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

November 2015

Abstract

In this paper we address the netwrok design of a responsive supply chain consisting of make-to-order (make-to-assemble) facilities facing stochastic demand from customers residing at the nodes of a network. Each facility has a certain finite processing capacity and thus the stochasticity of demand may lead to congestion delays at the facilities. Here we intend to determine the number, locations and service capacities of the facilities to minimize the total netwrok cost. We consider two sets of problems. In the first problem we intend to minimize the total netwrok cost while maintaining an acceptable response time with no delays. In the second problem we allow the facilities to fail on time delivery with a penalty. The penalty cost is a function of responsiveness. We assume that no penalty is charged if the order is delivered on or before its acceptable response time. If the order is delivered after its acceptable lead time, the network will be charged a penalty for every unit of time that the order is late from the acceptable delivery date.

1 Introduction

Make-to-Order (MTO) and Make-to-Assemble (MTA) systems are successful strategies to manage 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. Although an MTO strategy gets away with finished goods inventories and reduces a firm's exposure to the risk of obsolescence, it usually has a longer response time when compared to Make-to-Stock (MTS) strategies. In this paper we address the netwrok 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 certain finite processing capacity and thus the stochasticity of demand may lead to congestion delays at the facilities. Here we intend to determine the

number, locations and service capacities of the facilities to minimize the total netwrok 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 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 the paper is Vidyarthi et al. (2009), where they 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 through a direct relationship with the average waiting time, which is not really practical. In this paper, we consider two sets of problems. In the first problem we intend to minimize the total netwrok 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 the second set of problems we allow the facilities to fail on time delivery by paying a penalty. The penalty cost is a function of network's response time. We assume that no penalty is charged if the order is delivered on or before an acceptable target response time. If the order is delivered after the acceptable target response time, the network will be charged a penalty or a for every unit of time that the order is late from the acceptable delivery date. These set of problems are highly non-linear, but to solve them in an efficient manner, we use 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 $M = \{1, 2, .., m\}$ of potential facility locations, a discrete set $N = \{1, 2, ..., n\}$ of customer locations, and a distance metric d_{ij} for $i, j \in M \cup N$. Without loss of generality, we will assume $M \subset N$. Depending on the application, N could represent nodes of a network (in which case d_{ij} is the shortest path distance between $i, j \in N$), 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 rate $\lambda_i \geq 0$. Suppose there is a facility at $j \in M$ and that all customers at $i \in N$ use the services of facility j. Let E_j be the set of all demand points served by facility$j$. Then Λ_j, the total 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 will consider an $M/M/1$ single-channel Markovian service queue for the facilities where the service capacity level of each facility can be adjusted to a set of desired levels, the service rate $\mu_j \geq 0$ of the single server at facility j is a decision variable (note that μ_j can be chosen from a finite set of values). 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 W_j being more than a specific time are available; we assume Markovian service mainly for the ease of exposition. For an an $M/M/1$ queuing system, the probability thatW_j is more than t can be computed by:

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

We assume a fixed location cost f_j for locating a (zero-capacity) facility at $j \in M$ We assume that the set $G = \{g_1, g_2, ..., g_q\}$ represents the set of q service capacities which is available to obtain for each facility such that $\mu_j \in G$ and $H = \{h_1, h_2, ..., h_q\}$ represents the set of the corresponding costs to obtain each service capacity in G (i.e. h_r is the cost to obtain a service capacity of g_r for $r \in \{1, 2, ..., q\}$). Then $F_{jr} = f_j + h_r$ can be defined as the cost of locating a facility at $j \in M$ with a service capacity g_r for $r \in \{1, 2, ..., q\}$Define a binary decision variable x_j, $j \in M$ to be 1 if a facility is opened at j and 0 otherwise, and define binary decision variable z_{jr} to be 1 if service capacity g_r ($r \in \{1, 2, ..., q\}$) 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 netwrok cost while maintaining a target response time acceptable to its customers. Here we assume that the network uses a third party logistics (3PL) company (e.g. UPS) to ship the completed orders to the customers. To deliver the order to a customer, the netwrok has the option to use the standard ground shipping or a menu of expedited shipping times offered by the 3PL company (e.g. next day delivery, 2nd day delivery, etc.) at higher costs. Network assumes the shipping cost and the shipping times are guaranteed by the 3PL company. We also assume that the 3PL picks up the orders at the end of the day to deliver to customers even if they are completed early in the day. Orders completed during the day could be delivered the next day at earliest. Define K_{ij} to be a set of finite shipping times from facility j to customer i. Define c_{ijk} to be the shipping cost for shipping an order from facility j to customer i in $k \in K_{ij}$ days. Let the binary decision variable y_{ijk} be 1 if the order for customer $i \in N$ is shipped from facility $j \in M$ in $k \in K_{ij}$ days and 0 otherwise. Then the network cost

is given by

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

We also assume that the network's responsiveness is measured by (L, α), where L is the network's target response time to customer orders in days and α is the probability that the response time will not exceed L. In other words the probability that a customer order is not met by its deadline is less than or equal to $1 - \alpha$.

