CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

Hub-and-Spoke System Design for Freight Transportation with Priority Consignment Classes

Navneet Vidyarthi
Sachin Jayaswal
Rajesh Tyagi

November 2013

CIRRELT-2013-68

Bureaux de Montréal :
Université de Montréal
Pavillon André-Aisenstadt
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

Bureaux de Québec :
Université Laval
Pavillon Palasis-Prince
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

Hub-and-Spoke System Design for Freight Transportation with Priority Consignment Classes

Navneet Vidyarthi[1,*], Sachin Jayaswal[2], Rajesh Tyagi[3]

[1] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT) and Department of Supply Chain and Technology Management, John Molson School of Business, Concordia University, 1455, De Maisonneuve Blvd. West, Montreal, Canada, H3G 1M8

[2] Indian Institute of Management, Vastrapur, Ahmedabad, Gujarat, 380 015 India

[3] Department of Logistics and Operations Management, HEC Montréal, 3000, Côte-Sainte-Catherine, Montréal, Canada H3T 2A7

Abstract. Hub-and-spoke systems are widely used in freight transportation. In freight transportation, hub operations such as scanning, unloading sorting, weighting, and loading of consignments takes significant time. Motivated by the significance of service level at the hubs in hub-and-spoke systems, this paper presents a model for designing a capacitated single allocation *p*-hub location with stochastic demand and time-based service level constraints at the hubs. The proposed model seeks to simultaneously determine the location and the capacity of hubs and allocate flows to hubs so as to minimize the fixed cost of locating hubs and equipping them with processing capacity and the variable transportation costs subject to the service level constraints. The problem is setup as a network of spatially distributed hubs modelled as priority queues with Poisson arrival rate and exponential service times. We present a matrix geometric approach to deal with the service level constraints associated with priority flow classes. The numerical experiments are conducted with Civil Aeronautics Board (CAB) data sets. Through a numerical example, we illustrate the impact of service-level constraints on the configuration of the network. We show that substantial improvement in service level can be achieved with small increase in total costs in the design of hub-and-spoke systems.

Keywords: Hub location, capacity selection, network design, service-level constraints, priority queues, matrix geometric method.

Acknowledgements. This research was supported by the Natural Science and Engineering Research Council of Canada (NSERC) grant to the first author, and by the Research & Publication Grant, Indian Institute of Management Ahmedabad to the second author. The authors would like to acknowledge Mr. Kartikeya Mohan Sahai (Indian Institute of Technology, Guwahati) for writing the code and conducting the experiments.

---

* Corresponding author: Navneet.Vidyarthi@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2013

1. Introduction

Freight transportation has emerged as one of the most critical part of the global economy with the rising volume of trade across the globe. Freight transportation through hub-and-spoke network involves operations at hubs such as scanning, unloading sorting, weighting, and loading of consignments. For example, companies such as FedEx, UPS, DHL, and the United States Postal Service (USPS) receive and deliver millions of packages every day. UPS handles 2.3 million packages a day, accounting for 22% of its domestic revenues. Every night, at the FedEx Memphis hub, 2.2 million packages are scanned, sorted, weighted, and routed. These packages are not homogenous and have different service delivery requirements. With such large volumes of packages transported between many different origin-destination points, many sorting hub centers are involved. While some of the packages are for regular delivery that takes a week or more (low priority consignment), others are time-sensitive and have to be expedited (high priority consignments) for delivery within 24 hours. The configuration of the hub-and-spoke network plays an important role in the distribution of such large volumes of packages transported between many different origin-destination points. It is paramount that the delivery networks are designed and operated efficiently so as to be able to meet the time-based service delivery guarantees. For example, FedEx has strategically designed its hub-and-spoke network and located its hub at Memphis, Tennessee in order to serve the entire United States by providing overnight service to the entire nation and serving 95% of the global economy (220 countries on six continents) customers within 24-48 hours.

Hub-and-spoke network design problems have been widely studied. The first mathematical formulation of discrete hub location problems was proposed by O'Kelly (1987). Since then several variants and extensions have been formulated and studied. Amongst the hub location problems, the one that has received the most attention in literature is the p-hub median problem (Alumur and Kara, 2008). In the classical p-hub median problem, the objective is to locate p hubs from a set of n nodes in a graph, establishing a complete subgraph of these hubs by allocating spoke to the hubs and directing flows between pairs of origins-destinations through the hubs in the network. Single allocation p-hub median problem, implies that all the incoming and outgoing traffic of every demand node is routed through a single hub; whereas in multiple allocation, each demand node can receive and send flows through more than one hub. The single allocation p-hub median problem and its variants have been studied by O'Kelly (1987), Campbell (1994b), O'Kelly et al. (1995), Campbell (1996), Ernst and Krishnamoorthy (1996), Skorin-Kapov et al. (1996), Smith et al. (1996), Song and Park (2000), Abdinnour-Helm (2001), Ebery (2001), Yaman (2009), and Elhedhli and Wu (2010) among others. Solution methods for hub location problems include GRASP (greedy randomized adaptive search procedure) (Klincewicz, 1992), tabu search (Klincewicz, 1992), simulated annealing (Ernst and Krishnamoorthy, 1996), genetic algorithm (Topcuoglu et al., 2005; Kratica et al., 2007), evolutionary algorithms (Koksalan and Soylu, 2010), Neural networks (Smith et al., 1996), Lagrangean relaxation (Elhedhli and Hu, 2005; Contreras et al., 2009; Elhedhli and Wu, 2010), Benders decomposition (Camargo et al., 2008, 2009; Contreras et al., 2012), branch and bound (Ernst and Krishnamoorthy, 1996; Ebery, 2001),

among others. For more details, interested readers are referred to surveys by Campbell et al. (2002), Alumur and Kara (2008) and Campbell and O'Kelly (2012).

Marianov and Serra (2003) present models for location of hubs in airline networks, where hubs are modelled as M/D/c queues. Their model considers probabilistic service level constraints, which limit the number of planes in the queue. The model is solved using a tabu search based heuristic approach. Sim et al. (2009) present the stochastic p-hub center problem and use a chance-constrained formulation to model the minimum service-level requirement. The model takes into account the variability in travel times when designing the hub network so that the maximum travel time through the network is minimized. They present a linear MIP formulation for the problem, under the assumption that travel times on the arcs are independent normal random variables. Yang (2009) presents a two-stage stochastic programming model for hub location and route planning in air freight transportation under seasonal demand variations. Contreras et al. (2011) study hub location problems in which uncertainty is associated with demand and transportation costs. They show that stochastic models with uncertain demand or dependent transportation costs are equivalent to their associated deterministic expected value problem (EVP), in which random variables are replaced by their expected values. However, in case of uncertain independent transportation costs, the corresponding stochastic model is not equivalent to its EVP, for which they present a solution methodology based on Monte-Carlo simulation that integrates a sample average approximation scheme with a Benders decomposition algorithm.

The objective of this paper is to study the effect of incorporating service level constraint for the different classes of flows (customers) on the configuration of the hub-and-spoke network and to analyze the tradeoffs between hub location and capacity acquisition cost and transportation cost. We present a model that simultaneously determines the location and the capacity of hubs and allocate flow to these hubs such that the service-level constraints for the different customer classes are met. The service level constraint is defined at the fraction of flow served at a hub within a sojourn time (waiting in queue + service time) and hence is based on the complete distribution of sojourn time, and not just their averages. The problem is setup as a network of spatially distributed hubs modelled as priority queues with Poission arrival rate and exponential service times. We present a matrix geometric approach to deal with the service level constraints associated with priority flow classes. To the best of our knowledge, this is the first paper to model hubs as priority queues to account for non-homogenous consignment classes that have different delivery time requirements.

The remainder of the paper is organized as follows. In §2, we present the description of the problem and the mathematical formulation. In section 3, we present the solution methodology. §4 presents the sensitivity analysis and managerial insights using an illustrative example. Finally, we conclude with future research directions in §5.

2. Problem Description and Model Formulation

We extend the formulation of the basic uncapacitated single-allocation p-hub median problem (USApHMP) proposed by Skorin-Kapov et al. (1996) to build our model. For that, consider a graph $G = (N, A)$, where N be the set of nodes that exchange traffic and are potential hub locations. Let k and m be indices for potential hub locations and i and j be indices for the origin and destination nodes. If λ_{ij} is the amount of flow to be shipped from origin node i to destination j, then the total transportation cost of routing the flow from origin i to destination j routed via hubs k and m is given by: $C_{ijkm} = \lambda_{ij}(\chi c_{ik} + \beta c_{km} + \delta c_{mj})$, where χ is coefficient of the collection cost (per unit flow) from any origin to any hub node, δ is the coefficient of the collection cost (per unit flow) from any hub node to any destination, β is the discount factor from any hub node to any other hub node, and c_{ij} is the transportation cost per unit of flow from node i to node j. Let z_{ik} be 1 if node i is allocated to hub k and 0 otherwise; in particular, $z_{kk} = 1$ implies that node k is selected as a hub. The routing variable x_{ijkm} equals 1 if the total flow from node i to node j is routed via hubs located at k and m, in that order. With these notations, the formulation of (USApHMP) is as follows:

$$
\begin{aligned}
[\text{USA}p\text{HMP}]: \quad \min \quad & \sum_i \sum_j \sum_k \sum_m C_{ijkm} x_{ijkm} && && (1)\\
\text{s.t.} \quad & \sum_k z_{ik} = 1 && \forall i && (2)\\
& z_{ik} \leq z_{kk} && \forall i, k && (3)\\
& \sum_k z_{kk} = p && && (4)\\
& \sum_m x_{ijkm} = z_{ik} && \forall i, j, k && (5)\\
& \sum_k x_{ijkm} = z_{jm} && \forall i, j, m && (6)\\
& x_{ijkm}, z_{ik} \in \{0, 1\} && \forall i, j, k, m && (7)
\end{aligned}
$$

The objective function (1) minimizes the sum of total transportation cost of flow of commodities between all the origin-destination node pairs. Constraint set (2) ensures that every node is assigned to exactly one hub. Constraint set (3) guarantees that a node will be assigned to a open hub. Constraint (4) ensures that exactly p hubs are opened in the network. Constraint sets (5) and (6) ensure that all the traffic between an origin-destination pair have been routed via a hub sub-network.

