Modeling Congestion and Service Time in Hub Location Problems

Sibel A. Alumur[a], Stefan Nickel[b,c,*], Brita Rohrbeck[b],
Francisco Saldanha-da-Gama[d]

sibel.alumur@uwaterloo.ca, stefan.nickel@kit.edu, brita.rohrbeck@kit.edu, fsgama@fc.ul.pt

[a] Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada
[b] Institute for Operations Research, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[c] Forschungszentrum Informatik (FZI), Karlsruhe, Germany
[d] DEIO-CIO, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

Abstract

In this paper, we present a modeling framework for hub location problems with a service time limit considering congestion at hubs. Service time is modeled taking the traveling time on the hub network as well as the handling time and the delay caused by congestion at hubs into account. We develop mixed-integer linear programming formulations for the single and multiple allocation versions of this problem. We further extend the multiple allocation model with a possibility of direct shipments. We test our models on the well-known AP data set and analyze the effects of congestion and service time on costs and hub network design. We introduce a measure for the value of modeling congestion and show that not considering the effects of congestion may result in increased costs as well as in building infeasible hub networks.

1 Introduction

A hub location problem consists of selecting a subset of nodes of a network to become hubs and thus to consolidate and redistribute many-to-many traffic originated at and destined to the nodes of the network. By consolidating flow at the hubs it is possible to exploit economies of scale usually associated with the inter-hub traffic and thus to reduce transportation costs while routing demand between origin-destination pairs.

*corresponding author

Hub location is an area that has grown significantly in the past decades as we can observe in the book chapter by [1] as well as in the review papers by [2], [3], and [4]. The success of this area is much due to the fact that it has many applications in transportation, logistics, and telecommunications (the reader can refer to the above mentioned works and to the references therein).

It is possible to organize the area of hub location according to different aspects or features. One such aspect distinguishing the problems concerns the allocation pattern. In a single-allocation hub location problem each terminal node must be allocated (for sending and receiving traffic) to exactly one node; in turn, if multiple allocation is considered, no limit exists for the number of hubs to which a terminal is allocated. Recently, [5] unified both allocation patterns in a so-called r-allocation pattern, where r stands for the maximum number of nodes a terminal can be allocated to.

Another important feature in a hub location problem concerns the existence of capacity constraints. In this case, capacity may refer to the edges or to the hubs ([3] and [1]). When capacity is associated with hubs, a further distinction can be made. In some cases, capacity refers to non-processed incoming flow to the hubs; in others capacity refers to all the flow going through a hub.

In the beginning, motivated by some applications (see, e.g., the seminal papers by [6], [7], for air traffic) researchers focused on hub location problems in which the hub-level network is complete. However, other applications, such as logistics, called for incomplete hub networks. This is the case in the works by [8], [9], [10], [11], [12], [13], [14], [15], and [16], to mention a few.

With the need to consider progressively more comprehensive problems, several other variants and extensions have been considered. It is worth mentioning the research directions focusing uncertainty in the demands and/or in the costs ([17], [18], [19], [20], and [21]), time-dependent aspects ([22], [23]), and choice of transportation mode ([24]).

Finally, it is worth mentioning that in most of the works focusing on hub location problems it is assumed that direct shipments between terminals is not possible although some authors have considered that possibility such as [25], [26], [27], [14], [28], and [29].

Most of the above references show that research on hub location predominantly focuses on the minimization of costs, thus ignoring service level objectives such as delivery time. Nevertheless, the early work by [30] compares cost-minimal solutions with solutions balanced regarding their activity levels. Later, [31] introduced center and covering type hub location problems focusing exactly on service level objectives. Hub center problems seek the minimization of the maximum service time between origin-destination pairs (see, e.g., [32]). In covering problems, on the other

hand, the demand between an origin-destination pair is covered if it can be provided service within a given (service) time. Different notions of coverage are presented in the literature ([31]). Moreover, we can find the so-called set covering and maximal covering versions of a hub covering problem. The former minimizes the number of hubs to ensure a full demand coverage (see, e.g., [33]). The latter maximizes the amount of demand covered by a predetermined number of hubs to locate (e.g., [34]). One major drawback of center and covering type hub location problems is that transportation costs are not considered in these problems at all.

Cost and delivery time are two conflicting objectives that are present in many real-life service networks. This bi-objective nature of hub location problems is usually overlooked. In many applications, there is a promised service time, such as the overnight or 2-day delivery time guarantee in express shipment networks. In these applications, service time is a hard constraint rather than a component of an objective function to be minimized. If service time is overlooked while designing a hub network, the resulting hub network may not be feasible in terms of a desirable service level in the future. For instance, [35] points out that service level constraints are necessary in configuring networks for time definite motor carriers. We consider a similar setting in this paper as our starting point. In particular, we impose a limit on the service time for deliveries while minimizing total cost.

Service time has two components: the traveling time through the network and the time spent at hubs for processing the flow. We refer to the latter as the handling time at hubs. Typically, hubs have limited handling or sorting capacities. Furthermore, delays may occur in handling time, which is particularly true when a hub is operating close to its full capacity. In this case, we say that the hub is congested. The level of congestion may, in fact, vary depending on the amount of flow to be processed and the maximum capacity of the hub. For example, there may be no delay if a hub is working with less than 70% of its capacity, whereas a significant delay may occur if a hub is working over 90% of its capacity.

For the above reasons, in this work, we also model congestion at hubs since this is a means to account for the delay in handling time due to a high capacity utilization. Congestion is driven by two main factors: the amount of flow that needs to be processed by the hub and the capacity of the hub. We consider a problem suiting a pro-active decision maker who wishes to determine the capacities of hubs, thus influencing the congestion levels at the hubs and consequently the service time. In synthesis, we focus in this study on cost minimization hub location problems with capacity decisions and a service time limit reflecting the effects of congestion.

In his seminal paper on hub location, [6] was the first to emphasize possible negative repercussions of strongly utilized hubs. He suggested the minimization of the variability in terms

of hub usage. This idea was further considered by [36] using a simulation approach for ensuring a solution avoiding (as much as possible) an unbalanced usage of hubs. Later on, effects of congestion in hub location problems were addressed in the literature generally from a cost perspective. [20] viewed hubs as M/D/c-queuing systems. They proposed a capacity constraint imposing the probability of having more than a given number of airplanes in the queue to be smaller than or equal to a given value. [37] modeled congestion in the objective function by a convex non-linear cost function which increases exponentially in the flows assigned to hubs. [38] studied the multiple allocation version of this problem. [39] expressed the congestion at hubs as the ratio of total flow compared to the surplus capacity by treating the hub-and-spoke system as a network of M/M/1-queues. [40] compared the results of a single allocation model with two different cost functions for congestion: a power-law function and a Kleinrock average delay function. [41] focused on solving large-scale instances of their previously introduced models and presented a generalized Benders decomposition and an outer approximation method. [42] introduced the perspective of a network user, and compared their solutions to those from the perspective of the network owner. [43] followed up on this idea and solved the problem with a row generation procedure. None of the above studies considered the effects of congestion on delivery time.

[44] modeled hub operations as a G/G/1-queuing system and analyzed its effects on the design of inter-modal LTL logistics networks under service time constraints. They integrated steady state approximations of this queuing model within a p-hub median framework. The authors showed through computational analysis that available resources at a hub significantly affects both the location and number of hubs.

[45] consider a single allocation hub covering problem assuming a so-called M/M/x-queuing model for the operation at hubs. The third parameter indicates that the number of servers at a hub is to be determined. Additionally, the authors consider that the capacity at a hub induces a queue length.

[46] also investigate a hub covering location problem modeling congestion via a M/M/c-queuing model assuming an exogenous capacity for the queue in each hub. The authors consider the bi-objective nature of the problem and adopt a bi-objective modeling framework. In other words, they do not consider a maximum service time but assume that time is to be determined as part of the solution.

For the problem we investigate in this paper, we discretize congestion and include it in the discrete optimization models. In order to emphasize the strategic nature of a hub location problem, we are avoiding an explicit queuing model. The motivation is very similar to location

problems, where also the routing part is not solved explicitly. This allows us to tackle larger problem instances. We introduce novel mathematical formulations incorporating the effects of congestion in time definite delivery. We study the single and multiple allocation patterns. We report a set of computational experiments and discuss the relevance of capturing congestion and time in hub location problems.

The remainder of the paper is organized as follows: In Section 2, we give all the details concerning the problems we are investigating and we introduce some notation to be used in the following sections. Subsequently, we present a mixed-integer programming formulation for the single allocation version of the problem. We present two mixed-integer programming models for the multiple allocation version in Section 4. In Section 5, we test and evaluate our models on instances derived from the AP data. Section 6 provides an overview of the work done.

2 Problem definition and notation

In hub location problems investigated in this work we seek to (i) determine the location of hubs, (ii) their capacity, (iii) the allocation of non-hub nodes to hubs, and (iv) the routes of flow through the network. When making these decisions, we ensure that delivery between all pairs of nodes is within the given service time limit. We model both the single and the multiple allocation versions of the problem. For the latter, we introduce two variants, which are distinguished by whether or not direct shipment between non-hubs is possible.

The features of our problems can be itemized as follows:

- There is a network underlying the problems such that traffic should be shipped between pairs of nodes.
- There is full-cross traffic demand, i.e., each node sends flow to any other node in the network.
- Unless direct shipment between non-hub nodes is allowed, every shipment is routed via at least one hub.
- All hubs should be interconnected, i.e., the hub-level network should induce a complete graph.
- Hubs are capacitated. We assume that such capacity refers to the amount of non-processed inflow at the hubs. This is particularly relevant in many applications such as those emerging in postal delivery or more generally in logistics distribution where the capacity is

associated with sorting of unprocessed incoming traffic at the hubs (see, e.g., [47]; [48]). Furthermore, capacities are assumed to be modular, e.g., sorting lines (as in [49] and [22]).

- There is a service time limit for sending flow between all pairs of nodes within the network, i.e., each origin-destination (O-D) pair must receive service within this given service time limit.
- There is a single time limit for the whole network.
- Service time between two nodes is calculated by considering the travel time on the network as well as handling times at hubs.
- Handling time at a hub is dependent on the hub's capacity and its congestion level.
- The congestion level at a hub depends on the percent utilization of its capacity.
- Congestion levels are discretized. Each possibility induces some delay in terms of the handling time at a hub. In particular, the more congested a hub is, the higher the unitary handling time at the hub.
- Costs are accounted for (i) locating hubs, (ii) setting up their operating capacities, and (iii) routing the flow through the network.
- The transportation cost between adjacent nodes is scaled using different factors depending on the specific leg on the network they refer to ([50]). The transportation cost from a non-hub node to a hub (a collection leg) is scaled using a collection factor. Analogously, the transportation cost from a hub to a non-hub (a distribution leg) is scaled using a distribution factor. Finally, the transportation cost between two hubs (a inter-hub connection) is scaled using an appropriate discount factor.
- Transportation costs are assumed to satisfy the triangle inequality. This aspect together with the complete graph induced by the hubs ensures that all traffic is routed via at most two hubs.
- Travel times are symmetric and also satisfy the triangle inequality.

