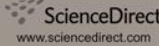


Volume 114, Issue 1	July 2008	ISSN 0925-5273
	international journal of production economics Manufacturing Systems, Strategy & Design	
Editors Editor-in-Chief R.W. GRUBBSTROM North-American Editor P. KELLE Asian-Pacific Editor T.C.E. CHENG Editorial Board P.J. AGRELL S. AXSÄTER W.L. BERRY L.E. CARDENAS BARRÓN N.S. CHEN A. CHIKAN A.H. CHRISTER H. DING S.M. DISNEY Th. DURAND S.E. ELMAGHRABY W.G. FERRELL B.E. FLORES J.R. FREELAND L.F. GELDERS T.N. GOH M. GREGORY A. GUNASEKARAN H.H. HINTERHUBER K. HITOMI R.H. HOLLIER T. ICHIMURA K. INDERFURTH K. ISHII U.S. KARMARKAR B.M.T. LIN R.J. LINN S. MINNER T. MORTON J.A. MUCKSTADT D.N.P. MURTHY Ch. O'BRIEN S. PARK L. PECCATI J.M. PROTH D.S. REMER B.H. RHO D.A. SAMSON B.R. SARKER C.A. SNYDER R. STEINBERG M.T. TABUCANON J.M.A. TANCHOCO D.R. TOWILL M. TUOMINEN L.N. VAN WASSENHOVE D.C. WHYBARK J. WILNGAARD S. WU H. YAMASHINA C.A. YANO PH. ZIPKIN	Available online at  www.sciencedirect.com	
CONTENTS		
Special Section on Competitive Advantage through Global Supply Chains Edited by: C.S. Lalwani and K.S. Pawar		
Editorial	C.S. Lalwani and K.S. Pawar	1
On the impact of order volatility in the European automotive sector	P. Childerhouse, S.M. Disney and D.R. Towill	2
Strategic adaptivity in global supply chains—Competitive advantage by autonomous cooperation	M. Hülsmann, J. Grapp and Y. Li	14
Supply chain management for servitised products: A multi-industry case study	M. Johnson and C. Mena	27
Linkages between service sourcing decisions and competitive advantage: A review, propositions, and illustrating cases	F. Nordin	40
Supply chain contracts with capacity investment decision: Two-way penalties for coordination	P.P. Mathur and J. Shah	56
Mass customised distribution	R. Mason and C. Lalwani	71
Regular Papers		
Analysis of efficiency of the European postal sector	M.J. Irurolde and C. Quirós	84
Business process perspectives: Theoretical developments vs. real-world practice	K. Vergidis, C.J. Turner and A. Tiwari	91
Self-assessment exercises: A comparison between a private sector organisation and higher education institutions	J.J. Tari	105
Inventory/distribution control system in a one-warehouse/multi-retailer supply chain	C. Monthatipkul and P. Yenradee	119
<i>Contents continued on back cover</i>		

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Inventory positioning, scheduling and lead-time quotation in supply chains

Philip Kaminsky^{a,*}, Onur Kaya^b

^a*Industrial Engineering and Operations Research, University of California Berkeley, CA 94720-1777, USA*

^b*Department of Industrial Engineering, Koc University, Istanbul, 34450, Turkey*

Received 15 November 2006; accepted 8 February 2008

Available online 19 February 2008

Abstract

We consider supply chain networks composed of several centrally managed production facilities as well as external suppliers. We design effective heuristics for inventory positioning, order sequencing, and short and reliable due-date quotation for this supply chain. We perform extensive computational testing to assess the effectiveness of our algorithms, and we explore the impact of supply chain topology on inventory costs and effective due-date quotation.

© 2008 Elsevier B.V. All rights reserved.

Keywords: Supply chain; Inventory; Scheduling; Lead time

1. Introduction

In complex supply chains where products and components of products are manufactured in many different facilities, inventory costs make up a significant proportion of total network costs. One way to manage inventory for whatever product or component is produced in a facility is to wait for specific orders to arrive before starting to manufacture; a facility managed in this way is called a make-to-order (MTO) facility. Alternatively, components can be manufactured ahead of time in anticipation of demand, and held in inventory—a so-called make-to-stock (MTS) facility. Clearly, end customers prefer to either immediately find what

they are looking for in the firm's finished goods inventory or to order items and receive them quickly. If manufacturing times for a particular product are short and the production network is relatively uncongested, the firm may not have to hold inventory, and can instead use a MTO approach to achieve the desired short lead time while minimizing inventory holding cost. Unfortunately, the total required time to manufacture a product is frequently greater than the acceptable delivery lead time for that product, so firms must start manufacturing in anticipation of specific orders, and keep some inventory in order to be able to meet customer demand in an acceptable amount of time. Keeping finished goods inventory, however, is not the only available option for reducing delivery lead times. Instead, firms may be able to store intermediate inventory at some facilities in the network (either at suppliers or at

*Corresponding author. Tel.: +1 510 642 4927.

E-mail addresses: kaminsky@ieor.berkeley.edu
(P. Kaminsky), okaya@ku.edu.tr (O. Kaya).

intermediate manufacturing stages). Thus, a key question that arises when managing inventory in complex multi-facility supply chains is *where to keep safety stock*. In other words, which facilities should produce to stock, and which should produce to order, in order to deliver good customer service at a reasonable cost.

Once the question of where to MTS and where to MTO is answered, the level of inventory that should be maintained at these facilities must be determined. In addition to inventory decisions, production of orders at MTO facilities needs to be sequenced, and lead times need to be quoted to customers whose demand is not met out of finished goods inventory. Indeed, firms need to quote short and reliable lead times to their customers to remain competitive in the market, and companies with multiple different products need to assess the impact of production sequencing decisions for one product on the effective lead times of all other products. In spite of this, most research on inventory positioning in supply chains ignores the intricacies of scheduling and lead-time quotation, typically assuming that lead times are exogenously determined, and that orders are processed in the sequence in which demand arrives.

