IIE Transactions (2009) **41**, 448–466

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/07408170802382741

Response time reduction in make-to-order and assemble-to-order supply chain design

NAVNEET VIDYARTHI[1,*], SAMIR ELHEDHLI[2] and ELIZABETH JEWKES[2]

[1]*Department of Decision Sciences and MIS, John Molson School of Business, Concordia University, Montreal, Quebec, Canada, H3G 1M8*
E-mail: navneetv@jmsb.concordia.ca
[2]*Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1,*
E-mail: {elhedhli, emjewkes}@uwaterloo.ca

Received December 2006 and accepted June 2008.

Make-to-order and assemble-to-order systems are successful business strategies in managing responsive supply chains, characterized by high product variety, highly variable customer demand and short product life cycles. These systems usually spell long customer response times due to congestion. Motivated by the strategic importance of response time reduction, this paper presents models for designing make-to-order and assemble-to-order supply chains under Poisson customer demand arrivals and general service time distributions. The make-to-order supply chain design model seeks to simultaneously determine the location and the capacity of distribution centers (DCs) and allocate stochastic customer demand to DCs by minimizing response time in addition to the fixed cost of opening DCs and equipping them with sufficient assembly capacity and the variable cost of serving customers. The problem is setup as a network of spatially distributed $M/G/1$ queues, modeled as a non-linear mixed-integer program, and linearized using a simple transformation and a piecewise linear approximation. An exact solution approach is presented that is based on the cutting plane method. Then, the problem of designing a two-echelon assemble-to-order supply chain comprising of plants and DCs serving a set of customers is considered. A Lagrangean heuristic is proposed that exploits the echelon structure of the problem and uses the solution methodology for the make-to-order problem. Computational results and managerial insights are provided. It is empirically shown that substantial reduction in response times can be achieved with minimal increase in total costs in the design of responsive supply chains. Furthermore, a supply chain configuration that considers congestion is proposed and its effect on the response time can be very different from the traditional configuration that ignores congestion.

Keywords: Response time, make to order, assemble to order, supply chain design, cutting plane method, Lagrangean relaxation

1. Introduction

Make-to-Order (MTO) systems are successful business strategies to manage responsive supply chains that are characterized by high product variety, highly variable customer demand and short product life cycles. Because of mass customization and competition on product variety, many firms adopt an MTO strategy to offer a variety of products and deal with product proliferation. Dell's manufacturing and distribution of Personal Computers (PCs) is an excellent example of an MTO supply chain (Margretta, 1998; Dell, 2000). Dell typically offers several lines of product, with each allowing at least dozens of "features" from which customers can select when placing an order—different combinations of CPU, hard drive, memory and other peripherals. In Dell's supply chain, multiple components are procured and kept in inventory at various assembly facilities, from which they are assembled into a wide variety of finished products in response to customer orders. Whereas each of these components takes a substantial lead time to manufacture, the time to assemble all these components into a PC is low, provided there is sufficient assembly capacity and the components are available. In traditional Make-To-Stock (MTS) supply chains, the customer orders are met from stocks of an inventory of finished products that are kept at various points of the network. This is done to reduce the delay in fulfilling customer orders, increase sales and avoid stockouts. However, the problems associated with holding inventory of finished products may outweigh the benefits, especially when those products become obsolete as technology advances or fashion changes. While an MTO strategy eliminates finished goods inventories and reduces a firm's exposure to the risk of obsolescence, it usually spells long customer response time (Gupta and Benjaafar, 2004).

*Corresponding author

In order to reconcile the dual needs of a quick response time and high product variety, many firms such as General Electric, American Standard, Compaq, IBM, BMW and National Bicycle use a hybrid strategy (i.e., mix of MTO and MTS) called the Assemble-To-Order (ATO) strategy, in which a subassembly, or a number of common subassemblies used in several products, are assembled and placed in inventory until an order is received for the finished product (Song and Zipkin, 2003). This allows the firm to customize the orders by having the product ready using the MTO strategy, while taking advantage of the economies of scale using the MTS strategy. Also, investment in the semi-finished product inventory is smaller compared to the option of maintaining a similar amount of finished goods inventory. Furthermore, demand pooling benefits can be realized. Although, maintaining a semi-finished product inventory in ATO systems lowers the customer response time as compared to a pure MTO system, it can be further reduced by minimizing congestion at the point of differentiation. Naturally, the response time to deliver the product is critical and forms the basis for competition. Consumers' willingness to pay a premium for a shorter response time provides further incentives for firms to reduce response time in MTO and ATO supply chains.

Although various integrated models of supply chain design have been proposed in recent years to support lead time reduction, these models have continued to be largely guided by more traditional concerns of efficiency and cost in MTS settings, where the primary focus is on minimizing the fixed cost of facility location and the variable transportation cost under fairly stable and deterministic customer demand settings. This approach is personified by the work of Dogan and Goetschalckx (1999), Vidal and Goetschalckx (2000), Teo and Shu (2004), Shen (2005), Eskigun *et al.* (2005) and Elhedhli and Gzara (2008). For example, 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 that the demand is deterministic. 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. Their solutions prescribe locating facilities whose capacity utilization is very high, resulting in an excessively long response time when subjected to variability in service times and randomness in customer orders. Reviews by Vidal and Goetschalckx (1997), Erengüç *et al.* (1999) and Sarmiento and Nagi (1999) also point out that most of the existing supply chain design models do not consider measures of customer service such as response time in making location/allocation decisions. Also, refer to the recent review by Klose and Drexl (2005). This is not surprising given the complexity of the model and the interplay of locational and queueing aspects of the problem. To the best of our knowledge, Huang *et al.* (2005) is one of the first to model the effect of congestion in the design of distribution networks. They model capacity using the mean and variance of the Distribution Centers (DCs) as continuous variables, whereas our model considers capacity as a set of discrete options with known means and variances. They propose solution procedures based on outer approximation and Lagrangean relaxation, and tested on small instances of the problem.

Another growing body of literature that is related to our work and accounts for congestion and its effect on response time in strategic planning is models for facility location with immobile servers, stochastic demand and congestion (such as location of emergency medical facilities, fire stations, telecommunication network design, automated teller machines or internet mirror site location). For an extensive review, refer to Berman and Krass (2002). Due to the complexity of the underlying problem, most papers in this area make very strong assumptions: (i) either the number or capacity of the facilities (or both) are assumed to be fixed; (ii) the facilities are assumed to be identical; (iii) the demand arrival process is assumed to be Poisson; and (iv) the service process is usually assumed to be exponential (see, Amiri (1997), Marianov and Serra (2002), Wang *et al.* (2003) and Elhedhli (2006) and references therein). Despite that, most of the techniques proposed to date to solve these problems, with the exception of Elhedhli (2006), are either approximate or heuristic based. Our work is also similar in spirit to models for capacity planning with congestion effects, for which only heuristic solution procedures have been reported; see Rajagopalan and Yu (2001) and references therein.

The objective of this paper is to model the effect of congestion on the response time and analyze the tradeoff among response time costs, facility location and capacity acquisition costs, and outbound transportation costs in the design of supply chain networks. More specifically, we present a model to determine the configuration of an MTO supply chain, where 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 (DCs) under stochastic customer demand settings. The DCs are modeled as spatially distributed queues with Poission arrivals and general service times to capture the dynamics of the response time. The model is formulated as a non-linear Mixed-Integer Programming (MIP) problem and is linearized using piecewise linear functions. We present a cutting plane algorithm that provides the optimal solution to the problem. Furthermore, we present a Lagrangean relaxation heuristic procedure for solving large-scale instances of such integrated models. Then, we present a model for the two-echelon ATO supply chain design problem, where a set of plants and DCs are to be established to distribute various finished products to a set of customers with stochastic demand. DCs act as assembly facilities, where semi-finished products, procured from plants are held in inventories, from which they are assembled into a wide variety of finished products in response to customer demands. We propose a Lagrangean relaxation heuristic that exploits the echelon structure of the problem and uses the solution methodology proposed above for the MTO problem. Explicit consideration of congestion effects and their impact on response time in making location, capacity and allocation decisions in supply chains distinguishes this work from most other supply chain design models.

The rest of the paper is organized as follows. Section 2 provides a non-linear MIP formulation of the MTO supply chain design problem, a piecewise linearization and an exact solution approach based on the cutting plane method. The simplifications resulting from assuming exponentially distributed service times ($M/M/1$ case) and deterministic service times ($M/D/1$ case) are also explicitly described. In Section 3, we present the formulation of the two-echelon ATO supply chain design problem and a Lagrangean heuristic. Computational results and managerial insights are reported in Section 4. Finally, Section 5 concludes with some directions for future research.

2. MTO supply chain design

Consider the problem of designing an MTO supply chain, where a set of DCs are to be established and equipped with sufficient capacity to serve a set of customers. Sufficient capacity here implies being able to obtain service without waiting for an excessively long time after the order is placed. The DCs maintain inventory of multiple components and facilitate the assembly and shipment of a wide variety of finished products in a timely fashion without carrying expensive finished-goods inventory and incurring a long response time. Response time refers to the interval between the placing of an order and receipt of the ordered product. In MTO supply chains, because a customer order triggers the assembly of finished product from components, the response time consists of the assembly lead time and the delivery lead time. The delivery time between individual DCs and customers is relatively constant compared to the order fulfilment time at DCs in such settings. Moreover, it can further be reduced (using alternative transportation modes or expedited delivery services) to respond quickly to customer orders on a short-term basis. However, the assembly lead time is highly dependent on the DC capacity and the allocated workload and is difficult to change (on a short-term basis) once the DC is established.

2.1. *Model formulation*

We consider the setting depicted in Fig. 1. We assume that the demand for each product from each customer is independent and occurs according to a Poisson process. Once the demand for a product is realized at the customers' end, the order is placed at the DCs. DCs will act as assembly facil-

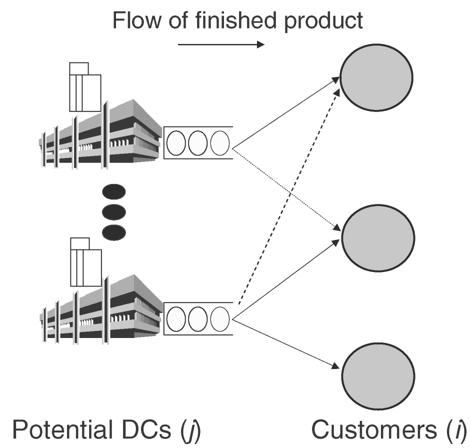

Fig. 1. An MTO supply chain network.

ities and the customers' orders arriving at the DCs are met on a First-Come First-Serve (FCFS) basis. We assume that each DC operates as a single flexible-capacity server with infinite buffers to accommodate customer orders waiting for service.

Under these assumptions, the MTO supply chain is modeled as a network of independent $M/G/1$ queues in which the DCs are treated as servers with service rates proportional to their capacity levels, where the capacity levels are discrete. We also assume that there is an unlimited supply of components and their inventory holding costs at the DCs are insignificant. Hence, the model formulated below simultaneously determines the location and capacity of DCs and the assignment of customer to DCs by minimizing the response time costs in addition to the fixed location and capacity acquisition costs, the assembly and transportation costs from DCs to customers. Besides capacity restrictions (steady-state conditions) at the DCs, and the demand requirements, there are constraints which ensure that at most one capacity level is selected at the DCs. To model this problem, we define the following notation.

Indices and parameters:

i = index for customers, $i = 1, 2, \ldots, I$;
j = index for potential DCs, $j = 1, 2, \ldots, J$;
k = index for potential capacity level at DCs, $k = 1, 2, \ldots, K$;
f_{jk} = fixed cost of opening DC j and acquiring capacity level k (\$/period);
c_{ij} = unit cost of serving customer i from DC j (\$/unit);
t = mean response time cost per unit time per customer (\$/period/customer);
λ_i = mean demand rate for the product from customer i (units/period);
μ_{jk} = mean *service rate* at DC j, if it is allocated capacity level k (units/period);
σ_{jk}^2 = variance of *service times* at DC j, if it is allocated capacity level k.

