KEDGE BUSINESS SCHOOL
ILS2016
Information Systems Logistics and Supply Chain
6th International Conference
université de BORDEAUX

Responsive Make-to-Order Supply Chain Network Design

Robert Aboolian
Department of Information Systems and Operations Management,
California State University San Marcos,
San Marcos, California 92096, USA

Oded Berman
Rotman School of Management, University of Toronto
105 St. George Street, Toronto, Ontario, Canada M5S 3E6

Jiamin Wang
College of Management, Long Island University
C.W. Post Campus 720 Northern Boulevard, Brookville, New York 11548, USA

March 2016

Abstract

In this paper we address the network design of a responsive supply chain consisting of make-to-order (make-to-assemble) facilities facing stochastic demand from customers residing at nodes of a network. Each facility has a finite (processing) capacity and thus the stochasticity of demand may lead to congestion delays at the facilities. The objective is to determine the number, locations and capacities of the facilities so as to minimize the total network cost. We consider three problems. In the first, we minimize the total network cost which includes delivery and capacity costs while maintaining an acceptable response time to customers. In the second, a penalty is charged on the number of units that are delivered later than the targeted response time. In the third the penalty charged depends also on the number of days that the delivery is late. In both problems 2 and 3 the penalty cost is a function of network's response time.

1 Introduction

Make-to-Order (MTO) and Make-to-Assemble (MTA) systems are successful strategies in managing supply chains that use mass customization and compete on product variety. Dell's manufacturing and distribution of Personal Computers (PCs) is an excellent example of an MTO supply chain. The typical response time in Make-to-Stock (MTS) systems is much shorter than in MTO and MTA systems. Therefore a reduction of response time (delivery

time) of orders to customers is a major issue in both MTO and MTA systems. In this paper we address the network design of a responsive supply chain consisting of MTO or MTA facilities facing stochastic demand from customers residing at the nodes of a network. Each facility has a finite (processing) capacity and thus the stochasticity of demand may lead to congestion delays at the facilities. We intend to determine the number, locations and capacities of the facilities to minimize the total network cost. While many supply chain design models have been proposed to support a reduction in response time, these models are more concerned with the efficiency and cost in MTS supply chains under a deterministic customer demand settings. Vidal and Goetschalckx (2000) present a model that captures the effect of a change in transportation lead time and demand on the optimal configuration of the global supply chain network, assuming a deterministic customer demand. Eskigun et al. (2005) incorporate delivery lead time and the choice of transportation mode in the design of a supply chain under a deterministic demand setting. These models tend to ignore congestion at the facilities and its effect on response time. The closest work to our paper is Vidyarthi et al. (2009), who present a model to determine the configuration of an MTO supply chain. In this paper the emphasis is on minimizing the customer response time through the acquisition of sufficient assembly capacity and the optimal allocation of workload to the assembly facilities. They model the cost for response time using a direct relationship with the average waiting time, which is not really practical.

In this paper, we consider three problems. In the first, we minimize the total network cost while maintaining an acceptable response time with an agreed upon probability for delays. Although originally formulated as a non-linear integer program, we show that this problem can be reformulated to a Mixed Integer Program (MIP). In both the second and third problems, if the order is delivered after the targeted response time, the network will be charged a penalty for late delivery. In the second problem a penalty is charged on the number of units that are delivered later than the targeted response time. In the third problem the penalty charged depends also on the number of days that the delivery is late. Although the problems are highly non-linear, we managed to solve them in an efficient manner using the Tangent Line Approximation (TLA) technique, developed in Aboolian et al. (2007b) to linearize the non-linearity in these models.

2 Problem Formulation

We consider a discrete set $M = \{1, 2, \ldots, m\}$ of potential facility locations, a discrete set $N = \{1, 2, \ldots, n\}$ of customer locations. Without loss of generality, we assume $M \subset N$. Depending on the application, N could represent nodes of a network or a set of points on a plane. A certain number of facilities offering a pre-specified set of services is to be located in M. The facilities provide make-to-order service, i.e., each facility can be thought of as a queuing system.