Computing service times between O-D pairs requires the knowledge about the so-called ready times. Hubs consolidate the flow and this can only be accomplished when all the incoming flow at the node is in fact at the hub. Accordingly, ready time is a value that we can associate with each hub representing the latest arrival time of all flow arriving at a hub from all the non-hub

nodes that are allocated to it ([51]). Ready time can alternatively be referred to as the collection time of a hub. The distribution time of a hub, on the other hand, is the longest travel time it takes to deliver flow from a hub to the non-hub nodes allocated to it.

Figure 1 illustrates this situation for a single allocation pattern. In this case, three nodes are allocated to hub k and we can see the corresponding travel times. In our setting, we assume that the hub needs to wait for the incoming flow from all three nodes to arrive to consolidate the flow. The latest flow arrives to hub k in four hours. If we suppose now that this traffic needs two hours to be sorted and consolidated with the other flow (the time depicted above the hub), plus three hours to travel to hub l, plus three hours at hub l for processing, then the flow is ready to be shipped from hub l only after 12 hours. Note from Figure 1 that it takes five hours to reach the furthest non-hub node that is allocated to hub l; i.e., the distribution time of hub l. We may then conclude that the maximum travel time in this network is 17 hours. We would like to remark that if the travel times are symmetric, the collection time is equal to the distribution time for each hub in the single allocation setting.

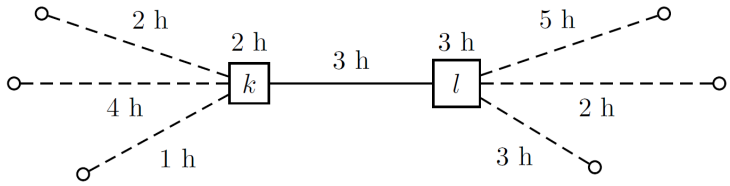

Figure 1: Illustration of ready times and service time for single allocation.

In a multiple allocation setting, the collection time at a hub is not necessarily equal to its distribution time. Moreover, collection and distribution times are dependent on hub-to-hub connections since a non-hub node may be allocated to more than one hub. These additional difficulties are well illustrated in Figure 2. We return to this figure in Section 4.

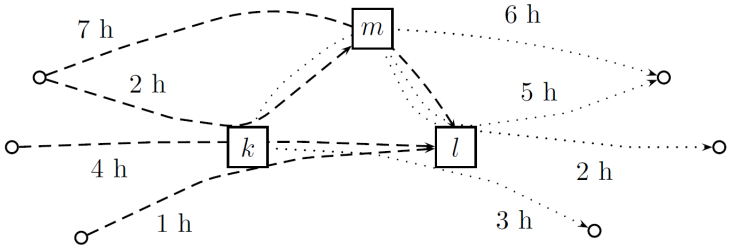

Figure 2: Illustration of ready times and service time for multiple allocation.

Independent from the allocation pattern we are studying, we must ensure that the maximum travel time between any two nodes is at most some given value. The objective is to find the network design and transportation plan for all traffic, ensuring the imposed service time and

minimizing the total costs involved.

Taking into account the common features of the different problems that we detail in the following sections, we directly introduce the common notation to be used hereafter:

Sets:

- N, set of nodes;
- C, set of modules available for hubs (each one having a certain capacity);
- G, set of congestion levels considered at hubs.

Parameters (flows and capacities):

- w_{ij}, flow to be sent from node $i \in N$ to node $j \in N$; from these values we can compute:
  $O_i = \sum_{j \in N} w_{ij}$, the flow originated in node $i \in N$, and
  $D_j = \sum_{i \in N} w_{ij}$, the flow destined to node $j \in N$;
- Γ^c, capacity of a hub with capacity level $c \in C$;
- γ^g, upper limit of the capacity utilization (percentage) for a hub congested at level $g \in G$

Parameters (costs):

- f_k^c, fixed cost for locating a hub with capacity level $c \in C$ at node $k \in N$;
- c_{ij}, transportation cost between nodes i and j $(i, j \in N)$;
- α, discount factor for inter-hub connections;
- χ, scaling factor for collection costs;
- δ, scaling factor for distribution costs.

Parameters (time):

- t_{ij}, travel time between nodes i and j $(i, j \in N)$;
- Δ^c, handling time at a hub with capacity level $c \in C$;
- τ_g, congestion factor, i.e., relative delay in time determined by congestion level $g \in G$;
- $\mathcal{T}$, service time limit.

3 Single allocation hub location with service time and congestion

In this section, we first present a mathematical formulation for the single allocation version of the problem. In Section 3.2, we discuss some features of an optimal solution to the problem.

3.1 An optimization model

An optimization model can be proposed for our single allocation hub location model with congestion (cong-SAHLP) that is primarily based on the formulation by [47]. This is a formulation well-known for its trade-off between size and tightness (see, for instance [49, 52]). In fact, it is a 3-index formulation that can efficiently be solved using general-purpose solvers.

In the model we are proposing next, we make use of the same allocation and flow variables as follows:

$$x_{ik} = \begin{cases} 1, & \text{if node } i \text{ is allocated to hub } k, \\ 0, & \text{otherwise.} \end{cases} \qquad (i, k \in N)$$

y_{kl}^i = Amount of flow originated in node i that is routed from hub k to hub l $(i, k, l \in N)$.

We additionally introduce two new sets of variables: one is associated with congestion, the other refers to ready times.

$$z_k^{cg} = \begin{cases} 1, & \text{if a hub with capacity level } c \text{ and congestion level } g \text{ is located at } k, \\ 0, & \text{otherwise.} \end{cases}$$

$$(c \in C,\, g \in G,\, k \in N)$$

r_k = ready time of hub k, $(k \in N)$.

Remark 1 *Since we are assuming that the travel times are symmetric, in the case of single allocation, recall that the collection time is equal to the distribution time for each hub. Hence, it suffices to consider the collection times (ready times) at hubs to ensure delivery within the service time limit between all pairs of nodes. This fact saves us from defining additional decision variables in the formulation of the single allocation problem.*

Remark 2 *For modeling purposes, instead of using the exact values of the ready times we will use upper bounds on them that are denoted by $\tilde{r}_k$, $k \in N$. This helps making the model readable and does not prevent it from obtaining an optimal solution to the problem. The exact ready times can be easily computed from an optimal solution to the model as we explain in the next section.*

Our problem can now be formulated as follows:

$$(P_{SA}) \text{ minimize} \quad \sum_{i\in N}\sum_{k\in N}(c_{ik}\chi O_i + c_{ki}\delta D_i)x_{ik} + \sum_{i\in N}\sum_{k\in N}\sum_{l\in N}\alpha c_{kl}y_{kl}^i + \sum_{k\in N}\sum_{c\in C}\sum_{g\in G} f_k^c z_k^{cg}, \quad (1)$$

$$\text{subject to} \quad \sum_{k\in N} x_{ik} = 1, \qquad i \in N, \tag{2}$$

$$\sum_{c\in C}\sum_{g\in G} z_k^{cg} = x_{kk}, \qquad k \in N, \tag{3}$$

$$\sum_{l\in N} y_{kl}^i \le O_i x_{ik}, \qquad i,k \in N, \tag{4}$$

$$\sum_{l\in N} y_{kl}^i - \sum_{l\in N} y_{lk}^i = O_i x_{ik} - \sum_{j\in N} w_{ij} x_{jk}, \qquad i,k \in N, \tag{5}$$

$$\sum_{i\in N} O_i x_{ik} \le \sum_{c\in C}\sum_{g\in G} \Gamma^c \gamma^g z_k^{cg}, \qquad k \in N, \tag{6}$$

$$\tilde{r}_k \ge t_{ik} x_{ik}, \qquad i,k \in N, \tag{7}$$

$$\tilde{r}_k + \sum_{c\in C}\sum_{g\in G} \Delta^c \tau_g z_k^{cg} + t_{kl} + \sum_{c\in C}\sum_{g\in G} \Delta^c z_l^{cg} + \tilde{r}_l \le \mathcal{T}, \qquad k,l \in N, \tag{8}$$

$$x_{ik} \in \{0,1\}, \qquad i,k \in N, \tag{9}$$

$$y_{kl}^i \ge 0, \qquad i,k,l \in N, \tag{10}$$

$$z_k^{cg} \in \{0,1\} \qquad c \in C,\ g \in G,\ k \in N. \tag{11}$$

The objective function (1) represents the overall cost to be minimized. It consists of transportation costs and costs for installing hubs of different capacities. Constraints (2) assure that every node is assigned to exactly one hub. Constraints (3) state that when a hub is established it should be set up with exactly one capacity level, which, in turn, induces some congestion level. Constraints (4) were introduced by [52] to ensure correct routing of flow through the hub network. Equations (5) state the flow balance constraints. Constraints (6) impose the capacity restrictions. These constraints concurrently determine the congestion level at each hub. For each hub, the higher the percentage utilization of its capacity, i.e., the higher the inflow it processes, the higher the congestion level it has. Constraints (7) determine upper bounds for the ready times. This bound for each hub is such that it is greater than the longest collection time from all the nodes that are allocated to it. To assure that the service time limit $\mathcal{T}$ is satisfied, we introduce Constraints (8). In these constraints we can observe the two components of the total service time: transportation time (independent from the congestion level at the hubs) and handling time at the hubs (depending on congestion). The travel time between an O-D pair is computed by adding up the upper bounds on the ready times at each hub, the travel time between the hubs, and the handling time at hubs (reflecting congestion delay). Note that delay due to congestion is adhered only at the first hub on the route since capacity restrictions are

only imposed on the inflow. Lastly, (9) to (11) are the domain constraints.

3.2 Ready times and congestion levels in an optimal solution

As noted in the previous section, in the model presented above we use upper bounds for the ready time, $\tilde{r}_k$, instead of the actual ready times, r_k. This makes it easier to formulate the problem since it becomes possible that the values of the ready times are inflated in an optimal solution of (P_{SA}) as long as the service time limit is satisfied for all pairs of nodes.