We extend the model to account for stochastic demand and service level constraints for the priority consignment classes. Without the loss of generality, the model remains valid for multiple consignment classes, $N \geq 2$. Let h represent the high priority class customers and l represent the lower priority class customers. We assume that the arrival rate of flow (demand) from high and low priority class customers that has to be shipped from origin i to destination j is an independent random variable that follows a Poisson process with mean λ_{ij}^h and λ_{ij}^l respectively. In that case, the aggregate flow rate through hub k is also a random

variable that follows a Poisson process with mean $\Lambda_k = \sum_i \sum_j (\lambda_{ij}^h + \lambda_{ij}^l) z_{ik}$ (due to the superposition of Poisson processes). Let us model service times at hubs as random variables that follow an exponential distribution with mean μ_{kl}, where l is the index for capacity level. The service rate reflects the server capacity or essentially the units of flow a hub can serve in a given time period. For every hub node k, we allow the model to select one of the discrete capacity levels, $\mu_{k1}, \mu_{k2}, ..., \mu_{kN}$ with fixed costs $F_{k1}, F_{k2}, ..., F_{kN}$ respectively. The fixed cost refers to the cost of using the hub (airport) amortized over the planning period. Each hub can be modelled as an M/M/1 queue, where the mean service rate of hub k, if it is allocated capacity level l, is given by $\mu_k = \sum_{l=1}^{L} \mu_{kn} y_{kn}$. In steady-state, the stability condition of the queueing system implies that at every hub node, the arrival rate of any consignment is less than the total service rate: $\Lambda_k < \mu_k$, $\forall k$. Alternatively, if ρ_k is the utilization of hub k ($\rho_k = \Lambda_k / \mu_k$), then $\rho_k < 1$.

We specify the service level constraints as the fraction of flow served within a specified sojourn time (waiting in queue + service time). This can expressed as the probability that a flow spends more than τ time units in service at hubs does not exceed α for some finite τ and $\alpha \in (0, 1)$. For a given flow routing $\mathbf{x}$ and location and capacity level $\mathbf{y}$, let the arrival rates and service rate at a hub k be denoted by $\Lambda_k^h(\mathbf{x})$, $\Lambda_k^l(\mathbf{x})$ and $\mu_k(\mathbf{y})$. If we let $W_k^h(\Lambda_k, \mu_k)$ and $W_k^l(\Lambda_k, \mu_k)$ denote the total time spent by the two classes in the system at hub k and τ^h and τ^l are their target service times, then the service level can be expressed as follows:

$$P\{W_k^h\left(\Lambda_k^h(\mathbf{x}), \mu_k(\mathbf{y})\right) \leq \tau^h\} \geq \alpha_h z_{kk} \qquad \forall k$$
$$P\{W_k^l\left(\Lambda_k^l(\mathbf{x}), \mu_k(\mathbf{y})\right) \leq \tau^l\} \geq \alpha_l z_{kk} \qquad \forall k$$

The resulting integer programming model is as follows:

$$[\mathrm{P}]: \quad \min \quad \sum_i \sum_j \sum_k \sum_m C_{ijkm} x_{ijkm} + \sum_k \sum_{l \in L_k} F_{kl} y_{kl} \tag{8}$$

$$\text{s.t.} \quad (2) - (6)$$

$$\sum_i \sum_j (\lambda_{ij}^h + \lambda_{ij}^l) z_{ik} \leq \sum_{l \in L_k} \mu_{kl} y_{kl} \qquad \forall k \tag{9}$$

$$\sum_{l \in L_k} y_{kl} = z_{kk} \qquad \forall k \tag{10}$$

$$P\{W_k^h\left(\Lambda_k^h(\mathbf{x}), \mu_k(\mathbf{y})\right) \leq \tau^h\} \geq \alpha_h z_{kk} \qquad \forall k \tag{11}$$

$$P\{W_k^l\left(\Lambda_k^l(\mathbf{x}), \mu_k(\mathbf{y})\right) \leq \tau^l\} \geq \alpha_l z_{kk} \qquad \forall k \tag{12}$$

$$x_{ijkm}, y_{kl}, z_{ik} \in \{0, 1\} \qquad \forall i, j, k, m, l \in L_k, n \tag{13}$$

The underlying model is difficult to solve due to the lack of closed form expression for service-level constraint (12) for lower priority class.

3. Solution Methodology

3.1. Estimation of Service-Level Function of High Priority Class Customers

The tail of the sojourn time distribution $S_k^h(.)$ for *high priority customers* in a preemptive priority queue us known to be exponential **?** and is given by:

$$S_k^h(.) = P(W_k^h \leq \tau^h) = 1 - e^{-(\mu_k - \Lambda_k^h)W_k}$$

The service level constraint for the high priority customer (12) can be expressed as a linear constraint as follows:

$$\sum_{l \in L_k} \mu_{kl} y_{kl} - \sum_i \sum_j \lambda_{ij}^h z_{ik}^h \geq \frac{-\ln(1-\alpha_h)}{\tau_h} z_{kk} \qquad \forall k \qquad (14)$$

Proposition 1: *The sojourn time distribution of higher priority customers $S_k^h(\mathbf{x}, \mathbf{y}, \mathbf{z}, \tau^n, \alpha_n)$ is (i) concave in μ_k and (ii) concave in Λ_k^h.*

Differentiating $S_k^h(.)$ w.r.t. μ_k twice, we get $\frac{\delta^2 S_k^h(.)}{\delta^2 \mu_k} < 0$, which proves that the function is concave in μ_k. Similarly, differentiating $S_k^h(.)$ w.r.t. Λ_k^h twice, we get $\frac{\delta^2 S_k^h(.)}{\delta^2 \Lambda_k^h} < 0$, which proves that the function is concave in Λ_k^h.

3.2. Estimation of Service-Level Function of Low Priority Class Customers

In this subsection, we describe a procedure based on matrix geometric methods for estimating the service level function of low priority class customers $S_k^l(.)$ and its subgradients. Details regarding this method can be found in Neuts (1981) and Latouche and Ramaswami (1999). The matrix-geometric methods can provide *near-exact* estimates of service level function in some cases.

3.2.1. Joint Stationary Queue Length Distribution

Let us determine the joint distribution of queue lengths. For that, let the number of high and low priority customers in the system (including the one in the service) at time t be denoted by $N_h(t)$ and $N_l(t)$, respectively. We assume that $N_l(t) \geq 0$ (infinite low priority class buffer size) whereas $0 \leq n_h \leq M$ (the buffer size of the high priority customers in the system be M). No other state variables are required to model the system since the service is exponential and it is not necessary to keep track of which type of customer the server is attending to. As long as there is at least one high priority customer present in the system, the system must be busy attending to high priority queue. Therefore, $N_h(t)$ and $N_l(t)$ are the state variables representing the number of high and low priority customers in the system at time t, and $\{\mathbf{N}(t)\} := \{N_l(t), N_h(t), t \geq 0\}$ is a continuous-time two-dimensional Markov chain with state space $\{\mathbf{n} = (n_l, n_h) | n_l \geq 0, \;\; 0 \leq n_h \leq M\}$. The key idea we employ here is that $\{\mathbf{N}(t)\}$ is a *quasi-birth-and-death* (QBD) process, which allows us to develop a matrix geometric solution for the joint distribution of the number of customers of each class in the

system.

In the Markov process $\{\mathbf{N(t)}\}$, a transition can occur only if a customer of either class arrives or a customer of either class is served. For example, with the arrival of a high priority customer with rate λ_h, the system transits from state $\mathbf{n} = \{(n_l, n_h)\}$ to $\mathbf{n'}$, where $\mathbf{n'} = \{(n_l, n_h{+}1\}$ and with the arrival of a low priority customer with rate λ_l, the system transits from state $\mathbf{n}$ to $\mathbf{n''} = \{(n_l{+}1, n_h)\}$. Similarly, with the service of a high priority customer with rate μ, the system transits from state $\mathbf{n} = \{(n_l, n_h)|n_l \geq 0, n_h > 0\}$ to $\dot{\mathbf{n}} = \{(n_l, n_h{-}1)\}$ and with the service of a low priority customer with rate μ, the system transits from state $\mathbf{n} = \{(n_l, n_h)|n_l \geq 0, n_h = 0\}$ to $\ddot{\mathbf{n}} = \{(n_l - 1, n_h)\}$. We order the states of the system lexicographically, i.e. $(0,0), (0,1), (0,2), \ldots, (0,M)$; $(1,0), (1,1), (1,2), \ldots, (1,M)$; $\ldots$; $(i,0), (i,1), (i,2), \ldots, (i,M)$, and define $\pi_{(i,s)}$ to be the stationary probability of the state (i,s). With n_l serving as the level and n_h as the sub-level, the infinitesimal generator of the chain $\{\mathbf{N(t)}\}$ for $n_l = 0, 1, 2$ and $n_h = 0, 1, ..., M$ is given by:

$$Q = \left(\begin{array}{c|cccc|cccc|cccc}
 & (0,0) & (0,1) & (0,...) & (0,M) & (1,0) & (1,1) & (1,...) & (1,M) & (2,0) & (2,1) & (2,...) & (2,M) \\
(0,0) & -\delta_1 & \lambda_h & & & \lambda_l & & & & & & & \\
(0,1) & \mu & -\delta_2 & \lambda_h & & & \lambda_l & & & & & & \\
(0,...) & & \mu & -\delta_2 & \lambda_h & & & \lambda_l & & & & & \\
(0,M) & & & \mu & -\delta_3 & & & & \lambda_l & & & & \\
\hline
(1,0) & \mu & & & & -\delta_2 & \lambda_h & & & \lambda_l & & & \\
(1,1) & & & & & \mu & -\delta_2 & \lambda_h & & & \lambda_l & & \\
(1,...) & & & & & & \mu & -\delta_2 & \lambda_h & & & \lambda_l & \\
(1,M) & & & & & & & \mu & -\delta_3 & & & & \lambda_l \\
\hline
(2,0) & & & & & \mu & & & & -\delta_2 & \lambda_h & & \\
(2,1) & & & & & & & & & \mu & -\delta_2 & \lambda_h & \\
(2,...) & & & & & & & & & & \mu & -\delta_2 & \lambda_h \\
(2,M) & & & & & & & & & & & \mu & -\delta_3
\end{array}\right)$$

where $\delta_1 = \lambda_h + \lambda_l$, $\delta_2 = \lambda_h + \lambda_l + \mu$, and $\delta_3 = \mu + \lambda_l$.