In this paper, we focus on integrating safety stock placement decisions along with scheduling and lead-time quotation into the determination of optimal safety stock placement in supply chains whose component facilities may MTS or MTO. Specifically, we present an effective approach for (1) determining locations at which to store inventory in this supply chain, (2) sequencing specific jobs at specific facilities, and (3) quoting lead times, so that system-wide costs are minimized, quoted lead times are relatively short, and these quoted lead times are typically met. As this is a very difficult problem (and indeed, the individual problems of inventory management, scheduling, and lead-time quotation are by themselves difficult problems), we are motivated to develop a series of heuristics for the integrated management of supply chain inventory, scheduling and lead time for a variety of different supply chain configurations. We computationally test these heuristics, and analyze the impact of system parameters including congestion level, the supply chain structure, and number of jobs, on the performance of the system and on the effectiveness of our heuristics.

There is a vast amount of literature on inventory placement models for multi-stage systems that is applicable to supply chains. Axsater (1993), Feder-

gruen (1993), Inderfurth (1994) and Diks et al. (1996) survey these models in detail. Several researchers including Inderfurth (1991, 1993), Inderfurth and Minner (1998) and Minner (1997) considered the problem of optimizing safety stock placement in supply chains based on the framework of Simpson (1958), who analyzed a serial supply chain to determine the optimal safety stock placement and found that the optimal solution is an “all or nothing” strategy for that model. Graves and Willems (1996, 2000, 2003) extended the results of Simpson (1958) to assembly, distribution and spanning tree network structures. Lee and Billington (1993), Glasserman and Tayur (1995) and Ettl et al. (2000) examined the determination of optimal base-stock levels in a supply chain and developed algorithms to optimize the safety stock placement. Magnanti et al. (2006) model the problem of inventory placement in supply chains as a nonlinear program and use successive piecewise linear approximation to obtain a tight approximation for the problem.

In Kaminsky and Kaya (2008a), we analyze pure MTO supply chains and design effective scheduling and due-date quotation algorithms for the centralized and decentralized versions of those systems. We show that these algorithms are asymptotically optimal (i.e. go to the optimal solution as the number of orders $n \rightarrow \infty$) for the minimization of a function of lead time related costs and tardiness related costs (or more specifically, for the function

$$Z_n = \sum_{i=1}^n (c^d d_i + c^T T_i),$$

where d_i is the quoted due-date for job i , $T_i = (C_i - d_i)^+$ is the tardiness of job i and c^d and c^T are the unit due-date and tardiness costs for the model). In Kaminsky and Kaya (2008b), we integrate inventory decisions, scheduling and due-date quotation issues for combined make-to-order/make-to-stock (MTO–MTS) systems for a two-facility supply chain. We develop models that provide guidance in deciding when to employ MTS and when to use MTO approaches, and how to effectively operate the system to minimize system-wide costs. We also quantify the value of centralization and information in these systems by building decentralized and centralized models, obtaining good solutions to these models, and designing computational experiments to explore the effectiveness of our algorithms and to compare the centralized and decentralized systems.

In this paper, we design effective algorithms for scheduling, lead-time quotation and inventory decisions to minimize total costs of more complex multi-facility supply chains. We utilize the general insights from the results related to scheduling and lead-time quotation in Kaminsky and Kaya (2008a,b), and extend these approaches to multi-facility systems under a variety of conditions. Unfortunately, the inventory calculations for the two-facility problem in Kaminsky and Kaya (2008b) do not generalize well to the multi-facility cases considered in this paper, so we develop a new heuristic strategy using linear programming models to determine appropriate target inventory levels at each facility in the supply chain.

In the next section, we present our model in detail, and then in Section 3, we develop a solution approach for this model. In Section 4, we present the results of our computational analysis.

2. The model

We consider a manufacturing firm facing stationary stochastic demand, whose supply chain consists of a single downstream stage (we call this stage the manufacturer) that receives orders from customers, and a series of stages (suppliers) upstream from this manufacturer. The suppliers can be internal or external suppliers depending on whether or not the managing firm controls these stages. The manufacturer receives customer orders over time, fills these orders immediately if the product ordered is in inventory at the manufacturer, and quotes lead times for the orders if they are for items that are not in inventory at the manufacturer. There are a total of N facilities in the supply chain (that is, the manufacturer and $N - 1$ suppliers), and K product types are offered by this firm to its customers. We assume a stationary stochastic demand process, where demand arrives at the manufacturer with known and possibly different arrival rates for each product. In particular, orders arrive at rate $\lambda = 1/D$ where D is the mean inter-arrival time and each arriving order is for product type i with probability δ_i , $i = 1, 2, \dots, K$. To facilitate our analysis, we assume stationary and independent inter-arrival times, so each order for product type i arrives at rate $\lambda_i = \lambda\delta_i$. Each product must be processed at a specific subset of the facilities in the supply chain, in a specified order, with specified processing and transshipment times (where a transshipment time is

the transportation time between two specific supply chain stages). Each product type thus has a predefined routing through the supply chain.

We model this supply chain as a network with nodes representing facilities in the production network at which specific operations (manufacturing, assembly, etc.) take place, and with arcs representing the flow of components. This network is thus represented by a directed and acyclic graph G , and we assume that each product type i has a known production network which is a subgraph of the entire supply network. In particular, the production operations/locations of each order of type i can be represented by a subgraph \bar{G}_i composed of a specified subset of the nodes and arcs in graph G . For ease of exposition, we assume that only one unit of upstream component is required per any downstream unit. As mentioned above, we assume that a single firm owns the downstream manufacturer and the internal suppliers over whom it has complete control, but there also may be external independent suppliers over whom the manufacturer has no control. An example is presented in Fig. 1, where the internal supplier nodes are represented with rectangles, and the external supplier nodes are represented with circles.