Decision variables:

x_{ij} = fraction of customer i's demand served by DC j ($0 \leq x_{ij} \leq 1$);

$$y_{jk} = \begin{cases} 1, & \text{if DC } j \text{ is opened and capacity level } k \text{ is acquired} \\ 0, & \text{otherwise.} \end{cases}$$

Let the demand for the product at customer location i be an independent random variable that follows a Poisson process with mean λ_i. If x_{ij} is the fraction of customer i's demand served by DC j, then the aggregate demand arrival rate at DC j is also a random variable that follows a Poisson process with mean $\lambda_j = \sum_{i=1}^{I} \lambda_i x_{ij}$, due to the superposition of Poisson processes. If the service times at each DC follow a general distribution and each DC is modeled as an $M/G/1$ queue, then the mean service rate of DC j, if it is allocated capacity level k, is given by $\mu_j = \sum_{k=1}^{K} \mu_{jk} y_{jk}$ and the variance in service times is $\sigma_j^2 = \sum_{k=1}^{K} \sigma_{jk}^2 y_{jk}$. This service rate reflects the server capacity or essentially the number of MTO products a DC can assemble and ship in a given time period. Let τ_j represent the mean service time at DC j ($\tau_j = 1/\mu_j$), ρ_j be the utilization of DC j ($\rho_j = \lambda_j/\mu_j$) and CV_j^2 be the squared coefficient of variation of service times ($CV_j^2 = \sigma_j^2/\tau_j^2$). Under steady-state conditions ($\lambda_j < \mu_j$) and FCFS queuing discipline, the expected *average* waiting time (including the service time) at DC j is given by the Pollaczek–Khintchine (PK) formula:

$$\begin{aligned} E[W_j(M/G/1)] &= \left(\frac{1 + CV_j^2}{2}\right) \frac{\tau_j \rho_j}{1 - \rho_j} + \tau_j \\ &= \left(\frac{1 + CV_j^2}{2}\right) \frac{\lambda_j}{\mu_j(\mu_j - \lambda_j)} + \frac{1}{\mu_j} \quad \forall j, \end{aligned}$$

and the expected *total* waiting time for DC j is obtained by multiplying the waiting time at DC j by the expected demand as

$$\left(\frac{1 + CV_j^2}{2}\right) \frac{\lambda_j^2}{\mu_j(\mu_j - \lambda_j)} + \frac{\lambda_j}{\mu_j} \qquad \forall j.$$

The expected *total* waiting time for the entire system is given by

$$E[W(M/G/1)] = \sum_{j=1}^{J} \left[\left(\frac{1 + CV_j^2}{2}\right) \frac{\lambda_j^2}{\mu_j(\mu_j - \lambda_j)} + \frac{\lambda_j}{\mu_j} \right],$$

and can be written as:

$$\frac{1}{2} \sum_{j=1}^{J} \left[\left(1 + CV_j^2\right) \frac{\lambda_j}{\mu_j - \lambda_j} + \left(1 - CV_j^2\right) \frac{\lambda_j}{\mu_j} \right].$$

This is equivalent to

$$\begin{aligned} \frac{1}{2} \sum_{j=1}^{J} \Bigg[&\left(1 + \sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right) \frac{\sum_{i=1}^{I} \lambda_i x_{ij}}{\sum_{k=1}^{K} \mu_{jk} y_{jk} - \sum_{i=1}^{I} \lambda_i x_{ij}} \\ &+ \left(1 - \sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right) \frac{\sum_{i=1}^{I} \lambda_i x_{ij}}{\sum_{k=1}^{K} \mu_{jk} y_{jk}} \Bigg]. \end{aligned} \tag{1}$$

The resulting non-linear MIP formulation is

$$(\mathrm{P_N}) : \min \sum_{j=1}^{J} \sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{i=1}^{I} \sum_{j=1}^{J} c_{ij} \lambda_i x_{ij} + t E[W(M/G/1)], \tag{2}$$

subject to

$$\sum_{i=1}^{I} \lambda_i x_{ij} \leq \sum_{k=1}^{K} \mu_{jk} y_{jk} \qquad \forall j, \tag{3}$$

$$\sum_{k=1}^{K} y_{jk} \leq 1 \qquad \forall j, \tag{4}$$

$$\sum_{j=1}^{J} x_{ij} = 1 \qquad \forall i, \tag{5}$$

$$0 \leq x_{ij} \leq 1, \quad y_{jk} \in \{0, 1\} \qquad \forall i, j, k. \tag{6}$$

The first term in the objective function (2) represents the fixed cost (amortized over the planning period) of locating DCs and equipping them with adequate assembly capacity. The second term accounts for the variable cost of assembly and shipment of products from DCs to customers. The third term is the *expected total response time cost* or the lost sales due to excessive response times between DCs and customers. The expected total response time cost is expressed as a product of average response time cost per unit time that one of its customers spends in the system and expected total waiting time in the system. In this paper, for simplicity and tractability reasons, we assume that the cost of the response time is linearly proportional to the waiting time. However, the model can be extended to other cost functions such as a piecewise linear function or a cost function proportional to $\max\{0, E[W(M/G/1)] - W_0\}$, where W_0 is the maximum tolerated waiting time for a customer. Moreover, the average response time cost t may vary from customer to customer (t_{ij}) if desired, but we assume for simplicity that it is the same across customers. It can be interpreted as a penalty function that reflects the true cost of not fulfilling customer orders in the committed lead time. In practice, determining the values of average response time cost can be challenging, however, one can rely on techniques outlined in Rao *et al.* (2000). To accurately reflect lost sales due to unacceptable response time, Rao *et al.* (2000) surveyed Caterpillar dealers to determine the percent of customers who would renege if a product was not available immediately, after 2 weeks, and after 4 weeks. A lower bound on the response time cost can be provided by the inventory holding costs (Eskigun *et al.*, 2005). Furthermore, the problems can be solved iteratively with different values of t to obtain a trade-off curve from which decision makers may choose a solution based on their preference between location and capacity acquisition cost, transportation cost, and response time costs.

Constraints (3) ensure that the steady-state conditions ($\lambda_j \leq \mu_j$) at the DCs are met. Constraint set (4) ensures that at most one capacity level is selected at a DC, whereas constraint set (5) ensures that the total demand is met.

Constraints (6) are non-negativity and binary restrictions. The formulation can easily handle single-sourcing requirements by imposing binary restrictions on x_{ij}. This would restrict the assignment of customer i's demand to one and only one DC j.

The non-linearity in (P_N) arises due to the expression of the total waiting time at the DCs, $E[W(M/G/1)]$. It can be shown that $E[W(M/G/1)]$ is convex in aggregate arrival rate λ_j, for a fixed value of μ_j and convex in service rate μ_j, for a fixed value of λ_j, where $\lambda_j = \sum_{i=1}^{I} \lambda_i x_{ij}$ and $\mu_j = \sum_{k=1}^{K} \mu_{jk} y_{jk}$. Intuitively, one would expect that the waiting time increases with increasing marginal returns as the arrival rate increases and decreases with decreasing marginal returns as the service rate increases. In the next section, we deal with the non-linearity due to the expression of the total average waiting time using a linearization based on a simple transformation and a piecewise linear approximation. We also present an exact solution procedure based on the cutting plane algorithm. Our cutting plane algorithm is similar to the outer-approximation algorithm (Duran and Grossman, 1986).

2.2. *Linearization and cutting plane method*

In order to linearize Equation (1), let us define non-negative auxiliary variables R_j, such that:

$$R_j = \frac{\lambda_j}{\lambda_j - \mu_j} = \frac{\sum_{i=1}^{I} \lambda_i x_{ij}}{\sum_{k=1}^{K} \mu_{jk} y_{jk} - \sum_{i=1}^{I} \lambda_i x_{ij}} \qquad \forall j$$

which implies that

$$\sum_{i=1}^{I} \lambda_i x_{ij} = \frac{R_j}{1+R_j} \sum_{k=1}^{K} \mu_{jk} y_{jk} = \rho_j \sum_{k=1}^{K} \mu_{jk} y_{jk} = \sum_{k=1}^{K} \mu_{jk} z_{jk} \tag{7}$$

where

$$z_{jk} = \begin{cases} 0 & \text{if } y_{jk} = 0 \\ \rho_j & \text{if } y_{jk} = 1 \end{cases} \qquad \forall j, k \tag{8}$$

and $\rho_j = R_j/(1 + R_j)$ is the server (DC) utilization.

Since there is at most one k' with $y_{jk'} = 1$ while $y_{jk} = 0$ for all other $k \neq k'$, the expression $z_{jk} = \rho_j y_{jk}$ can be ensured by adding the following constraints:

$$z_{jk} \le y_{jk} \qquad \forall j, k,$$

$$\sum_{k=1}^{K} z_{jk} = \rho_j \qquad \forall j.$$

The function $\rho_j = R_j/(1 + R_j)$ is concave. Given a set of points R_j^h indexed by H, ρ_j can be approximated by an infinite set of piecewise linear functions that are tangent to ρ_j at points R_j^h (as shown in Fig. 2) that is

$$\rho_j = \min_{h \in H} \left\{ \frac{1}{\left(1+R_j^h\right)^2} R_j + \frac{\left(R_j^h\right)^2}{\left(1+R_j^h\right)^2} \right\} \qquad \forall j$$

or

$$\rho_j \le \frac{1}{\left(1+R_j^h\right)^2} R_j + \frac{\left(R_j^h\right)^2}{\left(1+R_j^h\right)^2} \qquad \forall j,\ h \in H.$$

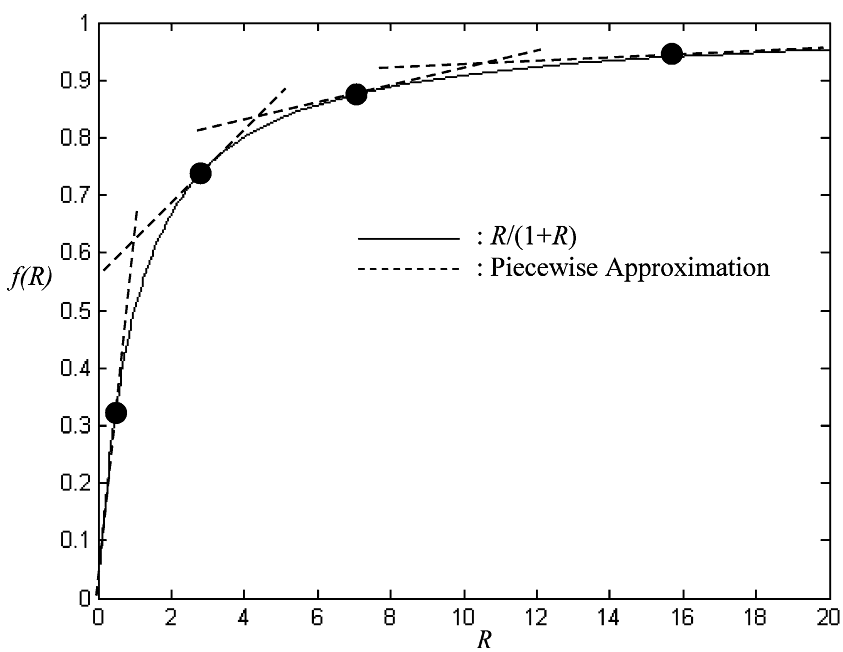

Fig. 2. A piecewise linear approximation of $R_j/(1 + R_j)$.

The expression for $E[W(M/G/1)]$ reduces to

$$E[W(M/G/1)] = \frac{1}{2}\sum_{j=1}^{J}\left\{\left(1+\sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right)R_j + \left(1-\sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right)\rho_j\right\}$$
$$= \frac{1}{2}\sum_{j=1}^{J}\left(R_j + \sum_{k=1}^{K} CV_{jk}^2 w_{jk} + \rho_j - \sum_{k=1}^{K} CV_{jk}^2 z_{jk}\right)$$

where

$$w_{jk} = \begin{cases} 0 & \text{if} \quad y_{jk}=0 \\ R_j & \text{if} \quad y_{jk}=1 \end{cases} \quad \text{and} \quad z_{jk} = \begin{cases} 0 & \text{if} \quad y_{jk}=0 \\ \rho_j & \text{if} \quad y_{jk}=1 \end{cases} \quad \forall j,k.$$

Similarly, because there exists at most one k' with $y_{jk'}=1$ while $y_{jk}=0$ for all other $k \neq k'$, the expression $w_{jk}=R_j y_{jk}$ can be ensured by adding the following constraints:

$$w_{jk} \leq M y_{jk} \qquad \forall j,k,$$
$$\sum_{k=1}^{K} w_{jk} = R_j \qquad \forall j,$$

where M is the usual Big-M.
The resulting linear MIP formulation is

$$(\mathrm{P}_{\mathrm{L}(H)}) : \min \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i x_{ij} + \frac{t}{2}\sum_{j=1}^{J}\left\{R_j + \rho_j + \sum_{k=1}^{K} CV_{jk}^2 (w_{jk} - z_{jk})\right\}, \tag{9}$$

subject to

$$\sum_{i=1}^{I}\lambda_i x_{ij} - \sum_{k=1}^{K}\mu_{jk} z_{jk} = 0 \qquad \forall j, \tag{10}$$
$$\sum_{k=1}^{K} y_{jk} \leq 1 \qquad \forall j, \tag{11}$$
$$\sum_{j=1}^{J} x_{ij} = 1 \qquad \forall i, \tag{12}$$
$$z_{jk} - y_{jk} \leq 0 \qquad \forall j,k, \tag{13}$$
$$\rho_j - \frac{1}{\left(1+R_j^h\right)^2}R_j \leq \frac{\left(R_j^h\right)^2}{\left(1+R_j^h\right)^2} \qquad \forall j, h \in H, \tag{14}$$
$$\rho_j - \sum_{k=1}^{K} z_{jk} = 0 \qquad \forall j, \tag{15}$$
$$w_{jk} - M y_{jk} \leq 0 \qquad \forall j,k, \tag{16}$$
$$\sum_{k=1}^{K} w_{jk} - R_j = 0 \qquad \forall j, \tag{17}$$
$$y_{jk} \in \{0,1\};\ 0 \leq x_{ij}, z_{jk} \leq 1;\quad \rho_j, R_j, w_{jk} \geq 0; \quad \forall i,j,k. \tag{18}$$

The steady-state conditions ($\lambda_j < \mu_j$) translate into capacity constraints, and are enforced by the constraints (10) and (13) and forced to "$<$" by the term R_j in the objective. $(\mathrm{P}_{\mathrm{L}(H)})$ is a minimization problem, at least one of the constraints in Equation (14) will be binding. This implies that

$$\rho_j = \min_{h\in H}\left[\frac{1}{\left(1+R_j^h\right)^2}R_j + \frac{\left(R_j^h\right)^2}{\left(1+R_j^h\right)^2}\right] \quad \forall j \ \text{ when } y_{jk}=1.$$

