capacity is continuously adjustable, this utilization can be achieved exactly, whereas the discrete nature of the multiple-server case may lead to a lower utilization level.

For many other service types, $\gamma(N)$ that solves (22) can be found numerically for each $N > \lambda$ and used to find the smallest N such that (23) holds.

3.4. Determining the Optimal Number of Facilities

Suppose that for any value of M we can solve the ELP(M) to find the optimal maximal arrival rate $\lambda^{\max}(M)$ and then apply the results of the previous subsection to find the optimal service capacity $\mu(\lambda^{\max}(M))$ or $N(\lambda^{\max}(M))$. In the current section we discuss how to search for the optimal value of M, thus completing the solution of the original SCFLP model.

For a particular $M > 0$, let

$$Z(M) = C_M + \tau c_M \mu(\lambda^{\max}(M)) + (1-\tau)c_M N(\lambda^{\max}(M));$$

if ELP(M) is not feasible for this M, we set $Z(M) = \infty$. Similarly,

$$Z^E(M) = C_M + \tau c_M \mu(\Lambda/M) + (1-\tau)c_M N(\Lambda/M).$$

Note that when the EFC location vector exists, $Z^E(M)$ provides the corresponding value of the objective function. We will represent the optimal number of facilities with M^*, and use M^E to represent the value of M for which $Z^E(M)$ is minimized. (Clearly, we only consider values of M for which ELP(M) is feasible.) Before presenting the algorithm for determining M^*, we make several useful observations.

Lemma 1.

1. $Z^E(M) \le Z(M)$ *for any* $M > 0$.
2. *Let*

$$M^0 = \min\{M > 0 \mid ELP(M) \text{ is feasible}\}. \tag{25}$$

Then, the problem ELP(M) (*and SCFLP*(M)) *is feasible if, and only if,* $M \ge M^0$.

3. *Suppose SCFLP*(M) *is feasible for a certain* M. *Define*

$$M^1(M) = \arg\min\{M' > M\colon C_{M'} > Z(M)\}. \tag{26}$$

Then $M^* < M^1(M)$.

We note that to find the minimal M^0 for which ELP(M) is feasible in Part 2, it suffices to solve the min-cover problem with radius r over space P, for which several algorithmic approaches are available—see Suzuki and Drezner (1996) and Plastria (2002).

We are now ready to present the algorithm for determining M^* (as well as the optimal location vector and optimal capacity of facilities). In addition to the properties listed above, the algorithm also uses the fact that evaluation of $Z^E(M)$ is cheaper computationally than the evaluation of $Z(M)$, since the latter requires us to solve ELP(M), as well as (19), while only the solution of (19) is required for the former (since $\lambda^{\max} = \Lambda/M$ in this case). We use M^E—the optimal number of facilities when EFC exists—as a starting point for our search, and M^0, $M^U = M^1(M^0)$ as lower and upper bounds, respectively.

Algorithm 1.

Step 1. *Initial Guess*: *Determine* M^E. Compute M^0 by (25) and $M^U = M^1(M^0)$ by (26). Let $M^E = \arg\min_{M^0 \le M < M^U} Z^E(M)$. Set $M^* = M^E, Z^* = Z(M^*)$. Find $M^1(M^*)$ from (26) and set $M^U = M^1(M^*)$.

Step 2. *Main Search; Upper Branch*

Step 2.1. Set $M = M^E + 1$.

Step 2.2. Set $Z_{\mathrm{LB}} = \min_{M \le L < M^U} Z^E(L)$. If $Z_{\mathrm{LB}} \ge Z^*$, proceed to Step 3.

Step 2.3. If $Z(M) < Z^*$ then set $Z^* = Z(M)$, $M^* = M$, $M^U = M^1(M^*)$).

Step 2.4. If $M \ge M^U - 1$ go to Step 3, Else set $M = M + 1$ and repeat Step 2.2.

Step 3. *Main Search; Lower Branch*

Step 3.1. Set $M = M^E - 1$. If $M = M^0$, STOP.

Step 3.2. Set $Z_{\mathrm{LB}} = \min_{M^0 \le L \le M} Z^E(L)$. If $Z_{\mathrm{LB}} \ge Z^*$, STOP.

Step 3.3. If $Z(M) < Z^*$, set $Z^* = Z(M)$, $M^* = M$.

Step 3.4. If $M = M^0$, STOP. Else, set $M = M - 1$ and repeat Step 3.2.

Upon exiting the algorithm, M^* contains the optimal number of facilities and Z^* the optimal value of the objective function. The algorithm proceeds as follows. In Step 1, the initial upper and lower bounds are computed, as well as the optimal number of facilities M^E for the EFC case. We then re-adjust the upper bound (note that upper bound can be tightened whenever an improved solution is found) and initialize the search for the optimal solution, using M^E as the first guess. The search is split up into two intervals: the upper branch (Step 2) searches over the $M^E, \ldots, M^U - 1$ values, while the lower branch searches over $M^0, \ldots, M^E - 1$. In either case, Part 1 of Lemma 1 implies that the search can stop as soon as $Z_{\mathrm{LB}} \ge Z^*$, where Z_{LB} represents the lowest value of Z^E over the remaining search interval, and Z^* is the value of the best feasible solution found so far. Note that the optimal location vector and the optimal capacity are computed implicitly—in the process of determining $Z(M^*)$.

Our experience with Algorithm 1 (as reported in §5) indicates that M^E usually provides an excellent approximation to the optimal number of facilities M^*. Thus, the optimal number of facilities is found quite quickly, the bulk of the running time is spent in proving that this value is, in fact, optimal. The proof process would be much faster if we could establish that $Z(M)$ is unimodal. (In that case, the searches in the upper and lower branches could stop as soon as the values of $Z(M)$ begin to increase.) In fact, even unimodality of $Z^E(M)$ would help because it would simplify Step 1, and allow us to simply set $Z_{\mathrm{LB}} = Z^E(M)$ in Steps 2 and 3. In §3.5.1, we further discuss the unimodality of $Z(M)$ and $Z^E(M)$.

3.5. Further Results for the Single-Server Case

For the single-server case we discuss the unimodality of $Z^E(M)$ and $Z(M)$ and show that the smallest feasible number of facilities is optimal when the per-unit capacity costs are increasing. Generalizing these results to the multiple-server case requires a careful consideration of the integrality in the number of servers and adds no insight, thus we do not pursue it.

3.5.1. Unimodality of the Objective Functions for the Single-Server Case. In our computational results we used the cost function

$$Z(M) = CM^{\theta} + cM^{\beta}\mu \tag{27}$$

for some $\theta, \beta \in (0, 1)$. This cost function presents increasing returns to scale in the number of facilities. We will assume this cost structure for the remainder of §3.5.

For every problem instance solved for the single-server case, we observed that $Z^E(M)$ was unimodal;

when the arrival at each point has an infinitesimal rate, $Z(M)$ was also unimodal. Nevertheless, it appears to be very hard to prove unimodality of either $Z^E(M)$ or $Z(M)$. We therefore state the following conjecture.

Conjecture: For the single-server case, $Z^E(M)$ is unimodal in M. Moreover, when the arrival process at each point has an infinitesimal rate, $Z(M)$ is also unimodal.

This conjecture likely holds under more general conditions on the cost function than (27). Proposition 3 substantiates this conjecture for $Z^E(M)$ for the single-server case with exponential service.