In our model, we assume that each facility observes demand from its downstream stage and places orders to its suppliers in response to the observed demand. We assume that there is no time delay in ordering and there are no fixed ordering costs, and thus we employ a base-stock policy, so that every time a customer order arrives at the manufacturer, the external demand propagates immediately up the supply chain.

Since the demand is stochastic, the supply chain may need to carry inventory in order to meet customer demand sufficiently rapidly. Each stage in the network is a potential location for holding

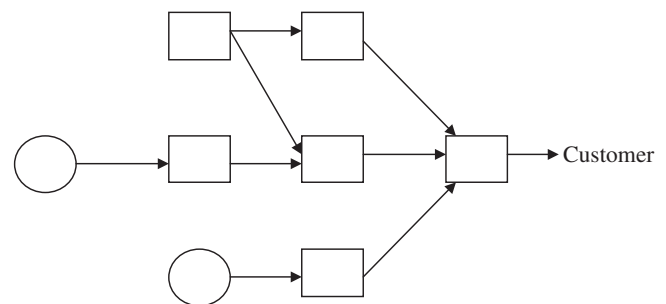


Fig. 1. Supply chain structure.

inventory of the item processed at that stage, so the firm can choose to stock intermediate inventory in this system instead of choosing to stock only finished goods at the downstream facility. As mentioned above, we assume that each stage in the model employs a base-stock inventory policy (possibly with a base-stock level of 0), and every time an order arrives, each facility starts the production of that type immediately, either to satisfy that order or to replenish inventory if that order is satisfied from the existing inventory.

Throughout this paper, for clarity of exposition we denote inventory in terms of the unit of time (e.g. days, hours, etc.) that the inventory covers. The exact amount of inventory can be determined using this time length, the exogenously specified service level, and the demand distribution. For example, 4 days worth of inventory at a 95% service level where demand is normally distributed with mean 2 and variance 1 per day is equal to $2 * 4 + \sqrt{4} * z_{0.95} = 11.92$. Holding cost is assumed to be in appropriate units for this notation, and inventory levels at a stage will be zero if the product is purely made-to-order at that stage.

Recall that in our model, each facility receives materials from upstream facilities, produces, and transfers finished items to downstream facilities. Since there are transshipment times between facilities, the finished product of facility k might be stored at the facility k as a finished product or it might be stored at downstream facility j as the component required for production at facility j . In addition, in our model, products might share common components, i.e. a component produced by a facility might be required downstream for the production of two or more different types of products.

We use the following additional notation throughout the paper:

- i : subscript used for product type.
- j, k : subscripts used for the facilities in the supply chain starting with $j = 1$ denoting the manufacturer.
- S : set of facilities that belong to the manufacturer.
- E : set of external suppliers.
- F_i : set of facilities that are in the production network of product type i .
- P_{ij} : set of facilities that are immediate predecessors to facility j in the production network of type i .

- r_o : arrival time of an external order o to the system.
- p_{ij} : processing time of the component required for product type i at facility j .
- t_{ikj} : shipment time of the component required for product type i from facility k to j .
- h_{ikj}^1 : unit raw material inventory holding cost at facility j for the component produced at the upstream facility k required for product type i .
- h_{ij}^2 : unit inventory holding cost for the finished component produced at facility j for product type i .
- c_i^d : cost of a unit increase in response time to orders for product type i .
- c_i^T : unit tardiness cost for product type i .
- f_{ij} for $\forall j \in E_i$: committed response time of orders from external supplier $j \in E$ for product type i .
- f_{ij} for $\forall j \in S$: aggregate lead time of orders for product type i up to completion of processing at facility $j \in S$, i.e. the lead time between the arrival of a customer of type i to the system and the completion of the component of that order at facility j .
- x_{ikj} : expected time that the safety stock of components stored at facility j that are received from upstream facility k for the production of type i will last.
- y_{ij} : expected time that the inventory of finished goods at facility j required for the production of product type i will last.

Note that when we consider single product cases of the model, we drop the subscript i .

Our goal is to quote short and reliable lead times without holding excessive inventory in the system. Thus, we attempt to minimize an objective function

$$\sum_{i=1}^K \left\{ \left(\sum_{j=1}^N h_{ij} E[I_{ij}] \right) + c_i^d E[d_i] + c_i^T E[W_i - d_i]^+ \right\} \quad (2.1)$$

consisting of inventory costs, lead-time costs, and tardiness costs. In particular, $E[I_{ij}]$, $E[d_i]$ and $E[W_i - d_i]^+$ denote the average inventory level at facility j , the average quoted lead time, and the average tardiness (the positive component of waiting time minus lead time) for product type i . For each product type i , the average cost of inventory at facility j , $h_{ij} E[I_{ij}]$ represents both components and

finished goods inventory so that

$$h_{ij}E[I_{ij}] = h_{ij}^2 v_{ij} + \sum_{k \in P_{ij}} h_{ikj}^1 x_{ikj}.$$

Clearly, to minimize (2.1), we need to coordinate lead-time quotation, sequencing and inventory management, and thus an optimal solution to this model would require simultaneous consideration of these three issues. However, as discussed above, this is a very difficult problem to solve, so we have elected to take a different approach. Observe that in an optimal offline solution to this model (that is, a solution to the version of this model when all problem data are deterministically known ahead of time), lead times are exactly equal to actual waiting times of jobs in the system, because in an offline model, we can solve the entire problem ahead of time and set due dates equal to completion times. Thus, the problem becomes equivalent to minimizing

$$\sum_{i=1}^K \left\{ \left(\sum_{j=1}^N h_{ij} E[I_{ij}] \right) + c_i^d E[W_i] \right\}.$$

Of course, in the online version of the problem (where data about job arrivals are not known until the jobs arrive), it is impossible to both minimize this function and set due dates equal to completion times, since due dates are assigned without knowledge of future arrivals, some of which may have to complete before jobs that have already arrived in an optimal schedule. (Note that the completion time of an order is equal to the waiting time of that order in the system plus the arrival time of that order to the system, and that the due-date of an order is equal to the lead time of that order plus the arrival time of that order to the system.) In our approach, to minimize (2.1), we first propose a scheduling approach designed to try to minimize the waiting time component of (2.1), and then, based on that schedule, we set inventory levels to minimize

$$\sum_{i=1}^K \left\{ \left(\sum_{j=1}^N h_{ij} E[I_{ij}] \right) + c_i^d E[W_i] \right\}.$$

Finally, for this schedule and set of inventory levels, we design a lead-time quotation approach that

generates lead-time estimates that are in general close to the actual waiting times, using only the information available at the arrival times of the orders.