The entries of the infinitesimal generator matrix can be grouped into blocks to form a block-tridiagonal matrix as follows:

$$Q = \begin{pmatrix}
B_0 & A_0 & & & \\
A_2 & A_1 & A_0 & & \\
 & A_2 & A_1 & A_0 & \\
 & & \ddots & \ddots & \ddots
\end{pmatrix}$$

where B_0, A_0, A_1, A_2 are square matrices of order $M + 1$. These matrices can be easily constructed using the transition rates described above.

$$A_0 = \begin{pmatrix} \lambda_l & & & & \\ & \lambda_l & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \lambda_l \end{pmatrix}; \quad A_2 = \begin{pmatrix} \mu & & & & \\ & 0 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 0 \end{pmatrix}; \quad B_0 = \begin{pmatrix} * & \lambda_h & & & \\ \mu & * & \lambda_h & & \\ & \mu & * & \lambda_h & \\ & & \ddots & \ddots & \ddots \\ & & & \mu & * \end{pmatrix}$$

where $*$ is such that $A_0\mathbf{e} + B_0\mathbf{e} = \mathbf{0}$. $A_1 = B_0 - A_2$.

The matrix B_0 contains all transitions when no low priority customers are present in the system and the server is devote to serving high priority customers. A_1 contains all transitions that represents arrivals of low priority customers, whereas A_{-1} contains transitions corresponding to the service of low priority customer. Since n_l can only change by ± 1, the only non-zero matrices are A_1, A_0, and A_{-1}.

We denote the steady-state probability vector of $\{\mathbf{N}(t)\}$ by $\pi \equiv (\underline{\pi}_0, \underline{\pi}_1, \underline{\pi}_2, \underline{\pi}_3, ...)$, where $\underline{\pi}_i \equiv (\pi_{(i,0)}, \pi_{(i,1)}, \pi_{(i,2)}, \pi_{(i,3)}, ..., \pi_{(i,M)})$. The vector $\mathbf{x}$ can be partitioned by levels into sub vectors $\mathbf{x}_i$, $i \geq 0$, where $\mathbf{x}_i = [x_{i0}, x_{i1}, \ldots, x_{iM}]$ is the stationary probability of states in level i ($n_l = i$). Thus, $\mathbf{x} = [\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \ldots]$. $\mathbf{x}$ can be obtained using a set of balance equations, given in matrix form by the following standard relations (Latouche and Ramaswami, 1999; Neuts, 1981):

Let $\mathbf{x} = [\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \ldots]$ be the stationary distribution of N(t), where $\mathbf{x}_k$ is the stationary probabilities of states in level k ($n_2 = k$). $\mathbf{x}$ can be solved using the balance equations, given in matrix form by:

$$\pi Q = \mathbf{0}; \quad \pi_{i+1} = \pi_i R$$

where R is the minimal non-negative solution to the matrix quadratic equation:

$$A_0 + RA_1 + R^2 A_2 = \mathbf{0}$$

The matrix R can be computed using well known methods (Latouche and Ramaswami, 1999). A simple iterative procedure often used is:

$$R(0) = 0\ ; \quad R(n+1) = -\left[A_0 + R^2(n)A_2\right] A_1^{-1}$$

The probabilities π_0 are determined from:

$$\pi_0(B_0 + RA_2) = \mathbf{0}$$

subject to the normalization equation:

$$\sum_{i=0}^{\infty} \pi_i \mathbf{e} = \pi_0 (I - R)^{-1} \mathbf{e} = 1$$

where $\mathbf{e}$ is a column vector of ones of size $M+1$.
These steady state probabilities will be used in estimating the service-level for low priority customers.

Note that this matrix geometric procedure is very efficient for obtaining the near-exact performance measures through judicious choice of the number of states. The computational implementation of the matrix geometric method, however, requires the number of states in the QBD process to be finite. For this, we treat the queue length of high priority customers (including the one in service) to be of finite size M, but of size large enough for the desired accuracy of our results. Since high priority customers are always served in priority over low priority customers, it is reasonable to assume that its queue size will always be bounded by some large number. However, the computational effort grows rapidly with the number of states and customer classes.

3.2.2. Estimation of Service-Level for Low Priority Class $S_l(\cdot)$

We derive the distribution of sojourn time of low priority customers. The sojourn time of a low priority customer W_j^l is the time between its arrival to the hub till it completes its service (i.e. waiting time in queue plus the time in service). It may be *preempted* by one or more of the high priority customers for service. Hence it is difficult to characterize the distribution of the service-level $S_k^l(.)$. However, Ramaswami and Lucantoni (1985) present an efficient algorithm for the derivation of complementary distribution of stationary waiting times in phase-type and QBD processes. Leeman (2001) uses the same approach to derive the complementary distribution of stationary waiting times in more complex queuing system. We adopt their approach to derive the distribution of sojourn time of low priority customers.

Let us consider a tagged low priority customer entering the system. The time spent by the tagged customer depends on the number of customers of either class already present in the system ahead of it, and also on the number of subsequent high priority arrivals before it completes its service. All subsequent low priority arrivals, however, have no influence on its time spent in the system. The tagged customer's time in the system is, therefore, simply the time until absorption in a modified Markov process $\{\tilde{\mathbf{N}}(t)\}$, obtained by setting $\lambda_l = 0$. Consequently, matrix $\tilde{A}_0$, representing transitions to a higher level, becomes a zero matrix. We define an *absorbing* state, call it state $0'$, as the state in which the tagged customer has finished its service. The infinitesimal generator for this process can be represented as:

$$\tilde{Q} = \left(\begin{array}{c|cccccc} 0 & 0 & 0 & 0 & 0 & \cdots \\ \hline b_0 & \tilde{B}_0 & 0 & & & \\ 0 & A_2 & \tilde{A}_1 & 0 & & \\ 0 & & A_2 & \tilde{A}_1 & 0 & \\ \vdots & & & \ddots & \ddots & \ddots \end{array}\right)$$

where, $\tilde{B}_0 = B_0 + A_0$; $\tilde{A}_1 = A_1 + A_0$; and $b_0 = [\mu \quad 0 \quad \cdots \quad 0]_{M+1}^T$. The first row and column in $\tilde{Q}$ corresponds to the absorbing state $0'$. The time spent in system by the tagged customer, which is the time until absorption in the modified Markov process with rate matrix

$\tilde{Q}$, depends on the arrival rates λ_h and λ_l and the service rate μ. For given arrival rates (λ_h, λ_l) and service rate μ, the distribution of the time spent by a low priority customer in the system is $S_l^j(\tau) = 1 - \overline{S_l^j}(\tau)$, where $\overline{S_l^j}(\tau)$ is the stationary probability that a low priority customer spends more than y units of time in the system. Further, let $\overline{S_{li}^k}(\tau)$ denote the conditional probability that a tagged customer, who finds i low priority customers ahead of it, spends a time exceeding y in the system. The probability that a tagged customer finds i low priority customers is given, using the PASTA property, by $\mathbf{x}_i = \mathbf{x}_0 R^i$. $\overline{S_l^k}(\tau)$ can be expressed as:

$$\overline{S_l^j}(\tau) = \sum_{i=0}^{\infty} \mathbf{x}_i \overline{S_{li}^k}(y)\mathbf{e} \tag{15}$$

$\overline{S_{li}^k}(y)$ can be computed more conveniently by uniformizing the Markov process $\{\tilde{\mathbf{N}}(t)\}$ with a Poisson process with rate γ, where

$$\gamma = \max_{0 \le i \le M} (-\tilde{A}_1)_{ii} = \max_{0 \le i \le M} -(A_0 + A_1)_{ii}$$

so that the rate matrix $\tilde{Q}$ is transformed into the discrete-time probability matrix:

$$\hat{Q} = \frac{1}{\gamma}\tilde{Q} + I = \left(\begin{array}{c|ccccc} 1 & 0 & 0 & 0 & 0 & \cdots \\ \hline \hat{b_0} & \hat{B}_0 & 0 & & & \\ 0 & \hat{A}_2 & \hat{A}_1 & 0 & & \\ 0 & & \hat{A}_2 & \hat{A}_1 & 0 & \\ \vdots & & & \ddots & \ddots & \ddots \end{array}\right)$$

where $\hat{A}_2 = \frac{A_2}{\gamma}$, $\hat{A}_1 = \frac{\tilde{A}_1}{\gamma} + I$, $\hat{b_0} = \frac{b_0}{\gamma}$. In this uniformized process, points of a Poisson process are generated with a rate γ, and transitions occur at these epochs only. The probability that n Poisson events are generated in time y equals $e^{-\gamma y}\frac{(\gamma y)^n}{n!}$. Suppose the tagged customer finds i low priority customers ahead of it. Then, for its time in the system to exceed y, at most i of the n Poisson points may correspond to transitions to lower levels (i.e., service completions of low priority customers). Therefore,

$$\overline{S_{li}^k}(y) = \sum_{n=0}^{\infty} e^{-\gamma y}\frac{(\gamma y)^n}{n!} \sum_{v=0}^{i} G_v^{(n)}\mathbf{e}, \qquad i \ge 0 \tag{16}$$

where, $G_v^{(n)}$ is a matrix such that its entries are the conditional probabilities, given that the system has made n transitions in the discrete-time Markov process with rate matrix $\hat{Q}$, that v of those transitions correspond to lower levels (i.e., service completions of low priority customers). Substituting the expression for $\overline{S_{li}^k}(y)$ from (16) into (15), we obtain:

$$\overline{S_l^k}(y) = \sum_{n=0}^{\infty} d_n e^{-\gamma y}\frac{(\gamma y)^n}{n!} \tag{17}$$

where, d_n is given by: $d_n = \sum_{i=0}^{\infty} \mathbf{x}_0 R^i \sum_{v=0}^{i} G_v^{(n)}\mathbf{e}, \quad n \ge 0.$