In order to deal with the infinite number of constraints (14) in the linear MIP model $(\mathrm{P}_{\mathrm{L}(H)})$, we use a cutting plane algorithm, described as follows. For an initial and finite set of points $(R_j^h)_{\bar{H}\subset H}$, $(\mathrm{P}_{\mathrm{L}(\bar{H})})$ is a relaxation of the full problem $(\mathrm{P}_{\mathrm{L}(H)})$, hence a lower bound to $(\mathrm{P}_{\mathrm{L}(H)})$ or $(\mathrm{P}_{\mathrm{N}})$ is provided by the optimal objective function value $v((\mathrm{P}_{\mathrm{L}(\bar{H})}))$, where

$$v((\mathrm{P}_{\mathrm{L}(\bar{H})})) = \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}\bar{y}_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i \bar{x}_{ij} + \frac{t}{2}\sum_{j=1}^{J}\left\{\bar{R}_j + \bar{\rho}_j + \sum_{k=1}^{K} CV_{jk}^2(\bar{w}_{jk} - \bar{z}_{jk})\right\},$$

where $(\bar{x}, \bar{y}, \bar{R}, \bar{\rho}, \bar{w}, \bar{z})$ is the solution of $(\mathrm{P}_{\mathrm{L}(\bar{H})})$. Furthermore, $(\bar{x}, \bar{y})$ is feasible to $(\mathrm{P}_{\mathrm{N}})$ and hence:

$$\sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}\bar{y}_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i\bar{x}_{ij} + \frac{t}{2}\sum_{j=1}^{J}\left[\left(1+\sum_{k=1}^{K} CV_{jk}^2\bar{y}_{jk}\right)\frac{\sum_{i=1}^{I}\lambda_i\bar{x}_{ij}}{\sum_{k=1}^{K}\mu_{jk}\bar{y}_{jk} - \sum_{i=1}^{I}\lambda_i\bar{x}_{ij}} + \left(1-\sum_{k=1}^{K} CV_{jk}^2\bar{y}_{jk}\right)\frac{\sum_{i=1}^{I}\lambda_i\bar{x}_{ij}}{\sum_{k=1}^{K}\mu_{jk}\bar{y}_{jk}}\right]$$

provides an upper bound to $(\mathrm{P}_{\mathrm{L}(\bar{H})})$ and $(\mathrm{P}_{\mathrm{N}})$. If the best known upper bound coincides with the lower bound at a given iteration, then the optimal solution is obtained and the algorithm is terminated. If not, a new set of cuts (14) are generated using $(\bar{R}_j)$ and appended to $(\mathrm{P}_{\mathrm{L}(\bar{H})})$ and the procedure is repeated. The computational performance of the algorithm is reported in Section 4.

2.3. *Special cases*

In this section, we examine two special cases that are commonly looked at in the literature.

2.3.1. *Systems with exponential service times (M/M/1 case)*

The exponential processing and assembly time is a reasonable assumption in cases where there is high variability in setup times and processing times, e.g., in semiconductor wafer fabrication (Kim and Tang, 1997). Also, this is more reasonable than deterministic processing and assembly times for MTO products with very high product variety and varying batch sizes. For exponentially distributed service times at the DCs, the total expected waiting time for the entire system is given by

$$E[W(M/M/1)] = \sum_{j=1}^{J} \frac{\lambda_j}{\mu_j - \lambda_j} = \sum_{j=1}^{J} \left(\frac{\sum_{i=1}^{I} \lambda_i x_{ij}}{\sum_{k=1}^{K} \mu_{jk} y_{jk} - \sum_{i=1}^{I} \lambda_i x_{ij}} \right) = \sum_{j=1}^{J} R_j.$$

The resulting linear MIP model is

$$\left(\mathrm{P}_{\mathrm{L}(H)}^{M/M/1}\right): \min \sum_{j=1}^{J} \sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{i=1}^{I} \sum_{j=1}^{J} c_{ij} \lambda_i x_{ij} + t \sum_{j=1}^{J} R_j \tag{19}$$

subject to Equations (10) to (15),

$$y_{jk} \in \{0, 1\}; \quad 0 \le x_{ij}, z_{jk} \le 1; \quad \rho_j, R_j \ge 0 \quad \forall i, j, k.$$

The model is structurally identical to the service system design model presented in Amiri (1997) and Elhedhli (2006).

2.3.2. *Systems with deterministic service times (M/D/1 case)*

In many cases, the processing/assembly of finished products at the DCs often involves repeated steps without much variation (Kim and Tang, 1997). This is particularly true for MTO products with limited options and batch size of one such as Dell's PCs. For deterministic service times, the expected waiting time for the entire system is given by

$$E[W(M/D/1)] = \frac{1}{2} \sum_{j=1}^{J} \left(\frac{\lambda_j}{\mu_j - \lambda_j} + \frac{\lambda_j}{\mu_j} \right) = \frac{1}{2} \sum_{j=1}^{J} (R_j + \rho_j).$$

The resulting linear MIP model is as follows:

$$\left(\mathrm{P}_{\mathrm{L}(H)}^{M/D/1}\right): \min \sum_{j=1}^{J} \sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{i=1}^{I} \sum_{j=1}^{J} c_{ij} \lambda_i x_{ij} + \frac{t}{2} \sum_{j=1}^{J} (R_j + \rho_j), \tag{20}$$

subject to Equations (10) to (15),

$$y_{jk} \in \{0, 1\}; \quad 0 \le x_{ij}, z_{jk} \le 1; \quad \rho_j, R_j \ge 0 \quad \forall i, j, k.$$

The cutting plane algorithm described above can be used to solve these models to optimality. Alternatively, one can rely on the Lagrangean heuristics. Some computational results are provided in Section 4.

3. ATO supply chain design

In this section, we consider the design of an ATO supply chain, where we seek to locate a set of plants and DCs to distribute a product with a non-trivial bill of materials to a set of customers with stochastic demand. The DCs will act as intermediate facilities between the plants and the customers and facilitate the shipment of products between the two echelons, as shown in Fig. 3.

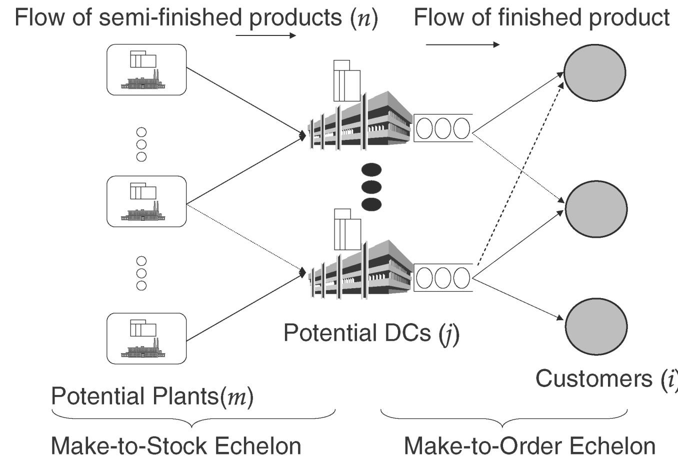

Fig. 3. An ATO supply chain network.

The semi-finished products are produced at the plants and shipped to the DCs. Once the demand is realized at the customers' end, the order is placed to the DC and the final product is assembled and the demand is met. Hence, the supply chain network is a combination of the MTS echelon (plant-DC echelon) and the MTO echelon (DC-customer echelon). The problem environment is characterized by a stochastic customer demand that has to be satisfied from a set of DCs and where sufficient capacity has to be acquired in order to avoid long response times. To model this problem, we define the following additional notation.

m = index for potential plants, $m = 1, 2, \ldots, M$;
n = index for semi-finished products, $n = 1, 2, \ldots, N$;
g_m = fixed cost of opening a plant at location m (\$/period);
P_m = maximum available capacity of plant m (units);
c'_{jmn} = unit production and transportation cost for semi-finished product n from plant m to DC j;
η_n = number of units of semi-finished product n required to make one unit of finished product;
u_m = decision variable that equals one, if plant m is opened; zero, otherwise;
v_{jmn} = number of units of semi-finished product n produced from plant m and shipped to DC j.

Under the assumption that the demand at customer i is an independent random variable that follows a Poisson process with mean λ_i and the service time at each DC follows a general distribution, each DC is modeled as an $M/G/1$ queue, whose mean service rate, if it is allocated capacity level k, is given by $\mu_j = \sum_{k=1}^{K} \mu_{jk} y_{jk}$ and the variance in service times is given by $\sigma_j^2 = \sum_{k=1}^{K} \sigma_{jk}^2 y_{jk}$. Under steady-state conditions ($\lambda_j < \mu_j$) and FCFS queuing discipline, the *total average waiting time* for the entire system (service plus queuing time) is given by Equation (1). The resulting non-linear MIP formulation that simultaneously determines the location and capacity of plants and DCs, the shipment levels from plants to DCs, and allocation of customers to DCs by minimizing response time costs in addition to fixed facility location and capacity acquisition costs, production and transportation costs between echelons is as follows:

$$\begin{aligned}(\mathrm{P_{ATO}}): \min \sum_{m=1}^{M} g_m u_m + \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N} c'_{jmn} v_{jmn} \\ + \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i x_{ij} + \frac{t}{2}\sum_{j=1}^{J}\left(1+\sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right) \\ \times \frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk} - \sum_{i=1}^{I}\lambda_i x_{ij}} \\ + \frac{t}{2}\sum_{j=1}^{J}\left(1-\sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk}}, \qquad (21)\end{aligned}$$

subject to

$$\sum_{j=1}^{J}\sum_{n=1}^{N} v_{jmn} \le P_m u_m \qquad \forall m, \qquad (22)$$

$$\sum_{m=1}^{M} v_{jmn} = \sum_{i=1}^{I} \eta_n \lambda_i x_{ij} \qquad \forall j, n, \qquad (23)$$

$$\sum_{i=1}^{I} \lambda_i x_{ij} \le \sum_{k=1}^{K} \mu_{jk} y_{jk} \qquad \forall j, \qquad (24)$$

$$\sum_{k=1}^{K} y_{jk} \le 1 \qquad \forall j, \qquad (25)$$

$$\sum_{j=1}^{J} x_{ij} = 1 \qquad \forall i, \qquad (26)$$

$$0 \le x_{ij} \le 1, \quad y_{jk}, u_m \in \{0,1\}, \quad v_{jmn} \ge 0 \quad \forall i,j,k,m,n. \qquad (27)$$

The objective function (21) consists of the fixed cost of opening plants, the fixed cost of locating DCs and equipping DCs with the required capacity level, the variable cost of producing and procuring semi-finished product, the variable cost of serving customers from DCs and the total waiting time costs at the DCs. Constraints (22) are capacity restrictions on the opened plants and permit the use of opened plants only. Constraints (23) are commodity flow conservation equations at the DCs. Constraints (24) ensure that the steady-state conditions at the DCs are met. Constraints (25) ensure that at most one capacity level is selected at a DC whereas constraints (26) ensure that the total demand is met. Constraints (27) are non-negativity and binary constraints.

3.1. *Lagrangean relaxation*

There are number of ways in which the model can be relaxed in Lagrangean fashion (see Klose and Drexl (2005) and reference therein). In this paper, we exploit the echelon structure of the ATO supply chain using Lagrangean relaxation to decompose the model into two subproblems. Note that in ($\mathrm{P_{ATO}}$), constraints (22) relate to the MTS echelon, and constraints (24) to (26) relate to the MTO echelon, whereas constraints (23) are the flow conservation constraints that link the two echelons. Upon relaxing the flow conservation constraints (23) with dual multipliers β_{jn}, the problem decomposes into two subproblems:

$$(\mathrm{SP_{MTS}}): \min \sum_{m=1}^{M} g_m u_m + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N}(c'_{jmn} - \beta_{jn}) v_{jmn}, \qquad (28)$$

subject to

$$\sum_{j=1}^{J}\sum_{n=1}^{N} v_{jmn} \le P_m u_m \qquad \forall m, \qquad (29)$$

$$u_m \in \{0,1\}, \quad v_{jmn} \ge 0 \qquad \forall j, m, n. \qquad (30)$$

$$\begin{aligned}(\mathrm{SP_{MTO}}): \min \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{N}(c_{ij} + \eta_n\beta_{jn})\lambda_i x_{ij} \\ + \frac{t}{2}\sum_{j=1}^{J}\left(1+\sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk} - \sum_{i=1}^{I}\lambda_i x_{ij}} \\ + \frac{t}{2}\sum_{j=1}^{J}\left(1-\sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk}}, \qquad (31)\end{aligned}$$

subject to

$$\sum_{i=1}^{I} \lambda_i x_{ij} \le \sum_{k=1}^{K} \mu_{jk} y_{jk}, \qquad (32)$$

$$\sum_{j=1}^{J} x_{ij} = 1 \qquad \forall i, \qquad (33)$$

$$\sum_{k=1}^{K} y_{jk} \le 1 \qquad \forall j, \qquad (34)$$

$$0 \le x_{ij} \le 1, \quad y_{jk} \in \{0,1\} \qquad \forall i,j,k. \qquad (35)$$

Subproblem ($\mathrm{SP_{MTS}}$) is a linear MIP model that determines the location of plants and the flow of semi-finished products into the DCs, whereas the subproblem ($\mathrm{SP_{MTO}}$) is a non-linear MIP model that provides the location and capacity level of DCs and the allocation of customers to DCs. Note that the subproblem ($\mathrm{SP_{MTO}}$) is the MTO supply chain design model presented in the previous section and hence we use the proposed cutting plane algorithm to solve it.