3. Analysis and results

In previous work (Kaminsky and Kaya, 2008a,b) we considered significantly simpler serial two-stage supply chain structures, and developed effective due-date quotation and sequencing approaches, which we explain and generalize below for the relatively rich model presented in this paper. However, previous approaches developed for inventory control in these simpler models (Kaminsky and Kaya, 2008b) do not generalize well to this complex setting, so we instead use a different approach to approximate the optimal inventory levels that should be stored at each facility for each product type. In the following sections, we present our approach for finding these inventory levels, and the algorithms we use for scheduling and lead-time quotation. We first present our results for a model with a single product type, and then extend this approach to a multiple product type setting.

3.1. Single product type model

In this section, we assume that the firm produces only one type of product. For more than one product, to achieve our objective we first attempt to sequence the jobs to minimize the total waiting times of the products. However, since there is only one type of product in this case, we can simply schedule jobs in the order in which they arrive, the so-called first come first serve (FCFS) approach at all facilities.

In most of the literature of inventory placement in supply chains, in order to simplify the analysis, researchers assume deterministic lead times for the production of components at all the facilities (e.g. Magnanti et al., 2006; Graves and Willems, 2000). To approximate the optimal inventory levels, we initially make the same assumption. That is, we assume that processing times are deterministic and each server has no capacity restriction, so that any number of jobs can be processed simultaneously at any facility in the system. Thus, although the demand is stochastic, the orders do not effect each other and the waiting time of the jobs at all of the facilities are deterministic and known in advance.

Later, we will relax this assumption and generalize our results to facilities with capacity restrictions at each facility.

For an uncapacitated single stage system, the lead times, inventory values and processing times satisfy the relation $l = \max\{p - x, 0\}$ where p is the deterministic processing time required to process an order, l is the lead time and x is the length of time that the available safety stock will last. Since the production lead time is p and the demand can be satisfied from the inventory for x time units, the first order that arrives after the inventory is depleted at time x , can only be satisfied at time p , leading to a lead time $l = p - x$ assuming $p > x$. If $p < x$, then $l = 0$ since, in this case, we can satisfy all the demand from the inventory and replenish the inventory before it is depleted.

Using this relationship, we can write a linear program to approximate the optimal inventory levels at each facility. The objective function to be minimized is composed of the total inventory plus the lead-time costs, and the constraints ensure that the lead time for the production of a component at a facility is no less than the lead time required for the components to arrive at that facility plus the required processing time minus the safety stock values. We write the LP formulation of this model below, with x_{kj} , y_j and f_j for $\forall j \in S$ as decision variables, and with the downstream manufacturer as facility 1:

$$\begin{aligned} \text{Min} \quad & \sum_{j=1}^N \left[\sum_{k=1}^N \{h_{kj}^1 x_{kj}\} + h_j^2 y_j \right] + c^d f_1 \\ \text{s.t.} \quad & \max \left\{ \max_{k \in P_j} \{f_k + t_{kj} - x_{kj}, 0\} + p_j - y_j, 0 \right\} \\ & \leq f_j \quad \text{for } \forall j \in S, \\ & \text{All variables} \geq 0. \end{aligned} \tag{3.1}$$

Note that if there is a desired target lead time, instead of minimizing the total inventory plus lead-time costs, we can fix the lead time f_1 and minimize the total inventory costs using the LP model (3.1) above with the fixed lead time as a parameter instead of a variable.

If there are capacity constraints that limit the production at a stage, then there will be queues at each stage in the system and the lead times (i.e. the waiting times of orders in the system) will be

affected by the demand process even though the production times are deterministic. We assume that there is a single server (or $K < \infty$ servers) at facility j , and we model the operations at that facility as an M/D/1 queue and approximate the waiting time of orders at facility j by determining the expected waiting time of a job in the system with arrival rate λ_j and processing time p_j . The mean waiting time of jobs at facility j can be calculated by employing the fact that in an M/D/1 queue with an FCFS schedule, the expected waiting time of a job at the queue of facility j is $E[W_j] = p_j / (2(1 - p_j D_j)) + p_j / 2$. These values can then be used instead of p_j in the LP formulation (3.1).

After inventory levels at each facility are determined using the LP (3.1), lead time for an order o can be quoted using only the information available at the time of the arrival of that order, r_o . Observe that if there is inventory at a facility, the demand for components from that facility is immediately satisfied and the jobs do not have to wait for the processing at or before that facility. Also, note that since there are inventories in the system, another job l that is already in the system at time r_o might be used to satisfy order o . Thus, we need to find the completion time of the job l that will be used to satisfy order o to quote the lead time for order o .

For this model, since jobs are sequenced in the order in which they arrive at the system (FCFS), there is no need to consider future arrivals, and lead times can be accurately quoted by considering only the state of the system at the time of the arrival of an order. In other words, an order only has to wait for the processing of orders that are already at the system when that order arrives. Thus, in Algorithm 1, to quote a lead time for order o , we first observe the system at time r_o , check for inventories, find the job l that will be used to satisfy order o , and then calculate the remaining processing time of the job l . For this purpose, in Algorithm 1, we form a subgraph G' by including only those facilities in G where job l still requires processing. In Algorithm 1, w_{lj} denotes the estimated waiting time and d_{lj} denotes the estimated completion time of job l at facility j . Also, U_{lj} denotes the set of orders that arrived at facility j before order l , but have not yet been delivered to the successor of facility j at time r_o .