We illustrate this aspect complementing the analysis of Figure 1 as depicted in Figure 3. As we already observed, the ready time or collection time at hub k (r_k) is 4 h. Similarly, the

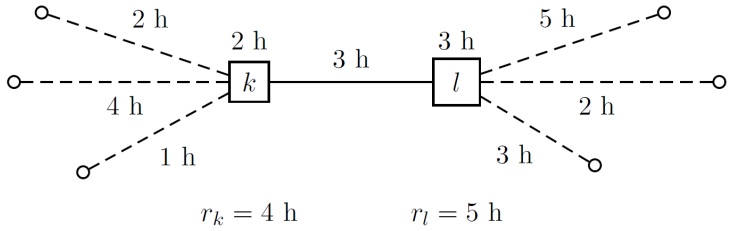

Figure 3: Illustration of ready times for single allocation.

ready time or distribution time r_l of hub l is 5 h. This results in a longest service time on this sub-network of 17 h. Suppose that the service time limit is 24 hours. In this case, any combination of the values for $\tilde{r}_k$ and $\tilde{r}_l$ is feasible for the model as long as $\tilde{r}_k \geq 4$, $\tilde{r}_l \geq 5$, and $\tilde{r}_k + \tilde{r}_l \leq 16$. For example, $\tilde{r}_k = 7$ and $\tilde{r}_l = 9$ would result in a feasible solution. On a similar note, $\tilde{r}_k$ variables may attain positive values for some non-hub nodes as well. Note, however, that this does not affect the optimality of a solution to the proposed model.

Similar to the inflation of ready time values, congestion levels at hubs may also be inflated in an optimal solution as long as the service time limit is satisfied without influencing the cost of a solution. Note that the congestion level is bounded from below by the capacity Constraints (6) and from above by the service time Constraints (8). When the service time constraint is non-binding and the capacity of a hub is not tight, a higher congestion level than the actual one might be selected by the model. In other words, similar to the ready times, setting the congestion levels at a higher level than it is necessary to assure the feasibility of the optimal min-cost solution renders a different solution that is also optimal from a cost point of view. In fact, the fixed costs for locating hubs depend only on their capacities and not on their congestion levels.

In conclusion, both the ready times and the congestion levels may be overestimated in an optimal solution to model (P_{SA}) without producing any effect on the optimal location and allocation decisions (unless there are alternative optimum solutions). Nevertheless, it is still of

interest to know the actual values of the ready times and congestion levels. This can be done by determining an optimal solution that minimizes the ready times and congestion levels.

This discussion sets our problem within a bi-objective setting. In fact, we can look at the minimization of ready times, the minimization of congestion levels, and the minimization of total cost as different objectives in a multicriteria problem. For a review on multicriteria facility location problems, we refer to [53].

In order to find the actual ready times and congestion levels we can simply employ a lexicographic approach ([54]) to our problem: Initially, we solve the model (P_{SA}) to optimality and obtain the optimal objective function value, i.e., the minimum total cost denoted by $\mathcal{V}(P_{SA})$. Then, we solve the following problem:

$$(P^{SA}_{con,r}) \text{ minimize} \quad \sum_{k\in N}\sum_{c\in C}\sum_{g\in G}\gamma^g z_k^{cg} + \sum_{k\in N} r_k, \tag{12}$$

$$\text{subject to} \quad \sum_{i\in N}\sum_{k\in N} c_{ik}(\chi O_i + \delta D_i)x_{ik} + \sum_{i\in N}\sum_{k\in N}\sum_{l\in N}\alpha c_{kl} y_{kl}^i + \sum_{k\in N}\sum_{c\in C}\sum_{g\in G} f_k^c z_k^{cg} \leq \mathcal{V}(P_{SA}), \tag{13}$$

$$(2) - (11).$$

The objective function (12) sums the congestion levels and the ready times and is to be minimized. It is relevant to note that the sum of the ready times and sum of the congestion levels are not conflicting objectives. Hence, we can sum these up and use a single objective function rather than solving each problem individually.

Constraint (13) ensures that total cost is less than or equal to the minimum feasible cost. Through this constraint, we obtain a solution which is as good as the solution of the original problem (P_{SA}) with respect to costs. However, it is possible that the optimal solution of $(P^{SA}_{con,r})$ results in different location or allocation decisions than (P_{SA}). In that case, $(P^{SA}_{con,r})$ identifies an alternative optimum solution of (P_{SA}) with better values of ready times and congestion levels. Such a solution may be preferred and considered as more robust against potential uncertainty in demands and travel times.

Lastly, we would like to point out that one may also adopt a goal programming approach and use a higher cost value than $\mathcal{V}(P_{SA})$ on the right-hand-side of Constraint (13), relaxing the target value for total cost. In this case, it may be possible to achieve even smaller ready times and congestion levels.

4 Multiple allocation hub location with service time and congestion

We now consider the multiple allocation case. Like the single allocation setting, we start by assuming that direct shipments between non hub nodes are not possible. Later, we relax this assumption. The reason for considering this possibility only in multiple allocation is that we are assuming that a non-hub node can be allocated to more than one (hub) node. Accordingly, the possibility of making direct shipments emerges as a more natural extension in a multiple allocation setting.

4.1 An optimization model

Again, our starting point is a well-known model for the capacitated multiple allocation hub location problem proposed by [48]. This is a 4-index formulation making use of decision variables that provide information about the entire path of a delivery. This makes it easier to compute the ready times, as we will see. In particular, we also consider the following variables:

y_{ij}^{kl} = fraction of flow from node i to node j that is shipped from hub k to hub l $(i, j, k, l \in N)$.

In addition to the above variables, we consider a set of binary variables for tracking the path of the flow from one node to another:

$$\hat{y}_{ij}^{kl} = \begin{cases} 1, & \text{if any traffic from } i \text{ to } j \text{ is routed from hub } k \text{ to } l, \\ 0, & \text{otherwise.} \end{cases}$$

$(i, j, k, l \in N)$

In order to account for the capacity and congestion levels of the hubs we use the same variables z_k^{cg} $(k \in N,\, c \in C,\, g \in G)$ as for the single allocation problem.

As we noted in Section 2, in a multiple allocation setting not only does the collection time of a hub not coincide with its distribution time but also these times can be different for each hub-to-hub connection. Accordingly, in order to determine the ready times at the hubs we must go deeper in terms of the variable definitions and include one extra index in comparison to what we did for single allocation:

r_{kl}^{in} = collection time at hub k with respect to the traffic that is going to be sent to hub l $(k, l \in N)$;

In other words, by using the above variables we are considering the latest arrival time of the flow from every non-hub node allocated to hub k that is destined to be transferred to hub l.

Likewise, we define:

r_{kl}^{out} = distribution time at hub l with respect to the flow coming from hub k, i.e., the longest time it takes to deliver the flow to the non-hub nodes allocated to hub l that is arriving from hub k ($k, l \in N$).

The above decision variables are illustrated in Figure 4.

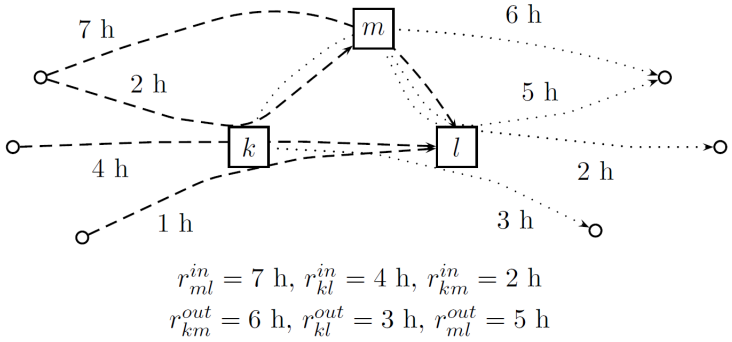

Figure 4: Illustration of ready times for multiple allocation.

Similar to the single allocation case, for modeling purposes we consider an upper bound on the ready times and leave the determination of the corresponding exact times for after an optimal solution has been obtained. The same can be done for the delivery times. These upper bounds will be denoted by $\tilde{r}_{kl}^{in}$, and $\tilde{r}_{kl}^{out}$, respectively, for the collection time at hub k for traffic going to hub l and for the delivery time at hub l w.r.t. the traffic coming from hub k ($k, l \in N$).

We can finally introduce an optimization model for the multiple allocation version of our problem (cong-MAHLP):

$$(P_{MA}) \quad \text{minimize} \quad \sum_{i\in N}\sum_{j\in N}\sum_{k\in N}\sum_{l\in N}(\chi c_{ik} + \alpha c_{kl} + \delta c_{lj})w_{ij}y_{ij}^{kl} + \sum_{k\in N}\sum_{c\in C}\sum_{g\in G} f_k^c z_k^{cg}, \tag{14}$$

$$\text{subject to} \quad \sum_{k\in N}\sum_{l\in N} y_{ij}^{kl} = 1, \qquad i, j \in N, \tag{15}$$

$$\sum_{c\in C}\sum_{g\in G} z_k^{cg} \leq 1, \qquad k \in N, \tag{16}$$

$$\sum_{l\in N} y_{ij}^{kl} \leq \sum_{c\in C}\sum_{g\in G} z_k^{cg} \qquad i, j, k \in N, \tag{17}$$

$$\sum_{k\in N} y_{ij}^{kl} \leq \sum_{c\in C}\sum_{g\in G} z_l^{cg} \qquad i, j, l \in N, \tag{18}$$

$$\sum_{i\in N}\sum_{j\in N} w_{ij}\sum_{l\in N} y_{ij}^{kl} \leq \sum_{c\in C}\sum_{g\in G} \gamma^g \Gamma^c z_k^{cg}, \qquad k \in N, \tag{19}$$

$$\hat{y}_{ij}^{kl} \geq y_{ij}^{kl}, \qquad i, j, k, l \in N, \tag{20}$$

$$\tilde{r}_{kl}^{in} \geq t_{ik}\hat{y}_{ij}^{kl}, \qquad i,j,k,l \in N, \tag{21}$$

$$\tilde{r}_{kl}^{out} \geq t_{lj}\hat{y}_{ij}^{kl}, \qquad i,j,k,l \in N, \tag{22}$$

$$\tilde{r}_{kl}^{in} + \sum_{c\in C}\sum_{g\in G}\Delta^c t^g z_k^{cg} + t_{kl} + \sum_{c\in C}\sum_{g\in G}\Delta^c z_l^{cg} + \tilde{r}_{kl}^{out} \leq \mathcal{T} \quad k,l \in N, \tag{23}$$

$$y_{ij}^{kl} \geq 0, \qquad i,j,k,l \in N, \tag{24}$$

$$\hat{y}_{ij}^{kl} \in \{0,1\}, \qquad i,j,k,l \in N, \tag{25}$$

$$z_k^{cg} \in \{0,1\}, \qquad c \in C, g \in G, k \in N. \tag{26}$$

The objective function (14) represents the sum of the transportation and opening costs, which is to be minimized. With Constraints (15) we guarantee that the whole fraction of demand between each pair of nodes is shipped via hubs. Constraints (16) state that every hub can have at most one capacity and one congestion level. With Constraints (17) and (18) no flow can be routed via a node that is not a hub. These constraints together with (16) ensure that every hub has exactly one capacity and one congestion level. Similar to the single allocation version, Constraints (19) bound the incoming flow of a hub according to its capacity and also determine the congestion level of the hub. Constraints (20) relate the fractional flow variables with binary path variables. The binary variables are needed to determine the collection and distribution times at hubs. Accordingly, Constraints (21) and (22) calculate the collection and distribution times, respectively, of an inter-hub link. Constraints (23) limit the maximum service time on the network. Lastly, Constraints (24)–(26) are the domain constraints.