Proposition 3. *Suppose the cost function is given by* (27), *the service time is exponential, and the large-deviation bound* (20) *is used to determine the optimal capacity. Then,* $Z^E(M) = CM^\theta + cM^\beta\mu(\Lambda/M)$ *is unimodal, and its unique minimum* $\widehat{M}$ *satisfies the first-order condition* (*FOC*)

$$0 = C\theta M^{\theta-1} + c(\beta-1)\Lambda M^{\beta-2} - c\beta\frac{\ln\alpha}{d}M^{\beta-1}. \quad (28)$$

Note that the minimizer $\widehat{M}$ in the preceding result may not be an integer; the optimal number of facilities M^E (provided EFC exists) is obtained by checking the two integer values $\lfloor\widehat{M}\rfloor$ and $\lceil\widehat{M}\rceil$. As will be discussed, a closed form solution to FOC (28) can be obtained in some cases, yielding interesting insights into the behavior of the optimal number of facilities and the optimal service capacity.

Using $\varepsilon = \theta - \beta$ we can rewrite (28)

$$C(\beta+\varepsilon)M^{\varepsilon+1} - c\beta\frac{\ln\alpha}{d}M + c(\beta-1)\Lambda = 0. \quad (29)$$

In general, the closed-form solution for (29) does not exist and the root $\widehat{M}$ must be found numerically. However, in the case of $\beta = \theta$ (i.e., $\varepsilon = 0$), the closed form solution for $\widehat{M}$ is

$$\widehat{M} = \frac{cd(1-\beta)\Lambda}{\beta(dC - c\ln\alpha)}. \quad (30)$$

Since the optimal number of facilities in the EFC case, M^E, is within 1 of $\widehat{M}$, and the optimal number of facilities M^* is usually well approximated by M^E, the expression above yields some insights into the sensitivity of the optimal number of facilities to the problem parameters when the cost function is given by (27) and $\beta \approx \theta$.

It is not hard to see that $\widehat{M}$ is linearly increasing with the total arrival rate Λ, and linearly decreasing with the facility cost parameter C. The behavior with respect to other problem parameters is less obvious, but after taking derivatives it can be verified that $\widehat{M}$ is increasing with the capacity cost c, threshold value d, and the service measure α, and is decreasing with the returns-to-scale parameter $\beta = \theta$ (in all cases the rate of change is less than linear). We can conclude that the optimal number of facilities should be fairly robust with respect to changes in the required service level and the cost of service capacity (particularly since M^* and M^E are both integers, and thus small-scale changes in $\widehat{M}$ have no effect on them).

From (19) and (30) we can also obtain a closed-form expression for the (approximate) optimal capacity when $\beta \approx \theta$

$$\mu^* \approx \mu(\lambda^{\max}(\widehat{M})) = \frac{\beta(dC - c\ln\alpha)}{cd(1-\beta)} - \frac{\ln\alpha}{d}.$$

It can be seen that μ^* is independent of Λ; decreasing with c, d, and α; linearly increasing with C; and increasing with $\beta = \theta$. Thus, changes in the service-level requirements and capacity costs are likely to be reflected in the optimal capacity level, whereas changes in the total demand rate Λ will lead to opening or closing of the facilities, but will not affect their capacity. Changes in the facility cost C may lead to both changes in the number of facilities and in their capacity.

3.5.2. Optimal Number of Facilities When Per-Unit Capacity Costs Are Increasing. To simplify the notation throughout this section, we use $\mu(M) \equiv \mu(\lambda^{\max}(M))$. Observe that the main trade-off represented by the objective function of the SCFLP is between the facility cost C_M that grows as the number of facilities is increased, and the system capacity cost $c_M\mu(M)$ that could decrease because as more facilities are added, less capacity μ is required at the busiest facility. Note, however, that the system capacity cost is a function of total system capacity $M\mu(M)$. In fact, we may think of the total capacity cost as $(c_M/M) * (M\mu(M))$, where the first term represents per-unit capacity cost and the second term represents overall system capacity. As discussed below, it can be

shown in some cases that the overall system capacity $M\mu(M)$ is nondecreasing with M. Thus, if the per-unit capacity cost c_M/M is also nondecreasing, then it can never be optimal to increase the number of facilities beyond the minimal level required for feasibility. Therefore,

Observation. Let M^0 be the minimal number of facilities for which ELP(M) is feasible. Suppose the system capacity $M\mu(M)$ and the per-unit capacity cost c_M/M are both nondecreasing in M. Then the optimal number of facilities is $M^* = M^E = M^0$.

It is difficult to draw conclusions about the system capacity in the general case: since the determination of $\mu(M)$ requires the solution of ELP(M) to find $\lambda^{\max}(M)$. However, the situation is different when EFC location vectors exist (i.e., with respect to the determination of M^E), since we know that $\lambda^{\max}(M) = \Lambda/M$ in this case. Indeed, for the case of exponential service, if the large-deviation bound (20) is used to determine system capacity, we observe that

$$M^E\mu(M^E) = M^E\Lambda + M^E\gamma^* \le (M^E+1)\Lambda + (M^E+1)\gamma^* = (M^E+1) * \mu(M^E+1),$$

showing that the system capacity is nondecreasing in M. Proposition 4 extends this result to the more general case.

Proposition 4. *Suppose the large-deviation bound* (20) *is used to determine the optimal capacity $\mu(M)$. Let $\mu^E(M)$ be the optimal server capacity when $\lambda^{\max} = \Lambda/M$. Then the system capacity $M\mu^E(M)$ is nondecreasing in M.*

The previous two results now lead to Corollary 3.

Corollary 3. *Suppose the large-deviation bound is used to determine the optimal system capacity and the per-unit capacity cost c_M/M is nondecreasing in M. Then $M^E = M^0$, where M^0 is the smallest M for which ELP(M) is feasible.*

Note that it is not necessary to solve ELP(M) to find M^0. As discussed above, it suffices to solve the (often much simpler) min-cover problem. Since M^E usually provides a good approximation to M^*, it follows that when the per-unit capacity costs are nondecreasing, the approximate solution to the SCFLP can often be obtained quite easily, without resorting to Algorithm 1. Moreover, if there is a feasible EFC vector for $M = M^E$, then $M^* = M^E$.

Intuitively, nondecreasing per-unit capacity cost implies that there are no economies of scale with respect to the capacity cost. The results of this section suggest that, in this case, it is typically optimal to keep the number of facilities as small as possible.

4. ELP and Existence of EFC

We now take a closer look at the ELP model defined in (10) of §3.1. Recall that for a given number of facilities M, this model identifies facility locations that minimize the demand assigned to the busiest facility (i.e., the arrival rate at the busiest facility). The general ELP model (i.e., defined over an arbitrary space P and with a general demand distribution $\lambda(\mathbf{x})$) appears to be quite difficult to solve. However, when specialized to particular topologies P (and, possibly, with certain additional assumptions with respect to $\lambda(\mathbf{x})$), ELP leads to an interesting family of problems that, to the best of our knowledge, are largely unexplored in location literature. In this section, we discuss two members of this family: the discrete location case where P is assumed to be a finite set, and a linear case where P is a line segment. In the latter case, we also present some results related to the existence of the EFC vector. The solution of the ELP over a finite plane is discussed in the follow-up paper by Baron et al. (2007). Remember that the results presented in this section hold for both the single- and multiple-server cases due to the decomposition of the SCFLP discussed in §3.

4.1. Discrete ELP

Suppose the set of customer demand points is discrete, i.e., $\lim_{\epsilon\to 0}\int_x^{x+\epsilon} dF_\Lambda(x) > 0$, with $P = \{1, \ldots, |P|\}$. Assume that a facility can be located at any point in P and that the distance between any two points of P is at least ϵ (i.e., any M-dimensional subset of P can serve as a valid facility set). For a customer demand point $k \in P$, let $\lambda(k)$ be the arrival rate of calls for service from k, with $\sum_{k\in P}\lambda(k) = \Lambda$.