Algorithm 1:

Step 1: Form a new subgraph, G' , of the supply chain network G accounting for system inventories at time r_o using the following subroutine.
 Denote the manufacturer by node j and form the subgraph G' by adding node j and all of its incoming arcs (k,j) to G' .
 While \exists arc $(k,j) \in G'$ and node $k \notin G'$
 If node j has inventory of finished components:
 Set $p_j = 0$ and delete all the incoming arcs (k,j) to it in G' .
 Else if there is a job l in process or waiting in the queue to be processed to replenish inventory at facility j instead of to satisfy a previous order:
 Set $p_j = q_j^l$ where q_j^l is the remaining time of job l at facility j and delete all the incoming arcs (k,j) to that node in G' .
 Else
 For all the predecessors k of node j
 If node j has inventory of raw materials required from node k
 Delete arc (k,j) from G'
 End For
 End Else
 If \exists arc $(k,j) \in G'$ and node $k \notin G'$
 Set $j = k$ and add that node j and all of its incoming arcs into G'
 End While.
 Step 2: For $\forall j \in G'$: if $j \in S$, then put facility j in set F
 Step 3: For $\forall j \in G'$
 If $j \in E$ then $d_{lj} = f_j$
 If $j \in S$ and j has no predecessor, then
 $d_{lj} = \sum_{m \in U_{lj}} p_m + p_j$ and delete facility j from set F
 Step 4: For $\forall j \in G'$, if $j \in F$ and $k \notin F$ for $\forall k \in P_j$, then delete facility j from set F and calculate
 $w_{lj} = \max\{\sum_{m \in U_{lj}} p_m - \max_{k \in P_j}\{d_{lk} + t_{kj}\}, 0\} + p_j$
 $d_{lj} = \max_{k \in P_j}\{d_{lk} + t_{kj}\} + w_{lj}$
 Step 5: Stop if set F is empty and set $d_o = d_{l1}$ as the lead time for order o .
 Return to step 4 otherwise.

3.2. Multiple product types model

In this section, we consider a firm that produces multiple products with different characteristics (i.e. different processing times, arrival rates and supply chain architecture). As there are different products, in addition to inventory level and lead-time determination, we also design an algorithm to sequence jobs at each facility. Since products have different characteristics, the processing sequence will impact completion and lead times.

Recall that we represent the supply chain network by a directed and acyclic graph G with nodes representing stages of the production operations of the components and the arcs representing the flow of components. Since each product must be processed at a specified subset of the facilities in

the supply chain, in a specified order, we represent the production operations of each of the product i by a subgraph \tilde{G}_i .

For the multiple product model, linear program (3.2) is used to determine inventory levels. The notation is the same as in the previous linear program, but in this case, every variable/parameter also includes the subscript i to denote the product type. Each component is denoted by (i,j) , the facility j that the component is produced at and the product subscript i that the component is used in. However, recall that two different products i and l might require the same component that is produced at facility j . Thus, the notation (i,j) and (l,j) might denote the same component. We denote the set of these components by set C and the elements of this set are denoted in terms of (i,l,j) .

The second and the third constraints in the LP formulation (3.2) ensure that the amount of inventory for the components (i, j) and (l, j) are equal if $(i, l, j) \in C$ since they denote the same component. Note that if there are more than two products that require the same component, we can arbitrarily pick one of the products and equate all the others to that one when we write the last two constraints in the LP formulation (3.2).

$$\begin{aligned} \text{Min} \quad & \sum_{i=1}^K \sum_{j \in F_i} \left[\sum_{k \in F_i} \{h_{ikj}^1 x_{ikj}\} + h_{ij}^2 y_{ij} \right] + c_i^d f_{i1} \\ \text{s.t.} \quad & \max \left\{ \max_{k \in P_{ij}} \{ \max \{ f_{ik} + t_{ikj} - x_{ikj}, 0 \} \} + p_{ij} - y_{ij}, 0 \right\} \\ & \leq f_{ij} \quad \text{for } \forall i = 1 \dots K \text{ and } \forall j \in F_i, \\ & y_{ij} = y_{lj} \quad \text{for } \forall (i, l, j) \in C, \\ & x_{ikj} = x_{lkj} \quad \text{for } \forall (i, l, j) \in C \text{ and } \forall k \in P_{ij}, \\ & \text{All variables } \geq 0. \end{aligned} \tag{3.2}$$

In this case, if there are no capacity restrictions at any of the supply chain stages, there is no need for a scheduling algorithm since there is no capacity constraint and each arriving job is immediately processed at each stage without waiting. Thus, each product type can be analyzed separately and the problem will be equivalent to the single product type case.

However, when we model restricted capacity and assume that there is a single server (or $M < \infty$ servers) at a facility, there will be congestion and queues in the system. In this case, the waiting time of orders at facility j can be approximated using an approach similar to the one employed for the single product model. We once again find the expected waiting time of an order type i in the queue at facility j with arrival rate λ_{ij} and processing time p_{ij} . In this case, the expected waiting time for type i at facility j , $E[W_{ij}]$, depends on both the other product types at j , and the schedule used at that facility. The mean waiting time of jobs of type i at facility j , $E[W_{ij}]$, is determined, and used in the place of p_{ij} in the LP formulation (3.2). Depending on the particular scheduling approach used, $E[W_{ij}]$ is approximated using different existing results for expected waiting times of jobs in an M/D/1 queue with multiple product types. For example, see Wallstrom (1980) and Conway et al. (1967) for the calculation of the expected waiting times in the system in an M/G/1 queue with multiple product types for the FCFS and shortest processing time available (SPTA) scheduling rules. (In the SPTA

rule, whenever the server is available, the shortest job is taken from the queue and processed.)