4.2 The multiple allocation problem with direct shipments

In this section, we relax the basic assumption in hub location stipulating that all deliveries must be shipped via at least one hub. Having direct connections between non-hub nodes, for example for having full-truckload shipments or direct airline connections, may be practical in some situations. The reader may refer to [55], [27], and [56] for some implementations.

In order to derive a model for this extension of the problem we can consider the same sets of decision variables already introduced for the multiple allocation problem. Additionally, we introduce the following ones:

$$s_{ij} = \begin{cases} 1, & \text{if all demand from node } i \text{ to node } j \text{ is shipped directly,} \\ 0, & \text{otherwise.} \end{cases} \qquad (i,j \in N)$$

In case a direct shipment is chosen, we consider a scaling factor $\mu \geq 1$ for the corresponding

transportation costs.

The multiple allocation problem with congestion and direct shipment (cong-DMAHLP) can then be stated as follows:

$$(P_{MA-D}) \quad \text{minimize} \quad \sum_{i\in N}\sum_{j\in N}\sum_{k\in N}\sum_{l\in N}(\chi c_{ik}+\alpha c_{kl}+\delta c_{lj})w_{ij}y_{ij}^{kl}+\sum_{i\in N}\sum_{j\in N}\mu c_{ij}w_{ij}s_{ij} + \sum_{k\in N}\sum_{c\in C}\sum_{g\in G} f_k^c z_k^{cg}, \tag{27}$$

$$\text{subject to} \quad (16)-(26),$$

$$\sum_{k\in N}\sum_{l\in N} y_{ij}^{kl}+s_{ij}=1, \qquad i,j\in N, \tag{28}$$

$$t_{ij}s_{ij}\le \mathcal{T}, \qquad i,j\in N, \tag{29}$$

$$s_{ij}\le 1-\sum_{c\in C}\sum_{g\in G} z_i^{cg}, \qquad i,j\in N, \tag{30}$$

$$s_{ij}\le 1-\sum_{c\in C}\sum_{g\in G} z_j^{cg}, \qquad i,j\in N, \tag{31}$$

$$s_{ij}\in\{0,1\}, \qquad i,j\in N. \tag{32}$$

The costs for the direct shipments are added to the transportation and fixed costs in the objective function (27). Constraints (28) result from modifying Constraints (15) to ensure that all the traffic to be sent from one node to another is either routed via some hubs or shipped directly. To guarantee that the service time limit is also adhered for direct deliveries, Constraints (29) are included in the model. Constraints (30) and (31) guarantee that direct shipments are allowed only between non-hub nodes. Finally, Constraints (32) provide the domain for the binary direct-shipment variables.

4.3 Ready times and congestion levels in an optimal solution

As we observed before for the single allocation problem, collection and distribution times as well as congestion levels can have slack in the optimal solutions to both versions of the multiple allocation problem presented above. This may result in some inflated values for these decision variables in a given solution. Similar to the approach we adopted for the single allocation version, we propose to use a lexicographical method in order to obtain the actual values of the collection and distribution times as well as of the congestion levels.

For the multiple allocation problem cong-MAHLP, the following model can be used with this

purpose:

$$(P^{MA}_{con,r}) \quad \text{minimize} \quad \sum_{k\in N}\sum_{c\in C}\sum_{g\in G}\gamma^g z_k^{cg} + \sum_{k\in N}\sum_{l\in N} r_{kl}^{in} + \sum_{k\in N}\sum_{l\in N} r_{kl}^{out}, \tag{33}$$

$$\text{subject to} \quad \sum_{i\in N}\sum_{j\in N}\sum_{k\in N}\sum_{l\in N}(\chi c_{ik} + \alpha c_{kl} + \delta c_{lj}) w_{ij} y_{ij}^{kl} + \sum_{k\in N}\sum_{c\in C}\sum_{g\in G} f_k^c z_k^{cg} \leq \mathcal{V}(P_{MA}), \tag{34}$$

$$(15) - (26),$$

where $\mathcal{V}(P_{MA})$ represents the optimal value resulting from solving model (P_{MA}).

When direct shipments are considered, the following model can be used for determining the actual values of the ready times and congestion levels in an optimal solution:

$$(P^{MA-D}_{con,r}) \quad \text{minimize} \quad \sum_{k\in N}\sum_{c\in C}\sum_{g\in G}\gamma^g z_k^{cg} + \sum_{k\in N}\sum_{l\in N} r_{kl}^{in} + \sum_{k\in N}\sum_{l\in N} r_{kl}^{out} \tag{35}$$

$$\text{subject to} \quad \sum_{i\in N}\sum_{j\in N}\sum_{k\in N}\sum_{l\in N}(\chi c_{ik} + \alpha c_{kl} + \delta c_{lj}) w_{ij} y_{ij}^{kl} + \sum_{i\in N}\sum_{j\in N}\mu c_{ij} w_{ij} s_{ij} + \sum_{k\in N}\sum_{c\in C}\sum_{g\in G} f_k^c z_k^{cg} \leq \mathcal{V}(P_{MA-D}), \tag{36}$$

$$(16) - (26), (28) - (32),$$

where $\mathcal{V}(P_{MA-D})$ denotes the optimal objective function value of (P_{MA-D}).

5 Computational tests

We tested our formulations on the Australia Post (AP) data set [57]. The AP data set was first introduced by [50] and comprises instances with up to two-hundred nodes. It contains the coordinates of each node as well as the demand between each pair of nodes. There are two different sets for fixed costs and capacities in the AP data set, referred to as loose (L) and tight (T). The abbreviations LL, LT, TL, and TT are used to refer to these fixed cost and capacity settings specifically in this order. Transportation costs are assumed to depend linearly on the Euclidean distance between the nodes. Values of the scaling factors for collection, transfer, and distribution costs (χ, α, and δ) are provided [57]. In addition to $\alpha = 0.75$, we also tested the values 0.25 and 0.5 to observe the effects of the economies of scale factor on the solutions.

In our computational tests, we allow three differently sized hubs for each capacity set (loose and tight): small (S), medium (M), and large (L). We use the average values from the AP data to determine a medium capacity level for the hubs. We then deviate these average values by 20% to obtain the capacities and costs of small and large hubs.

Handling time in a medium sized hub is set to 2.5 hours. This time is also varied by 20% for hubs of small and large capacities. Travel times between the nodes are linearly dependent on the distance. We used a scaling factor t for travel times which is dependent on the number of nodes in the network to ensure that a service time limit of around 21 hours is feasible. This time limit is motivated mainly from express shipment networks where it is appropriate to assume deliveries to be executed within 24 hours. In other contexts, other values for t may be reasonable.

We tested different values for the service time limit. The values ranged from $\mathcal{T}_{min}$ to 100 hours, where $\mathcal{T}_{min}$ is the smallest feasible service time limit rounded up to the next half-an-hour. For each instance, we tested the values $\mathcal{T}_{min}$, 24, 30, and 100 h. A time value of 100 h is a big enough value and it practically means no time restrictions. We particularly tested this instance as a reference point for minimum costs. In general, we tested integer values of time, however, for some specific instances, we observed the solutions with half-an-hour increments in service time.

We considered three different congestion levels using the values 70%, 85%, and 100%. These percentages (γ^g) represent upper bounds on the capacity utilization of a hub for each congestion level g. The congestion factor τ indicates by how much the handling time in a hub increases as a result of a higher congestion level. For a capacity utilization percentage between 70% to 85%, processing times in hubs are multiplied by $\tau^g = 1 + \tau$. For utilization rates higher than 85%, τ^g is chosen to be $(1 + \tau)^2$. In general, we tested three different values for τ: 0%, 30%, and 100%. In some instances, we additionally tested $\tau \in \{10\%, 50\%, 200\%\}$.

All the settings we used in our computational tests with the AP data set are summarized in Table 1.

We ran our tests on a 64-bit Windows 7 Enterprise PC with a 2.6 GHz Intel(R) Xeon(R) processor and 48 GB RAM. To solve our models, we used IBM ILOG CPLEX optimization studio 12.6.1. We limited CPLEX run time to 6 hours. We additionally set an optimality gap of 1% for some difficult instances. CPU time requirements by CPLEX are detailed in Section 5.6. All of the presented CPU times refer to the basic model without the explicit minimization of ready times and congestion levels unless stated otherwise.

Description	Parameter	Value
Sets:		
Number of nodes	$\|N\|$	10, 20, 40
Capacity levels	C	S, M, L
Number of congestion levels	$\|G\|$	3
Flows and Costs:		
Flows	w_{ij}	flows of AP data
Fixed costs: medium hub T	f_k^M	average of tight costs of AP data
Fixed costs: medium hub L	f_k^M	average of loose costs of AP data
Fixed costs: small hub T, L	f_k^S	$0.8 f_k^M$
Fixed costs: large hub T, L	f_k^L	$1.2 f_k^M$
Transportation cost per unit of flow	c_{ij}	costs of AP data
Discount factor for inter-hub connections	α	0.25, 0.5, 0.75
Scaling factor for collection costs	χ	3
Scaling factor for distribution costs	δ	2
Scaling factor for direct shipments	μ	4
Capacities:		
Medium hub capacity T	Γ_k^M	average of tight values of AP data
Medium hub capacity L	Γ_k^M	average of loose values of AP data
Small hub capacity T, L	Γ_k^S	$0.8\Gamma_k^M$
Large hub capacity T, L	Γ_k^L	$1.2\Gamma_k^M$
Time:		
Handling time: medium hub	Δ^M	2.5 h
Handling time: small hub	Δ^S	2 h
Handling time: large hub	Δ^L	3 h
Transportation time	t_{ij}	$t \cdot d_{ij}$
Distances	d_{ij}	Euclidean distances of AP data
Service time limit	$\mathcal{T}$	various depending on feasibility
Congestion:		
UBs on capacity utilization for congestion level g	γ^g	$70\%, 85\%, 100\%$
Relative time delay for congestion level g	τ_g	$(1+\tau)^{g-1}$, $\tau \in \{0\%, 10\%, 30\%, 50\%, 100\%, 200\%\}$

Table 1: Parameter settings.