From model ($\mathrm{P_{ATO}}$), we can derive some valid constraints. For example, consider the following set of constraints:

$$\sum_{m=1}^{M} P_m u_m \geq \sum_{i=1}^{I} \lambda_i, \tag{36}$$

$$\sum_{m=1}^{M} v_{jmn} \leq \eta_n\big(\max_k \mu_{jk}\big) \qquad \forall j, n, \tag{37}$$

$$\sum_{j=1}^{J}\sum_{m=1}^{M} v_{jmn} \geq \eta_n \sum_{i=1}^{I} \lambda_i \qquad \forall n. \tag{38}$$

Constraints (36) are aggregate capacity constraints for the MTS echelon. Constraints (37) are derived from Equations (23) and (24) and constraints (38) follow from Equations (23) and (26). Constraints (37) imply that the total flow of semi-finished products through a DC should not exceed the DC's maximum throughput capacity, whereas constraints (38) ensure that the flow of every semi-finished product from plants to DC is at least equal to the bill of materials multiplied by the demand of that product from all the customers. These constraints are redundant in the original MIP formulation, but they improve the quality of the subproblem solutions in terms of the feasibility to the original problem upon relaxing the flow conservation constraints. This results in better heuristic solutions. Therefore, we add these set of constraints to ($\mathrm{SP_{MTS}}$) as follows:

$$(\mathrm{SP_{MTS}}): \min \sum_{m=1}^{M} g_m u_m + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N}(c'_{jmn} - \beta_{jn}) v_{jmn}, \tag{39}$$

subject to

$$\sum_{j=1}^{J}\sum_{n=1}^{N} v_{jmn} \leq P_m u_m \qquad \forall m, \tag{40}$$

$$\sum_{m=1}^{M} P_m u_m \geq \sum_{i=1}^{I} \lambda_i, \tag{41}$$

$$\sum_{m=1}^{M} v_{jmn} \leq \eta_n\big(\max_k \mu_{jk}\big) \qquad \forall j, n, \tag{42}$$

$$\sum_{j=1}^{J}\sum_{m=1}^{M} v_{jmn} \geq \eta_n \sum_{i=1}^{I} \lambda_i \qquad \forall n, \tag{43}$$

$$u_m \in \{0, 1\}, \quad v_{jmn} \geq 0 \qquad \forall j, m, n. \tag{44}$$

3.1.1. *The lower bound*

The Lagrangean lower bound is given by the solution of the Lagrangean dual problem, $\max_{\beta}[v(\mathrm{SP_{MTS}}) + v(\mathrm{SP_{MTO}})]$ which is equivalent to

$$\max_{\beta}\Bigg\{\min_{h\in I_{u,v}} \sum_{m=1}^{M} g_m u_m^h + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N}(c'_{jmn} - \beta_{jn}) v_{jmn}^h$$
$$+ \min_{h\in I_{x,y}} \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk}^h + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{N}(c_{ij} + \eta_n\beta_{jn})\lambda_i x_{ij}^h$$
$$+ \frac{t}{2}\sum_{j=1}^{J}\left(1 + \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^h\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}^h}{\sum_{k=1}^{K}\mu_{jk} y_{jk}^h - \sum_{i=1}^{I}\lambda_i x_{ij}^h}$$
$$+ \frac{t}{2}\sum_{j=1}^{J}\left(1 - \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^h\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk} y_{jk}}\Bigg\}.$$

Thus can be explicitly written as

$$(\mathrm{MP}): \max_{\beta}\ \theta_1 + \theta_2,$$

subject to

$$\theta_1 + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N} v_{jmn}^h \beta_{jn} \leq \sum_{m=1}^{M} g_m u_m^h + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N} c'_{jmn} v_{jmn}^h \quad h \in I_{u,v},$$

$$\theta_2 - \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{N}\big(\eta_n\lambda_i x_{ij}^h\big)\beta_{jn} \leq \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk}^h + \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i x_{ij}^h$$
$$+\frac{t}{2}\sum_{j=1}^{J}\left(1 + \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^h\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}^h}{\sum_{k=1}^{K}\mu_{jk} y_{jk}^h - \sum_{i=1}^{I}\lambda_i x_{ij}^h}$$
$$+\frac{t}{2}\sum_{j=1}^{J}\left(1 - \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^h\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk} y_{jk}} \quad h \in I_{x,y},$$

where $I_{u,v}$ is the index set of feasible points of the set:

$\{(u_m, v_{jmn})$: Equations (40) to (43); $u_m \in \{0, 1\}; v_{jmn} \geq 0, \forall j, m, n\}$,

and $I_{x,y}$ is the index set of feasible points of the set:

$\{(x_{ij}, y_{jk})$: Equations (32) to (34); $x_{ij} \geq 0; y_{jk} \in \{0, 1\}, \forall i, j, k\}$.

We use Kelley's cutting plane method, in which the point $\bar{\beta}$ is the solution of the relaxed master problem (RMP), defined on subsets $\bar{I}_{u,v} \subset I_{u,v}$ and $\bar{I}_{x,y} \subset I_{x,y}$ (Kelley, 1960). This $\bar{\beta}$ from (RMP) is used to solve the subproblems ($\mathrm{SP_{MTS}}$) and ($\mathrm{SP_{MTO}}$), and generate two cuts of the form:

$$\theta_1 + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N} v_{jmn}^{\bar{i}}\beta_{jn} \leq \sum_{m=1}^{M} g_m u_m^{\bar{i}} + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N} c'_{jmn} v_{jmn}^{\bar{i}}, \tag{45}$$

$$\theta_2 - \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{N}\big(\eta_n\lambda_i x_{ij}^{\bar{i}'}\big)\beta_{jn} \leq \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk}^{\bar{i}'}$$
$$+ \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i x_{ij}^{\bar{i}'} + \frac{t}{2}\sum_{j=1}^{J}\left(1 + \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^{\bar{i}'}\right)$$
$$\times \frac{\sum_{i=1}^{I}\lambda_i x_{ij}^{\bar{i}'}}{\sum_{k=1}^{K}\mu_{jk} y_{jk}^{\bar{i}'} - \sum_{i=1}^{I}\lambda_i x_{ij}^{\bar{i}'}} + \frac{t}{2}\sum_{j=1}^{J}\left(1 - \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^{\bar{i}'}\right)$$
$$\times \frac{\sum_{i=1}^{I}\lambda_i x_{ij}^{\bar{i}'}}{\sum_{k=1}^{K}\mu_{jk} y_{jk}^{\bar{i}'}}. \tag{46}$$

The index sets $\bar{I}_{u,v}$ and $\bar{I}_{x,y}$ are updated as $\bar{I}_{u,v} \cup \{\bar{i}\}$ and $\bar{I}_{x,y} \cup \{\bar{i}'\}$, respectively, as the algorithm proceeds through the iterations.

3.1.2. *The heuristic: finding a feasible solution*

The first subproblem (SP_{MTS}) provides the location of plants (u_m) and the flow of semi-finished products into the DCs (v_{jmn}), whereas the second subproblem (SP_{MTO}) provides the location and the capacity decisions of the DCs (y_{jk}), the assignment of customers to DCs (x_{ij}). Note that the link between the two subproblems is the flow balance of products in and out of DCs. Hence, a feasible solution to problem (P_{ATO}) can be constructed by solving (SP_{MTS}) with the additional set of constraints $\sum_{m=1}^{M} v_{jmn} = \sum_{i=1}^{I} \eta_n \lambda_i \overline{x}_{ij}$, where $\overline{x}_{ij}$ is obtained from the solution of (SP_{MTO}). The overall procedure is shown in Fig. 4. Computational results are provided next.

4. Computational results and insights

In this section, we report our computational experiences with the proposed solution methodologies and present some insights. All the proposed solution procedures were coded in C and the MIP problems were solved using ILOG CPLEX 10.1 (using the Callable Library) on a Sun Blade 2500 workstation with 1.6-GHz UltraSPARC IIIi processors. In the implementation of the iterative cutting plane algorithm and after the solution of the relaxed MIP, we use the procedure `CPXaddrows()` to append the cuts generated and exploit warm starting.

4.1. *Test problems*

The test problems are derived from the 2000 census data consisting of 150 largest cities in the continental United States (see Daskin (2004)). We generate nine sets of test problems by setting the number of customers (I) to the 50, 100 and 150 largest cities, and the potential DC locations (J) to the five, ten and 20 most populated cities. The mean customer demand rates λ_i are obtained by dividing the population of those cities by 10^3. The unit transportation costs c_{ij} are obtained by dividing the great-circle distance between the customer i and the potential DC location j by 100. The service rate of DC j equipped with capacity level k, is set to $\mu_{jk} = \beta_k \sum_i \lambda_i$ (where $\beta_k = 0.15, 0.20, 0.45$ for $I = 50$; $\beta_k = 0.10, 0.20, 0.30$ for $I = 100$; $\beta_k = 0.10, 0.15, 0.20, 0.30, 0.45$ for $I = 150$). The fixed costs, f_{jk} are set to $100 \times \sqrt{\mu}_{jk}$ to

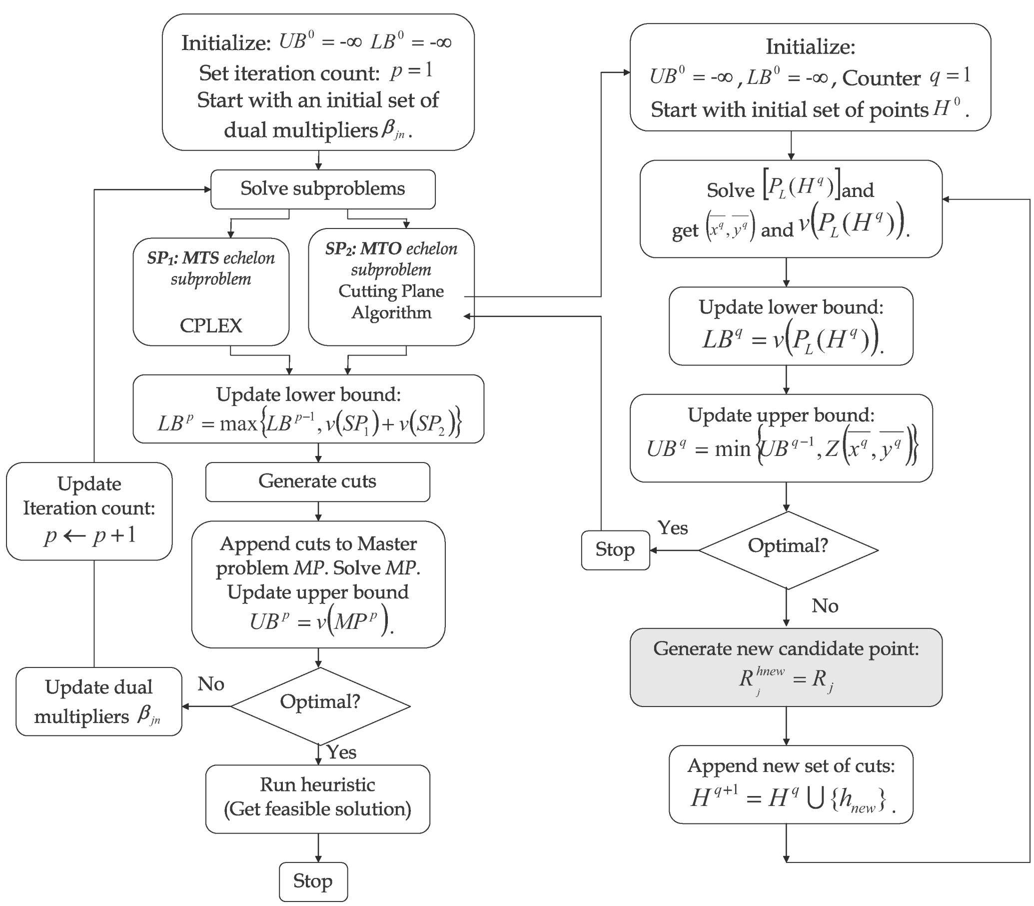

Fig. 4. Solution procedure for two-echelon ATO supply chain design.

reflect economies of scale. For the $M/M/1$ case, the variance of service times σ_{jk}^2 are set to the mean service rates, μ_{jk} whereas for the $M/G/1$, the σ_{jk}^2 is obtained by setting coefficient of variation (CV) to 1.5. The average response time cost t is set to $\theta \times (\sum_i \sum_j (\lambda_i c_{ij})/(I \times J))$, where θ is the response time cost multiplier and $\lambda_i c_{ij}$ denotes the total production and transportation cost associated with the order from the ith customer served by the jth DC. In order to explore the sensitivity of the solution to different levels of response time costs, the multiplier θ is tested for 0.1, 1, 5, 10, 50, 100 and 200 (higher value of θ models the situation in which losing a customer order due to high expected waiting time is extremely costly). For the ATO supply chains, the instances are generated by setting the capacities of plants to: $P_m = U[0.1, 0.5] \times \sum_{i=1}^{I} \lambda_i$ whereas their fixed cost are set to: $g_m = U[1000, 2000] \times \sqrt{(P_m)}$. In order to compare the performance of the cutting plane method and the Lagrangean heuristic for different values of the ratio of plant capacities to total demand (r), the instances are generated by setting the capacities of plants to: $P_m = (r \times \sum_{i=1}^{I} \lambda_i)/M$ and the fixed costs to: $g_m = U[100, 110] \times \sqrt{(P_m)}$. The production coefficients (bill of materials) η_n were randomly generated in the range $U[1, 5]$ and rounded up to the nearest integer value.

4.2. *An illustrative case study*

Using the first set of test problems (where $I = 50$, $J = 5$ and $K = 3$), we illustrate that the MTO supply chain configuration that considers congestion and its effect on response time can be different from the configuration that ignores congestion. Furthermore, we show empirically that substantial reduction in response times can be achieved with minimal increase in total costs in the design of responsive supply chains.