A lead-time quotation approach similar to the one presented in Algorithm 1 is utilized. However, in this case, the jobs are not necessarily sequenced FCFS at each facility. Recall that, in order to minimize objective function (2.1), we would first like to minimize the waiting time component by employing an effective scheduling algorithm. Thus, we would like to find a scheduling algorithm to minimize the total waiting times, and since completion time of a job is equal to the arrival time of that job to the system plus the waiting time in the system, we would like to minimize the total completion times. Although the simplest case of this problem, the problem of minimizing total completion times at a single facility is known to be NP-Hard, Kaminsky and Simchi-Levi (2001) show that the SPTA rule is asymptotically optimal (i.e. optimal as the number of jobs go to ∞) for this problem. Under the SPTA heuristic, each time a job completes processing, the shortest available job which has yet not been processed is selected for processing. Also, note that this approach to sequencing does not take quoted lead times into account, and is thus easily implemented.

In Kaminsky and Kaya (2008a,b), SPTA at the supplier and FCFS at the manufacturer is found to be effective in minimizing total completion times for a two-facility supply chains. Indeed, this works was motivated by results of Xia et al. (2000) and Kaminsky and Simchi-Levi (1998), who proved that for a flow shop with m machines, if the processing times of a job on each of the machines are independent and exchangeable, processing the jobs according to the shortest total processing time

$$p_i = \sum_{j=1}^m p_{ij}$$

at the first facility and on an FCFS basis at the others is asymptotically optimal if all the release times are 0.

Thus, we consider sequencing orders using a mix of FCFS and SPTA sequences. If FCFS schedule were used at all facilities, then Algorithm 1 would be effective for quoting lead times. However, if an SPTA based schedule is used at a facility, then future arrivals have to be accounted for when quoting lead times, as these future arrivals might be processed before jobs currently in the queue, and delaying the delivery times of those jobs.

Based on these ideas, and motivated by our sequencing approach in Kaminsky and Kaya (2008a,b), we present the following scheduling algorithm for this system. For each product type i , on the production graph \bar{G}_i , we find the longest path from the manufacturer to the end supplier, where arc lengths between two nodes j and k is $l_{ijk} = p_{ij} + t_{ijk}$ and then we find the total processing time of each product type i by summing all the arc lengths on this path. Then, we schedule the jobs according to shortest total processing times at the facilities that do not have any internal suppliers in the whole supply chain network and use an FCFS schedule at all the other facilities. Based on this schedule, we design a lead-time quotation algorithm using the approach introduced in Kaminsky and Kaya (2008a).

We detail this scheduling and lead-time quotation approach in Algorithm 2. Recall that, as in Algorithm 1, since there are inventories in the system, another job l of type i that is already in the

system at time r_o might be used to satisfy order o of type i . We use the following additional notation for this case:

- n : number of jobs that will arrive to the system.
- M_i : set of product types that are going to be scheduled before product type i .
- $\psi_i = \sum_{k \in M_i} \delta_k$: probability that an arriving job is going to be scheduled before job type i .
- $\mu_{ij} = \sum_{k \in M_i} \{\delta_k p_{kj}\}$: expected processing time of a job that is going to be scheduled before job type i at facility j .
- U_{lj} : set of jobs in front of job l in the queue of facility j at time r_o .
- N_{lj} : approximated number of orders that will arrive after order o but will be scheduled before job l at facility j .
- w_{lj} : approximated waiting time of job l at facility j .
- d_{lj} : approximated completion time of job l at facility j .

Algorithm 2:

Scheduling:

Step 1: Find the total time required to process each product type i (i.e. the longest path from the suppliers at the end of the chain to the manufacturer for product type i where arc length between two nodes j and k is $l_{ijk} = p_{ij} + t_{ijk}$) and denote it by T_i .

Step 2: Define set L to be set of facilities that have no internal supplier.

Whenever a facility $j \in L$ is available, process the job type i with shortest T_i and use FCFS schedule at other facilities $j \notin L$.

Lead-Time Quotation:

Step 3: Form the subgraph G'_i of \bar{G}_i as in Algorithm 1 and set $N_{lj} = 0$ for $\forall j$

Step 4: For $\forall j \in G'_i$: if $j \in S$, then put facility j in set F

Step 5: For $\forall j \in G'_i$

If $j \in E$ then $d_{lj} = f_j$

If $j \in S$ and j has no predecessor in the production network of type i , then delete facility j from set F and calculate

$$w_{lj} = \sum_{m \in U_{lj}} p_{mj}$$

$$\text{slack}_{lj}^i = \begin{cases} \min \left\{ \frac{w_{lj} \psi_i^i \mu_{ij}}{D - \psi_i \mu_{ij}}, (n - l) \psi_i \mu_{ij} \right\} & \text{if } D - \psi_i \mu_{ij} > 0 \\ (n - l) \psi_i \mu_{ij} & \text{otherwise} \end{cases}$$

$$N_{lj} = \text{slack}_{lj}^i / \mu_{ij}$$

$$d_{lj} = p_{ij} + w_{lj} + \text{slack}_{lj}^i$$

Step 6: For $\forall j \in G'_i$, if $j \in F$ and $k \notin F$ for $\forall k \in P_{ij}$, then delete facility j from set F and if facility $j \in L$, then calculate

$$w_{lj} = \max \left\{ \sum_{m \in U_{lj}} p_{mj} - \max_{k \in P_{ij}} \{d_{lk} + t_{lk}\}, 0 \right\}$$

$$\text{slack}_{lj} = \begin{cases} \min \left\{ \frac{w_{lj}\psi^i\mu_{ij}}{D - \psi_i\mu_{ij}}, (n - l)\psi_i\mu_{ij} \right\} & \text{if } D - \psi_i\mu_{ij} > 0 \\ (n - l)\psi_i\mu_{ij} & \text{otherwise} \end{cases}$$

$$N_{lj} = \text{slack}_{lj} / \mu_{ij}$$

$$d_{lj} = \max_{k \in P_{ij}} \{d_{lk} + t_{lk}\} + p_{ij} + w_{lj} + \text{slack}_{lj}$$

If facility $j \notin L$

$$N_{lj} = \max_{k \in P_{ij}} N_{lk}$$

$$w_{lj} = \max \left\{ \sum_{m \in U_{lj}} p_{mj} + \max_{k \in P_{ij}} \{N_{lk}\}\mu_{ij} - \max_{k \in P_{ij}} \{d_{lk} + t_{lk}\}, 0 \right\}$$

$$d_{lj} = \max_{k \in P_{ij}} \{d_{lk} + t_{lk}\} + p_{lj} + w_{lj}$$

Step 7: Stop if set F is empty and set $d_o = d_{l1}$ as the lead time for order o .