5.1 Influence of service time and congestion on hub networks

To illustrate the effects of service time limit and congestion on hub network design we provide the following solution from the 10-node AP data set. Figure 5 depicts a single allocation instance with a congestion factor τ of 200%. Fixed costs and capacities for hubs are both tight (TT) at this instance. Hubs are represented with squares in this drawing, where the size of the square portrays the capacity of the hub. Moreover, blank squares represent no congestion (capacity utilization of less than 70%) whereas gray squares represent hubs which are lightly congested (capacity utilization between 70% to 85%).

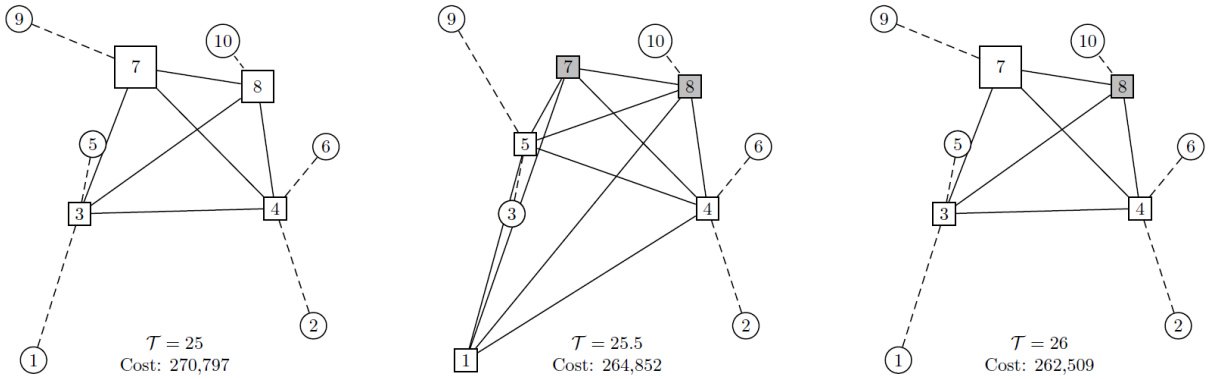

Figure 5: Solutions demonstrating the effects of service time and congestion on hub networks.

Each network in Figure 5 corresponds to an optimal solution with a different service time limit, incremented by half-an-hour. As expected, optimal total costs decrease with an increase in the service time limit $\mathcal{T}$. Note that locations, capacities, and congestion levels of hubs change with different service levels. When the service time is relaxed from 25 to 25.5 hours, it becomes optimal to open five small hubs, two of which are lightly congested, rather than opening four hubs with larger capacities as in $\mathcal{T} = 25$. When $\mathcal{T}$ is increased further to 26 hours, again only four hubs are opened, but one of them is now congested. The only difference between the resulting hub networks with $\mathcal{T} = 25$ and 26 is that the capacity of hub node 8 is smaller in the latter. When service time is relaxed, it becomes feasible to open a smaller hub at node 8 to reduce total costs allowing to spend more time in it due to congestion.

These example solutions clearly demonstrate the effect of considering service time limit and congestion on the design of hub networks. In the next section, we analyze the change in total costs.

5.2 Costs of service time and congestion

As already observed through the solutions presented in the previous section, adhering a certain service time limit comes at a price. Taking the effects of congestion into account may increase total costs even more. The decision maker certainly needs to know the trade-off between cost and maximal service time. In this section, we analyze this trade-off under different models and parameter settings.

Figures 6 to 8 present the changes in total costs under different service time limits with varying congestion factor (τ), cost and capacity sets (loose & tight) for 10-node instances.

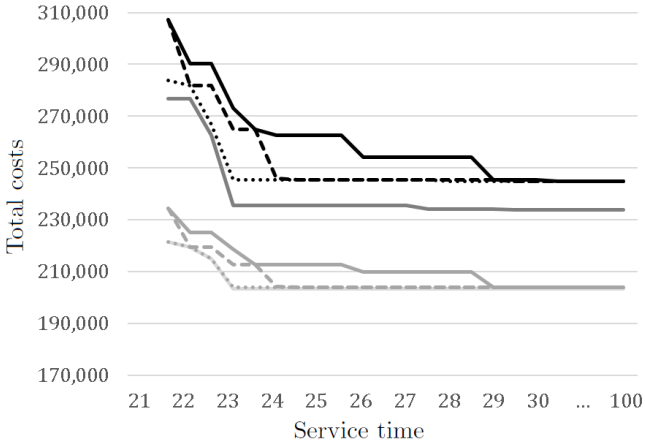

Figure 6: Total costs for single allocation.

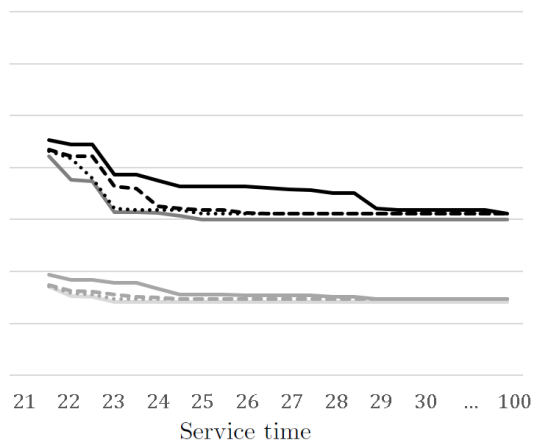

Figure 7: Total costs for multiple allocation.

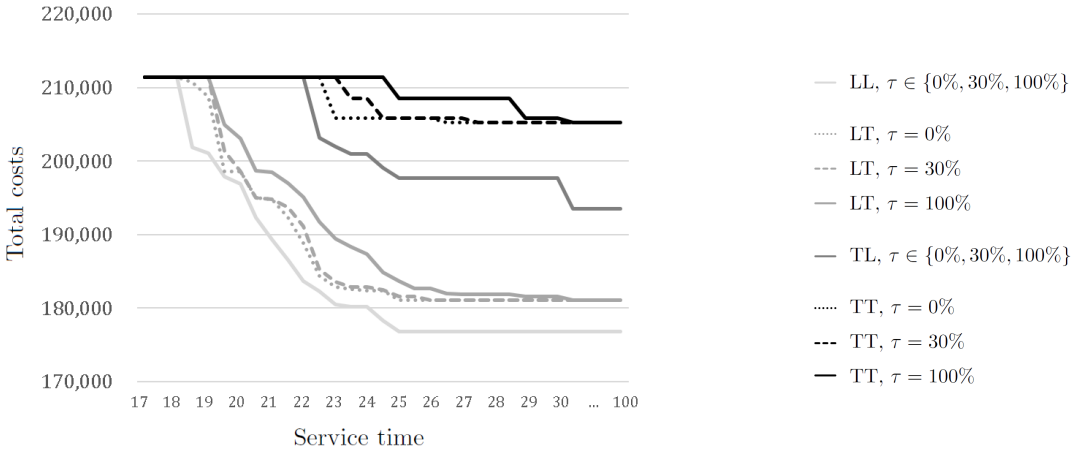

Figure 8: Total costs for multiple allocation with direct shipment.

In Figures 6–8, for each set of cost and capacity values, and different congestion factors, the x-axis lists the service time limit and the y-axis shows the corresponding optimal total cost value.

We tested service times starting from the minimum feasible value ($\mathcal{T}_{min}$) up to 30 hours. As noted before, we additionally tested a service time limit of 100 hours to demonstrate a solution when the service time constraint is non-binding. Lower service time limits become feasible when direct shipments are allowed. For instance, a time limit of 17 hours is not feasible with both single and multiple allocation without direct shipment, however, this service time is feasible with direct shipments.

Observe from Figures 6–8 that the single allocation problems with a service time limit of 21 hours result in the highest total cost. For the same set of parameter values, multiple allocation problems result in lower total cost values than single allocation problems, as expected. Note that single allocation solutions provide an upper bound for the multiple allocation problem as the solution to a single allocation problem is always feasible to the multiple allocation variant. The total cost of the direct shipment problems, on the other hand, is bounded above by the cost of shipping all the demand directly between each origin-destination pair. This cost value corresponds to about 211,000.

The congestion factor does not have any influence on the set of instances with the loose capacity set (LL and TL). In these instances, with each of the models, different congestion factors (0%, 30%, and 100%) resulted in the same total cost. Hence, these instances were represented with a single line in Figures 6–8. This results from the fact that the loose capacity set of the AP data provides extremely generous capacities for hubs. Even in the instances when only a few hubs were opened, congestion was rarely observed. For tight capacities, however, the congestion factor has a clear effect on total costs.

The service time limit has a more dominant effect on total cost than the congestion factor. If the decision maker wants to provide service within the smallest feasible service time, then s/he needs to bear higher costs in building the hub network. For each allocation rule, we calculated the highest percent difference in total cost between $\mathcal{T}_{min}$ and $\mathcal{T} = 100$. For single allocation with no congestion ($\tau = 0\%$), this highest percentage is 18.4%, meaning that the total cost of providing service in 21 hours is up to 18.4% higher than building the hub network with no service time limit. The corresponding values are 10.7% and 19.5% for multiple allocation and multiple allocation with direct shipment models, respectively. As expected, service time has a more severe effect on total cost with single allocation compared with multiple allocation.

The congestion factor enhances the effect of service time on costs. When $\tau = 100\%$, the percent difference in total cost between $\mathcal{T}_{min}$ and 100 hours goes up to 25.4% for single allocation. Similarly, with $\tau = 100\%$, the percent increase in total costs are 12.3% and 19.5% for multiple allocation and multiple allocation with direct shipment problems, respectively.

In practice, the network does not have to be built for the smallest feasible service time limit, but for a reasonable value of $\mathcal{T}$. With loose service time limits the increase in total cost compared with no service time limit ($\mathcal{T} = 100$) is less. When $\mathcal{T} = 24$ and $\tau = 100\%$ for example, the increase is still remarkable, but at most 7.2%, 5.6%, and 3.9%, for single, multiple, and multiple allocation with direct shipment models, respectively.

It is also possible to calculate the difference in total costs between two target service time limits. Consequently, our proposed hub location models with congestion can provide a good decision support in determining the delivery time promises to customers while designing hub networks.

We also analyzed the isolated effects of congestion on total costs. For this analysis, we set the service time limit to 24 hours and tested the three models under different parameter settings. Figure 9 depicts the values of overall costs with congestion factor $\tau \in \{0\%, 10\%, 30\%, 50\%, 100\%, 200\%\}$ for 20-node instances.