The first set of test problems consists of 50 customers, five potential DC locations ($j = 1$, New York, NY; $j = 2$, Los Angeles, CA; $j = 3$, Chicago, IL; $j = 4$, Houston, TX; $j = 5$, Philadelphia, PA) that can be equipped with three capacity levels ($k = 1$, small; $k = 2$, medium; $k = 3$, large). The problem is solved to optimality (with a gap of 10^{-6}) using the cutting plane algorithm. Table 1 summarizes the results for different values of the response time cost by setting θ to 0, 0.1, 1, 10, 100 and 1000 under four different cases: $M/G/1$ case with $CV = 1.5$, $M/M/1$ case, $M/G/1$ case with $CV = 0.5$, and $M/D/1$ case. The table shows the total objective function value (TC), fixed cost (FC), variable cost (VC), total response time cost (RC), total expected waiting time ($E(W)$), average DC utilization ($\bar{\rho}$), DCs opened and their capacity levels (Open DCs (y_{jk})), expected waiting time at the DCs (W_j), DC utilization (ρ_j), the number of cuts generated (CUT), the number of iterations required (ITR) and the CPU time in seconds (CPU(s)) under different scenarios. The supply chain network configuration for two extreme values of response time cost ($\theta = 0$ versus $\theta = 1000$) are shown in Fig. 5. Figure 6(a) shows the effect of changing response time cost on the total expected waiting time ($E(W)$), and Fig. 6(b) shows the effect of changing the total expected waiting time $E(W)$ on the sum of the fixed and transportation costs. The observations are as follows.

1. Figure 5 shows that the supply chain configuration that ignores congestion opens four DCs (medium size DCs in New York, Los Angeles and Chicago and a small DC in Houston) whereas the configuration that considers congestion opens five DCs, all with capacity level 3. Also, there is substantial reallocation of customer demand among the DCs in order to balance the workload to reduce the DC utilization and overall response time in the system. Hence, the supply chain configuration that considers congestion and its effect on response time can be different from the traditional configuration that ignores congestion. However, we observe that at very high values of θ, the configurations (location, capacity and demand allocations) are not significantly different among the $M/G/1$, $M/M/1$ and $M/D/1$ cases.
2. As we see from Table 1, even with very small values of average response time cost, $t = 0.1 \times \text{mean}_{i,j}(\lambda_i c_{ij})$ or $t = 1 \times \text{mean}_{i,j}(\lambda_i c_{ij})$, substantial improvement in the total expected waiting time ($E[W]$) can be achieved over $t = 0$. For example, for the $M/G/1$ case ($CV = 1.5$), $E[W]$ decreases from 77.41 units to 48.37 units for $\theta = 0.1$ and 1 respectively. This is due to the even distribution of demand among DCs. Figure 6(a) also shows that the substantial reduction in response time can be achieved with a small value of the response time costs. This is because, as we increase the magnitude of the response time cost, DCs with higher capacity are used and/or the number of DCs opened increases, average DC utilization decreases, thereby reducing congestion and improving average response time. From Fig. 6(b), we see that the left portion of the curves are quite flat, indicating that substantial improvement (decrease) in response time can be achieved with a small increase in fixed and transportation costs.
3. As the response time cost becomes dominant compared to other cost components, DCs with higher capacity levels are opened, and average DC utilization decreases, thereby improving (decreasing) the response time. For example, in Table 1, in the $M/G/1$ case ($CV = 1.5$), for $\theta = 1$, four medium size DCs are opened, whereas for $\theta = 1000$, five large size DCs are opened. The average DC utilization decreases from 0.83 to 0.44 and the expected waiting time decreases from 48.37 to 5.44.
4. As we increase the magnitude of the response time cost, the transportation cost decreases initially and then increases. For example, in Table 1, in the $M/G/1$ case ($CV = 1.5$), for $\theta = 0$, TC = 103, 577; for $\theta = 1$, TC = 101, 849; and for $\theta = 1000$, TC = 106, 029. This is due to the reallocation of customer demand among DCs in an attempt to reduce the total expected waiting time.

Table 1. Comparison of the MTO supply chain network configurations for $M/G/1$, $M/M/1$ and $M/D/1$ cases: an illustrative example

		M/G/1 (CV = 1.5)	*M/M/1 (CV = 1)*	*M/G/1 (CV = 0.5)*	*M/D/1 (CV = 0)*
$\theta = 0$	TC	146, 965	146, 965	146, 965	146, 965
	FC	43, 388	43, 388	43, 388	43, 388
	VC	103, 577	103, 577	103, 577	103, 577
	RC	0	0	0	0
	$E(W)$	∞	∞	∞	∞
	$\bar{\rho}$	0.96	0.96	0.96	0.96
	Open DCs	1(2), 2(2), 3(2), 4(1)	1(2), 2(2), 3(2), 4(1)	1(2), 2(2), 3(2), 4(1)	1(2), 2(2), 3(2), 4(1)
	W_j	[138.79, ∞, 5.27, ∞]	[138.79, ∞, 5.27, ∞]	[138.79, ∞, 5.27, ∞]	[138.79, ∞, 5.27, ∞]
	ρ_j	[0.99, 1.00, 0.84, 1.00]	[0.99, 1.00, 0.84, 1.00]	[0.99, 1.00, 0.84, 1.00]	[0.99, 1.00, 0.84, 1.00]
	CUT	0	0	0	0
	ITR	1	1	1	1
	CPU(s)	0.14	0.14	0.14	0.14
$\theta = 0.1$	TC	148, 567	148, 333	148, 042	146, 724
	FC	46, 816	43, 388	43, 388	43, 388
	VC	101, 494	104, 235	104, 093	104, 033
	RC	257	710	560	503
	$E(W)$	77.41	214.05	169.00	152.40
	$\bar{\rho}$	0.83	0.96	0.96	0.96
	Open DCs	1(2), 2(2), 3(2), 4(2)	1(2), 2(2), 3(2), 4(1)	1(2), 2(2), 3(2), 4(1)	1(2), 2(2), 3(2), 4(1)
	W_j	[10.09, 33.51, 2.49, 2.83]	[33.93, 98.34, 7.24, 74.55]	[42.56, 124.89, 6.66, 94.06]	[48.12, 139.37, 6.43, 107.05]
	ρ_j	[0.99, 0.97, 0.71, 0.74]	[0.97, 0.99, 0.88, 0.99]	[0.98, 0.99, 0.87, 0.99]	[0.98, 0.99, 0.87, 0.99]
	CUT	4	20	20	20
	ITR	2	6	6	6
	CPU(s)	0.35	0.88	1.69	0.9
$\theta = 1$	TC	150, 269	149, 683	149, 286	149, 139
	FC	46, 816	46, 816	46, 816	46, 816
	VC	101, 849	101, 744	101, 649	101, 609
	RC	1604	1,124	822	714
	$E(W)$	48.37	33.9	24.79	21.55
	$\bar{\rho}$	0.83	0.83	0.83	0.83
	Open DCs	1(2), 2(2), 3(2), 4(2)	1(2), 2(2), 3(2), 4(2)	1(2), 2(2), 3(2), 4(2)	1(2), 2(2), 3(2), 4(2)
	W_j	[10.09, 15.08, 2.49, 3.39]	[10.09, 18.11, 2.49, 3.21]	[10.09, 22.02, 2.49, 3.06]	[10.09, 24.19, 2.49, 3.00]
	ρ_j	[0.91, 0.94, 0.71, 0.77]	[0.91, 0.95, 0.71, 0.76]	[0.91, 0.96, 0.71, 0.75]	[0.91, 0.96, 0.71, 0.75]
	CUT	4	8	16	12
	ITR	2	3	5	4
	CPU(s)	0.33	0.41	0.91	0.45
$\theta = 10$	TC	157, 274	155, 674	154, 324	153, 874
	FC	52, 078	49, 447	49, 447	49, 447
	VC	101, 407	101, 640	101, 638	101, 616
	RC	3789	4,587	3,240	2811
	$E(W)$	11.43	13.84	9.77	8.48
	$\bar{\rho}$	0.67	0.75	0.75	0.75
	Open DCs	1(3), 2(3), 3(2), 4(2)	1(2), 2(3), 3(2), 4(2)	1(2), 2(3), 3(2), 4(2)	1(2), 2(3), 3(2), 4(2)
	W_j	[1.75, 2.27, 2.10, 1.94]	[6.63, 2.27, 2.54, 2.39]	[6.63, 2.27, 2.58, 2.36]	[6.78, 2.27, 2.65, 2.28]
	ρ_j	[0.64, 0.69, 0.68, 0.66]	[0.87, 0.69, 0.72, 0.71]	[0.87, 0.69, 0.72, 0.70]	[0.87, 0.69, 0.73, 0.69]
	CUT	12	12	8	12
	ITR	4	4	3	4
	CPU(s)	0.65	0.62	0.49	0.60
$\theta = 100$	TC	182, 451	176, 255	172, 486	170, 940
	FC	57, 340	57, 340	57, 340	54, 709
	VC	101, 987	101, 494	101, 494	101, 557
	RC	23, 124	17, 421	13, 650	14, 675
	$E(W)$	6.98	5.26	4.12	4.43
	$\bar{\rho}$	0.56	0.56	0.56	0.61
	Open DCs	1(3), 2(3), 3(3), 4(3)	1(3), 2(3), 3(3), 4(3)	1(3), 2(3), 3(3), 4(3)	1(3), 2(3), 3(3), 4(2)
	W_j	[1.45, 1.68, 0.96, 1.05]	[1.54, 1.84, 0.91, 0.97]	[1.54, 1.84, 0.91, 0.97]	[1.75, 2.15, 1.00, 1.54]
	ρ_j	[0.59, 0.63, 0.49, 0.51]	[0.61, 0.65, 0.48, 0.49]	[0.61, 0.65, 0.48, 0.49]	[0.64, 0.68, 0.50, 0.61]
	CUT	16	4	4	16
	ITR	5	2	2	5
	CPU(s)	0.77	0.26	0.35	0.93

(Continued on next page)

Table 1. Comparison of the MTO supply chain network configurations for $M/G/1$, $M/M/1$ and $M/D/1$ cases: an illustrative example *(Continued)*

		$M/G/1$ $(CV = 1.5)$	$M/M/1$ $(CV = 1)$	$M/G/1$ $(CV = 0.5)$	$M/D/1$ $(CV = 0)$
$\theta = 1000$	TC	358, 156	316, 316	289, 863	280, 873
	FC	71, 675	71, 675	71, 675	71, 675
	VC	106, 029	102, 974	99, 683	99, 460
	RC	180, 453	141, 666	118, 506	109, 738
	$E(W)$	5.44	4.27	3.57	3.31
	$\bar{\rho}$	0.44	0.44	0.44	0.44
	Open DCs	1(3), 2(3), 3(3), 4(3), 5(3)	1(3), 2(3), 3(3), 4(3), 5(3)	1(3), 2(3), 3(3), 4(3), 5(3)	1(3), 2(3), 3(3), 4(3), 5(3)
	W_j	[0.64, 1.39, 0.78, 0.84, 0.55]	[0.66, 1.51, 0.77, 0.83, 0.50]	[0.68, 1.67, 0.76, 0.84, 0.43]	[0.72, 1.67, 0.76, 0.85, 0.41]
	ρ_j	[0.39, 0.58, 0.44, 0.46, 0.35]	[0.40, 0.60, 0.44, 0.45, 0.34]	[0.40, 0.63, 0.43, 0.46, 0.30]	[0.42, 0.63, 0.43, 0.46, 0.29]
	CUT	35	35	25	10
	ITR	8	8	6	3
	CPU(s)	1.66	1.45	1.2	0.39

4.3. *Performance of the cutting plane method for MTO supply chain design*

Table 2 displays the performance of the cutting plane method for MTO supply chain design problems under $M/G/1$ $(CV = 1.5)$, $M/M/1$ $(CV = 1.0)$ and $M/D/1$ $(CV = 0)$ cases for varying problem sizes. The columns marked FC, VC and RC represent the fixed costs, the variable production and transportation, and the response time costs respectively, expressed as a percentage of total costs (TC). The columns marked $E(W)$ are the total average waiting time in the system, $\overline{\rho}$ is the average DC utilization and