For $k \in P$, we define the following set of facility locations from which k can be covered

$$R_k = \{j \in P \mid d(j,k) \le r\},$$

where $d(j,k)$ is the distance between j and k. Let $x_j = 1$ if a facility is located at point j and 0 otherwise

for $j \in P$, and let $y_{kj} = 1$ if customer k is assigned to facility j and 0 otherwise for $k \in P$ and $j \in R_k$. The discrete version of the ELP can now be formulated as follows (we will refer to this model as ELPd).

$$\min \ \lambda^{\max}$$

$$\text{subject to } \lambda^{\max} \geq \sum_{\{k \in P \mid j \in R_k\}} \lambda(k) y_{kj}, \quad j \in P \tag{31}$$

$$y_{kj} \leq x_j \quad k \in P,\ j \in R_k \tag{32}$$

$$\sum_{j \in P} x_j = M \tag{33}$$

$$\sum_{j \in R_k} y_{kj} = 1, \quad k \in P \tag{34}$$

$$\sum_{\{j' \in R_k \mid d(j',k) \leq d(k,j)\}} y_{kj'} \geq x_j, \quad kj' \in P \tag{35}$$

$$y_{kj}, x_j \in \{0,1\}, \quad k, j \in P.$$

Constraint (31) defines $\lambda^{\max}$ as the maximum arrival rate to any facility. Constraint (32) ensures that a customer can only be assigned to an open facility. Constraint (33) sets the total number of facilities to M, and (34) ensures that a customer is assigned to exactly one facility. Constraint (35) enforces the requirement that a customer must be assigned to the closest open facility. (See Gerrard and Church 1996 and Berman et al. 2006 for alternative forms of closest assignment constraints.)

To get some idea of the difficulty of ELPd we ran problem pmed1—the smallest-size problem from Beasley (1990) for the p-median problem, which is a network with 100 nodes and five facilities. We used CPLEX 9.1 to solve the problem on a Pentium 4 computer with 256 MB RAM. We vary the radius r from 127 to 133 because there is no feasible solution for problem with $r \leq 126$, and problems with $r \geq 134$ could not be solved in less than three hours. As can be seen in Table 1, none of the instances can be solved in less than 35 minutes.

Table 1 Solution of ELPd with a Network of 100 Nodes and Five Facilities

n	p	Radius (dist)	CPU	$\lambda^{\max}$
100	5	127	1,260.91	1,454.29
100	5	128	4,407.61	1,385.58
100	5	129	6,058.45	1,163.19
100	5	130	5,129.23	1,162.34
100	5	131	10,327.46	1,162.34
100	5	132	4,323.96	1,162.34
100	5	133	5,248.87	1,081.48

We believe that the difficulty in solving the ELPd problem can be in part attributed to the closest assignment constraints. We suggest development of a special algorithm for this problem in future research. A possible direction can be to exploit the similarity of the problem to the maximal weight cover problem, e.g., Church and ReVelle (1982); and the M-center problem, e.g., Courrent et al. (2002).

4.2. Linear ELP

In this section, we assume that P is a line segment with $\lim_{\epsilon \to 0} \int_x^{x+\epsilon} dF_\Lambda(x) = 0$ for each $x \in P$. Without loss of generality, we assume P to be of unit length—with the corresponding rescaling of the coverage radius r and the minimal separation distance ε. We begin by formulating the version of ELP for this case.

As before, we assume that there are M facilities to locate and denote the location of the jth facility by x_j for each $j = 1, \ldots, M$, where $x_j < x_{j+1}$ for all $j < M$. The no collocation constraints are

$$x_j - x_{j-1} \geq \varepsilon \quad \forall\, j = 2, \ldots, M.$$

Since customers must be assigned to the closest open facility, customers from $y \in [0, (x_1 + x_2)/2]$ are assigned to the facility at x_1, customers from $y \in [(x_1 + x_2)/2, (x_2 + x_3)/2]$ are routed to the facility at x_2, and so on. With this assignment rule, the total arrival rate to the jth facility is

$$\lambda^{x_j} = \Lambda\left[F_\Lambda\left(\frac{x_j + x_{j+1}}{2}\right) - F_\Lambda\left(\frac{x_j + x_{j-1}}{2}\right)\right] \quad \forall j = 1, \ldots, M, \tag{36}$$

where we defined $x_0 = -x_1$ and $x_{M+1} = 2 - x_M$.

Using (36), the ELP for the linear case (ELPl) is

$$\min \ \lambda^{\max}$$

$$\text{s.t. } \Lambda\left[F_\Lambda\left(\frac{x_j + x_{j+1}}{2}\right) - F_\Lambda\left(\frac{x_j + x_{j-1}}{2}\right)\right] \leq \lambda^{\max} \quad \forall\, j = 1, \ldots, M \tag{37}$$

$$0 \leq x_1, x_M \leq 1$$

$$x_j - x_{j-1} \geq \varepsilon \quad \forall\, j = 2, \ldots, M \tag{38}$$

$$x_{j+1} - x_j \leq 2r \quad \forall\, j = 0, \ldots, M \tag{39}$$

$$x_0 = -x_1, \quad x_{M+1} = 2 - x_M. \tag{40}$$

Constraint (38) ensures that the minimal distance between facilities is at least ε. Constraint (39) enforces the coverage requirement: Due to the closest assignment rule, ensuring that every point in $[0, 1]$ is within distance r of a facility is equivalent to ensuring that no two facilities are more than $2r$ apart. Constraint (40) defines x_0 and x_{M+1} in a way that makes the preceding constraints applicable for x_1 and x_M.

Observe that (ELPl) cannot be feasible unless the following basic feasibility condition holds

$$\frac{1}{\varepsilon} \geq M \geq \frac{1}{2r}. \tag{41}$$

We note that Constraint (37) is typically nonlinear. Moreover, investigation of the Hessian matrix of (37) shows that these constraints are only convex if $F_\Lambda(x)$ is uniform, in which case the problem can be solved easily without the use of mathematical programming techniques (see Example 1 below). Therefore, (ELPl) should be solved using nonconvex optimization techniques.

This discussion demonstrates that ELP is nontrivial even in the linear case. (We note that, as discussed in §5, standard nonlinear solvers appear to be able to handle this problem successfully for some common distributions $F_\Lambda(x)$ and small-to-moderate values of M.)

In the remainder of this section, we suggest some alternative solution approaches that may be used when $F_\Lambda(x)$ satisfies certain constraints. Recall that if a feasible EFC location vector exists, then it must be optimal. In some cases, candidate EFC vectors are fairly easy to characterize, as demonstrated in the following example.

Example 1 (Uniformly Distributed Demand). Suppose the demand distribution density $F_\Lambda(x)$ is uniform, i.e., $F_\Lambda(x) = x$ for $x \in [0, 1]$. Then it is clear that an EFC vector is given by

$$x_j = \frac{2j-1}{2M} \quad \text{for } j = 1, \ldots, M, \tag{42}$$

with the resulting $\lambda^{\max} = \Lambda/M$. Note, however, that the location vector given above is not the only EFC vector in this case. In fact, there are an infinite number of EFC vectors that can be characterized as follows

$$x_j = \frac{2j-1}{2M} + (-1)^j\delta \quad \text{for } j = 1, \ldots, M, \text{ and } \delta \in \left(-\frac{1}{2M}, \frac{1}{2M}\right). \tag{43}$$

An EFC vector is feasible only if it satisfies the coverage and minimal separation constraints, i.e., if

$$\varepsilon \leq x_{j+1} - x_j \leq 2r \quad \text{for all } j \in \{0, \ldots, M\}, \tag{44}$$

where, as before, we set $x_0 = -x_1$, and $x_{M+1} = 2 - x_M$. It is easy to see that the EFC vector given by (42), corresponding to setting $\delta = 0$ in (43), has the property that the maximum distance between any two adjoining facilities is as small as possible and the minimum distance between any two adjoining facilities is as large as possible. Thus, if this vector fails to satisfy the feasibility condition (44), no feasible EFC vector exists for the current value of M.