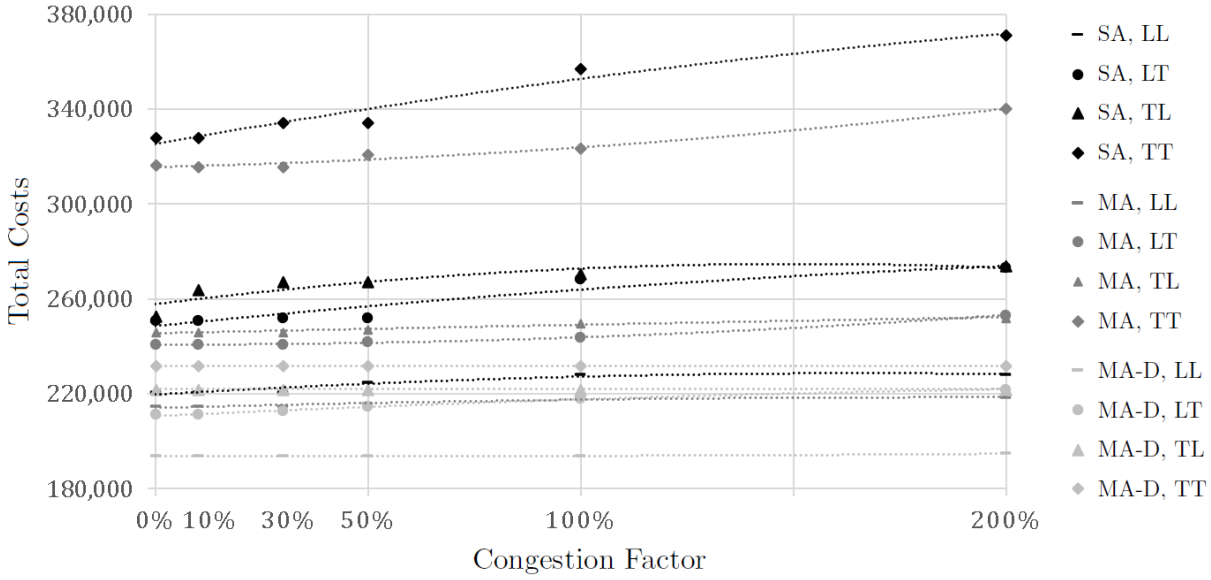

Figure 9: Total costs with varying congestion factor when $\mathcal{T} = 24$.

Figure 9 clearly shows the increase in total costs with an increasing congestion factor. The increase in total costs become more significant with tight capacities. For those instances, where the capacities are tight, there is an increase in costs by up to 30 to 50% compared with no congestion ($\tau = 0\%$).

A congestion factor of 10% results in an increase in total costs of around 1% compared with having no congestion effect. A small congestion factor may not result in a big difference in the overall costs and sometimes does not result in a different solution at all. However, the network

can be made resilient against congestion with a reasonable increase in total cost. In Section 5.4, we specifically analyze the value of considering congestion in hub location problems.

5.3 Number of hubs and ratio of direct shipment

Total cost of a solution depends on the number of opened hubs and their sizes. In general, the number of hubs in a solution decreases with an increased, and, hence, a more relaxed, service time limit. With loose service times, demand nodes can be allocated to distant hubs and more flow can be routed via hubs to exploit economies of scale.

Table 2 gives an overview on the number of hubs that are opened for the instances with 20 nodes. For each fixed cost and capacity combination, when the value of the congestion factor τ is increased, the number of hubs that are opened either increases or remains the same.

Single Allocation												
	$\tau = 0\%$				$\tau = 30\%$				$\tau = 100\%$			
$\mathcal{T}$	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT
22	4	5	4	5	5	5	4	5	5	6	4	6
24	3	5	2	4	3	5	2	5	3	6	2	5
30	3	5	2	4	3	5	2	4	3	5	2	5
100	3	5	2	4	3	5	2	4	3	5	2	4

Multiple Allocation												
	$\tau = 0\%$				$\tau = 30\%$				$\tau = 100\%$			
$\mathcal{T}$	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT
22	3	5	3	5	3	5	3	5	3	5	3	5
24	2	5	2	4	2	5	2	4	3	5	2	5
30	2	5	2	4	2	5	2	4	2	5	2	4
100	2	5	2	4	2	5	2	4	2	5	2	4

Multiple Allocation with Direct Shipment												
	$\tau = 0\%$				$\tau = 30\%$				$\tau = 100\%$			
$\mathcal{T}$	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT
22	2	2	2	0	2	2	2	0	2	3	2	0
24	2	2	1	0	2	2	1	0	2	3	1	0
30	2	2	1	0	2	2	1	0	2	2	1	0
100	2	2	1	0	2	2	1	0	2	2	1	0

Table 2: Number of hubs that are opened for $|N| = 20$.

For some instances with direct shipments, observe from Table 2 that, no hubs are established. This is the case with tight fixed costs and capacities (TT). In such instances, demand is satisfied only through direct shipments between the non-hub nodes.

We also observed the capacities of the hubs that are opened in the solutions. With tight capacity sets, the models result in opening more large-sized hubs as expected. Apart from this observation, however, there is not a clear trend in the solutions with regard to the hub

capacities. The solutions illustrated in Section 5.1 also demonstrate that differently sized hubs can be opened in different problem instances.

We analyzed the congestion levels at hubs in the solutions as well. Hubs tend to have higher utilization rates with tight capacities. The utilization rate of the hubs decreases, however, with an increasing congestion factor. This is because the relative time delay is dependent on the congestion factor.

Congestion in the hubs can be avoided by using direct shipments. Figure 10 depicts the ratio of direct shipments for 20-node instances with $\tau = 100\%$. This ratio is calculated by dividing the total amount of flow on direct shipments by the total flow on the network.

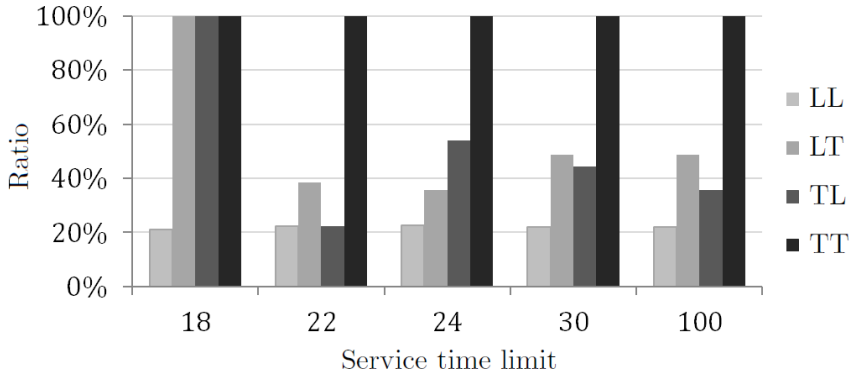

Figure 10: Ratio of direct shipment.

With tight fixed cost and capacities (TT), observe from Figure 10 that all flow (100%) is shipped through direct connections under every service time limit. Hence, no hubs are opened in those instances. When the service time limit is very tight ($\mathcal{T} = 18$), the optimal strategy is again to ship all flow directly between the nodes, except for the set of loose costs and capacities (LL). With loose service time limits and capacities, we see that the ratio of direct shipments ranges from 20% to 50%. We may conclude from this analysis that the service time limit has a major effect on the design of hub networks with direct shipments. In the next section, we define and analyze the value of considering congestion in hub location.

5.4 The value of considering congestion

To evaluate the significance of considering congestion, we define a measure named the *Value of Congestion* (VoC). The VoC is the relative additional cost the decision maker has to bear to adapt a solution that was built without considering congestion to a scenario with congestion. The idea is similar to the value of the multi-period solution introduced in [22].

In order to calculate the VoC, the models are initially solved without considering any delay

due to congestion. Then, the obtained locations and sizes of the hubs are set as input parameters in the models that include congestion, and the allocations, flows, ready times, and congestion levels are re-optimized using the proposed models. The resulting objective function value shall be denoted by f^{voc}. It is compared with the optimal objective function value f^* of the model with congestion, and the Value of Congestion is defined as follows:

$$VoC(\%) = \frac{f^{voc} - f^*}{f^*} \times 100\%$$

If the hub locations and sizes obtained without considering any delay are infeasible, then the solution cannot be adapted and, hence, no f^{voc} can be determined. In this case, VoC is defined as infinity.

We calculated the value of congestion using all the three models that we developed under different service time limits and congestion factors. We report results corresponding to two service time limits and five congestion factors for 20-node instances in Table 3.

Single Allocation

	$\tau = 10\%$				$\tau = 30\%$				$\tau = 50\%$				$\tau = 100\%$				$\tau = 200\%$			
$\mathcal{T}$	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT
22	0.28	0	0.23	0	0.75	0.82	0.60	0.62	2.29	$*$	1.81	$*$	6.34	$*$	5.02	$*$	6.34	$*$	5.02	$*$
24	0	0	5.09	0	0	2.30	$*$	$*$	1.65	$*$	$*$	$*$	3.62	$*$	$*$	$*$	3.62	$*$	$*$	$*$

Multiple Allocation

	$\tau = 10\%$				$\tau = 30\%$				$\tau = 50\%$				$\tau = 100\%$				$\tau = 200\%$			
$\mathcal{T}$	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT
22	0	0.46	0	0.31	0	2.78	0	2.05	0	3.46	0	2.57	0.06	$*$	0.05	$*$	0.99	$*$	0.82	$*$
24	0	0	0	0	0.01	0.05	0.01	$*$	2.99	0.38	0.49	$*$	1.69	1.70	1.48	$*$	2.99	$*$	2.61	$*$

Multiple Allocation with Direct Shipment

	$\tau = 10\%$				$\tau = 30\%$				$\tau = 50\%$				$\tau = 100\%$				$\tau = 200\%$			
$\mathcal{T}$	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT	LL	LT	TL	TT
22	0	0	0.01	0	0.49	0.19	0.56	0	0.73	3.04	0.77	0	0.97	4.60	0.97	0	0.97	7.88	0.97	0
24	0	0	0	0	0	0.48	0.02	0	0.48	1.68	0.02	0	0	4.62	0.02	0	0.48	6.85	0.02	0

Table 3: Value of congestion in percentages, $|N| = 20$.

For each model and input parameter, Table 3 reports the value of congestion in percentages. The zero entries correspond to instances where the problems resulted in an optimal solution in which there is no value of considering congestion. That is, in these instances, the models resulted in exactly the same hub network with and without considering congestion. The starred entries ($*$) in Table 3, on the other hand, correspond to instances where the VoC is infinity as defined above.

With the single allocation model, among the 40 instances listed in Table 3, 17 of them resulted in an infeasible solution when the congestion effects is not considered. With the multiple

allocation model, the number of instances that resulted in an infeasible solution drops down to 9 over 40. Obviously, congestion has a more severe effect on the single allocation instances compared with multiple allocation. With the possibility of direct shipments, on the other hand, the solutions obtained without considering any delays do not lead to an infeasible hub network for the original problem.

The effect of neglecting congestion may result in building hub networks with total costs of up to 6.34% more with single allocation. This percentage is even more with the multiple allocation with direct shipment model. In these instances, the cost of not considering congestion can be up to 7.88% more than the optimal cost. To conclude, this analysis shows that if the decision maker does not consider the effects of congestion, then the total cost of building the hub network can be up to 7.88% more than the optimal value, or the solution is not even adaptable and hence infeasible.