We assume that customers at $i \in N$ generate a stream of Poisson demands with homogeneous daily rate $\lambda_i > 0$. Consider a facility at $j \in M$ and let E_j be the set of all demand points served by facility j. Then Λ_j, the

total daily demand at facility j, is given by

$$\Lambda_j = \sum_{i \in E_j} \lambda_i \quad \text{for} \quad j \in M. \tag{1}$$

We consider an $M/M/1$ single-channel Markovian service queue for the facilities where the capacity level of each facility $\mu_j > \Lambda_j$ can be chosen from a finite set of desired levels. The service rate μ_j of the single server at facility j, is a decision variable in our problems. Define W_j as the total time an order spends at the facility including waiting and service time. We note that in the models that follow, no significant complications arise when non-Markovian service is allowed (i.e. $M/G/1$ disciplines), as long as formulas for the probability of $P(W_j > t)$ are available. For an $M/M/1$ queuing system,

$$P(W_j > t) = e^{-(\mu_j - \Lambda_j)t} \quad \text{for } j \in M. \tag{2}$$

We assume a fixed location cost f_j for locating a (zero-capacity) facility at $j \in M$. Let the set $G = \{g_1, g_2, \ldots, g_q\}$ represents q service capacities available for the facilities and $H = \{h_1, h_2, \ldots, h_q\}$ represents the set of the corresponding costs (i.e. h_r is the cost to obtain a service capacity of g_r for $r \in \{1, 2, \ldots, q\}$). Let $F_{jr} = f_j + h_r$ be the cost of locating a facility at $j \in M$ with a service capacity g_r for $r \in \{1, 2, \ldots, q\}$. Define a binary decision variable x_j, $j \in M$ to be 1 if a facility is opened at j and 0 otherwise, and a binary decision variable z_{jr} to be 1 if service capacity g_r ($r \in \{1, 2, \ldots, q\}$) is assigned to facility $j \in M$ and 0 otherwise. Then

$$\mu_j = \sum_{r=1}^{q} g_r z_{jr} \quad \text{for } j \in M. \tag{3}$$

We first consider a problem, in which we intend to minimize the total network cost while maintaining an acceptable response time to its customers. Here we assume that the network uses a third party logistics which we call 3PL (e.g. UPS) to deliver the completed orders to the customers. To deliver the order to a customer, the network has the option to use the standard ground delivery time or a menu of delivery times offered by 3PL (e.g. next day delivery, 2nd day delivery, etc.) at higher costs. It is assumed that the delivery cost and the delivery times are guaranteed by 3PL which picks up the orders at the end of the day. (Orders completed during the day could be delivered at the end of the next day at the earliest.)

Let L_i be the targeted response time to customers from node i in days and α_i be the probability that the response time to node i will not exceed L_i. For simplicity of presentation we use L and α instead of L_i and α_i. The responsiveness constraint is the probability that an order is not delivered in L days is less than or equal to $1 - \alpha$.

Define K_{ij} to be the set of finite delivery times from facility j to customer i and c_{ijk} to be the cost for delivering an order from facility j to customer i in k days. The costs c_{ijk} may be dependent on shortest distances $d_{ij}, i, j \in M \cup N$ (in which case we assume that d_{ij} is the shortest path between i and j). We also assume that L is larger than all delivery times in $\bigcup_{i,j} K_{ij}$. Let the binary decision variable y_{ijk} be 1 if the order for customer $i \in N$ is delivered from facility $j \in M$ in $k \in K_{ij}$ days and 0 otherwise. Then the network cost is given by the

sum of delivery and capacity costs:

$$Z = \sum_{j\in M}\sum_{i\in N}\sum_{k\in K_{ij}} \lambda_i c_{ijk} y_{ijk} + \sum_{j\in M}\sum_{r=1}^{q} F_{jr} z_{jr}. \tag{4}$$