Denote $i(M) \in M$ to be the facility which serves customer $i \in N$. We also assume that R_i, network's response time to customer $i \in N$, is the sum of $t_{i,i(M)}$, the certain shipping time from facility $i(M) \in M$ to customer $i \in N$ and, $W_{i(M)}$, the uncertain time an order spends at facility $i(M)$, then $R_i = t_{i,i(M)} + W_{i(M)}$ is also uncertain and the network's responsiveness constraint for customers at $i \in N$ can be presented as

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

Thus, the network responsiveness conditions can be presented as:

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

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

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

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

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

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

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

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

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

Constraints (8,9) are the standard constraints in location models enforcing the connections between the decision variables and making sure that each customer is assigned to one facility with one shipping time. Constraints (10) are the ensure that no service capacity is assigned at facility $j \in M$ when $x_j = 0$

and only one of service capacities in G is assigned at 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. The constraints (12) ensure the supply chain's responsiveness requirements.

The main difficulty in solving the model above is, clearly, the supply chain's responsiveness requirements (12), which are non-linear. In the next section we will explore how we could solve this problem by linearizing constraints (12).

Next, instead of setting a service level of α for the network, we will look to find the optimal level for α for the network. Here, we assume that if an order is not fulfilled on or before the target response time L it would be considered a backorder and a penalty should be paid for each unit which is backordered. Define ϕ to be the penalty that the network has to pay for each unit of the backorder if the order is delivered later than the target response time of L days. Then the network cost is given by

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

or

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

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

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

$$\sum_{j\in M}\sum_{k\in K_{ij}} y_{ijk} = 1, \quad i \in N, \tag{17}$$

$$y_{ijk} \le x_j, \quad i \in N, j \in M, k \in K_{ij}, \tag{18}$$

$$\sum_{r=1}^{q} z_{jr} = x_j, \ j \in M, \tag{19}$$

$$\sum_{r=1}^{q} g_r z_{jr} \ge \sum_{i\in N}\sum_{k\in K_{ij}} \lambda_i y_{ijk}, \ j \in M, \tag{20}$$

$$x_j \in \{0,1\}, y_{ijk} \in \{0,1\}, z_{jr} \in \{0,1\}, \ \ i \in N, j \in M, r \in \{1,2,...,q\}, k \in K_{ij}. \tag{21}$$

again here constraints (17,18) are the standard constraints in location models enforcing the connections between the decision variables and making sure that each customer is assigned to one facility with one shipping time. Constraints (19) are to ensure that no service capacity is assigned at facility $j \in M$ when $x_j = 0$ and only one of service capacities in G is assigned at facility $j \in M$ when $x_j = 1$ and constraints (20) ensure the stability of the queuing system at each open facility.

The main difficulty in 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.

Next we consider the the case in which late delivery penalty depends on the number of days that the network is late to deliver the customer order. Define ϕ_l to be the penalty that the network has to pay for each unit of the backorder if the order is delivered $l \in \{1, 2, .., l^{\max}\}$ days later than the target response time of L (in $l + L$ days), where $\phi_{l^{\max}}$ is the penalty paid when the order is late $l^{\max}$ days.

Please note that we define an order to be $l \in \{1, 2, .., l^{\max} - 1\}$ 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 the definition of $i(M) \in M$ which is the facility which serves customer $i \in N$ and R_i, network's response time to customer $i \in N$. Then the probability that the netwrok is late to deliver the order to customer i by $l \in \{1, 2, .., l^{\max}-1\}$ days equals $P(L+l-1 < R_{i(M)} \leq L+l) = P(L+l-1-t_{i,i(M)} < W_{i(M)} \leq L + l - t_{i,i(M)}) = P(W_{i(M)} > L + l - 1 - t_{i,i(M)}) - P(W_{i(M)} > L + l - t_{i,i(M)})$ and the probability that the netwrok is late to deliver the order to customer i by $l^{\max}$ days equals $P(W_{i(M)} > L + l^{\max} - 1 - t_{i,i(M)})$.