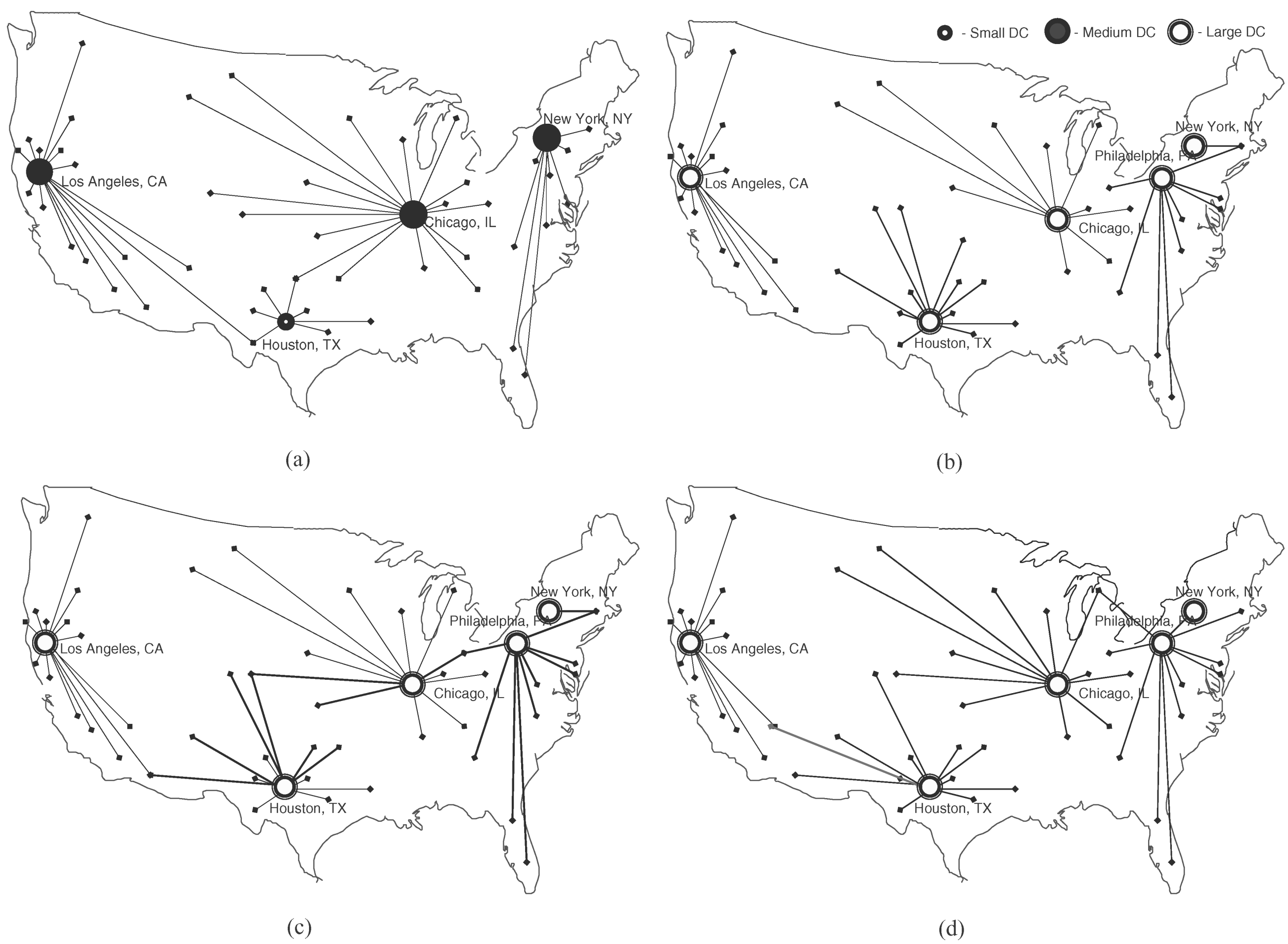

Fig. 5. Effect of changing response time cost on the supply chain network configuration: (a) $\theta = 0$; (b) the $M/D/1$ case, $\theta = 1000$; (c) the $M/M/1$ case, $\theta = 1000$; and (d) the $M/G/1$ case $(CV = 1.5)$, $\theta = 1000$.

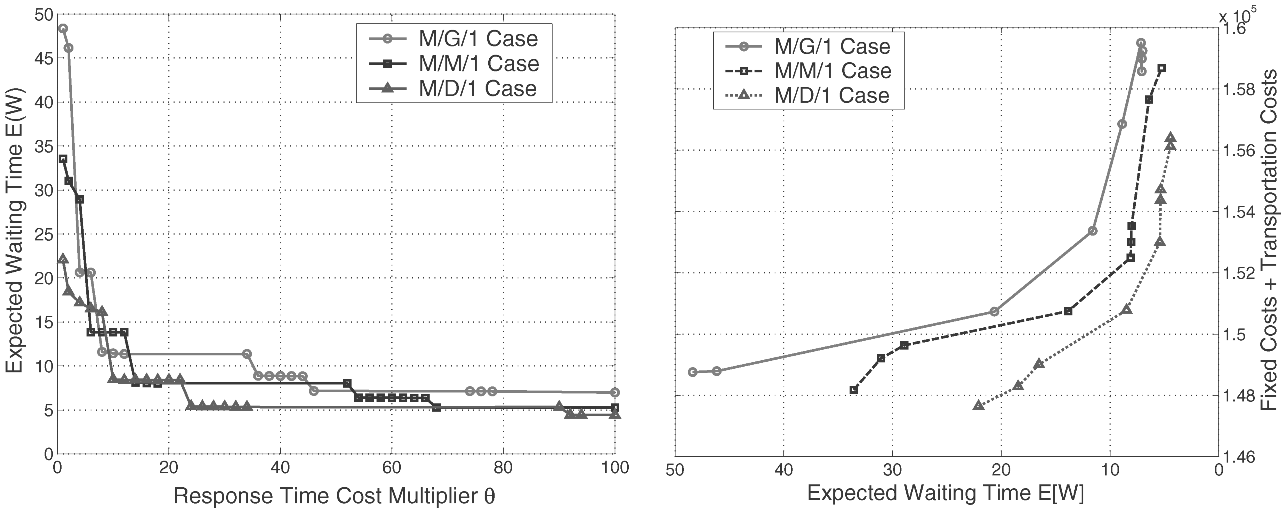

Fig. 6. (a) Effect of changing response time cost (t) on the total expected waiting time $E[W]$; and (b) effect of changing expected waiting time $E[W]$ on the fixed costs plus the transportation costs.

DC represents the number of DCs opened. The table also displays the number of constraints generated (CUT), the number of iterations of the algorithm (ITR) and the total CPU time in seconds required to obtain the optimal solution.

The results show that the average CPU time for $M/G/1$, $M/M/1$ and $M/D/1$ cases are 48, 21 and 9 seconds respectively, whereas the average number of cuts required are 26, 25 and 25 respectively. Also, the maximum CPU time for $M/G/1$, $M/M/1$ and $M/D/1$ cases are 1147, 410 and 107 seconds respectively, whereas the maximum number of cuts required are 66, 60 and 50 respectively. The computation times reveal the stability and the efficiency of the cutting plane algorithm for different percentages of fixed, variable and response time costs, whereas the number of iterations imply that only a fraction of the constraints in ($P_L(H)$) is required. As the magnitude of response time cost (t) increases, the percentage of response time cost becomes more significant with respect to other cost components and the algorithm seems to require more CPU time and iterations as large number of cuts are generated. Furthermore, in almost all of the instances, the $M/G/1$ case requires more cuts, and hence more CPU time to solve than the $M/M/1$ and $M/D/1$ cases. This is attributed to the non-linearity in the expression of expected waiting time for $M/G/1$ queues. It is also worthwhile noting that the computational times for the second set of instances ($I = 50$, $J = 20$ and $K = 3$) are comparatively higher than others because the optimal solution has highly congested DCs. In the model, this corresponds to the value of $R/(1 + R)$ approaching one. At the flat portion of $R/(1 + R)$, a higher number of cuts is needed to close the gap.

4.4. *Performance of the Lagrangean heuristic for ATO supply chain design*

The computational performance of the Lagrangean heuristic for the two-echelon ATO supply chain model for the $M/G/1$, $M/M/1$ and $M/D/1$ cases is shown in Table 3. The second subproblem pertaining to the MTO echelon was solved to optimality using the cutting plane approach. In all these test problems, the heuristic is activated at the final iteration of the Lagrangean procedure. The Lagrangean bound (LAG-H) is expressed as the percentage of heuristic solution and the quality of the heuristic solution (GAP) is expressed as: $100 \times$ (Heuristic Solution − LAG-H)/LAG-H. The table shows the computational time of the subproblems (SP), the master problem (MP) and the heuristic (H) expressed as a percentage of the total computational time (CPU) for various single ($L = 1$) and multiple ($L = 3$ and 5) product instances. From these results, it is evident that the proposed heuristic succeeds in finding feasible solutions that are within an average of 2.81, 2.58 and 2.99% of the Lagrangean bound in reasonable computational time: 427, 390 and 345 seconds for the $M/G/1$, $M/M/1$ and $M/D/1$ cases respectively. The total computational time can be as high as 1038 seconds in some cases. In terms of the size of the test problems, the heuristic is able to solve problems with up to 35 plants, 20 DCs and 150 customers, five products, five capacity levels and five semi-finished products within a maximum of a 6% gap from the optimal solution. Table 4 shows that the solution of subproblem 2 accounts for most of the computational time, 89.08, 89.50 and 89.93% for $M/G/1$, $M/M/1$ and $M/D/1$ cases respectively. The master problem accounts for 9.87, 9.36 and 9.03%, whereas the

Table 2. Computational performance of the cutting plane method: MTO supply chain design

					M/G/1 case (CV = 1.5)										*M/M/1 case*										*M/D/1 case*									
No.	*I*	*J*	*K*	θ	*FC* (%)	*VC* (%)	*RC* (%)	*TC*	*E(W)*	$\overline{\rho}$	*DC*	*CUT*	*ITR*	*CPU*	*FC* (%)	*VC* (%)	*RC* (%)	*TC*	*E(W)*	$\overline{\rho}$	*DC*	*CUT*	*ITR*	*CPU*	*FC* (%)	*VC* (%)	*RC* (%)	*TC*	*E(W)*	$\overline{\rho}$	*DC*	*CUT*	*ITR*	*CPU*
1	50	10	3	0.1	40	60	0	133, 713	221	0.93	6	24	5	2	40	60	0	133, 594	152	0.93	6	18	4	1	40	60	0	133, 484	95	0.92	6	18	4	1
				1	39	59	2	136, 202	198	0.93	6	30	6	2	39	59	1	135, 181	126	0.93	6	24	5	1	40	60	1	134, 356	66	0.93	6	24	5	1
				5	39	58	3	140, 635	60	0.85	6	24	5	2	40	58	2	139, 148	41	0.85	6	18	4	2	40	59	1	137, 862	26	0.86	6	30	6	3
				10	40	56	3	143, 888	35	0.77	6	24	5	3	39	57	4	141, 829	37	0.85	6	30	6	3	40	58	2	139, 498	23	0.85	6	18	4	2
				50	42	52	6	158, 723	14	0.61	6	6	2	2	40	53	8	153, 355	17	0.70	6	18	4	3	39	54	6	148, 628	14	0.77	6	18	4	2
				100	39	49	12	168, 467	14	0.61	6	18	4	5	40	51	9	162, 917	10	0.61	6	6	2	2	39	52	9	156, 215	11	0.7	6	18	4	3
2	50	20	3	0.1	59	41	0	122, 775	419	0.92	11	44	5	10	59	41	0	122, 625	296	0.92	11	44	5	8	59	41	0	122, 483	188	0.92	11	44	5	7
				1	56	43	1	124, 843	198	0.91	10	40	5	10	56	43	1	124, 153	139	0.91	10	40	5	7	59	41	1	123, 482	99	0.92	11	44	5	7
				5	57	39	4	129, 842	117	0.86	11	55	6	24	58	39	3	127, 937	80	0.87	11	55	6	12	55	42	2	126, 045	63	0.91	10	40	5	8
				10	57	37	6	134, 288	89	0.83	11	55	6	44	57	38	5	131, 372	74	0.86	11	55	6	22	58	39	3	128, 366	45	0.87	11	44	5	10
				50	51	37	12	154, 123	41	0.71	10	50	6	541	53	37	10	147, 618	33	0.74	10	50	6	281	54	38	8	140, 915	24	0.76	10	40	5	50
				100	52	33	16	170, 689	29	0.64	11	66	7	1147	49	35	15	160, 786	27	0.71	10	60	7	410	52	37	12	150, 932	20	0.74	10	50	6	107
3	100	5	3	1	27	73	0	196, 048	51	0.83	4	12	4	1	27	73	0	195, 869	37	0.83	4	8	3	1	27	73	0	195, 713	21	0.83	4	12	4	1
				5	27	72	1	197, 594	41	0.83	4	4	2	0	27	73	1	196, 932	26	0.83	4	4	2	0	27	73	0	196, 389	16	0.83	4	12	4	1
				10	26	72	2	199, 377	37	0.83	4	16	5	1	27	72	1	198, 142	26	0.83	4	12	4	1	27	73	1	197, 085	15	0.83	4	4	2	0
				50	28	69	3	206, 964	12	0.67	4	16	5	1	27	70	3	204, 858	13	0.75	4	12	4	1	26	71	3	202, 211	13	0.83	4	8	3	1
				100	28	67	5	212, 399	12	0.67	4	16	5	1	28	68	4	209, 168	8	0.67	4	12	4	1	27	69	4	206, 235	8	0.75	4	12	4	1
				200	29	65	6	220, 249	7	0.56	4	16	5	1	28	66	5	216, 430	6	0.61	4	8	3	1	28	68	5	211, 637	6	0.67	4	12	4	1
4	100	10	3	1	47	52	0	165, 402	134	0.74	8	32	5	8	34	66	1	178, 665	284	0.93	6	30	6	5	34	66	0	178, 135	153	0.93	6	24	5	3
				5	53	47	0	166, 564	32	0.65	9	18	3	6	37	62	1	180, 104	57	0.86	7	21	4	3	38	62	0	179, 576	38	0.86	7	21	4	3
				10	52	47	1	167, 152	31	0.65	9	27	4	6	37	62	1	181, 077	49	0.85	7	28	5	3	37	62	1	180, 216	32	0.86	7	21	4	2
				50	51	46	3	171, 453	28	0.65	9	36	5	9	38	60	3	186, 480	26	0.77	7	28	5	4	37	61	2	184, 155	25	0.85	7	28	5	3
				100	50	44	6	176, 749	28	0.65	9	27	4	12	37	58	5	191, 084	24	0.77	7	28	5	5	38	59	3	187, 494	15	0.77	7	28	5	4
				200	48	46	6	184, 572	14	0.55	8	8	2	12	35	59	6	198, 182	16	0.70	6	18	4	4	34	61	5	192, 882	13	0.77	6	24	5	4
5	100	20	3	1	47	52	0	165, 218	159	0.74	8	32	5	9	47	52	0	165, 061	119	0.74	8	32	5	6	48	52	0	164, 903	79	0.74	8	24	4	5
				5	47	52	1	166, 275	86	0.74	8	32	5	9	47	52	0	165, 867	65	0.74	8	24	4	5	47	52	0	165, 458	44	0.74	8	32	5	6
				10	53	47	0	166, 751	32	0.65	9	18	3	5	53	47	0	166, 503	23	0.65	9	18	3	4	47	52	1	165, 939	36	0.74	8	24	4	5
				50	52	46	2	169, 629	29	0.65	9	36	5	7	52	47	1	168, 556	20	0.65	9	27	4	5	52	47	1	167, 667	13	0.65	9	27	4	5
				100	51	45	4	173, 133	28	0.65	9	36	5	10	51	46	3	170, 995	20	0.65	9	27	4	5	52	46	2	169, 281	13	0.65	9	27	4	5
				200	46	47	7	179, 780	25	0.65	8	32	5	21	50	45	6	175, 861	20	0.65	9	36	5	9	51	46	4	172, 442	13	0.65	9	27	4	5
6	150	5	5	1	24	76	0	225, 407	149	0.91	4	16	5	2	24	76	0	225, 124	114	0.91	4	16	5	2	24	76	0	224, 830	81	0.92	4	16	5	2
				5	24	75	1	227, 400	96	0.91	4	8	3	1	24	76	1	226, 563	64	0.91	4	12	4	1	24	76	0	225, 807	42	0.91	4	16	5	2
				10	25	74	1	229, 437	38	0.82	4	8	3	1	24	75	1	227, 963	59	0.91	4	16	5	2	24	76	1	226, 646	34	0.91	4	12	4	1
				50	25	73	2	234, 236	19	0.75	4	16	5	2	26	73	1	232, 794	13	0.75	4	4	2	1	25	74	1	230, 868	13	0.82	4	16	5	2
				100	25	72	3	238, 462	18	0.75	4	12	4	1	25	72	2	235, 758	13	0.75	4	16	5	2	26	73	2	233, 477	8	0.75	4	4	2	1
				200	26	70	4	243, 902	12	0.67	4	12	4	2	26	71	3	240, 689	8	0.67	4	8	3	1	25	72	3	237, 123	8	0.75	4	16	5	2
7	150	10	5	1	31	69	0	207, 164	294	0.92	6	24	5	6	31	69	0	206, 933	224	0.93	6	24	5	5	31	69	0	206, 707	144	0.93	6	24	5	6
				5	32	68	0	208, 307	87	0.88	6	6	2	3	34	66	0	207, 975	88	0.89	7	28	5	5	31	69	0	207, 474	80	0.92	6	24	5	5
				10	32	68	1	209, 108	81	0.88	6	18	4	5	32	68	1	208, 534	55	0.88	6	6	2	2	34	66	0	208, 034	47	0.89	7	28	5	6
				50	34	64	2	213, 571	54	0.83	7	28	5	10	34	64	2	211, 759	37	0.84	7	28	5	6	35	64	1	210, 229	23	0.84	7	21	4	4
				100	31	65	3	217, 917	37	0.8	6	24	5	13	34	63	3	215, 170	35	0.83	7	21	4	7	34	64	2	212, 314	22	0.84	7	21	4	5
				200	32	64	4	223, 722	25	0.73	6	18	4	21	31	65	4	220, 220	24	0.8	6	24	5	16	34	63	4	216, 271	21	0.83	7	21	4	7
8	150	20	5	1	48	52	0	177, 444	332	0.9	10	40	5	11	48	52	0	177, 277	245	0.9	10	40	5	8	48	52	0	177, 113	162	0.9	10	40	5	8
				5	45	54	1	178, 543	169	0.89	9	27	4	11	45	54	0	178, 141	115	0.89	9	27	4	7	48	52	0	177, 693	99	0.9	10	40	5	9
				10	45	54	1	179, 538	157	0.89	9	36	5	16	45	54	1	178, 818	103	0.89	9	27	4	8	45	54	0	178, 190	61	0.89	9	27	4	7
				50	46	52	2	184, 593	62	0.81	9	27	4	38	45	53	2	183, 147	64	0.85	9	36	5	27	45	53	2	180, 893	51	0.88	9	36	5	14
				100	45	51	4	188, 335	59	0.81	9	27	4	54	45	52	3	185, 827	41	0.81	9	27	4	27	45	52	2	183, 620	36	0.85	9	36	5	24
				200	49	47	5	194, 476	37	0.71	10	30	4	211	44	51	5	190, 699	39	0.81	9	27	4	68	45	52	3	186, 730	24	0.81	9	27	4	25
				min	24	33	0	122, 775	7	0.55	4	4	2	0	24	35	0	122, 625	6	0.61	4	4	2	0	24	37	0	122, 483	6	0.65	4	4	2	0
				max	59	76	16	243, 902	419	0.93	11	66	7	1147	59	76	15	240, 689	296	0.93	11	60	7	410	59	76	12	237, 123	188	0.93	11	50	6	107
				mean	41	56	3	183, 022	82	0.77	7	26	4	48	39	58	3	182, 995	67	0.80	7	25	4	21	39	59	2	186, 516	46	0.82	7	25	4	9