Return to step 6 otherwise.

4. Computational study

We perform a variety of computational experiments in order to evaluate the performance of our algorithms. However, to the best of our knowledge there is no previous computational study that considers the models that we consider, so we compare our results with lower bounds on optimal solutions for our models and with the traditionally used pure MTO and MTS strategies for these same problems. We complete an extensive simulation study utilizing a supply chain network with a variety of different processing times, transshipment times between facilities, unit holding costs and unit waiting costs and implement our heuristics in C++. Whenever needed, we solve the LP model (3.1) or (3.2) using ILOG AMPL/CPLEX 7.0.

We use the supply chain network as shown in Fig. 2. The meanings of the numbers in Fig. 2 are explained in Fig. 3. In Fig. 2, S_j denotes the facilities that belong to the same firm and E_j denotes the external suppliers.

4.1. Effect of inventory positioning for uncapacitated systems

We first consider the uncapacitated single product case where the waiting time of a job at facility j is deterministic and equal to p_j (e.g. an infinite server model). As we discussed in Section 2, since the demand is stochastic, the firm needs to keep some safety stock to achieve the desired service level. In Table 1, we compare the optimal objective values of LP (3.1) with inventory held at every facility as opposed to the same objective function when

holding no inventory at all or holding only finished good inventories. The ratios of the objective function values of LP (3.1) with the combined strategy over the costs with pure MTS and MTO strategies for different combinations of inventory holding cost at the manufacturer, h , and unit lead-time cost, c^d are shown in Table 1. The values in Table 1 illustrate the importance of effective inventory placement in supply chains, although these specific values clearly depend on the holding costs at each of the facilities. In the following sections, we make a similar comparison using a simulation of the system.

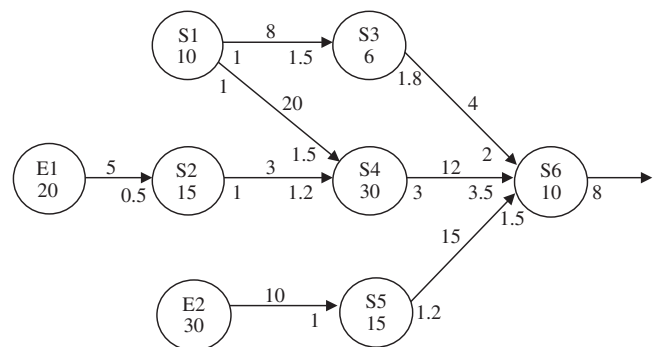


Fig. 2. Supply chain network example.

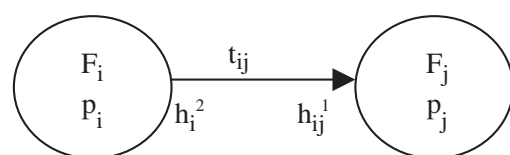


Fig. 3. Explanation of the values.

Table 1
Comparison of combined strategy with pure strategies for the incapacitated system

	$h = 8, c^d = 11$	$h = 8, c^d = 10$	$h = 8, c^d = 5$	$h = 4, c^d = 5$	$h = 2, c^d = 5$
$Z_{\text{MTO-MTS}}/Z_{\text{MTS}}$	0.519	0.519	0.444	0.808	0.892
$Z_{\text{MTO-MTS}}/Z_{\text{MTO}}$	0.377	0.415	0.710	0.647	0.357

Table 2
Comparison of combined strategy with pure strategies for a single server model

	$h = 8, c^d = 11$	$h = 8, c^d = 10$	$h = 8, c^d = 5$	$h = 4, c^d = 5$	$h = 2, c^d = 5$
$Z_{\text{MTO-MTS}}/Z_{\text{MTS}}$	0.476	0.476	0.422	0.784	0.901
$Z_{\text{MTO-MTS}}/Z_{\text{MTO}}$	0.346	0.381	0.675	0.628	0.360

As we have discussed, firms traditionally use either an MTO strategy (with no inventory) or an MTS strategy keeping only finished goods inventory. However, as we show in Table 1 using the objective function of LP (3.1), a combined strategy is clearly much better than these pure strategies for minimizing the total inventory plus lead-time costs for this uncapacitated system. For example, when holding costs are as given in Fig. 2 with the exception of holding cost for finished goods at the last facility, which is 4 ($h = 4$) and $c^d = 5$, with a pure MTS strategy (with lead time 0) we need to keep a finished goods inventory of 95 units with a cost of 380. With a pure MTO strategy, the lead time will be 95 and the cost is 475. However, with a combined strategy with $y_{S_1} = 10, y_{S_2} = 15, y_{S_4} = 12, y_{S_6} = 40, x_{E_1,S_2} = 25, x_{E_2,S_5} = 40, x_{S_1,S_4} = 20, x_{S_2,S_4} = 3$, the total cost will be 307.1. The cost of the combined strategy is significantly lower than the costs of either pure strategy.