Denote $j(i) \in M$ to be the facility that serves customer $i \in N$. Then R_i, the network's response time to customer $j(l) \in N$, is the sum of $t_{i,j(i)}$, the delivery time from facility $j(i) \in M$ to customer $i \in N$ and, $W_{i(j)}$, the uncertain time an order spends at facility $j(i)$, i.e. $R_i = t_{i,j(i)} + W_{j(i)}$. The responsiveness constraint for customers at $i \in N$ can be presented as

$$P(R_{j(i)} > L) = P(W_{j(i)} > L - t_{i,j(i)}) \leq 1 - \alpha \quad \text{for} \quad i \in N. \tag{5}$$

Thus, the network's responsiveness constraint can be presented as:

$$\sum_{k\in K_{ij}} P(W_j > L - k) y_{ijk} \leq 1 - \alpha \quad \text{for} \quad i \in Nj \in M. \tag{6}$$

We can now state the mathematical programming formulation of the Responsive Supply Chain Network Design (RSCND) as follows:

$$MinZ = \sum_{j\in M}\sum_{i\in N}\sum_{k\in K_{ij}} \lambda_i c_{ijk} y_{ijk} + \sum_{j\in M}\sum_{r=1}^{q} F_{jr} z_{jr} \tag{7}$$

$$\sum_{j\in M}\sum_{k\in K_{ij}} y_{ijk} = 1, \ i \in N, \tag{8}$$

$$y_{ijk} \leq x_j, \quad i \in N, j \in M, k \in K_{ij}, \tag{9}$$

$$\sum_{r=1}^{q} z_{jr} = x_j, \ j \in M, \tag{10}$$

$$\sum_{r=1}^{q} g_r z_{jr} \geq \sum_{i\in N}\sum_{k\in K_{ij}} \lambda_i y_{ijk}, \ j \in M, \tag{11}$$

$$\sum_{k\in K_{ij}} P(W_j > L - k) y_{ijk} \leq 1 - \alpha\,, \ i \in N, j \in M, \tag{12}$$

$$x_j, z_{jr}, y_{ijk} \in \{0,1\}, i \in N, j \in M, r \in \{1,2,\ldots,q\}, k \in K_{ij}. \tag{13}$$

Constraints (8, 9) are the standard constraints in location models enforcing the connections between the decision variables and making sure that only open facilities can serve customers and each customer is assigned to one facility at one delivery time. Constraints (10) ensure that when $x_j = 0$ no service capacity is assigned to facility $j \in M$ and only one service capacity in G is assigned to facility $j \in M$ when $x_j = 1$. The next set of constraints (11) ensure the stability of the queuing system at each open facility. Constraints (12) ensure the network's responsiveness requirements.

The main difficulty of solving the model above is, clearly, the responsiveness requirements (12), which are non-linear. In the next section we show how to linearize constraints (12).

In the second problem, instead of setting a service level of α, we look for the optimal level of α for the network. Here, we assume that if an order is not fulfilled on or before the targeted response time L a penalty should be

paid for each unit late. Define ϕ to be the penalty that the network has to pay for each unit delivered later than the targeted response time of L days. Then the network cost is given by

$$Z = \sum_{j\in M}\sum_{i\in N}\sum_{k\in K_{ij}} \lambda_i c_{ijk} y_{ijk} + \sum_{j\in M}\sum_{r=1}^{q} F_{jr} z_{jr} + \sum_{i\in N}\sum_{j\in M}\sum_{k\in K_{ij}} \lambda_i \phi P(W_j > L-k) y_{ijk}, \quad (14)$$

or

$$Z = \sum_{j\in M}\sum_{i\in N}\sum_{k\in K_{ij}} \lambda_i (c_{ijk} + \phi P(W_j > L-k)) y_{ijk} + \sum_{j\in M}\sum_{r=1}^{q} F_{jr} z_{jr}. \quad (15)$$

We can now state the mathematical programming formulation of the Responsive Supply Chain Network Design with Single Penalty Cost (RSCNDSPC) as follows:

$$MinZ = \sum_{j\in M}\sum_{i\in N}\sum_{k\in K_{ij}} \lambda_i (c_{ijk} + \phi P(W_j > L-k)) y_{ijk} + \sum_{j\in M}\sum_{r=1}^{q} F_{jr} z_{jr} \quad (16)$$

subject to Constraints (8)-(11) and (13).