5.5 The impact of the discount factor

The economies of scale discount factor (α) in the AP data set is provided as 0.75 [57]. To evaluate the impact of α on the solutions, we performed computational experiments with two additional values of α: 0.25 and 0.5. We compare total cost and the optimal hub locations under these three different α values in Table 4 for 10-node instances.

As expected, total cost increases with an increasing value of α. More importantly, Table 4 shows that the optimal locations of hubs are affected from the economies of scale parameter. Although there are some preferred locations of hubs in most of the instances, one can clearly observe the differences in the optimal hub locations with the change in the α value. In particular, observe that in some instances the models tend to locate fewer hubs with a higher α value.

We performed additional analysis with the three α values under different service time limits and congestion factors as well. The change in both of these parameters resulted in similar conclusions presented for Table 4. Moreover, we also calculated the value of congestion under different values of α. We cannot identify a clear trend in the VoC with the change in the economies of scale factor. We conclude that the economies of scale factor does not have a significant effect on the value of considering congestion.

5.6 Computation times

The computation times required by CPLEX to solve the models vary not only with the number of nodes, but also a lot with the values of costs and capacities as well as with the allocation patterns and the congestion factor. Generally, CPU time requirement is higher with tight fixed costs and

Allocation pattern	Cost & Capacity	α	Total cost	Locations of hubs
SA	LL	0.25	183,525	1,3,4,7,8
SA	LL	0.5	194,558	1,3,4,7,8
SA	LL	0.75	203,362	3,4,7,8
SA	LT	0.25	191,625	1,3,4,7,8
SA	LT	0.5	202,218	1,3,4,7,8
SA	LT	0.75	212,812	1,3,4,7,8
SA	TL	0.25	219,879	1,4,7
SA	TL	0.5	228,039	3,4,7
SA	TL	0.75	235,457	3,4,7
SA	TT	0.25	243,665	1,4,5,7,8
SA	TT	0.5	253,372	3,4,7,8
SA	TT	0.75	262,509	3,4,7,8
MA	LL	0.25	181,631	1,2,3,7,8
MA	LL	0.5	190,926	1,2,3,7,8
MA	LL	0.75	198,321	2,3,7,8
MA	LT	0.25	187,330	1,3,4,7,8
MA	LT	0.5	196,748	1,3,4,7,8
MA	LT	0.75	203,467	1,3,4,7,8
MA	TL	0.25	219,680	3,4,7
MA	TL	0.5	226,485	3,4,7
MA	TL	0.75	232,490	3,7,8
MA	TT	0.25	231,907	3,4,7,8
MA	TT	0.5	238,585	3,4,7,8
MA	TT	0.75	245,099	3,4,7,8
MA-D	LL	0.25	167,840	3,4,7
MA-D	LL	0.5	174,308	3,4,7
MA-D	LL	0.75	180,178	3,7
MA-D	LT	0.25	175,824	3,4,7,8
MA-D	LT	0.5	183,106	3,7,8
MA-D	LT	0.75	187,347	3,7,8
MA-D	TL	0.25	195,287	3,7
MA-D	TL	0.5	198,391	3,7
MA-D	TL	0.75	200,994	3,7
MA-D	TT	0.25	209,669	3,7,8
MA-D	TT	0.5	211,382	-
MA-D	TT	0.75	211,382	-

Table 4: Results with different α values when $\mathcal{T} = 24$ and $\tau = 100\%$.

capacities and also with a high congestion factor. Furthermore, the multiple allocation problem turned out to be more time consuming than its single allocation counterpart. The computation times increase even more if direct shipments are allowed. Figures 11 to 13 present computation times with CPLEX for increasing numbers of nodes, with two congestion factors, and different sets of cost and capacities.

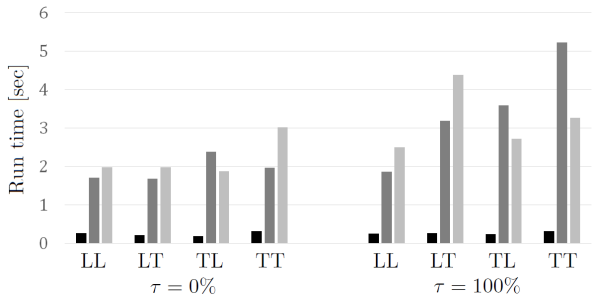

Figure 11: Average run time for $|N| = 10$.

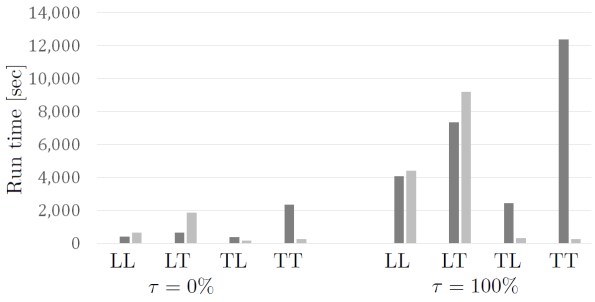

Figure 12: Average run time for $|N| = 20$.

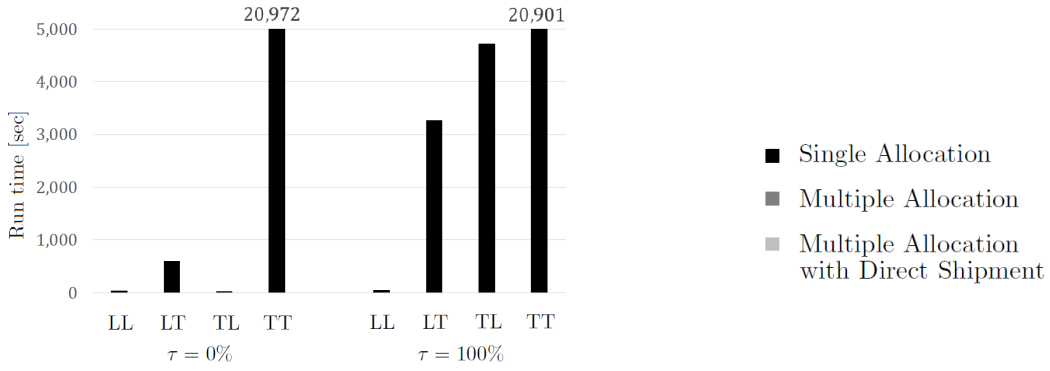

Figure 13: Average run time for $|N| = 40$ with single allocation.

For 10-node problems, all single allocation instances can be solved in less than a second. For both of the multiple allocation models, run times lie between 1.8 and 3 seconds if there is no delay due to congestion, and between 1.9 and 5.2 seconds if congestion doubles the handling time ($\tau = 100\%$).

All 20-node instances can be solved in at most forty minutes when there is no delay. With $\tau = 100\%$, run times increase to almost 3.5 hours. 20-node single allocation instances solved on the average in 10 seconds, hence, these instances cannot be observed due to the scaling in Figure 12. TL- and TT-instances with the multiple allocation with direct shipment model resulted in lower CPU time requirements as all shipments are made by direct connections in these instances.

40-node instances are much more challenging. All single allocation instances with 40 nodes can be solved optimally within 1.5 hours except for the tight fixed cost and capacity (TT) instances. Multiple allocation instances with 40 nodes, on the other hand, could not be solved within 6 hours even with a 1% optimality gap. Table 5 gives an overview of our results with a service time limit of 24 hours, considering a congestion factor of 100% and a discount factor α of 0.75.

Allocation Pattern	Number of Nodes	Cost & Capacity	LP-Relax. Gap (%)	Total Cost	Number of Hubs	CPU Time (Sec)	Gap (%)
SA	10	LL	1.4	203,362	4	0.3	0
SA	10	LT	4.1	212,812	5	0.3	0
SA	10	TL	1.6	235,457	3	0.4	0
SA	10	TT	8.1	262,509	4	0.3	0
SA	20	LL	4.6	228,103	3	1.4	0
SA	20	LT	8.9	268,404	6	12	0
SA	20	TL	7.3	270,328	2	2.2	0
SA	20	TT	11.8	357,164	5	13	0
SA	40	LL	2.2	232,986	4	58	0
SA	40	LT	7.6	291,523	7	7.0	0
SA	40	TL	7.1	360,447	3	73	0
SA	40	TT	11.6	593,189	7	21,600	4.7
MA	10	LL	0	198,321	4	1.3	0
MA	10	LT	1.6	203,467	4	4.8	0
MA	10	TL	1.1	232,490	3	2.8	0
MA	10	TT	4.5	245,099	4	7.9	0
MA	20	LL	1.6	218,147	3	575	0
MA	20	LT	1.2	243,714	5	6204	0
MA	20	TL	1.4	249,574	2	308	0
MA	20	TT	4.0	323,285	5	21,600	2.1
MA-D	10	LL	1.8	180,178	2	3.9	0
MA-D	10	LT	2.9	187,347	3	7.0	0
MA-D	10	TL	3.1	200,994	2	4.4	0
MA-D	10	TT	3.7	211,382	0	3.6	0
MA-D	20	LL	0.8	193,777	2	104	0.8
MA-D	20	LT	3.7	217,920	3	6125	1.0
MA-D	20	TL	2.6	222,023	1	434	0.5
MA-D	20	TT	0	231,645	0	251	0

Table 5: Results with $\mathcal{T} = 24$, $\tau = 100\%$, and $\alpha = 0.75$.

For each allocation pattern, problem size, cost and capacity set, Table 5 presents the percent gap of the LP relaxation from the best known solution, total cost of the best solution, number of hubs in the corresponding solution, CPU time requirement by CPLEX to obtain this solution,

and the gap reported by CPLEX at the end of the run time.

As can be seen from Table 5, the LP relaxations of all the three models are quite tight. The LP relaxation gap increases when costs and capacities get tighter. TT instances are among the hardest instances. CPLEX could not find the optimal solution for some of such instances even after 6 hours (21,600 seconds) of run time. The direct shipment model is also challenging, thus, an optimality gap of 1% was set when solving this model with 20 node instances.

To see the extent of the solution potential with CPLEX we additionally tested the single allocation model on 100-node AP instances as well. Even after 6 hours of run time, CPLEX resulted in a gap of up to 30.6%. After 24 hours of computation time, the maximum gap reduced to 18.6%. This shows that instances with 100 nodes are not solvable within an acceptable optimality gap in reasonable CPU times. Thus, there is an obvious need to develop tailored exact or heuristic solution methodologies for these problems in the future.

6 Conclusion

In this paper, we modeled service time constraints and congestion in hub location problems. Service time is calculated by considering both the travel time on the network connections and handling time at hubs. Congestion is taken into account in order to determine the delay caused by the increase in handling times at hubs.