Now,

$$\begin{aligned}
\sum_{i=0}^{\infty} R^i \sum_{v=0}^{i} G_v^{(n)}\mathbf{e} &= \sum_{i=0}^{n+1} R^i \sum_{v=0}^{i} G_v^{(n)}\mathbf{e} + \sum_{i=n+2}^{\infty} R^i \sum_{v=0}^{n} G_v^{(n)}\mathbf{e} && \left(\text{since} \quad G_v^{(n)} = 0 \quad \text{for} \quad v > n\right) \\
&= \sum_{v=0}^{n+1}\sum_{i=v}^{n+1} R^i G_v^{(n)}\mathbf{e} + (I-R)^{-1}R^{n+2}\mathbf{e} && \left(\text{since} \quad \sum_{v=0}^{n} G_v^{(n)}\mathbf{e} = \mathbf{e}\right) \\
&= \sum_{v=0}^{n+1}(I-R)^{-1}(R^v - R^{n+2})G_v^{(n)}\mathbf{e} + (I-R)^{-1}R^{n+2}\mathbf{e} \\
&= \sum_{v=0}^{n}(I-R)^{-1}R^v G_v^{(n)}\mathbf{e} + (I-R)^{-1}R^{n+1}G_{n+1}^{(n)}\mathbf{e} && \left(\text{since} \quad \sum_{v=0}^{n+1} G_v^{(n)}\mathbf{e} = \mathbf{e}\right) \\
&= \sum_{v=0}^{n}(I-R)^{-1}R^v G_v^{(n)}\mathbf{e} && \left(\text{since} \quad G_v^{(n)} = 0 \quad \text{for} \quad v > n\right) \\
&= (I-R)^{-1}H_n\mathbf{e} && n \geq 0
\end{aligned}$$

where, $H_n = \sum_{v=0}^{n} R^v G_v^{(n)}$.

Therefore,

$$S_l^k(L_l) = 1 - \overline{S_l^k}(L_l) = \sum_{n=0}^{\infty} e^{-\gamma L_l}\frac{(\gamma L_l)^n}{n!}\mathbf{x}_0(I-R)^{-1}H_n\mathbf{e} \tag{18}$$

H_n can be computed recursively as:

$$H_{n+1} = H_n\hat{A}_1 + RH_n\hat{A}_2; \quad H_0 = I$$

Therefore, for given prices (p_h^k, p_l^k) and service rate (μ^k), $S_l^k(\cdot)$ in (16) can be computed using (18).

Proposition 2: *The sojourn time distribution of lower priority customers $S_k^l(\mathbf{x}, \mathbf{y}, \mathbf{z}, \tau^n, \alpha_n)$ is (i) concave in μ_k and (ii) concave in Λ_k^l.*

The plots of $S_k^l(.)$ vs. μ_k and $S_k^l(.)$ vs. Λ_k^l obtained using the matrix geometric method show that the components of $S_k^l(.)$ are concave. Intuitively, one would expect that the sojourn time increases with decreasing marginal returns as the service rate increases. Furthermore, it should decrease with increasing marginal returns as the arrival rate increases.

3.3. Linearization of Service-Level Constraints

Under this concavity assumption, the service level for *low priority customers* $S_k^l(.)$ can be approximated by a set of supporting hyperplanes that are tangent to $S_k^l(.)$ at various points

$(\Lambda_k^{hq}, \Lambda_k^{lq}, \mu_k^q), \forall q \in Q$, that is

$$S_k^l(.) = \min_{q \in Q} \left\{ S_k^{lq}(.) + (\Lambda_k^l - \Lambda_k^{lq}) \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^l} + (\Lambda_k^h - \Lambda_k^{hq}) \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^h} + (\mu_k - \mu_k^q) \frac{\partial S_j^{lq}(.)}{\partial \mu_k} \right\} \ \forall k \quad (19)$$

where $S_k^l(.)$ denotes the value of function $S_k^l(.)$ at a fixed point $(\Lambda_k^{hq}, \Lambda_k^{lq}, \mu_k^q), \forall q \in Q$, and $\frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^h}$, $\frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^l}$ and $\frac{\partial S_k^{lq}(.)}{\partial \mu_k}$ are the subgradients of $S_k^{lq}(.)$ at points $(\Lambda_k^h,\ \Lambda_k^l,\ \mu_k)$.

This expression can be written as

$$S_k^l(.) \leq \quad S_k^{lq}(.) + (\Lambda_k^l - \Lambda_k^{lq}) \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^l} + (\Lambda_k^h - \Lambda_k^{hq}) \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^h} + (\mu_k - \mu_k^q) \frac{\partial S_k^{lq}(.)}{\partial \mu_k} \quad \forall k, q \in Q$$

This implies that the service level constraint for low priority customers can be expressed as a set of linear constraints as follows:

$$\frac{\partial S_k^{lq}(.)}{\partial \mu_k} \sum_{l \in L_k} \mu_{kl} y_{kl} + \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^l} \sum_i \sum_j \lambda_{ij}^l z_{ik} + \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^h} \sum_i \sum_j \lambda_{ij}^h z_{ik} \geq \left\{ \alpha^l - S_k^{lq}(.) + \Lambda_k^{lq} \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^l} + \Lambda_k^{hq} \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^h} + \mu_k^q \frac{\partial S_k^{lq}(.)}{\partial \mu_k} \right\} z_{kk} \quad \forall k, q \in Q \quad (20)$$

The resulting linear MIP model $[PL]$ is as follows:

$$\min \ \sum_i \sum_j \sum_k \sum_m C_{ijkm} x_{ijkm} + \sum_k \sum_{l \in L_k} F_{kl} y_{kl} \quad (21)$$

s.t. (2) − (6)

$$\sum_{l \in L_k} y_{kl} = z_{kk} \qquad \forall k \quad (22)$$

$$\sum_{l \in L_k} \mu_{kl} y_{kl} - \sum_i \sum_j \lambda_{ij}^h z_{ik} \geq \frac{-\ln(1-\alpha^h)}{\tau_k^h} z_{kk} \qquad \forall k \quad (23)$$

$$\frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^l} \sum_i \sum_j \lambda_{ij}^l z_{ik} + \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^h} \sum_i \sum_j \lambda_{ij}^h z_{ik} + \frac{\partial S_k^{lq}(.)}{\partial \mu_j} \sum_k \mu_{jk} y_{jk} \geq \left\{ \alpha^l - S_k^{lq}(.) + \Lambda_k^{lq} \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^l} + \Lambda_k^{hq} \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^h} + \mu_k^q \frac{\partial S_k^{lq}(.)}{\partial \mu_k} \right\} z_{kk} \quad \forall k, q \in Q \quad (24)$$

$$x_{ijkm}, y_{kl}, z_{ik} \in \{0, 1\} \qquad \forall i, j, k, m, l \quad (25)$$

In the next subsection, we describe the procedure for estimating the the subgradients of service level function $S_j^l(.)$.

3.4. Estimation of Subgradients of the Service Level Function

We use finite difference method as it has been shown to provide better estimates of gradients (Andradottir, 1998; Atlason et al., 2004, 2008). Gradient estimation through finite difference method can be obtained using forward differences, backward differences, or central

differences. We use central differences as they usually provide an estimate that has less bias than the forward or backward differences (Andradottir, 1998).

In order to estimate the subgradients of a function (i.e. partial derivatives with respect to a continuous variable) using central finite differences, the function is evaluated at two different points. Then an estimate of the partial derivative at a particular value can be found by linear interpolation. If the variable is integer, then the smallest difference between the two points is one. In our case, the arrival rates Λ_k^h and Λ_k^l are continuous variables as $0 \leq x_{ij} \leq 1$, and service rate μ_k is a discrete variable as $y_{jk} \in \{0, 1\}$.