Table 3. Computational performance of the Lagrangean heuristic: ATO supply chain design

								M/G/1 case(CV = 1.5)						*M/M/1 case*						*M/D/1 case*					
No.	*I*	*J*	*K*	*L*	*M*	*N*	θ	*LAG-H* (%)	*GAP* (%)	*SP* (%)	*MP* (%)	*H* (%)	*CPU(s)*	*LAG-H* (%)	*GAP* (%)	*SP* (%)	*MP* (%)	*H* (%)	*CPU(s)*	*LAG-H* (%)	*GAP* (%)	*SP* (%)	*MP* (%)	*H* (%)	*CPU(s)*
1	50	10	3	1	10	3	0.1	99.78	0.22	84.31	14.36	1.33	108	96.83	3.17	86.06	12.75	1.19	85	98.33	1.67	92.96	6.29	0.75	56
							1	95.84	4.16	84.55	13.51	1.94	111	99.59	0.41	85.21	12.96	1.83	85	94.58	5.42	91.34	8.06	0.60	62
							5	94.77	5.23	87.1	11.51	1.39	118	99.97	0.03	86.13	12.19	1.68	86	98.84	1.16	92.05	5.94	2.01	69
							10	98.33	1.67	83	14.98	2.02	121	95.77	4.23	93.14	5.53	1.33	88	96.94	3.06	88.43	10.90	0.67	71
							50	98.91	1.09	91.52	6.84	1.64	122	94.12	5.88	89.7	8.82	1.48	101	99.86	0.14	88.58	9.59	1.83	73
							100	97.20	2.80	88.78	11.13	0.09	127	95.40	4.6	91.73	7.01	1.26	104	98.65	1.35	90.91	7.93	1.16	75
2	50	20	3	1	10	3	0.1	97.07	2.93	87.85	11.17	0.98	191	99.66	0.34	85.99	12.41	1.6	176	98.23	1.77	93.59	5.44	0.97	131
							1	98.24	1.76	85.03	13.15	1.82	192	96.36	3.64	91.17	8.29	0.54	179	94.79	5.21	91.87	6.36	1.77	147
							5	99.32	0.68	85.22	13.04	1.74	199	99.07	0.93	92.56	6.14	1.3	181	95.92	4.08	94.09	5.39	0.52	148
							10	97.89	2.11	88.68	10.73	0.59	209	99.37	0.63	90.55	8.59	0.86	185	99.73	0.27	86.77	11.58	1.65	155
							50	94.06	5.94	94.66	5.17	0.17	211	99.02	0.98	87.18	12.46	0.36	186	97.71	2.29	91.02	6.97	2.01	158
							100	94.65	5.35	89.75	9.67	0.58	213	95.09	4.91	92.35	7.59	0.06	189	97.05	2.95	93.33	6.03	0.64	162
3	100	5	3	3	20	5	1	99.97	0.03	93.78	5.68	0.54	182	98.35	1.65	89.18	10.14	0.68	168	98.56	1.44	86.09	12.09	1.82	150
							5	96.58	3.42	85.45	14.39	0.16	185	99.75	0.25	92.9	6.7	0.4	169	95.33	4.67	88.18	11.27	0.55	159
							10	97.23	2.77	86.79	12.64	0.57	187	94.26	5.74	87.14	10.9	1.96	171	95.07	4.93	86.89	11.50	1.61	159
							50	97.41	2.59	93.22	5.83	0.95	190	96.71	3.29	92.66	5.43	1.91	176	99.89	0.11	86.5	12.72	0.78	161
							100	99.68	0.32	93.11	6.64	0.25	194	99.22	0.78	92.25	6.42	1.33	180	98.85	1.15	88.02	11.86	0.12	165
							200	95.64	4.36	93.61	5.06	1.33	196	99.55	0.45	88.63	10.35	1.02	175	94.97	5.03	85.38	12.94	1.68	165
4	100	10	3	3	20	5	1	94.47	5.53	91.73	8.09	0.18	238	94.01	5.99	87.35	11.95	0.7	217	95.40	4.60	91.29	8.25	0.46	197
							5	94.51	5.49	90.31	8.26	1.43	268	95.93	4.07	92.1	7.29	0.61	218	94.21	5.79	86.72	11.72	1.56	201
							10	98.83	1.17	90.63	9.29	0.08	277	98.77	1.23	87.36	11.18	1.46	222	97.99	2.01	92.51	5.88	1.61	202
							50	97.22	2.78	86.47	11.87	1.66	284	95.41	4.59	86.51	12.94	0.55	232	99.58	0.42	87.82	11.84	0.34	204
							100	97.40	2.60	91.17	7.41	1.42	296	98.45	1.55	90.21	9.1	0.69	234	96.62	3.38	85.71	12.99	1.30	210
							200	95.94	4.06	85.5	13.22	1.28	298	97.45	2.55	90.95	7.79	1.26	234	94.80	5.20	91.81	6.52	1.67	215
5	100	20	3	3	20	5	1	98.97	1.03	85.81	12.14	2.05	297	98.08	1.92	87.23	10.92	1.85	283	95.40	4.60	92.22	5.79	1.99	251
							5	97.61	2.39	92.9	5.91	1.19	298	94.37	5.63	92.15	7.57	0.28	285	98.65	1.35	93.11	5.43	1.46	268
							10	95.82	4.18	87.46	11.13	1.41	300	97.12	2.88	87.54	10.71	1.75	285	99.61	0.39	88.22	11.10	0.68	269
							50	95.55	4.45	91.68	7.47	0.85	309	97.03	2.97	89.47	9.11	1.42	286	94.92	5.08	86.04	12.13	1.83	271
							100	99.12	0.88	85.59	12.88	1.53	313	98.66	1.34	87.43	11.79	0.78	289	94.12	5.88	86.81	12.00	1.19	277
							200	95.46	4.54	89.1	9.64	1.26	348	96.44	3.56	92.22	7.55	0.23	293	96.12	3.88	91.16	8.60	0.24	278
6	150	5	5	5	35	5	1	98.67	1.33	89.13	9.93	0.94	551	99.94	0.06	86.61	12.26	1.13	486	95.08	4.92	92.21	6.98	0.81	423
							5	98.35	1.65	85.54	13.96	0.50	563	94.75	5.25	86.2	11.91	1.89	499	95.08	4.92	93.19	5.55	1.26	427
							10	94.73	5.27	93.29	6.17	0.54	576	96.05	3.95	90.43	9.19	0.38	513	98.69	1.31	87.81	10.86	1.33	444
							50	99.83	0.17	89.76	9.06	1.18	593	98.66	1.34	86.77	11.68	1.55	536	99.16	0.84	90.33	9.11	0.56	451
							100	99.01	0.99	85.81	13.66	0.53	599	96.90	3.1	92.06	7.38	0.56	544	97.22	2.78	91.59	7.71	0.70	484
							200	99.87	0.13	87.42	12.07	0.51	617	99.44	0.56	92.73	5.44	1.83	547	94.26	5.74	88.29	10.55	1.16	485
7	150	10	5	5	35	5	1	99.23	0.77	90.39	9.29	0.32	715	99.91	0.09	89.18	9.53	1.29	645	94.90	5.10	93.54	6.09	0.37	588
							5	97.23	2.77	91.5	7.63	0.87	717	99.00	1	89	9	2	690	96.47	3.53	90.24	9.10	0.66	598
							10	95.34	4.66	94.55	5.32	0.13	726	98.73	1.27	89.71	8.64	1.65	691	99.17	0.83	89.6	9.16	1.24	615
							50	97.13	2.87	90.74	7.41	1.85	741	98.15	1.85	91.6	7.61	0.79	697	95.27	4.73	88.82	10.47	0.71	624
							100	95.27	4.73	91.74	6.66	1.60	743	98.50	1.5	87.91	11.34	0.75	698	99.94	0.06	87.06	12.88	0.06	627
							200	98.12	1.88	91.44	7.15	1.41	747	96.97	3.03	91.45	7.14	1.41	706	97.41	2.59	92.49	6.96	0.55	632
8	150	20	5	5	35	5	1	98.31	1.69	91.9	6.98	1.12	981	95.47	4.53	87.58	11.25	1.17	922	95.07	4.93	94.1	5.13	0.77	765
							5	96.68	3.32	89.86	8.58	1.56	1000	99.81	0.19	91.22	7.21	1.57	935	95.90	4.10	85.95	12.40	1.65	800
							10	95.12	4.88	93.2	5.87	0.93	1004	96.02	3.98	87.96	10.64	1.4	945	98.35	1.65	92.89	6.46	0.65	810
							50	95.71	4.29	84.32	14.16	1.52	1007	99.50	0.5	91.95	7.83	0.22	955	94.35	5.65	86.9	13.03	0.07	871
							100	98.64	1.36	85.82	12.35	1.83	1019	94.38	5.62	89.96	9.18	0.86	972	99.94	0.06	90.32	9.60	0.08	890
							200	94.38	5.62	84.72	14.67	0.61	1038	94.09	5.91	88.84	9.49	1.67	976	99.72	0.28	92.09	6.17	1.74	904
							min	94.06	0.03	83.00	5.06	0.08	108	99.97	0.03	85.21	5.43	0.06	85	94.12	0.06	85.38	5.13	0.06	56
							max	99.97	5.94	94.66	14.98	2.05	1038	94.01	5.99	93.14	12.96	2.00	976	99.94	5.88	94.10	13.03	2.01	904
							mean	97.19	2.81	89.08	9.87	1.05	427	97.42	2.58	89.50	9.36	1.14	390	97.01	2.99	89.93	9.03	1.04	345