Since, for the EFC vector given by (42), $x_{j+1} - x_j = 1/M$, we observe that if (44) is not satisfied, then $1/(2r) > M$ or $M > 1/\varepsilon$, violating the basic feasibility condition (41). On the other hand, if this condition is satisfied with a strict inequality, it will also be satisfied by an infinite number of EFC vectors corresponding to a nonempty range of δ in (43)—in effect, the coverage and minimal distance constraint (44) can be interpreted as constraints on δ.

Thus, we conclude that in the case of uniform demand, the (ELPl) is quite easy to solve: As long as M satisfies the basic feasibility condition (41), the EFC vector given by (42) is feasible and optimal. Moreover, if (41) is satisfied as a strict inequality, the optimal location vector is not unique—there exist an infinite number of feasible EFC vectors.

This example gives the flavor of the results developed in the following section for general demand densities: Feasible EFC vectors, when they exist, are typically not unique, and the conditions for the existence of such vectors are relatively easy to check.

4.2.1. Conditions for the Feasibility of EFC Vectors. Let $y_0 \equiv 0$, $y_1, \ldots, y_{M-1}$ be breakpoints such that $F_\Lambda(y_{j+1}) - F_\Lambda(y_j) = \Lambda/M$ for $i = 1, \ldots, M-1$, and $y_M = 1$. Suppose $\mathbf{x}$ is an EFC vector. Then the coverage area of each facility x_j must be $[y_{j-1}, y_j]$, for $j = 1, \ldots, M$. By the closest assignment constraints, it follows that the conditions for $\mathbf{x}$ to be a feasible EFC are given by

$$y_{j-1} \leq x_j \leq y_j \quad j = 1, \ldots, M \tag{45}$$

$$y_j - x_j = x_{j+1} - y_j \quad j = 1, \ldots, M-1 \tag{46}$$

$$\begin{cases} \varepsilon/2 \le x_{j+1} - y_j \le r, \quad j = 1, \ldots, M-1 \\ x_1 \le r,\ x_M \ge 1 - r. \end{cases} \tag{47}$$

The first two conditions (45 and 46) ensure that $\mathbf{x}$ is an EFC vector, and condition (47) enforces the coverage and minimal separation constraints. In Proposition 5 we rewrite the conditions above in terms of x_1 only, yielding a set of inequalities in one variable that only need to be checked for consistency, which can be accomplished in $O(M^2)$ time.

Proposition 5. *A feasible EFC exists if, and only if,* $\exists x_1 \in [0, \min\{y_1, r\}]$ *such that*

$$y_j \le 2\sum_{k=1}^{j}(-1)^{j-k} y_k + (-1)^j x_1 \le y_{j+1}, \quad j = 0, \ldots, M-1 \tag{48}$$

and

$$\begin{cases} \varepsilon/2 \le y_j + 2\sum_{k=1}^{j-1}(-1)^{j-k} y_k + (-1)^j x_1 \le r, \\ \quad j = 1, \ldots, M-1 \\ 2\sum_{k=1}^{M-1}(-1)^{M-1-k} y_k + (-1)^{M-1} x_1 \ge 1 - r. \end{cases} \tag{49}$$

We will refer to the vector that satisfies (48) above as an EFC vector and the vector that satisfies both (48) and (49) as a feasible EFC vector. While it is not hard to see that an EFC vector must exist for $M = 2$, the existence is not ensured for $M = 3$ or higher, as demonstrated in the following example.

Example 2 (Existence of an EFC Vector in the Linear Case). Suppose $M = 3$ and $F_\Lambda(x) = Beta(a, a)$—which, for $0 < a < 1$, is a symmetrical U-shaped distribution with mean of 0.5. The parameter a is the shape parameter—the closer it is to zero, the steeper the U-shape.

Since $\lambda(x)$ is symmetrical on $[0, 1]$, we know that for breakpoints y_1 and y_2 we have $y_1 = 1 - y_2$. From (45) and (46) we know that $x_1 \le y_1$, $x_3 \ge y_2$, and x_2 must satisfy

$$\begin{cases} y_1 - x_1 = x_2 - y_1, \\ y_2 - x_2 = x_3 - y_2. \end{cases}$$

It is clear that this system has an infinite number of solutions if $y_1 \in (0, 0.25)$, a unique solution if $y_1 = 0.25$ (in this case, $x_1 = 0, x_2 = 0.5, x_3 = 1$), and no solutions if $y_1 \in (0.25, 1)$. Since $M = 3$, y_1 represents the 33rd percentile of the $Beta(a, a)$ distribution has the following properties

$$\begin{aligned} 0 < y_1 < 0.25 \quad &\text{if } a \in (0, 0.5); \\ y_1 = 0.25 \quad &\text{if } a = 0.5; \\ 0.25 < y_1 < 0.5 \quad &\text{if } 0.5 < a < 1. \end{aligned}$$

Thus, we conclude that EFC vector exists in this case if, and only if, $0 < a \le 0.5$. Moreover, for $a = 0.5$, this vector is unique, while for $0 < a < 0.5$, there are an infinite number of EFC vectors. (Of course, even when EFC vectors exist, they may not be feasible due to the coverage and minimal distance constraints.)

Motivated by the preceding example, it is natural to ask whether there are families of demand densities $F_\Lambda(x)$ for which the existence of EFC vectors is ensured. Theorem 3 establishes sufficient conditions that are satisfied by many common probability distributions.

Theorem 3. (1) *Suppose* $F_\Lambda(x)$ *is nonincreasing on* $[0, 1]$. *Then any* $x_1 \in [0, y_1)$ *satisfies* (48) *and thus can be used to generate an EFC vector.*

(2) *Suppose* $F_\Lambda(x)$ *is nondecreasing on* $[0, 1]$. *Let* $x_M \in (y_M, 1]$ *and define* $x_j = 2y_j - x_{j+1}$ *for* $j = M-1, \ldots, 1$. *The resulting vector is EFC.*

(3) *Suppose* $F_\Lambda(x)$ *is nondecreasing on* $[0, z]$ *and nonincreasing on* $[z, 1]$ *for some* $z \in [0, 1]$. *Let* $j' \in \{1, \ldots, M\}$ *be such that the difference* $y_{j'} - y_{j'-1}$ *is minimized. Let* $x_j \in (y_{j'-1}, y_{j'})$ *and define all other components of* $\mathbf{x}$ *by* (46). *Then the resulting location vector is an EFC.*

We observe that many common distributions—e.g., the Normal distribution (truncated to $[0, 1]$)—satisfy the conditions of Proposition 3. On the other hand, this result does not work for distributions where a decreasing part is followed by the increasing part—as demonstrated by Example 2.

The preceding results established some sufficient conditions for the existence of EFC vectors. What about the existence of feasible EFC vectors? Observe that the minimal distance constraints are satisfied by the EFCs in Theorem 3 as long as $x_{j'}$ is at least $\varepsilon/2$ away from both $y_{j'-1}$ and $y_{j'}$. On the other hand,

because the breakpoints are farthest apart at either end of the interval, as long as the coverage constraint holds there, it will hold everywhere else as well. This leads to Corollary 4.