If $\frac{\partial S_j^l(\Lambda_k^h, \Lambda_k^l, \mu_j)}{\partial \Lambda_j^h}$, $\frac{\partial S_k^l(\Lambda_k^h, \Lambda_k^l, \mu_j)}{\partial \lambda_j^l}$, and $\frac{\partial S_j^l(\Lambda_k^h, \Lambda_k^l, \mu_j)}{\partial \mu}$ denote the subgradient of $S_j^l\left(\Lambda_k^h, \Lambda_k^l, \mu_j\right)$, then the central finite difference estimate are obtained as follows:

$$\frac{\partial S_k^l(\Lambda_k^h, \Lambda_k^l, \mu_k)}{\partial \Lambda_k^h} \simeq \frac{\widehat{S}_k^l\left(\Lambda_k^h + d\Lambda_k^h, \Lambda_k^l, \mu_k\right) - \widehat{S}_k^l\left(\Lambda_k^h - d\Lambda_k^h, \Lambda_k^l, \mu_k\right)}{2d\Lambda_k^h}$$

$$\frac{\partial S_k^l(\Lambda_k^h, \Lambda_k^l, \mu_k)}{\partial \Lambda_k^l} \simeq \frac{\widehat{S}_k^l\left(\Lambda_k^h, \Lambda_k^l + d\Lambda_k^l, \mu_k\right) - \widehat{S}_k^l\left(\Lambda_k^h, \Lambda_k^l - d\Lambda_k^l, \mu_k\right)}{2d\Lambda_k^l}$$

$$\frac{\partial S_k^l(\Lambda_k^h, \Lambda_k^l, \mu_k)}{\partial \mu} \simeq \frac{\widehat{S}_k^l\left(\Lambda_k^h, \Lambda_k^l, \mu_k + d\mu_k\right) - \widehat{S}_k^l\left(\Lambda_k^h, \Lambda_k^l, \mu_k - d\mu_k\right)}{2d\mu}$$

where $d\Lambda_k^h$, $d\Lambda_k^l$ and $d\mu_k$ (referred as step size) are the incremental change in arrival rate of high priority, arrival rate of low priority and service rate at hub k respectively. Note that the symbols $\widehat{S}_k^l(.)$ denote the estimates of $S_k^l(.)$ obtained from numerical computations (using the matrix geometric method) for their corresponding parameter values. It is clear that we would conduct six computational runs to obtain these three estimates of subgradients at a point $(\Lambda_k^h, \Lambda_k^l, \mu_k)$ for every hub k that is selected open ($z_{kk} = 1$). These estimates of subgradients are used to generate the constraints of the form (24).

3.5. Solution Algorithm

The linear model $[PL]$ with infinite number of constraints is amenable to an iterative cutting plane method, where the service level and its subgradients are estimated using matrix geometric method. It differs from the traditional description of the algorithm only in that we use matrix geometric method to evaluate the service level function and its subgradients due to the lack of existence of an algebraic expression for the function. The idea is to optimize a relaxed version of the problem by generating cuts from the violated service-level constraints and adding corresponding linear constraints until the optimal solution of the relaxed problem is feasible for the original problem. For that, we relax the service-level constraints (24), and solve the linear IP model to obtain an initial solution (x^0, y^0, z^0). Using the flow allocation and the capacity level at the hubs, we compute the aggregate arrival rates (λ_j^n) and service rates (μ_j) at all the hubs selected open. We then use matrix geometric method with the arrival rates and service rates obtained from the solution to get the estimates of service level function S_j^l and its three subgradients. If these estimates satisfy the service-level constraints (24), then we stop with the optimal solution to model $[P]$, else we add a

set of linear constraints of the form (24) to the relaxed problem so that it will eliminate the current solution without eliminating any feasible solution. This procedure is repeated until all the service level constraints are satisfied. The convergence of this solution procedure can be proved along the lines of Atlason et al. (2004, 2008). The steps of the algorithm are summarized as follows:

Algorithm 1 Cutting Plane Algorithm

Ensure: $UB \leftarrow \infty; LB \leftarrow -\infty; q \leftarrow 0$

1: Solve $PL(H^q)$ without any service level constraints to obtain $(\overline{x}^q, \overline{y}^q, \overline{z}^q)$ and $(\Lambda_k^{hq}, \Lambda_k^{lq}, \mu_k^q)$.
2: Compute service-level function $S_k^{lq}(\Lambda_k^{hq}, \Lambda_k^{lq}, \mu_k^q)$ using the matrix geometric method.
3: **while** the service level constraint is not satisfied, i.e. $S_k^{lq}(.) < \alpha_l$, **do**
4: Compute gradients of service-level function: $\frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^l}$, $\frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^h}$, and $\frac{\partial S_k^{lq}(.)}{\partial \mu_k}$.
5: Generate new constraints:

$$\frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^l} \sum_i \sum_j \lambda_{ij}^l z_{ik} + \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^h} \sum_i \sum_j \lambda_{ij}^h z_{ik} + \frac{\partial S_k^{lq}(.)}{\partial \mu_j} \sum_k \mu_{jk} y_{jk} \geq$$
$$- \left\{ S_k^{lq}(.) + \Lambda_k^{lq} \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^l} + \Lambda_k^{hq} \frac{\partial S_k^{lq}(.)}{\partial \Lambda_k^h} + \mu_k^q \frac{\partial S_k^{lq}(.)}{\partial \mu_k} + \alpha^l \right\} z_{kk} \quad \forall k, q \in Q$$

6: Append new constraints: $H^{q+1} \leftarrow H^q \cup \{h_{new}\}$
7: $q \leftarrow q + 1$
8: Solve $PL(H^q)$ to obtain $(\overline{x}^q, \overline{y}^q, \overline{z}^q)$ and $(\Lambda_k^{hq}, \Lambda_k^{lq}, \mu_k^q)$.
9: Compute service-level function $S_k^{lq}(\Lambda_k^{hq}, \Lambda_k^{lq}, \mu_k^q)$ using the matrix geometric method.
10: **end while**

4. Computational Results and Analysis

In this section, we report the computational results and analysis with the proposed solution approaches and present some insights. The test problems are derived from U.S. Civil Aeronautics Board (CAB) data set (O'Kelly, 1987). The solution procedures were coded in C and the MIP problems were solved using CPLEX 11.2 on a Dell Intel Core PC with 2.40 GHz processor with 2 GB of RAM. The problems are solved to optimality (with an optimality gap of 10^-3).

4.1. Sensitivity Analysis and Observations

The sensitivity analysis is performed over an instance of CAB dataset with 15 Nodes and $p = 3$ hubs and the interdiscount factor β of 0.2, 0.4 and 0.8. The average total flow rate/demand λ_{ij} and the unit transportation cost c_{ij} between each pair of nodes (i, j) are obtained from the dataset. The collection and distribution cost coefficients are set to $\chi = \delta = 1$ per unit. For every potential hub, we generate three capacity levels - small (S), medium (M) and large (L), to choose from. The fixed costs are set to 150 (S), 200 (M) and 250 (L) and the capacity levels are set to $\frac{Cap}{p} + \beta \times A_l \times Cap$, where $Cap = \sum_i \sum_{j \neq k} \lambda_{ij}$, k is the hub in a one-hub network with nodes which sends the least total flow, p is the number of hubs to be opened, $\beta = 0.21, 0.22, 0.23, 0.24$ for 10, 15, 20, and 25 nodes respectively, A_l is a constant

that takes the value of -1, 0, and 1 for $l = 1(\text{S}), 2(\text{M})$, and $3(\text{L})$ respectively. In all the test problems, we consider two customer classes - the mean demand arrival rate of the high priority (λ_{ij}^h) and low priority (λ_{ij}^l) priority customer is obtained by multiplying the total flow rate (λ_{ij}) by 0.60 and 0.40 respectively. The service level requirements are varied as shown in the table. The threshold on the target time W_j is set to expected waiting time in an M/M/1 queue: $\min_j\{\lambda_j/\mu_j(\mu_j - \lambda_j) + 1/\mu_j\}$, where $\lambda_j = \lambda_j^h + \lambda_j^l$ and μ_j are the arrival and service rates of the hubs selected open at the first iteration of the solution procedure (i.e. based on initial solution - x^0, y^0). In the implementation of the matrix geometric procedure, number of labels of high and low priority customer classes are set to M = 100. This is based on our initial tests that shows that as the value of M increases, the error introduced due to the truncation decreases, hence we set M = 100. The step size for estimating the subgradient using the finite difference method are set to: $d\lambda_h = d\lambda_l = d\mu = 0.05$.

In order to analyze the impact of target times (τ^h, τ^l) and service levels (α_h, α_l) on the configuration of the hub-and-spoke network, we conduct six set of experiments described as follows: (i) Effect of changing service level of single customer class (Table 1); (ii) Effect of changing target time of single customer class (Table 2); (iii) Effect of changing target time of high priority class (Table 3); (iv) Effect of changing target time of low priority class (Table 4); (v) Effect of changing service level of high priority class (Table 5); and (vi) Effect of changing service level of low priority class (Table 6). These tables report the total cost (TOTC), fixed cost (FC), transportation cost (TC), hubs locations and capacity level, the total high priority (Λ_h) and low priority (Λ_l) routed, as well as the number of iteration (ITR) and the computation time (CPU) of the algorithm. Plots in Figure 1 show the effect of increasing service level and decreasing target time on the total cost. Plots in Figure 2 show the effect of service level and target time of high and low priority class customers on the total costs. Our observations are as follows:

- *The hub-and-spoke network configuration (hub locations, capacity selection, and allocation of nodes to hubs) that considers service-levels can be very different from the traditional configuration that ignores service-levels.* Tables 1-6 shows the configuration of hub-and-spoke network for different values of service levels and target times. From these table, we observe that the network configuration i.e. hub location, their capacity levels, and the total flow through the hubs are different for different values of service levels and target times. For example, the optimal network with a discount factor of 0.2 and 85% service level has hubs 4, 12, and 13, where the model recommends opening hubs 1, 4, and 12 at 98% service level. As the service level increases to 99% service level, the hubs opened are 4, 6, and 7. Therefore, as the service level increases, the congestion decreases through the reallocation of non-hub nodes among the hubs in order to balance the flow through hubs to reduce hub utilization and congestion in the system. Although large hub capacity might seem a uneconomical decision for a firm at the beginning due to higher fixed costs, it can provide the firm with the competitive advantage of routing the flows in a timely and responsive manner, thereby guaranteeing higher service levels. This provides decision makers with the insight that the hub

location, capacity selection, and allocation/routing of flows are interrelated decisions and should be made in conjunction rather than isolation.