As for RSCND the main difficulty of solving RSCNDMPC is the non-linearity of the objective function. In the next section we show how we could solve this problem by linearizing the objective function.

The penalty on late delivery depends also on the number of days that the network is late to deliver the customer order. Define ϕ_l to be the penalty charged for each unit late if the order is delivered $l \in \{1, 2, .., l^{\max}\}$ days later than the targeted response time of L (i.e. in $l + L$ days), where $\phi_{l^{\max}}$ is the penalty paid when the order is late $l^{\max}$ days.

Note that we define an order to be $l \in \{1, 2 \ldots, l^{\max} - 1\}$ days late if the response time is greater than $L+l-1$ days but less than or equal to $L + l$ days. We also define an order to be $l^{\max}$ days late if the response time is greater than $L + l^{\max} - 1$ days.

Recall that $j(i) \in M$ is the facility that serves customer $i \in N$ and R_i is the response time to customer $i \in N$. Then the probability that the order to customer i is late by $l \in \{1, 2, \ldots, l^{\max} - 1\}$ days is $P(L + l - 1 < R_{j(i)} \leq L + l) = P(L + l - 1 - t_{i,j(i)} < W_{j(i)} \leq L + l - t_{i,j(i)}) = P(W_{j(i)} > L + l - 1 - t_{i,j(i)}) - P(W_{j(i)} > L + l - t_{i,j(i)})$ and the probability that the network is late to deliver the order to customer i by $l^{\max}$ days equals $P(W_{j(i)} > L + l^{\max} - 1 - t_{i,j(i)})$.

Therefore, the expected penalty cost for network's late delivery to customer i is

$$\sum_{l=1}^{l^{\max}-1} \phi_l \left(P(W_{j(i)} > L + l - 1 - t_{i,j(i)}) - P(W_{j(i)} > L + l - t_{i,j(i)})\right) + \phi_{l^{\max}} P(W_{j(i)} > L + l^{\max} - 1 - t_{i,j(i)}).$$

Note that we can rewrite this cost as $\sum_{l=1}^{l^{\max}} (\phi_l - \phi_{l-1}) P(W_{j(i)} > L + l - 1 - t_{i,j(i)})$, where $\phi_0 = 0$. Then taking into consideration that delivery time between facility $j(i)$ and customer i is a value in K_{ij} chosen by the facility, the network cost is given by

$$Z = \sum_{j\in M}\sum_{i\in N}\sum_{k\in K_{ij}} \lambda_i c_{ijk} y_{ijk} + \sum_{j\in M}\sum_{r=1}^{q} F_{jr} z_{jr} + \sum_{i\in N}\sum_{j\in M}\sum_{k\in K_{ij}}\sum_{l=1}^{l^{\max}} \lambda_i (\phi_l - \phi_{l-1}) P(W_j > L + l - 1 - k) y_{ijk}, \quad (17)$$

or

$$Z = \sum_{j\in M}\sum_{i\in N}\lambda_i \sum_{k\in K_{ij}} (c_{ijk} + \sum_{l=1}^{l^{\max}}(\phi_l - \phi_{l-1})P(W_j > L + l - 1 - k))y_{ijk} + \sum_{j\in M}\sum_{r=1}^{q} F_{jr}z_{jr}. \tag{18}$$

We can now state the mathematical programming formulation of the Responsive Supply Chain Network Design with Multiple Penalty Costs (RSCNDMPC) as follows:

$$\min Z = \sum_{j\in M}\sum_{i\in N}\lambda_i \sum_{k\in K_{ij}} (c_{ijk} + \sum_{l=1}^{l^{\max}}(\phi_l - \phi_{l-1})P(W_j > L + l - 1 - k))y_{ijk} + \sum_{j\in M}\sum_{r=1}^{q} F_{jr}z_{jr} \tag{19}$$

subject to Constraints (8)-(11) and (13).