Table 4. Comparison between the cutting plane method and the Lagrangean heuristic for ATO supply chain design for $I = 100$, $J = 10$, $K = 3$, $M = 20$ and $N = 1$

	Cutting plane method									Lagrangean heuristic								
θ	*FC* (%)	*VC* (%)	*RC* (%)	*E(W)*	$\bar{\rho}$	*DCs*	*CUT*	*ITR*	*CPU* (*s*)	*FC* (%)	*VC* (%)	*RC* (%)	*E(W)*	$\bar{\rho}$	*DCs*	*LR* (%)	*GAP* (%)	*CPU* (*s*)
Tight capacities, $r = 3$																		
0.1	50	50	0	689.3	0.96	5	10	3	714	41	59	0	689.3	0.96	5	95.88	4.12	251
1	49	50	1	209.2	0.96	5	10	3	1029	41	59	1	218.2	0.96	5	96.78	3.22	268
5	48	50	2	169.3	0.95	5	10	3	1585	43	57	1	195.9	0.94	5	95.04	4.96	269
10	51	49	1	32.5	0.69	5	10	3	2180	42	56	2	51.84	0.76	5	95.81	4.19	271
50	51	47	2	14.51	0.6	5	5	2	1002	44	55	2	17.47	0.62	5	96.72	3.28	177
100	51	47	2	8.39	0.54	5	5	2	810	43	54	3	11.47	0.6	5	94.99	5.01	278
500	50	46	4	8.39	0.54	5	5	2	964	43	49	8	10.82	0.65	5	96.4	3.6	423
1000	49	44	7	6.21	0.44	5	5	2	1100	40	46	14	9.82	0.44	5	96.88	3.12	427
2000	46	41	13	5.99	0.44	5	5	2	1148	39	41	21	4.75	0.43	6	97.11	2.89	444
5000	38	32	30	3.95	0.32	7	28	5	1699	33	33	34	3.98	0.32	7	97.01	2.99	451
Moderate capacities, $r = 5$																		
0.1	46	54	0	261.9	0.84	6	12	3	1500	39	61	0	332.3	0.85	4	98.11	1.89	205
1	46	54	0	98.12	0.83	6	6	2	964	39	61	0	122.7	0.84	4	96.83	3.17	407
5	41	58	1	34.16	0.83	4	4	2	1019	38	61	1	40.72	0.84	4	96.69	3.31	325
10	41	58	1	34.16	0.83	4	4	2	898	38	61	1	40.72	0.84	4	95.52	4.48	280
50	41	56	4	21.84	0.76	4	4	2	1467	39	58	3	27.56	0.79	4	97.27	2.73	322
100	43	55	2	6.81	0.56	4	4	2	1495	40	57	2	7.74	0.58	4	97.11	2.89	333
500	42	54	4	6.81	0.56	4	4	2	1122	37	53	10	7.74	0.58	4	96.88	3.12	365
1000	40	50	10	6.81	0.56	4	4	2	1370	34	48	18	7.74	0.58	4	94.01	5.99	291
2000	41	44	15	5.52	0.44	5	10	3	2524	40	38	22	6.63	0.47	6	95.6	4.4	589
5000	30	29	41	4.55	0.37	6	12	3	1981	30	29	40	6.53	0.44	6	97.67	2.33	232
Loose capacities, $r = 10$																		
0.1	40	60	0	81.01	0.83	4	4	2	1103	40	60	0	81.66	0.83	4	96.57	3.43	546
1	40	60	0	62.75	0.83	4	4	2	955	39	60	0	63.85	0.83	4	96.65	3.35	435
5	40	60	1	40.27	0.83	4	4	2	906	39	60	1	42.23	0.83	4	98.91	1.09	487
10	39	59	1	40.27	0.83	4	4	2	875	39	60	1	41.78	0.83	4	94.41	5.59	327
50	40	58	2	11.71	0.67	4	4	2	979	40	58	2	12.97	0.67	4	96.19	3.81	354
100	40	56	4	11.71	0.67	4	4	2	1140	40	57	3	11.75	0.61	4	96.99	3.01	654
500	41	55	5	7.02	0.56	4	4	2	1080	38	52	11	8.96	0.56	4	97.27	2.73	765
1000	38	52	10	6.94	0.56	4	8	3	1765	34	47	19	7.84	0.56	4	98.1	1.9	642
2000	34	47	19	6.85	0.56	4	8	3	1884	41	38	21	7.54	0.37	6	95.59	4.41	822
5000	32	29	39	4.43	0.37	6	12	3	2148	31	29	39	4.43	0.37	6	96.37	3.63	984
Min	30	29	0	3.95	0.32	4	4	2	714	30	29	0	3.98	0.32	4	94.01	1.09	177
Max	51	60	41	689.3	0.96	7	28	5	2524	44	61	40	689.3	0.96	7	98.91	5.99	984
Mean	43	50	7	63.38	0.66	5	7	2	1314	39	52	9	64	0.64	5	97.51	3.49	421

heuristic accounts for 1.05, 1.14 and 1.04%, on average for $M/G/1$, $M/M/1$ and $M/D/1$ cases respectively.

In Table 4, we compare the performance of the cutting plane algorithm and the Lagrangean heuristic and report the results for one problem set ($I = 100$, $J = 10$, $K = 3$, $M = 20$ and $N = 1$) for different values of the ratio of plant capacities to total demand ($r = \sum_m P_m / \sum_i \lambda_i$). The results show that the Lagrangean heuristic outperforms the cutting plane method in terms of computational time. On average, the Lagrangean heuristic takes 421 seconds whereas the cutting plane method takes 1344 seconds.

5. Conclusions

In this paper, we modeled and analyzed the effect of response time consideration on the design of MTO and ATO supply chain networks. We presented an MTO supply chain design model that captures the trade-off among response time, the fixed cost of opening DCs and equipping them with sufficient capacity, and the transportation cost associated with serving customers. Under the assumption that the customer demand follows a Poisson process and service times follow general distribution, the DCs were modeled as a network of single-server queues, whose capacity levels and locations are decision variables. We presented a non-linear MIP formulation, a linearization procedure and an exact solution approach based on a cutting plane method. Our computational results indicate that the cutting plane algorithm provides optimal solution for moderate instances of the problem in few iterations and reasonable computation times.

We also presented a model for designing two-echelon ATO supply chain networks, that consists of plants and DCs serving a set of customers. Lagrangean relaxation was applied to decompose the problem by echelon—one for the MTS echelon and the other for the MTO echelon. While Lagrangean relaxation provides a lower bound, a heuristic is proposed that uses the solution of the subproblems to construct an overall feasible solution. Computational results reveal that the heuristic solution is on average within 6% of its optimal. We used the models to demonstrate empirically that substantial improvement (decrease) in response time can be achieved with a minimal increase in total cost associated with designing supply chains. Also, we showed that the supply chain configuration (DC location and capacity, and allocation of customers to DCs) obtained using the model that considers congestion can be very different from those obtained using the traditional models that ignores response time.

The focus of our ongoing research is to develop models for the design of MTO and ATO supply chains under more general settings: general demand arrivals, general service time distributions and multiple servers. Due to the lack of an exact expression for the waiting time in these settings, we plan to explore two approaches to model the problem. The first approach relies on approximations of expected waiting times whereas the second approach integrates simulation within an MIP framework.

Acknowledgement

The authors would like to thank two anonymous referees and the Associate Editor for helpful comments.

References

Amiri, A. (1997) Solution procedures for the service system design problem. *Computers and Operations Research*, **24**, 49–60.

Berman, O. and Krass, D. (2002) Facility location problems with stochastic demands and congestion, in *Facility Location: Applications and Theory*, Drezner, Z. and Hamacher, H.W. (eds.), Springer-Verlag, New York, pp. 329–369.

Daskin, M.S. (2004) SITATION—facility location software. Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL. Available at http://users.iems.nwu.edu/∽mdaskin.

Dell, M. (2000) *Direct from Dell: Stratagies that Revolutionized an Industry*, Harper Collins, New York, NY.

Dogan, K. and Goetschalckx, M. (1999) A primal decomposition method for the integrated design of multi-period production-distribution systems. *IIE Transactions*, **31**(11), 1027–1036.

Duran, M.A. and Grossman, I.E. (1986) An outer approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical Programming*, **36**, 307–339.

Elhedhli, S. (2006) Service system design with immobile servers, stochastic demand and congestion. *Manufacturing and Service Operations Management*, **8**(1), 92–97.

Elhedhli, S. and Gzara, F. (2008) Integrated design of supply chain networks with three echelons, multiple commodities, and technology selection. *IIE Transactions*, **40**(1), 31–44.

Erengüç, S.S., Simpson, N.C. and Vakharia, A.J. (1999) Integrated production/distribution planning in supply chains: an invited review. *European Journal of Operational Research*, **115**, 219–236.

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.

Gupta, D. and Benjaafar, S. (2004) Make-to-order, make-to-stock, or delay product differentiation? A common framework for modelling and analysis. *IIE Transactions*, **36**, 529–546.

Huang, S., Batta, R. and Nagi, R. (2005) Distribution network design: selection and sizing of congested connections. *Naval Research Logistics*, **52**, 701–712.

Kelley, J. E. (1960) The cutting plane method for solving convex programs. *Journal of the SIAM*, **8**, 703–712.

Kim, I. and Tang, C.S. (1997) Lead time and response time in a pull production control system. *European Journal of Operational Research*, **101**, 474–485.

Klose, A. and Drexl, A. (2005) Facility location models for distribution system design. *European Journal of Operational Research*, **162**, 4–29.

Margretta, J. (1998) The power of virtual integration: an interview with Dell computer's Michael Dell. *Harvard Business Review*, **76**(2), 73–84.

Marianov, V. and Serra, D. (2002) Location-allocation of multiple-server service centers with constrained queues or waiting times, *Annals of Operations Research*, **111**, 35–50.

Rajagopalan, S. and Yu, H.L. (2001) Capacity planning with congestion effects. *European Journal of Operational Research*, **134**, 365–377.

Rao, U., Scheller-Wolf, A. and Tayur, S. (2000) Development of a rapid-response supply chain at Caterpillar. *Operations Research*, **48**(2), 189–204.

Sarmiento, A.M. and Nagi. R. (1999) A review of integrated analysis of production-distribution systems. *IIE Transactions*, **31**, 1061–1074.

Shen, Z.J. (2005) A multi-commodity supply chain design problem. *IIE Transactions*, **37**(8), 753–762.

Song, J.-S. and Zipkin, P. (2003) Supply chain operations: assemble-to-order systems, in *Supply Chain Management: Design, Coordination, and Operation*, De Kok, T. and Graves, S. (eds), North-Holland, Amsterdam, The Netherlands, pp. 561–596.

Teo, C.-P. and Shu, J. (2004) Warehouse-retailer network design problem. *Operations Research*, **52**(3), 396–408.

Vidal, C.J. and Goetschalckx, M. (1997) Strategic production-distribution models: a critical review with emphasis on global supply chain models. *European Journal of Operational Research*, **98**, 1–18.

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

Wang, Q., Batta, R. and Rump, C.M. (2003) Facility location models for immobile servers with stochastic demand. *Naval Research Logistics*, **51**, 138–152.

Biographies

Navneet Vidyarthi is an Assistant Professor of Operations Management in the Department of Decision Sciences and MIS in the John Molson School of Business at Concordia University, Montreal, Canada. His current research interests are in supply chain and logistics management with methodological interests in large-scale optimization, simulation-based optimization and meta-heuristics. He has published in *Transportation Science*, *IIE Transactions*, *International Journal of Production Research* and *Managerial Auditing Journal*. He holds a Ph.D. degree in Management Sciences/Operations Research from the University of Waterloo, an M.A.Sc. degree in Industrial Engineering from the University of Windsor, and a B.Tech. degree in Mechanical Engineering from the North Eastern Regional Institute of Science and Technology, India.

Samir Elhedhli is an Associate Professor in the Department of Management Sciences at the University of Waterloo. He holds B.Sc. and M.Sc. degrees in Industrial Engineering from Bilkent University and a Ph.D. in Management from McGill University. He has research interests in large-scale optimization with applications in supply chain and service systems design, classification models and energy models. As a student, two of his papers won awards from the Turkish Operations Research Society for the best undergraduate research paper in 1994 and from CORS for the best graduate research paper in 2001. His work has appeared in scientific journals such as *Management Science, Mathematical Programming*, *Operations Research Letters*, *European Journal of Operational Research*, *Computers and Operations Research*, and several others. He is a member of CORS, INFORMS, SIAM and the Waterloo Management of Integrated Manufacturing Systems Research Centre.

Elizabeth Jewkes is Professor and Chair of the Department of Management Sciences at the University of Waterloo. Her research interests are in stochastic models of production and service supply chain systems. She has published in journals such as *Transportation Science*, *IIE Transactions*, *Stochastic Systems*, *Computers and Operations Research* and the *European Journal of Operational Research*. She is a co-author of the textbook *Engineering Economics in Canada*, which is currently in revision for its fourth edition.