- *Substantial improvement in service level can be achieved with a small increase in total cost (transportation cost + fixed cost).* As can be seen from the plots in Figures 1 and 2 and the results in Tables 1, 5 and 6, as the service level increases, the total cost increases, however the increase in the total cost is marginal initially. Hence, substantial improvement in service level can be achieved with a small increase in total cost. This is because, as we increase the magnitude of service level, hubs with higher capacity level are utilized, flow gets distributed more evenly across hubs, average hub utilization decreases, thereby reducing overall congestion. This results in increase in service level.

- *Significant improvement in target time can be achieved with a small increase in total cost (transportation cost + fixed cost).* Plots in Figures 1 and 2 and the results in Tables 2, 3, and 4 indicate that as the target time decreases, the total cost increases, however the increase in the total cost is marginal initially. This is because, as we decrease the target time, the flow gets distributed more evenly across hubs, hubs with higher capacity level are utilized, the average hub utilization decreases, thereby reducing overall congestion at the hubs and improving service level. Hence, the decision maker should incorporate effects of service level in the model while designing such systems, to hedge against the variability in the flow.

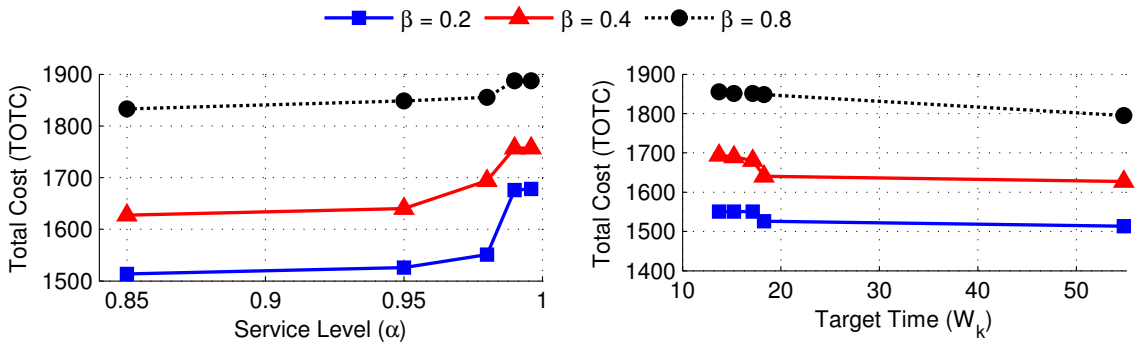

Figure 1: Effect of changing service level and target time on the total cost: The case of single customer class

5. Conclusion and Future Research

In this paper, we modelled and analyzed the effect of service-level constraints on the design of single allocation p-hub location problem. The model presented captures the tradeoff between the transportation cost savings induced by the economics of scale and the service level due to the variability of arrival rates of priority class customers and service rates of flow at hubs. Hubs are modelled as single server queues with Poisson arrivals and exponential service time distributions. The service level is measured using the number of customers served within a target time at hubs. For high priority class, we present a matrix geometric approach to evaluate the service-level constraints. Under concavity assumption, we linearize

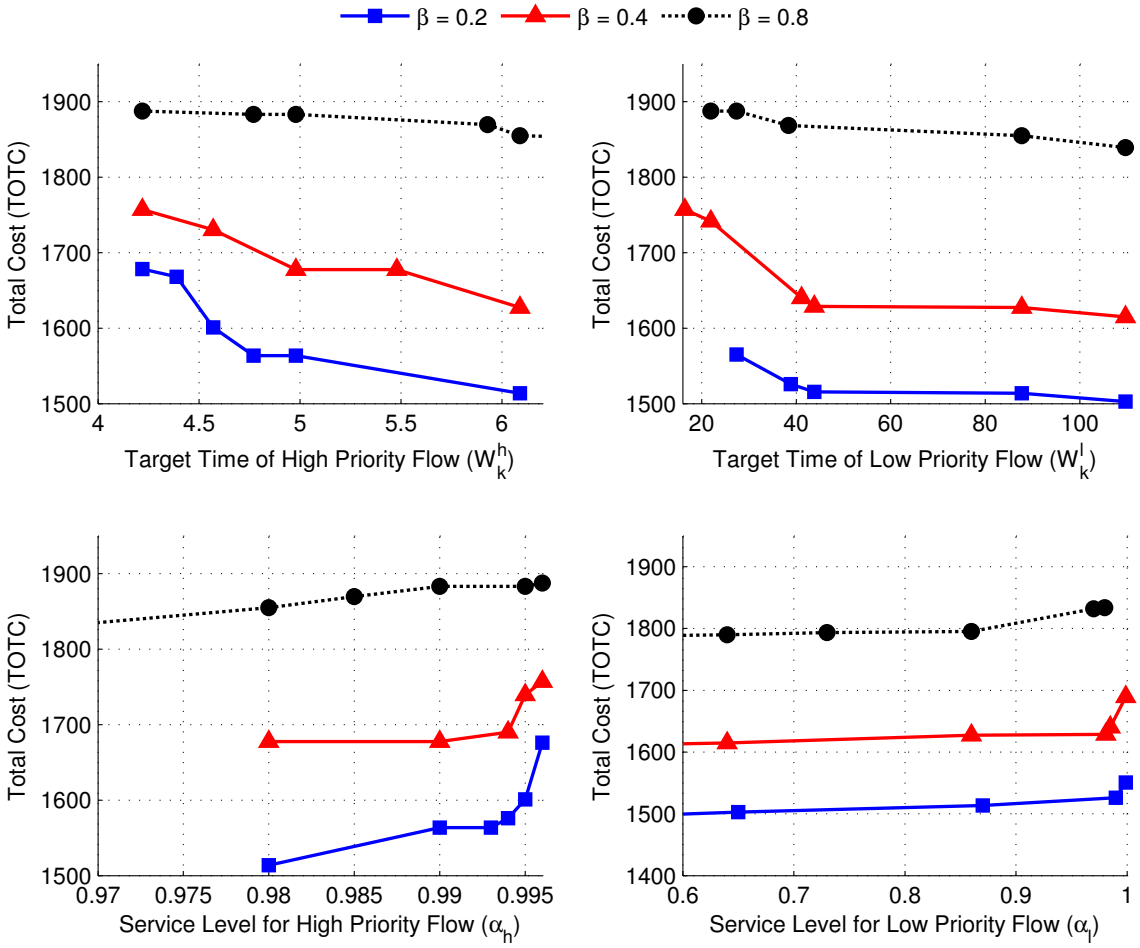

Figure 2: (a) Effect of changing target time of high priority flow on the total cost, (b) Effect of changing target time of low priority class on the total cost (c) Effect of service level of high priority flow on the total cost (d) Effect of service level of low priority flow on the total cost.

Table 1: Effect of Changing Service Level of Single Class on the Hub Location and Network Configuration: Example with 15 Nodes, 3 Hubs, $\tau_k = 13.71$

β	α	TOTC	FC	TC	Hubs	Capacity (μ)	Λ	CPU
0.2	0.85	1513.7	700	813.7	4,12,13	1275193, 768188, 1275193	1130983, 276108, 957851	268
	0.95	1526.0	700	826.0	4,12,13	1275193, 768188, 1275193	1050728, 276108, 1038106	239
	0.98	1550.8	700	850.8	1,4,12	1275193, 1275193, 768188	977583, 988123, 399236	233
	0.99	1676.1	750	926.1	4,6,7	1275193, 1275193, 1275193	746260, 878956, 739726	396
	0.996	1678.2	750	928.2	4,6,7	1275193, 1275193, 1275193	797140, 828076, 739726	281
0.4	0.85	1627.4	700	927.4	4,12,13	1275193, 768188, 1275193	1130983, 276108, 957851	298
	0.95	1640.2	700	940.2	4,12,13	1275193, 768188, 1275193	1050728, 276108, 1038106	244
	0.98	1694.0	700	994.0	1,4,12	1275193, 1275193, 768188	977583, 988123, 399236	385
	0.99	1757.0	750	1007.2	4,5,7	1275193, 1275193, 1275193	857469, 828124, 679349	392
	0.996	1757.2	750	1007.2	4,5,7	1275193, 1275193, 1275193	857469, 828124, 679349	421
0.8	0.85	1832.9	650	1182.9	4,11,13	1275193, 768188, 768188	1130983, 622351, 611608	376
	0.95	1848.5	700	1148.5	4,8,13	1275193, 768188, 1275193	1050728, 399236, 914978	315
	0.98	1855.4	700	1155.4	4,5,11	768188, 1275193, 1275193	441534, 952263, 971145	329
	0.99	1887.7	750	1137.7	4,5,11	1275193, 1275193, 1275193	754053, 828124, 782765	395
	0.996	1887.7	750	1137.7	4,5,11	1275193, 1275193, 1275193	754053, 828124, 782765	332

the model and use an cutting plane algorithm based procedure. In order to mitigate the effect of congestion and meet the service level constraints, the model prescribes redistribu-

Table 2: Effect of Changing Target Time of Single Customer Class on the Hub Location and Network Configuration: Example with 15 Nodes, 3 Hubs, $\alpha = 0.98$.