We developed mixed-integer linear programming formulations for the single and multiple allocation versions of this problem. We also modeled the possibility of direct shipments within the multiple allocation setting.

The models were tested on the well-known AP data set. The results were analyzed under different parameter settings including variations in service time limit, congestion factor, fixed costs, and capacities.

We showed that total costs increase with tighter service time requirements. Service time limit has a more dominant effect on total cost compared with the effect of the congestion factor. Nevertheless, the congestion factor enhances the effect of service time on costs. For all models, the increase in total costs due to either the service time limit or the congestion factor is more significant with tight hub capacities. Moreover, the effect of service time and congestion on costs is more severe on the single allocation instances compared with multiple allocation.

We defined a measure for the value of modeling congestion and showed that for a given service time limit, the decision maker may end up with higher costs or an infeasible hub network unless delay due to congestion is taken into account. The proposed models are valuable decision support tools in determining delivery time guarantees while considering congestion in the design

of hub networks.

A central aspect in our work concerns congestion and the way of capturing it. We discretized the congestion and embedded it in discrete optimization models that were tackled by a general-purpose solver. Having observed that even such a simplified way of capturing congestion may render solutions (and costs) quite different from those obtained if that aspect is ignored, we are encouraged to explore other ways (possibly closer to real-world needs) of capturing congestion like using piece-wise linear functions. Nevertheless, changing the way congestion is modeled leads to a complete change in the structure of the mathematical problems to tackle, which may call for appropriate valid inequalities like given in [58] orspecially tailored approaches for the resulting models. This is certainly an aspect that requires further research. Run times and currently solvable instance sizes additionally motivate developing heuristics or tailored exact methods to tackle the introduced problems.

References

[1] I. Contreras, Hub location problems, in: G. Laporte, S. Nickel, F. Saldanha-da-Gama (Eds.), Location Science, Springer, 2015, pp. 311–344.

[2] S. Alumur, B. Kara, Network hub location problems: the state of the art, European Journal of Operational research 190 (2008) 1–21.

[3] J. Campbell, A. Ernst, M. Krishnamoorthy, Hub location problems, in: Z. Drezner, H. W. Hamacher (Eds.), Facility Location: Applications and Theory, Springer, 2002, pp. 373–407.

[4] J. Campbell, M. O'Kelly, Twenty-five years of hub location research, Transportation Science 46 (2012) 153–169.

[5] H. Yaman, Allocation strategies in hub networks, European Journal of Operational Research 211 (2011) 442–451.

[6] M. O'Kelly, The location of interacting hub facilities, Transportation Science 20 (1986) 92–106.

[7] M. O'Kelly, A quadratic integer program for the location of interacting hub facilities, European Journal of Operational Research 32 (1987) 393–404.

[8] A. Alibeyg, I. Contreras, E. Fernández, Hub network design problems with profits, Transportation Research Part E: Logistics and Transportation Review 96 (2016) 40–59.

[9] S. Alumur, B. Kara, O. Karasan, The design of single allocation incomplete hub networks, Transportation Research Part B 43 (2009) 936–951.

[10] S. Alumur, B. Kara, O. Karasan, Multimodal hub location and hub network design, Omega 40 (2012) 927–939.

[11] J. Campbell, A. Ernst, M. Krishnamoorthy, Hub arc location problems: Part I - Introduction and Results, Management Science 51 (10) (2005) 1540–1555.

[12] J. Campbell, A. Ernst, M. Krishnamoorthy, Hub arc location problems: Part II - Formulations and Optimal Algorithms, Management Science 51 (10) (2005) 1556–1571.

[13] I. Contreras, E. Fernández, A. Marín, The tree of hubs location problem, European Journal of Operational Research 202 (2010) 390–400.

[14] S. Nickel, A. Schöbel, T. Sonneborn, Hub location problems in urban traffic networks, in: J. Niittymaki, M. Pursula (Eds.), Mathematics Methods and Optimization in Transportation Systems, Kluwer Academic Publishers, 2001, pp. 1–12.

[15] A.-K. Rothenbächer, M. Drexl, S. Irnich, Branch-and-price-and-cut for a service network design and hub location problem, European Journal of Operational Research 255 (2016) 935–947.

[16] H. Yaman, Star p-hub median problem with modular arc capacities, Computers & Operations Research 35 (2008) 3009–3019.

[17] S. Alumur, S. Nickel, F. Saldanha-da-Gama, Hub location under uncertainty, Transportation Research Part B: Methodological 46 (2012) 529–543.

[18] R. Bollapragada, J. Camm, U. Rao, J. Wu, A two-phase greedy algorithm to locate and allocate hubs for fixed-wireless broadband access, Operations Research Letters 33 (2005) 134–142.

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

[20] V. Marianov, D. Serra, Location models for airline hubs behaving as M/D/c queues, Computers & Operations Research 30 (2003) 983–1003.

[21] T. Sim, T. Lowe, B. Thomas, The stochastic p-hub center problem with service-level constraints, Computers & Operations Research 36 (12) (2009) 3166–3177.

[22] S. Alumur, S. Nickel, F. Saldanha-da-Gama, Y. Secerdin, Multi-period hub network design problems with modular capacities, Annals of Operations Research 246 (2016) 289–312.

[23] I. Contreras, J.-F. Cordeau, G. Laporte, The dynamic uncapacitated hub location problem, Transportation Science 45 (2011) 18–32.

[24] S. Alumur, E. Serper, The design of capacitated intermodal hub networks with different vehicle types, Transportation Research B 86 (2016) 51–65.

[25] T. Aykin, Lagrangian relaxation based approaches to capacitated hub-and-spoke network design problem, European Journal of Operational Research 79 (1994) 501–523.

[26] T. Aykin, The hub location and routing problem, European Journal of Operational Research 83 (1995) 200–219.

[27] T. Aykin, Networking policies for hub-and-spoke systems with application to the air transportation system, Transportation Science 29 (1995) 201–221.

[28] C. Sung, H. Jin, Dual-based approach for a hub network design problem under non-restrictive policy, European Journal of Operational Research 132 (2001) 88–105.

[29] B. Wagner, An exact solution procedure for a cluster hub location problem, European Journal of Operational Research 178 (2007) 391–401.

[30] M.E. O'Kelly, Activity levels at hub facilities in interacting networks, Geographical Analysis 18 (4) (1986) 343–356.

[31] J. Campbell, Integer programming formulations of discrete hub location problems, European Journal of Operations Research 72 (1994) 387–405.

[32] A. Ernst, H. Hamacher, H. Jiang, M. Krishnamoorthy, G. Woeginger, Uncapacitated single and multiple allocation p-hub center problems, Computers & Operations Research 36 (2009) 2230–2241.

[33] B. Wagner, Model formulations for hub covering problems, Journal of the Operational Research Society 59 (2008) 932–938.

[34] M. Peker, B. Kara, The p-hub maximal covering problem and extensions for gradual decay functions, Omega 54 (2015) 158–172.

[35] J. Campbell, Hub location for time definite transportation, Computers & Operations Research 36 (2009) 3107–3116.

[36] P. Grove, M. O'Kelly, Hub networks and simulated schedule delay, Papers of the Regional Science Association 59 (1986) 103–119.

[37] S. Elhedhli, F. Hu, Hub-and-spoke network design with congestion, Computers & Operations Research 32 (2005) 1615–1632.

[38] R. Camargo, G. Miranda, R. Ferreira, H. Luna, Multiple allocation hub-and-spoke network design under hub congestion, Computers & Operations Research 36 (2009) 3097–3106.

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

[40] R. Camargo, G. Miranda, R. Ferreira, A hybrid outer-approximation/benders decomposition algorithm for the single allocation hub location problem under congestion, Operations Research Letters 39 (2011) 329–337.

[41] R. Camargo, G. Miranda, Addressing congestion on single allocation hub-and-spoke networks, Pesquisa Operacional 32 (2012) 465–496.

[42] R. Camargo, G. Miranda, Single allocation hub location problem under congestion: Network owner and user perspectives, Expert Systems with Applications 39 (2012) 3385–3391.

[43] J. Meier, U. Clausen, Solving classical and new single allocation hub location problems on euclidean data, Tech. rep., Tech. rep., Optimization Online, http://www. optimization-online. org/DB_HTML/2015/03/4816. html (2015).

[44] R. Ishfaq, C. Sox, Design of intermodal logistics networks with hub delays, European Journal of Operational Research 220 (2012) 629–641.

[45] H. Hasanzadeh, M. Bashiri, A. Amiri, A new approach to optimize a hub covering location problem with a queue estimation component using genetic programming, Soft Computing-DOI: 10.1007/s00500-016-2398-1.

[46] Y. Rahimi, R. Tavakkoli-Moghaddam, M. Mohammadi, M. Sadeghi, Multi-objective hub network design under uncertainty considering congestion: An m/m/c/k queue system, Applied Mathematical Modelling 40 (2016) 4179–4198.

[47] A. Ernst, M. Krishnamoorthy, Solution algorithms for the capacitated single allocation hub location problem, Annals of Operations Research 86 (1999) 141–159.

[48] J. Ebery, M. Krishnamoorthy, A. Ernst, N. Boland, The capacitated multiple allocation hub location problem: Formulations and algorithms, European Journal of Operational Research 120 (2000) 614–631.

[49] I. Correia, S. Nickel, F. Saldanha-da-Gama, Single-assignment hub location problems with multiple capacity levels, Transportation Research Part B 44 (2010) 1047–1066.

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

[51] B. Kara, B. Tansel, The latest arrival hub location problem, Management Science 47 (2001) 1408–1420.

[52] I. Correia, S. Nickel, F. Saldanha-da-Gama, The capacitated single-allocation hub location problem revisited: A note on a classical formulation, European Journal of Operations Research 207 (2010) 92–96.

[53] Farahani, Reza Zanjirani and Steadie Seifi, Maryam and Asgari, Nasrin, Multiple criteria facility location problems: A survey, Applied Mathematical Modelling 34 (7) (2010) 1689–1709. doi:10.1016/j.apm.2009.10.005.

[54] M. Ehrgott, Multicriteria optimization, Springer Science & Business Media, 2013.

[55] R. Hall, Direct versus terminal freight routing on a network with concave costs, Transportation Research Part B 21 (1987) 287–298.

[56] A. Mahmutogulari, B. Kara, Hub location problem with allowed routing between nonhub nodes, Geographical Analysis 47 (2015) 410–430.

[57] J. Beasley, OR Library: Hub location (1990).
URL `http://people.brunel.ac.uk/~mastjjb/jeb/orlib/phubinfo.html`

[58] Isabel Correia and Stefan Nickel and Francisco Saldanha-da-Gama, Hub and spoke network design with single-assignment, capacity decisions and balancing requirements, Applied Mathematical Modelling 35 (10) (2011) 4841–4851. doi:10.1016/j.apm.2011.03.046.