COROLLARY 4. *Under the conditions of Theorem* 3, *suppose that* $y_{j'} - y_{j'-1} \geq \varepsilon$, *and* $y_1 \leq r$, $y_M \geq 1 - r$. *Then a feasible EFC vector is obtained by taking* $x_{j'} \in (y_{j'-1} + \varepsilon/2, y_{j'} - \varepsilon/2)$ *and generating all other components using* (46).

Theorem 3 and Corollary 4 demonstrate that, when P is a line segment, an EFC vector exists for many common distributions, and both the existence of an EFC vector and the existence of a feasible EFC vector are relatively easy to check. When an EFC (or a feasible EFC) vector exists, it is usually not unique—indicating that infinitely many optimal location vectors often exist in the linear case. The solvability of (ELPl) is further investigated in §5.

5. Computational Results: The Importance of Being Fair

In this section, we present the results of several sets of computational experiments. The first set investigates the solvability of (ELPl) and the properties of the optimal solutions for different spatial distributions of demand over the line segment $[0, 1]$. The second set investigates the overall performance of Algorithm 1 for the SCFLP, the efficiency of large-deviation bounds for estimating the capacity in the single-server case, and the properties of the optimal solution.

The solvability of ELP on the plane is discussed in Baron et al. (2007). We note that our experiments on the efficiency of large-deviation bounds for estimating capacity were focused on the single-server case. Similar experiments for the multiple-server case can be presented, although we believe they would not provide further insight. Furthermore, while a linear topology is assumed, the results related to the efficiency of capacity estimation are, in fact, applicable to any topology. Finally, we consider exponential, normal, and deterministic service distributions. For all these, (19) leads to a closed form expression for $\mu(\lambda)$.

In the first set of computational experiments we analyze three issues related to the linear ELP under different coverage radii and different spatial distributions of demand: (a) the solvability of the nonlinear programming formulation of ELPl given by (37–40), (b) the likelihood of obtaining a feasible EFC solution, and (c) the degree of deviations from the EFC solutions when a feasible EFC solution does not exist.

We set the total demand over the line to $\Lambda = 1$ and consider three spatial distributions of demand over the line segment $[0, 1]$: *Beta*$(2, 2)$—a symmetrical distribution centered at 0.5 with a Normal-like shape, *Beta*$(0.25, 2)$—a sharply decreasing distribution with 68% of demand falling between 0 and 0.1, and *Beta*$(0.5, 0.5)$—a deep bathtub-shaped distribution centered at 0.5. These shapes were designed to represent different levels of difficulty for ELPl. For comparison, we also computed the results for the uniform distribution, where a feasible EFC is guaranteed to exist (as discussed in Example 1).

The number of facilities M was set to 5, 10, and 20. Recall that the basic feasibility condition for ELPl is that the coverage radius r must satisfy $r \geq 1/(2M)$. In our experiments, we set $r = (1 + \delta)/2M$ for $\delta = 0.1, 0.2, 0.3, 0.5, 0.75, 1, 2, 3, 4, 5$—generating 10 different coverage radii for each value of M and each distribution shape (for a total of 120 instances); the coverage constraints are very tight for small values of δ and quite loose for large values of δ. The minimal required distance between facilities was set to $\epsilon = 1/(1{,}000M)$.

An ELPl model was formulated for each instance and solved using the Frontline Systems Premium NLP Solver for MS Excel. The results are presented in Table 2. The first column of the table contains the coverage radius. The next four columns contain the value of the optimal solution—i.e., the arrival rate at the busiest facility. The cases where feasible EFC exist can be identified by comparing the value of the optimal solution with the value for the uniform distribution (where a feasible EFC is guaranteed to exist). Recall that the capacity of the busiest facility has two components: $\lambda^{\max}$, and the safety capacity required to meet the service-level constraint. Assuming that the safety capacity is (nearly) identical in all cases, the gap between the optimal value for a given case and an EFC solution (which is always equal to $1/M$) can be used as a measure of the excess capacity of the busiest facility. Since all facilities are identical, the estimated total amount of excess capacity is given by

$$\text{EFC Gap} \equiv \frac{M * \lambda^{\max} - \Lambda}{\Lambda}, \tag{50}$$

Table 2 Experimental Results for Linear ELP

	$\lambda^{max}(M)$ Arrival rate for the busiest facility				Amount of excess capacity (vs. EFC case)		
Coverage radius	*Beta*(2, 2)	*Beta*(0.5, 0.5)	*Beta*(0.25, 2)	Uniform	*Beta*(2, 2)	*Beta*(0.5, 0.5)	*Beta*(0.25, 2)
$M = 5$ facilities							
0.110	0.253	0.271	0.718	0.2	0.263	0.353	2.590
0.120	0.245	0.244	0.654	0.2	0.227	0.221	2.271
0.130	0.240	0.225	0.571	0.2	0.200	0.126	1.853
0.150	0.228	0.205	0.365	0.2	0.140	0.024	0.824
0.175	0.217	0.200	0.311	0.2	0.084	0.000	0.554
0.200	0.205	0.200	0.277	0.2	0.023	0.000	0.385
0.300	0.200	0.200	0.232	0.2	0.000	0.000	0.159
0.400	0.200	0.200	0.226	0.2	0.000	0.000	0.132
0.500	0.200	0.200	0.220	0.2	0.000	0.000	0.099
0.600	0.200	0.200	0.211	0.2	0.000	0.000	0.055
$M = 10$ facilities							
0.055	0.132	0.158	0.599	0.1	0.324	0.575	4.987
0.060	0.125	0.128	0.527	0.1	0.254	0.282	4.273
0.065	0.122	0.109	0.406	0.1	0.219	0.086	3.062
0.075	0.115	0.101	0.263	0.1	0.146	0.010	1.632
0.088	0.112	0.100	0.207	0.1	0.121	0.000	1.073
0.100	0.110	0.100	0.178	0.1	0.098	0.000	0.779
0.150	0.102	0.100	0.136	0.1	0.019	0.000	0.360
0.200	0.100	0.100	0.122	0.1	0.000	0.000	0.224
0.250	0.100	0.100	0.115	0.1	0.000	0.000	0.147
0.300	0.100	0.100	0.112	0.1	0.000	0.000	0.121
$M = 20$ facilities							
0.028	0.077	0.081	0.232	0.05	0.535	0.627	3.645
0.030	0.071	0.062	0.144	0.05	0.418	0.241	1.878
0.033	0.057	0.055	0.112	0.05	0.145	0.099	1.245
0.038	0.055	0.050	0.088	0.05	0.098	0.008	0.768
0.044	0.053	0.050	0.088	0.05	0.063	0.000	0.768
0.050	0.052	0.050	0.088	0.05	0.049	0.000	0.768
0.075	0.051	0.050	0.088	0.05	0.011	0.000	0.768
0.100	0.050	0.050	0.088	0.05	0.000	0.000	0.768
0.125	0.050	0.050	0.088	0.05	0.000	0.000	0.768
0.150	0.050	0.050	0.088	0.05	0.000	0.000	0.768

which is presented in the last three columns of Table 2 (we omit the uniform case for which the gap is always zero). Note that the EFC Gap is normalized by Λ.

In all cases, the solution times were essentially instantaneous and thus are not presented. Although due to the nonlinearity and nonconvexity of the constraints the solutions may only be locally optimal, starting the solver from multiple starting points did not result in better solutions. Overall, it appears that ELPl is relatively easy to solve.