β	τ_k	TOTC	FC	TC	Hubs	Capacity (μ)	Λ	CPU
0.2	13.71	1550.8	700	850.8	1,4,12	1275193, 1275193, 768188	977583, 988123, 399236	215
	15.23	1550.8	700	850.8	1,4,12	1275193, 1275193, 768188	977583, 988123, 399236	262
	17.13	1550.8	700	850.8	1,4,12	1275193, 1275193, 768188	977583, 988123, 399236	294
	18.28	1526.0	700	826.0	4,12,13	1275193, 768188, 1275193	1050728, 276108, 1038106	276
	54.83	1513.7	700	813.7	4,12,13	1275193, 768188, 1275193	1130983, 276108, 957851	331
	164.48	1474.6	650	824.6	4,12,13	1275193, 768188, 768188	493760, 159694, 292523	319
0.4	13.71	1694.0	700	994.0	1,4,12	1275193, 1275193, 768188	977583, 988123, 399236	396
	15.23	1689.6	700	989.6	4,12,13	1275193, 768188, 1275193	1000329, 399236, 965377	353
	17.13	1680.4	700	980.4	4,12,13	1275193, 768188, 1275193	1043202, 276108, 1045632	381
	18.28	1640.2	700	940.2	4,12,13	1275193, 768188, 1275193	1050728, 276108, 1038106	236
	54.83	1627.4	700	927.4	4,12,13	1275193, 768188, 1275193	1130983, 276108, 957851	316
	0.00							
0.8	13.71	1855.4	700	1155.4	4,5,11	768188, 1275193, 1275193	441534, 952263, 971145	315
	15.23	1851.4	700	1151.4	1,4,11	768188, 1275193, 1275193	397765, 996032, 971145	319
	17.13	1851.4	700	1151.4	1,4,11	768188, 1275193, 1275193	397765, 996032, 971145	356
	18.28	1848.5	700	1148.5	4,8,13	1275193, 768188, 1275193	1050728, 399236, 914978	320
	54.83	1795.2	650	1145.2	4,5,7	768188, 1275193, 768188	672953, 1012640, 679349	370
	164.48	1778.8	650	1128.8	4,8,13	1275193, 768188, 768188	1234399, 399236, 731307	358

Table 3: Effect of Changing Target Time of High Priority Customer Class on the Hub Location and Network Configuration: Example with 15 Nodes, 3 Hubs, $\alpha_h = \alpha_l = 0.98$.

β	τ_k^h	τ_k^l	$\frac{\tau_k^l}{\tau_k^h}$	TOTC	FC	TC	Hubs opened	λ_h	λ_l	ITR	CPU(s)
0.2	4.22	54.83	13.0	1678.2	750	928.23	4 (L), 6(L), 7(L)	318856, 331230, 295890	478284, 496846, 443836	0	437.2
	4.39		12.5	1667.9	750	917.97	1(L), 4(L) , 7(L)	249665, 383206, 313106	374498, 574808, 469659	0	402.4
	4.57		12.0	1600.8	750	850.81	1(L), 4(L), 12(L)	391033, 395249, 159694	586550, 592874, 239542	0	405.6
	4.77		11.5	1563.7	750	813.71	4(L),12(L),13(L)	452393, 110443, 383140	678590, 165665, 574711	0	410.4
	4.98		11.0	1563.7	750	813.71	4(L),12(L),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	436.4
	6.09		9.0	1513.7	700	813.71	4(L),12(L),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	521.7
0.4	4.22	54.83	13.0	1757.2	750	1007.16	4(L),5(L),7(L)	342988, 331250, 271740	514481, 496874, 407609	0	416.1
	4.57		12.0	1730.4	750	980.38	4(L),12(L),13(L)	417281, 110443, 418253	625921, 165665, 627379	0	375.8
	4.98		11.0	1677.4	750	927.44	4(L),12(L),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	456.5
	5.48		10.0	1677.4	750	927.44	4(L),12(L),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	726.1
	6.09		9.0	1627.4	700	927.44	4(L),12(L),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	573.7
0.8	4.22	54.83	13.0	1887.7	750	1137.67	4(L),5(L),11(L)	301621, 331250, 313106	452432, 496874, 469659	0	310.7
	4.77		11.5	1883.3	750	1133.29	4(L),8(L),13(L)	452393, 159694, 333889	678590, 239542, 500834	0	360.2
	4.98		11.0	1883.3	750	1133.29	4(L),8(L),13(L)	452393, 159694, 333889	678590, 239542, 500834	1	561.4
	5.93		9.3	1869.7	700	1169.68	4(L),11(L),14(L)	430515, 438133, 77329.2	645772, 657199, 115994	1	658.4
	6.09		9.0	1854.8	700	1154.84	4(L),12(M),13(L)	449614, 110443, 385920	674421, 165665, 578879	1	580.1
	10.97		5.0	1831.7	650	1181.76	4(L),5(M),7(M)	444500, 254837, 246639	666751, 382256, 369959	2	993.9

tion of flow across the hubs to achieve more even hub utilization and/or the location of hubs with larger capacity to achieve higher relative difference of total flow through hub and its capacity. We used the models to demonstrate that substantial improvement in service level constraints can be achieved with a small increase in total costs associated with designing such networks. Also, we illustrated that the configuration of the hub-and-spoke network (hub location and capacity, and allocation of nodes to hubs) obtained using the model can be very different from those obtained using the traditional models that ignores service-levels. Our computational experiments also reveal that the cutting plane algorithm with matric geometric method proved to be an efficient solution procedure for finding the optimal solution to the problem in reasonable computation time.

There are many future research avenues for extending the queueing-based service-level

Table 4: Effect of Changing Target Time of Low Priority Customer Class on the Hub Location and Network Configuration: Example with 15 Nodes, 3 Hubs, $\alpha_h = \alpha_l = 0.98$.

β	τ_k^h	τ_k^l	$\frac{\tau_k^l}{\tau_k^h}$	TOTC	FC	TC	Hubs Opened	λ_h	λ_l	ITR	CPU(s)
0.2	6.09	109.65	18.0	1502.9	700	802.88	4(L),12(M),13(L)	481716, 110443, 353818	722574, 165665, 530726	0	313.8
		87.72	14.4	1513.7	700	813.71	4(L),12(M),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	653.4
		43.86	7.2	1515.7	700	815.68	4(L),12(M),13(L)	449614, 110443, 385920	674421, 165665, 578879	2	826.8
		38.38	6.3	1525.9	700	825.97	4(L),12(M),13(L)	420291, 110443, 415242	630437, 165665, 622864	2	811.6
		27.41	4.5	1564.9	700	864.94	4(L),12(M),13(L)	417281, 110443, 418253	625921, 165665, 627379	3	1212.8
0.4	6.09	109.65	18.0	1614.9	700	914.91	4(L),12(M),13(L)	481716, 110443, 353818	722574, 165665, 530726	0	293.3
		87.72	14.4	1627.4	700	927.44	4(L),12(M),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	596.6
		43.86	7.2	1628.7	700	928.73	4(L),12(M),13(L)	449614, 110443, 385920	674421, 165665, 578879	2	909.8
		41.12	6.8	1640.2	700	940.17	4(L),12(M),13(L)	420291, 110443, 415242	630437, 165665, 622864	2	899.7
		21.93	3.6	1741.6	750	991.61	4(L),12(L),13(L)	397352, 159694, 388930	596029, 239542, 583395	3	1310.4
		16.45	2.7	1757.0	750	1007.01	4(L),6(L),7(L)	322655, 351582, 271740	483982, 527374, 407609	1	650.0
0.8	6.09	109.65	18.0	1838.9	700	1138.98	4(L),12(M),13(L)	481716, 110443, 353818	722574, 165665, 530726	0	341.8
		87.72	14.4	1854.8	700	1154.84	4(L),12(M),13(L)	449614, 110443, 385920	674421, 165665, 578879	1	621.8
		38.38	6.3	1868.6	700	1168.59	4(L),12(M),13(L)	420291, 110443, 415242	630437, 165665, 622864	2	923.7
		27.41	4.5	1887.7	750	1137.67	4(L),5(L),11(L)	301621, 331250, 313106	452432, 496874, 469659	3	1267.1
		21.93	3.6	1887.7	750	1137.67	4(L),5(L),11(L)	301621, 331250, 313106	452432, 496874, 469659	2	918.0

Table 5: Effect of Changing Service Level of High Priority Customer Class on the Hub Location and Network Configuration: Example with 15 Nodes, 3 Hubs, $\tau_k^h = 6.09$, $\tau_k^l = 43.86$.

β	α_h	TOTC	FC	TC	Hubs Opened	λ_h	λ_l	ITR	CPU(s)
0.2	0.996	1676.1	750	926.105	4(L),6(L),7(L)	298504, 351582, 295890	447756, 527374, 443836	0	340.2
	0.995	1600.8	750	850.813	1(L),4(L),12(L)	391033, 395249, 159694	586550, 592874, 239542	0	331.7
	0.994	1575.9	750	825.965	4(L),12(L),13(L)	420291, 110443, 415242	630437, 165665, 622864	0	289.7
	0.993	1563.7	750	813.711	4(L),12(L),13(L)	452393, 110443, 383140	678590, 165665, 574711	0	247.9
	0.99	1563.7	750	813.711	4(L),12(L),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	519.6
	0.98	1513.7	700	813.711	4(L),12(L),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	796.3
0.4	0.996	1757.0	750	1007.01	4(L),6(L),7(L)	322655, 351582, 271740	483982, 527374, 407609	0	331.6
	0.995	1739.5	750	989.565	4(L),12(L),13(L)	400132, 159694, 386151	600197, 239542, 579226	0	445.6
	0.994	1690.2	750	940.172	4(L),12(L),13(L)	420291, 110443, 415242	630437, 165665, 622864	0	329.7
	0.99	1677.4	750	927.444	4(L),12(L),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	474.5
	0.98	971.2	700	927.444	4(L),12(M),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	593.7
0.8	0.996	1887.7	750	1137.67	4(L),5(L),11(L)	301621, 331250, 313106	452432, 496874, 469659	0	334.4
	0.993	1883.3	750	1133.29	4(L),8(L),13(L)	452393, 159694, 333889	678590, 239542, 500834	0	335.4
	0.99	1883.3	750	1133.29	4(L),8(L),13(L)	452393, 159694, 333889	678590, 239542, 500834	1	519.7
	0.985	1869.7	700	1169.68	4(L),11(L),14(M)	430515, 438133, 77329.2	645772, 657199, 115994	1	585.3
	0.98	1854.8	700	1154.84	4(L),12(M),13(L)	449614, 110443, 385920	674421, 165665, 578879	1	640.7
	0.95	1795.2	650	1145.19	4(M),5(L),7(M)	269181, 405056, 271740	403772, 607584, 407609	1	671.6

modelling framework. One of them is to extend the framework to other hub location formulation such as multiple allocation hub location models, p-center hub location models and hub covering models. Another promising avenue that can be explored is to extend the queueing-based service-level modelling framework to deal with congestion on links (and link capacity selection) in the hub-and-spoke network. One can also explore the use of embedding the proposed cutting plane based solution procedure within the Lagrangean relaxation/Benders decomposition framework to solve large-scale instances of problems.