Again, the main difficulty of solving the model above is the non-linearity of the objective function. In the next section we will explore how we could solve this problem by linearizing the objective function.

3 Solution Approaches

Here we present exact and approximate approaches for the above models.

3.1 Responsive Supply Chain Network Design

Given (1-3) , and (5), the network responsiveness conditions for customer $i \in N$ can be presented as:

$$e^{-\left(\mu_{j(i)} - \Lambda_{j(i)}\right)(L - t_{i,j(i)})} \leq 1 - \alpha \quad \text{for} \quad i \in N. \tag{20}$$

By taking logarithm of each side of (20), we could simplify it by

$$\mu_{j(i)} \geq \Lambda_{j(i)} - \frac{Ln(1-\alpha)}{L - t_{i,j(i)}} \quad \text{for} \quad i \in N. \tag{21}$$

Since the left and right sides of (11) are respectively $\mu_{j(i)}$ and $\Lambda_{j(i)}$ constraints (11) and (12) in RSCND could be replaced with linear constraints

$$\sum_{r=1}^{q} g_r z_{jr} \geq \sum_{l\in N}\sum_{s\in K_{lj}} \lambda_l y_{ljs} - \frac{Ln(1-\alpha)}{L-k} y_{ijk}, \text{ for } i \in N, j \in M, k \in K_{ij}. \tag{22}$$

Note that if $x_j = 0$ for $j \in M$, then given (9) and (10) $y_{ijk} = 0$, for $i \in N$, and $k \in K_{ij}$, and $z_{jr} = 0$ for $r \in \{1, 2, \ldots, q\}$. Therefore (22) will not be a constraint since $0 \geq 0$. If $x_j = 1$, then given (8) $\ni u \in N, v \in K_{vj}$, for which $y_{ujv} = 1$, and $y_{ijk} = 0$ for $i \in N - \{u\}$, and $k \in K_{ij} - \{v\}$. Therefore since $-\frac{Ln(1-\alpha)}{L-k}$ is positive, (22) turns into $\sum_{r=1}^{q} g_r z_{jr} \geq \sum_{l\in N}\sum_{s\in K_{lj}} \lambda_l y_{ljs} - \frac{Ln(1-\alpha)}{L-v}$. By replacing (22) with (11) and (12) in RSCND, we now have an MIP which is no longer non linear.

3.2 Responsive Supply Chain Network Design with Single Penalty Cost

Here we use the TLA technique to linearize the non-linearity in the objective function and solve the problems for a pre-specified negligible error.

3.3 Responsive Supply Chain Network Design with Multiple Penalty Cost

Here we use the relaxed linearized model with a relatively large error (to reduce the size of the problem) to find the lower bounds. We present a heuristic that is based on a neighborhood search over the location set from the solution to the relaxed model. We also develop an exact approach, which is based on successive lower and upper bound improvements for the original model.

References

[1] Aboolian, R., Berman. O. and Krass, D. (2007a). Competitive Facility Location and Design Problem. European Journal of Operational Research 182, 40-62.

[2] Eskigun, E., Uzsoy, R., Preckel, P.V., Beaujon, G., Krishnan, S. and Tew, J.D. (2005). Outbound supply chain network design with mode selection, lead times, and capacitated vehicle distribution centers, European Journal of Operational Research, 165(1), 182-206.

[3] Vidal,C.J. and Goetschalckx, M. (2000). Modeling the effect of uncertainties on global logistics systems, Journal of Business Logistics, 21(1), 95-120.

[4] Vidyarthi N., Elhedli, S., and Jewkies, E. (2009). Response time reduction in make-to-order and assemble-to-order supply chain design, IIE Transactions 41, 448-466.