It can be seen from Table 2 that the quality of the solutions is highly dependent on the shape of the spatial distribution of demand, with *Beta*(0.25, 2) (sharply decreasing shape) leading to worst-quality solutions, and normal-like *Beta*(2, 2) leading to the best solutions for the tight coverage radii (excluding the uniform case, of course). As expected, the solution quality decreases as the coverage radius gets smaller. When the coverage radius is tight, optimal solutions may require four to five times more system capacity than the EFC case. The feasibility of EFC locations depends on both the tightness of the coverage radius and the distribution shape. For the hardest cases (tight radii, *Beta*(0.25, 2) distribution, or both), EFC is never feasible. For the other two distributions, EFC tends to become feasible for larger coverage radii. It should be noted that the infeasibility in case of larger radii for *Beta*(0.25, 2) is usually caused

by the required minimal separation constraints. This is due to the rather extreme levels of concentration of demand for this distribution.

It is interesting to observe how the solution quality varies with the number of facilities (for the same spatial distribution and coverage radius). For both $Beta(2,2)$ and $Beta(0.5,0.5)$ cases, the total amount of excess capacity in the system tends to increase with the number of facilities. However, in the case of $Beta(0.25,2)$ distribution, the total system capacity is larger for $M=10$ than for $M=20$, indicating that a system with more facilities may be more efficient in this case.

Note that in the experiments described above the number of facilities M was fixed and the coverage radius r was adapted to M to generate more or less tight coverage constraints, which led to a lot of excess capacity in the system. However, in the process of solving the SCFLP using Algorithm 1, the opposite mechanism is in play: The coverage constraint is specified first, and then the algorithm has to select the optimal value of M. By electing to have a larger number of facilities for a given value of r it is always possible, in effect, to make the coverage constraint loose. This increases the likelihood of EFC location vectors to be feasible (at the cost of having more facilities). As will be seen next, this is the trade-off that appears to be nearly always optimal in the SCFLP.

To evaluate the solvability of the SCFLP, we generated a number of problem instances with different service time distributions, spatial demand distributions, coverage radii r, waiting time limits d, and cost functions. For each instance we let $\Lambda=100$ and applied Algorithm 1 with two different estimates of the required capacity $\mu(\lambda^{\max})$. The first estimate was based on the large-deviation bound (19). The second estimate was based on the optimal capacity (13) for the case of exponential service and, for nonexponential service, on estimated capacity based on an M/G/1 queuing simulator to determine the capacity. The test problem instances had the following characteristics:

Service distributions: Exponential(1), Deterministic(1), Normal$(1, 0.1^2)$, Normal$(1, 0.3^2)$; the first value in parentheses refers to the average service time, the second value (for Normal distributions) to the standard deviation of service times.

Demand and its spatial distributions: $Beta(2,2)$ and $Beta(0.5,0.5)$ distributions on $[0,1]$.

Coverage radii: r was set to 0.1 and 0.5; the minimal required separation of the facilities was set to $\epsilon=0.0001$.

Waiting time limits: d was set to 2 and 10; with $\alpha=0.05$ in all cases. Because the average service time is 1, the service-level constraint requires the waiting time to be no larger than two (or 10) service cycles with probability 95%.

Cost function: we used $Z(M)=CM^{\theta}+cM^{\beta}\mu$ form of the cost function (as in (27)), with $\theta=\beta=0.9$ and the ratio C/c set to 0.5, 1, and 2. Note that the larger the C/c ratio, the higher the fixed costs of opening a new facility compared to the cost of increasing capacity at the existing facilities.

Algorithm 1 was implemented in Visual Basic for MS Excel. The queuing simulator was written in C and called from MS Excel as a DLL object. The embedded ELPl instances were solved with Frontline Systems' Premium Solver for MS Excel. The algorithm was very fast, typically converging in fewer than 10 iterations and requiring fewer than two seconds of CPU time for the approximate case and exact case with exponential service. The running times were much slower for the nonexponential exact cases where the queuing simulator had to be used: ranging from four to seven hours of CPU time. A summary of the results is presented in Tables 3–5.

For each combination of waiting time bound d and spatial distribution of demand, Table 3 presents the average percentage difference in optimal costs

Table 3 **Percent Difference in Optimal Costs for the Capacity Computed Using Large-Deviation Bound vs. Optimal Costs for the Capacity Computed Using Exact Formula for the Exponential Service and Using a Queuing Simulator for the Nonexponential Service**

		Spatial distribution of demand		
	Waiting time bound (d)	$Beta(0.5,0.5)$ (%)	$Beta(2,2)$ (%)	Average (%)
Exponential service time	2	0.26	0.34	0.30
	10	0.04	0.03	0.04
Average		0.15	0.19	0.17
Nonexponential service time	2	−4	−3	−3
	10	−1	−1	−1
Average		−3	−2	−2

Table 4(a) Properties of the Optimal Solution (Averages) for Different Service Time Distributions and Waiting Time Bounds

Distribution of service times	Waiting time bound (d)	Opt. no. of fac. if EFC feasible M^E	Opt. no. of fac. M^*	Optimal cost $Z(M^*)$	Safety capacity (%)
Detm(1)	2	6.5	8.25	93.84	6
	10	10.667	12.08	89.42	2
EXP(1)	2	4.833	6.67	98.47	10
	10	9.333	10.25	90.7	3
Norm(1, 0.1)	2	6.5	9.17	94.6	7
	10	10.667	13	89.88	2
Norm(1, 0.3)	2	6.5	10.17	95.46	8
	10	10.5	12.83	90	2
Average		8.069	9.97	92.9	5

Table 4(b) Properties of the Optimal Solution (Averages) for Cost Ratios and Coverage Radii

Cost ratio C/c	Coverage radius (r)	Opt. no. of fac. if EFC feasible M^E	Opt. no. of fac. M^*	Optimal cost $Z(M^*)$	Safety capacity (%)
0.5	0.1	11.917	14.25	88.16	8
	0.5	12	12	87.5	5
1	0.1	7.667	11.75	92.98	6
	0.5	7.5	7.5	91.24	3
2	0.1	5	10	101.54	5
	0.5	4.333	4.33	95.98	2
Average		8.069	9.97	92.9	5

between the approximate solutions (with the facility capacity obtained using the large-deviation bound) and the optimal solutions. (For nonexponential service, the capacity was obtained via simulation as described above.) Because the large-deviation bound always overestimates the required capacity, the approximate solutions should always have higher costs. This is indeed the case for exponential service. However, the differences between optimal and approximate solutions are very small: less than 1% in all cases.

For nonexponential service, the differences between approximate and optimal costs are actually negative, because we used simulation-based estimates of optimal capacity. While the simulation was run for a long time (the stopping criterion was that the difference between the required and simulated service levels should be less than 0.05%), the quality of the resulting approximation is evidently not as good as that of the approximation provided by the large-deviation bound.

Table 5 EFC Properties of the Optimal Location Vector

Cost ratio C/c	Coverage radius (r)	Solution not EFC (% cases) (%)	Solution is EFC (% cases) (%)	EFC Gap (for non-EFC cases only)
0.5	0.1	0	100	N/A
	0.5	0	100	N/A
1	0.1	17	83	6%
	0.5	0	100	N/A
2	0.1	25	75	6%
	0.5	8	92	6%
Average		8	92	6%

Note. For the cases where the optimal vector is not EFC, EFC Gap measures the excess Structural Capacity in the system.

The same results were observed in further computational tests designed to evaluate the efficiency of the large-deviation bound. In these tests, we assumed existence of EFC and used $\Lambda \in \{50, 100, 200\}$, $d \in \{5, 10, 20\}$, $\alpha \in \{0.9, 0.95, 0.99\}$, $C/c \in \{0.5, 1, 2\}$, $\theta \in \{0.9, 0.95, 0.99\}$, and $\beta \in \{0.9, 0.95, 0.99\}$ (a total of 729 cases for each service distribution). We selected the following service time distributions with a coefficient of variation between zero and one: deterministic service, normal distributed service time with standard deviation $\sigma \in \{0.1, 0.2, 0.3\}$, and exponential distributed service time. While the detailed results of these tests are not shown, the difference in optimal costs between the solutions with capacity obtained using the large-deviation bound, and the optimal solutions (simulation based for nonexponential service) did not exceed 2.25% and was often negative for simulation-based solutions.