Table 6: Effect of Changing Service Level of Low Priority Customer Class on the Hub Location and Network Configuration: Example with 15 Nodes, 3 Hubs, $\tau_k^h = 6.09$, $\tau_k^l = 43.86$

β	α_h	α_l	TOTC	FC	TC	Hubs Opened	λ_h	λ_l	ITR	CPU(s)
0.2	0.90	0.2	1474.6	650	824.625	4(L),12(M),13(M)	493760, 159694, 292523	740639, 239542, 438784	1	559.1
		0.65	1502.9	700	802.88	4(L),12(M),13(L)	481716, 110443, 353818	722574, 165665, 530726	2	856.9
		0.87	1513.7	700	813.711	4(L),12(M),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	558.2
		0.99	1525.9	700	825.965	4(L),12(M),13(L)	420291, 110443, 415242	630437, 165665, 622864	2	473.3
		0.999	1550.8	700	850.813	1(L),4(L),12(M)	391033, 395249, 159694	586550, 592874, 239542	3	985.1
0.4	0.90	0.2	1600.8	650	950.848	4(L),12(M),13(M)	493760, 159694, 292523	740639, 239542, 438784	1	672.7
		0.64	1614.9	700	914.913	4(L),12(M),13(L)	481716, 110443, 353818	722574, 165665, 530726	2	971.8
		0.86	1627.4	700	927.444	4(L),12(M),13(L)	452393, 110443, 383140	678590, 165665, 574711	1	634.0
		0.981	1628.7	700	928.733	4(L),12(M),13(L)	449614, 110443, 385920	674421, 165665, 578879	2	883.8
		0.985	1640.2	700	940.172	4(L),12(M),13(L)	420291, 110443, 415242	630437, 165665, 622864	2	959.8
		0.999	1689.6	700	989.565	4(L),12(M),13(L)	400132, 159694, 386151	600197, 239542, 579226	3	1227.0
0.8	0.90	0.2	1778.8	650	1128.84	4(L),8(M),13(M)	493760, 159694, 292523	740639, 239542, 438784	1	619.3
		0.64	1790.0	650	1140.04	1(M),4(L),7(M)	183257, 490980, 271740	274885, 736471, 407609	2	893.3
		0.73	1793.6	650	1143.61	4(L),11(M),13(M)	481716, 201061, 263200	722574, 301591, 394800	2	985.5
		0.86	1795.2	650	1145.19	4(M),5(L),7(M)	269181, 405056, 271740	403772, 607584, 407609	1	615.2
		0.97	1831.7	650	1181.76	4(L),5(M),7(M)	444500, 254837, 246639	666751, 382256, 369959	2	935.9
		0.98	1833.9	700	1133.88	4(L),8(M),13(L)	449614, 159694, 336668	674421, 239542, 505003	2	954.0

Acknowledgment

This research was supported by the National Science and Engineering Research Council of Canada (NSERC) grant to the first author, and by the Research & Publication Grant, Indian Institute of Management Ahmedabad to the second author. The authors would like to acknowledge Mr. Kartikeya Mohan Sahai (Indian Institute of Technology, Guwahati) for writing the code and conducting the experiments.

References

Abdinnour-Helm, S. 2001. Using simulated annealing to solve the p-hub median problem. *International Journal of Physical Distribution and Logistics Management* **31**(3) 203–220.

Alumur, S., B.Y. Kara. 2008. Network hub location problems: The state-of-the-art. *European Journal of Operational Research* **190** 1–21.

Andradottir, S. 1998. Simulation optimization. J. Banks, ed., *Handbook of Simulation*. John Wiley and Sons, New York, New York, 307–333.

Atlason, J., M. A. Epelman, S. G. Henderson. 2004. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research* **127** 333–358.

Atlason, J., M. A. Epelman, S. G. Henderson. 2008. Optimizing call center staffing with simulation and analytic center cutting-plane methods. *Management Science* **54**(2) 295–309.

Camargo, R.S.de, G. Miranda Jr., H.P. Luna. 2008. Benders decomposition for the uncapacitated multiple allocation hub-and-spoke network design. *Computers and Operations Research* **35** 1047–1064.

Camargo, R.S.de, G. Miranda Jr., H.P. Luna. 2009. Benders decomposition for the hub location problems with economies of scale. *Transportation Science* **43**(1) 86–97.

Campbell, J.F. 1994b. Integer programming formulations of discrete hub location problems. *European Journal of Operational Research* **72** 387–405.

Campbell, J.F. 1996. Hub location and the p-hub median problem. *Operations Research* **44**(6) 1–13.

Campbell, J.F., A.T. Ernst, M. Krishnamoorthy. 2002. Hub location problems. Z. Drezner, H. Hamacher, eds., *Location Analysis: Theory and Applications*. Springer, Berlin, 373–408.

Campbell, J.F., M.E. O'Kelly. 2012. Twenty-five years of hub location research. *Transportation Science* **46**(2) 153–169.

Contreras, I., J.-F. Cordeau, G. Laporte. 2011. Stochastic uncapacitated hub location. *European Journal of Operational Research* **212**(3) 518–528.

Contreras, I., J.-F. Cordeau, G. Laporte. 2012. Benders decomposition for large-scale uncapacitated hub location problem. *Operations Research* **59**(6) 1477–1490.

Contreras, I., J. Diaz, A. Marin. 2009. Lagrangean relaxation for the capacitated hub location problem with single assignment. *OR Spectrum* **31**(3) 485–505.

Ebery, J. 2001. Solving large single allocation p-hub location problems with two or three hubs. *European Journal of Operational Research* **128** 447–458.

Elhedhli, S., F. X. Hu. 2005. Hub-and-spoke network design with congestion. *Computers and Operations Research* **32** 1615–1632.

Elhedhli, S., H. Wu. 2010. A Lagrangean heuristic for hub-and-spoke system design with capacity selection and congestion. *INFORMS Journal of Computing* **22**(2) 282–296.

Ernst, A.T., M. Krishnamoorthy. 1996. Efficient algorithms for the uncapacitated single allocation p-hub median problem. *Location Science* **4**(3) 139–154.

Klincewicz, J.G. 1992. Avoiding local optima in the p-hub location problem using tabu search and GRASP. *Annals of Operations Research* **40** 283–302.

Koksalan, M., B. Soylu. 2010. Bicriteria p-hub location problems and evolutionary algorithms. *INFORMS Journal of Computing* **22**(4) 528–542.

Kratica, J., Z. Stanimirovic, D. Tosic, V.Filipovic. 2007. Two genetic algorithms for solving the uncapacitated single allocation p-hub median problem. *European Journal of Operational Research* **182**(1) 15–28.

Latouche, G., V. Ramaswami. 1999. *An Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial and Applied Mathematics, Philadephia, PA, USA.

Leeman, H. T. 2001. Waiting time distribution in a two-class two-server hetereogeneous priority queues. *Performance Evaluation* **43** 133–150.

Marianov, V., D. Serra. 2003. Location models for airline hubs behaving as M/D/c queues. *Computer and Operations Research* **30** 983–1003.

Neuts, F. M. 1981. *Matrix Geometric Solutions in Stochastic Methods: An Algorithmic Approach*. Dover Publications, Mineola, NY, USA.

O'Kelly, M.E. 1987. A quadratic integer program for the location of interacting hub facilities. *European Journal of Operational Research* **32** 393–404.

O'Kelly, M.E., D. Skorin-Kapov, S. Skorin-Kapov. 1995. Lower bounds for the hub location problem. *Management Science* **41**(4) 713–721.

Ramaswami, V., D. M. Lucantoni. 1985. Stationary waiting time distribution in queues with phase type service and quasi-birth-and-death processes. *Communication in Statistics-Stochastic Models* **1**(2) 125–136.

Sim, T., T.J. Lowe, B.W. Thomas. 2009. The stochastic p-hub center problem with service level constraints. *Computers and Operations Research* **36** 3166–3177.

Skorin-Kapov, D., J. Skorin-Kapov, M.E. O'Kelly. 1996. Tight linear programming relaxations of uncapacitated p-hub median problems. *European Journal of Operational Research* **94** 582–593.

Smith, K., M. Krishnamoorthy, M. Palaniswami. 1996. Neural versus traditional approaches to the location of interacting hub facilities. *Location Science* **4**(3) 155–171.

Song, J., S. Park. 2000. The single allocation problem in the interacting three-hub network. *Network* **35**(1) 17–25.

Topcuoglu, H., F. Corut, M. Ermis, G. Yilmaz. 2005. Solving the uncapacitated hub location problem using genetic algorithms. *Computers and Operations Research* **32**(4) 967–984.

Yaman, H. 2009. The hierarchical hub median problem with single assignment. *Transportation Research Part B* **43** 643–658.

Yang, T.-H. 2009. Stochastic air freight hub location and flight routes planning. *Applied Mathematical Modelling* **33** 4424–4430.