Therefore, the expected penalty cost for network's late delivery to customer i equals $\sum_{l=1}^{l^{\max}-1} \phi_l \left(P(W_{i(M)} > L + l - 1 - t_{i,i(M)}) - P(W_{i(M)} > L + l - t_{i,i(M)})\right) + \phi_{l^{\max}} P(W_{i(M)} > L + l^{\max} - 1 - t_{i,i(M)})$.Please note that we can rewrite this cost as $\sum_{l=1}^{l^{\max}} (\phi_l - \phi_{l-1}) P(W_{i(M)} > L + l - 1 - t_{i,i(M)})$, where $\phi_0 = 0$. Then the network cost is given by

$$Z = \sum_{j\in M}\sum_{i\in N}\sum_{k\in K_{ij}} \lambda_i c_{ijk} y_{ijk} + \sum_{j\in M}\sum_{r=1}^{q} F_{jr} z_{jr} + \sum_{i\in N}\sum_{j\in M}\sum_{k\in K_{ij}}\sum_{l=1}^{l^{\max}} \lambda_i (\phi_l - \phi_{l-1}) P(W_j > L + l - 1 - k) y_{ijk}, \tag{22}$$

or

$$Z = \sum_{j\in M}\sum_{i\in N} \lambda_i \sum_{k\in K_{ij}} (c_{ijk} + \sum_{l=1}^{l^{\max}} (\phi_l - \phi_{l-1}) P(W_j > L + l - 1 - k)) y_{ijk} + \sum_{j\in M}\sum_{r=1}^{q} F_{jr} z_{jr.} \tag{23}$$

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

follows:

$$MinZ = \sum_{j\in M}\sum_{i\in N}\lambda_i \sum_{k\in K_{ij}} (c_{ijk} + \sum_{l=1}^{l^{\max}}(\phi_l - \phi_{l-1})P(W_j > L + l - 1 - k))y_{ijk} + \sum_{j\in M}\sum_{r=1}^{q} F_{jr}z_{jr} \tag{24}$$

$$\sum_{j\in M}\sum_{k\in K_{ij}} y_{ijk} = 1, \quad i \in N, \tag{25}$$

$$y_{ijk} \leq x_j, \quad i \in N, j \in M, k \in K_{ij}, \tag{26}$$

$$\sum_{r=1}^{q} z_{jr} = x_j, \ j \in M, \tag{27}$$

$$\sum_{r=1}^{q} g_r z_{jr} \geq \sum_{i\in N}\sum_{k\in K_{ij}} \lambda_i y_{ijk}, \ j \in M, \tag{28}$$

$$x_j \in \{0,1\}, y_{ijk} \in \{0,1\}, z_{jr} \in \{0,1\}, \ \ i \in N, j \in M, k \in K_{ij}, r \in \{1,2,...,q\}. \tag{29}$$

The main difficulty in solving the model above again 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

In this Section we will present exact and approximate approaches for the four models developed in the previous Section.

3.1 Responsive Supply Chain Network Design

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

$$e^{-\left(\mu_{i(M)} - \Lambda_{i(M)}\right)(L - t_{i,i(M)})} \leq 1 - \alpha \ \text{ for } \ i \in N \tag{30}$$

By taking logarithm of each side of (30), we could simplify it by

$$\mu_{i(M)} \geq \Lambda_{i(M)} - \frac{Ln(1-\alpha)}{L - t_{i,i(M)}} \ \text{ for } \ i \in N \tag{31}$$

Therefore constraints (11) and (12) in RSCND could be replaced with linear constraints

$$\sum_{r=1}^{q} g_r z_{jr} \geq \sum_{l \in N} \sum_{s \in K_{lj}} \lambda_l y_{ljs} - \frac{Ln(1-\alpha)}{L-k} y_{ijk}, \text{ for } i \in N, j \in M, k \in K_{ij}. \quad (32)$$

Please note that if $x_j = 0$, then given (9) and (10) $y_{ijk} = 0$, for $i \in N, j \in M, k \in K_{ij}$, and $z_{jr} = 0$ for $j \in M, r \in \{1, 2, ..., q\}$.Therefore (33) will not be a constraint since $0 \geq 0$. If $x_j = 1$, then given (8) $\ni u \in N, v \in K_{vj}$,in which $y_{ujv} = 1$, and $y_{ijk} = 0$ for $i \in N - \{u\}, k \in K_{ij}$, and $i \in N$, $k \in K_{ij} - \{v\}$.Therefore since $-\frac{Ln(1-\alpha)}{L-k}$ is positive, (33) 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 (33) 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

3.3 Responsive Supply Chain Network Design with Multiple Penalty Cost

4 Computational Results

5 Concluding Remarks

References

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

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.

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

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