We thus conclude that using Algorithm 1 together with the large-deviation bound (19) to estimate the required capacity is highly effective for the SCFLP model—the quality of the resulting solutions is excellent and the running times are quite small. Of course, if the exact expressions for the required capacity are available (as in the exponential case) there is no need to use the large-deviation bound. We emphasize that these results are independent of the linear topology used to generate demand: It appears that the large-deviation bound is a very accurate and efficient way to estimate the required capacity of the facilities.

Thus, the key determinant of how efficient Algorithm 1 is for the SCFLP is whether there exists an effective way of solving the ELP (which Algorithm 1

calls as a subroutine). While this is indeed the case for the linear ELP, the situation is quite different when the demand is distributed over a region of the plane; in this case, the ELP presents serious difficulties even for the uniform distribution of demand. We refer the reader to Baron et al. (2007) for further discussion.

The properties of the optimal solutions are summarized in Tables 4(a) and 4(b). Table 4(a) is stratified by the distribution of service times and the values of the waiting time limit d. For each combination of the service time distribution and d, we display M^E—the optimal number of facilities assuming the corresponding EFC vector is feasible (recall that M^E is used as the initial guess for the number of facilities in Algorithm 1), M^*—the optimal number of facilities, $Z(M^*)$—the corresponding optimal cost, and the

$$\textit{safety capacity } (\%) \equiv 100\frac{M^*\mu(M^*)-\Lambda}{\Lambda},$$

where $M^*\mu(M^*)$ is the total capacity in the system. Table 4(b) displays the same information stratified by the cost ratio C/c and the coverage radius r.

Overall, the results are quite intuitive: The optimal costs are generally driven by the service-level constraint (costs are higher for smaller value of d) and by the variability of service times. (For the same level of d, the costs are highest for the exponential case, which has the highest coefficient of variation, and lowest for the deterministic case.) The optimal number of facilities is larger and the safety capacity is smaller when the service-level constraint is loose (i.e., $d = 10$). When the service-level constraint is tight it is more economical to use the pooling effects by building fewer, but larger, facilities. Also, the number of facilities is larger when the C/c ratio is low (corresponding to lower fixed costs for opening new facilities) and when the coverage radius is small.

It is interesting to note two effects: (a) The value of the initial guess M^E is generally close to the value of M^*—validating the assumption made in Algorithm 1 that M^E is a good starting point for the search, and (b) the optimal number of facilities M^* is always substantially larger than the minimal number $1/(2r)$ required to satisfy the coverage constraint. In effect, the algorithm always selects the number of facilities for which the coverage constraint is quite loose.

This effect is further explored in Table 5, which is organized similarly to Table 4(b). For each combination of the cost ratio and coverage radius, we indicate the percentage of cases for which the optimal location vector is EFC (i.e., for which $\lambda^{\max} = \Lambda/M^*$). For the cases where the optimal location vector is not EFC, we compute the EFC gap from (50).

The obvious (and unexpected) observation is that in the majority of cases the optimal facility vector is EFC. Moreover, in cases where the optimal facility vector is not equitable, it is very close to EFC: The excess system capacity (measured by the EFC gap) is only 6%.

Recall from the results of the first set of experiments that a feasible EFC vector tends to exist only when the coverage constraint is loose. Thus, the optimization algorithm always loosens the coverage constraint by increasing the number of facilities to the point where an EFC (or nearly EFC) location vector is feasible. Moreover, as observed in Conjecture 1, the function $Z(M)$ tends to be unimodal with minimum at M^E, implying that the costs corresponding to EFC location vectors tend to grow with M for $M \geq M^E$. Thus, when the optimal location vector is EFC, it will correspond to the smallest value of M for which a feasible EFC vector exists. To summarize, the computational results suggest that the optimal solution to the SCFLP should generally be as fair as possible. This suggests the following very simple heuristic decision rule for the SCFLP:

Heuristic decision rule:

Step 1. Find the minimal number of facilities M for which a feasible EFC location vector exists in ELP(M). Set $M^* = M$ and location vector to EFC.

Step 2. Determine the capacity of each facility by using the large-deviation bound (19).

Note that this decision rule is very robust because it is completely independent of the facility costs. The only implementation requirement is to have an efficient way of solving the ELP(M) for different values of M (This may not be easy for general space P). While this decision rule has a lot of intuitive appeal, further computational and theoretical work is needed to substantiate it.

6. Extensions and Generalizations of the SCFLP

In this section, we discuss several generalizations and extensions of the SCFLP model.

Nonlinear capacity costs: The SCFLP model formulated here assumes that the capacity cost is linear for a fixed M. In fact, all of the results (with the exception of §3.5.2) continue to hold for a more general capacity cost, as long as it is nondecreasing in the facilities' capacity and in M. We chose to use the simplified form (5) of the objective function throughout the paper for ease of exposition.

SCFLP with general assignment rules: The SCFLP formulation used in this paper assumes that customers always obtain service from the closest facility. An alternative assumption is to consider directed or general assignments where a customer can be assigned to any facility located within radius r. We point out that this does not materially change our model—any assignment of customers to facilities can be used to define the maximum travel distances $R(x_j)$ in the formulation presented in §2. All of the results in §3 apply to the SCFLP with directed assignments. The $ELP(M)$ problem in §4 is simplified in the case of directed assignments. For example, in the linear case an EFC always exists and can be constructed by using the breakpoints $y_k, k = 1, \ldots, M$, defined in §4.2.1 as the facility locations. Of course, this EFC may not be feasible, but contrary to the closest assignment SCFLP, its existence is not an issue.

SCFLP with nonidentical facilities: Consider an SCFLP model where the requirement that all facilities have identical capacity is dropped. Let μ_j and N_j be the capacity of facility j, for the single- and multiple-server case, respectively, and replace Objective Function (5) with

$$\min_{M,\mathbf{x},\boldsymbol{\mu},\mathbf{N}} C_M + \tau c_M \sum_{j=1}^{M} \mu_j + (1-\tau) c_M \sum_{j=1}^{M} N_j. \qquad (51)$$

We show that, under some assumptions, if a feasible EFC vector exists, it is optimal for this model as well. We relax the problem in (51) by replacing the coverage constraint (4) and the no-collocation constraints with a single constraint that all customers are accepted to the system, i.e., $\sum_{j=1}^{M} \lambda^{x_j} = \Lambda$. It is clear that at each facility we will select the lowest capacity required to satisfy the service-level constraint, thus the results of Corollary 1 apply, and the functions $\mu(\lambda)$ and $N(\lambda)$, from (6) and (7), respectively, are well defined and nondecreasing in λ. Then

Proposition 6. *Consider the SCFLP with nonidentical facilities and with the number of facilities fixed to M. Suppose an EFC vector exists and is feasible for this M. Then, for the single-server case, if $\mu(\lambda)$ is differentiable and convex, the optimal location vector is given by EFC and all facilities in the optimal solution are identical with capacities set to $\mu(\Lambda/M)$. Similarly, for the multiple-server case, if $N(\lambda)$ is differentiable, the optimal location vector is given by EFC and all facilities in the optimal solution are identical with capacities set to $N(\Lambda/M)$.*

In view of the results presented in §5 showing that EFC are typically optimal for the SCFLP with identical facilities, Proposition 6 shows that these solutions would remain optimal even if facilities were allowed to be nonidentical (as long as the number of facilities M was unchanged). Obviously, if EFC exists and is feasible for every M (as would, for example, be the case for a demand distribution satisfying Theorem 3 and if the coverage constraints are loose), then the optimal solutions to SCFLPs with identical and nonidentical facilities are the same.

For the single-server case, the assumption that $\mu(\lambda)$ is differentiable and convex does not appear to be too stringent. In particular, it holds for the case of exponential service. Moreover, if the large-deviation bound (19) is used to set capacities of the facilities (which, according to our numerical results, provides near-optimal capacities) then this assumption holds as well. (This can be verified using the Implicit Function Theorem.) For the multiple-server case, $N(\lambda)$ is likely a step function that is almost always differentiable, which should be enough for Proposition 6 as well. Clearly, more work remains to be done on the SCFLP with nonidentical facilities for the cases where a feasible EFC may not exist.

SCFLP with different service-level measures: Our SCFLP model uses the probability of waiting time to exceed a given threshold as the service-level measure. However, there are alternative measures. We observe that for such measures it is enough to establish that Assumption 1 holds in order to establish that Theorem 1 holds. For example, consider a service-level constraint that requires that the mean waiting time will

not exceed a given threshold. Denoting the mean waiting time in a queue with arrival rate λ and service rate μ by $E(W(\lambda, \mu))$, we can define

$$\mu(\lambda) = \inf\{\mu > \lambda \mid E(W(\lambda, \mu)) > d\}.$$

Analogously, we define $N(\lambda)$, $\lambda(\mu)$, and $\lambda(N)$ similar to (7–9). We note that

$$E(W(\lambda, \mu)) = \int_0^\infty P(W(\lambda, \mu) > x)\, dx,$$

thus similar results to Propositions 1 and 2 follow. Thus,

Proposition 7. *Assumption* 1 *and Theorem* 1 *hold for the SCFLP with constraints* (4) *replaced by*

$$E(W^j) \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W_i^j \le d \quad j = 1, \ldots, M.$$

Given a method to optimally choose capacity for the SCFLP with constraints on the expected waiting time, the results of this paper substantially simplify this problem and similar problems with different service-level measures.

7. Summary

In this paper, we investigated the SCFLP—a version of the location problem with stochastic demand and congestion where customers travel to facilities to obtain service. Service level and maximum travel distance constraints are included to ensure that adequate service is provided to customers. Customers are assumed to travel to the closest facility. Facilities are modeled as single- or multiple-server queues with capacity as one of the decision variables in the model. Facilities are assumed to be identical.

We address this problem under a significantly more general setting than previously attempted in the literature: The service process is allowed to be general, the demand is assumed to be spatially distributed following a general renewal process or a Poisson process (for the discrete location version) over a bounded space with a norm, and facilities can be located anywhere in that space. We decompose the problem into several simpler subproblems, allowing us to develop an efficient algorithm for determining the optimal number, location, and capacity of the facilities.

Our results point to the importance of the equitable facility locations that ensure that each facility is facing approximately the same demand. This leads to the deterministic ELP and the associated concept of EFC, which is an ideal solution to the ELP. Our computational results indicate that the optimal solutions often possess EFC (or near-EFC) properties, suggesting a robust heuristic decision rule for the SCFLP. Moreover, we show that the equitable facility locations may be optimal even when the requirement that all facilities be identical is dropped.

This research can be extended in several ways. First, the ability to successfully solve the SCFLP critically depends on the ability to solve the ELP. The latter appears to be a new form of the deterministic facility location problem, with a clear relationship to the *p*-center problem. While we provide some analysis of the ELP and existence of EFC vectors for the case of linearly distributed demand, much work remains to be done in more general settings. A follow-up paper (Baron et al. 2007) is a first step in this direction. Second, while some generalizations of the SCFLP model have been presented here, several strong assumptions remain. In particular, more work is needed on relaxing the assumption that the fixed location costs are uniform, as well as on the more complete analysis of the problem with non-identical facilities. Third, as mentioned, a special purpose algorithm to solve ELPd is required.

Acknowledgments

The authors thank the referees and, in particular, the associate editor for their many helpful comments. This research was supported by National Sciences and Engineering Research Council (NSERC) grants to the three authors.

References

Abate, J., G. L. Choudhury, W. Whitt. 1995. Exponential approximation for tail probabilities in queues, I: Waiting times. *Oper. Res.* **43**(5) 885–901.

Aurenhammer, F. 1991. Voronoi diagrams: A survey of a fundamental data structure. *ACM Comput. Surveys* **23** 329–371.

Baron, O., O. Berman, D. Krass, Q. Wang. 2007. The equitable facility location problem on a plane. *Eur. J. Oper. Res.* **183**(2) 578–590.

Beasley, J. E. 1990. OR-library-distributing test problems by electronic mail. *J. Oper. Res. Soc.* **41** 1069–1072.

Berman, O., D. Krass. 2002. Facility location problems with stochastic demand and congestion. Z. Drezner, H. Hamacher, eds. *Facility Location: Applications and Theory*, Chapter 11. Springer-Verlag, Berlin, Germany.

Berman, O., D. Krass, J. Wang. 2006. Locating service facilities to reduce lost demand. *IIE Trans. Scheduling Logist.* **38** 933–946.

Church, R. L., C. ReVelle. 1982. The maximal covering location problem. *Papers Regional Sci. Assoc.* **32** 101–118.

Cohen, J. W. 1982. *The Single Server Queue*, 2nd ed. North Holland, Amsterdam, The Netherlands.

Corless, R. M., G. H. Gonett, D. E. G. Harc, D. J. Jeffrey, D. E. Knuth. 1996. On the Lambert W function. *Adv. Computational Math.* **5** 329–359.

Courrent, J. R., M. S. Daskin, D. A. Schilling. 2002. Discrete network location models. Z. Drezner, H. Hamacher, eds. *Facility Location: Applications and Theory*, Chapter 3. Springer-Verlag, Berlin, Germany.

Gallager, R. G. 1996. *Discrete Stochastic Processes*. Kluwer Academic Publishers, Boston, MA.

Gerrard, R. A., R. L. Church. 1996. Closest assignment constraints and location models: Properties and structure. *Location Sci.* **4** 251–270.

Marianov, V., M. Rios. 2001. A probabilistic quality of service constraint for a location model of switches in ATM communication networks. *Ann. Oper. Res.* **96** 237–243.

Marianov, V., D. Serra. 1998. Probabilistic maximal covering location-allocation for congested systems. *J. Regional Sci.* **38**(3) 401–424.

Marianov, V., D. Serra. 2002. Location-allocation of single and multiple server service centers with constrained queues or service times. *Ann. Oper. Res.* **111** 35–50.

Plastria, F. 2002. Continuous covering location problems. Z. Drezner, H. Hamacher, eds. *Facility Location: Applications and Theory*, Chapter 2. Springer-Verlag, Berlin, Germany.

Ross, S. M. 1974. Bounds on the delay distribution in Gi/G/1 queues. *J. Appl. Probab.* **11** 417–421.

Suzuki, A., Z. Drezner. 1996. The *p*-center location problem in an area. *Location Sci.* **4**(1/2) 69–82.

Suzuki, A., A. Okabe. 1995. Using Voronoi diagrams. Z. Drezner, ed. *Facility Location: A Survey of Application and Methods*, Chapter 6. Springer Series in Operations Research, New York.

Wang, Q., R. J. Batta, C. M. Rump. 2002. Algorithms for a facility location problem with stochastic customer demand and immobile servers. *Ann. Oper. Res.* **111** 17–35.

Weber, R. R. 1983. A note on waiting times in single server queues. *Oper. Res.* **31** 950–951.