We consider another example, in which the target lead time is 30 and we try to achieve this lead time by holding inventory. If we only keep finished goods inventory, then $y_{S_6} = 65$ and the total inventory cost is 260. However, by keeping inventory at other facilities, with the same lead time, the total inventory cost can be decreased to 187.1 with $y_{S_1} = 10, y_{S_2} = 15, y_{S_6} = 22, x_{E_1,S_2} = 25, x_{E_2,S_5} = 28, x_{S_1,S_4} = 20, x_{S_2,S_4} = 3$. In addition, we will be able to cut the lead time by half to 15 at a cost of 247.1 which is even less than the cost with the initial strategy.

We see that as the unit inventory holding cost at the manufacturer, h , increases or the unit lead-time cost, c^d , decreases, an MTO strategy gives results closer to those of the combined strategy, and as h

decreases or c^d increases, the MTS strategy becomes more effective. Also, increasing c^d (or h) beyond a certain level does not impact the system because the lead time (or finished goods inventory) in the optimal solution is optimally set to 0 for c^d (or h) high enough, so increasing c^d (or h) further will not affect the system (assuming that everything else remains the same).

4.2. Effect of inventory positioning with congestion effects for the single product type model

Next, we investigate the single product capacitated case assuming that there is a single server at each facility. To employ this approach, we first find the mean waiting time of a job at each facility and use these values in place of the deterministic value p_j in the LP formulation (3.1). To approximate the mean waiting times of jobs at each facility, we assume an M/D/1 queue. That is, the inter-arrival times are taken to be independent and exponentially distributed with mean 40 and the processing times and transshipment times are deterministic and given in Fig. 2. For this case, in Table 2 the ratios of the objective values in formulation (3.1) using the combined strategy are compared to those of pure strategies.

Note that these costs are found using our LP formulation (3.1) and the approximations described above. In reality, however, inventory values fluctuate over time due to the stochastic nature of demand, and each job has a different lead time, depending on the actual congestion in the system at the time of that job's arrival. Thus, we also simulate this stochastic system, assuming that there is a single server at each facility with deterministic processing

times and other values as given in Fig. 2, and independent and exponentially distributed inter-arrival times with mean 40. Using $n = 5000$ jobs in our heuristics, we compare the average of 10 runs of the objective values

$$\sum_{j=1}^N h_j E[I_j] + c^d E[d] + c^T E[W - d]^+$$

in Table 3 where $E[I_j]$, $E[d]$ and $E[W - d]^+$ denote the average inventory level at facility j , the average quoted lead time and the average tardiness, respectively.

For this case, we first find the initial inventory levels using our LP formulation and then complete our simulations starting with these inventory values. In our simulations, each time a customer order arrives at the system, we quote a lead time using Algorithm 1 and start the production of a new product either to satisfy that order or to replenish the inventory. We keep track of the fluctuation of the inventory levels over time at each facility and calculate the average inventory costs with the average quoted lead time and tardiness costs. We compare the objective functions

$$\sum_{j=1}^N h_j E[I_j] + c^d E[d] + c^T E[W - d]^+$$

with this combined model to the same objective function with pure MTO and MTS models. The initial inventories are all 0 for the MTO model and there is only finished goods inventory for the MTS model. The ratios of the costs are presented in Table 3 for different h and c^d combinations with $c^T = 12$.

To assess the effectiveness of lead-time quotation Algorithm 1, we also compare the lead times quoted for this single type system to the actual waiting times of the jobs in the system using $c^d = 5$ and $c^T = 7$. Let

$$Z_{LT}^n = \sum_{i=1}^n \{c^d d_i + c^T (W_i - d_i)^+\}$$

denote the total lead time plus tardiness costs,

$$Z_{DD}^n = Z_{LT} + \sum_{i=1}^n r_i$$

denote the total due dates plus tardiness costs (recalling that lead time plus release time equals due date),

$$Z_W^n = \sum_{i=1}^n \{c^d W_i\}$$

denote the total waiting times of the jobs in the system and

$$Z_C^n = Z_W + \sum_{i=1}^n r_i$$

denote the total completion times of the jobs (recalling that waiting time plus release time equals completion time). We present ratios for these values for different number of jobs, n , in Table 4. As we see in Table 4, the lead times quoted with Algorithm 1 are very close to the actual waiting times and Z_{DD}^n approach Z_C^n as n increases.

4.3. Effectiveness of the algorithms for multiple product types

We also simulate a system with multiple product types to assess the effectiveness of our algorithms. In addition to the single product type we considered in the previous section, now we use four additional product types with arrival probabilities and processing times as shown in Table 5. In our model, all of the products have the same supply chain architec-

Table 4
Comparison of lead times and due dates to actual waiting times and completion times

	$n = 10$	$n = 100$	$n = 1000$	$n = 5000$
Z_W^n / Z_{LT}^n	0.891	0.950	0.964	0.962
Z_C^n / Z_{DD}^n	0.925	0.967	0.992	0.996

Table 3
Simulation analysis of combined strategy compared to pure strategies

	$h = 8, c^d = 11$	$h = 8, c^d = 10$	$h = 8, c^d = 5$	$h = 4, c^d = 5$	$h = 2, c^d = 5$
$Z_{MTO-MTS} / Z_{MTS}$	0.833	0.833	0.810	0.892	0.944
$Z_{MTO-MTS} / Z_{MTO}$	0.723	0.785	0.871	0.827	0.651

ture and require processing at all of the facilities in Fig. 2. Also, we use equal transshipment times and inventory holding costs for each of the products (shown in Fig. 2). Inter-arrival times are independent and exponentially distributed with mean 40. We simulate 10 trials, each with $n = 5000$ jobs, and present the averages of these runs in the following tables.

We first consider three different product types using the first three product types in Table 5 and then five different product types considering all the products in Table 5. For the arrival probability of type $l = 1$, we use 0.55 in the three-product-type model instead of 0.2 so that the sum of the arrival probabilities of all types will be 1.

We use the scheduling and lead-time quotation approach detailed in Algorithm 2. To explore the effectiveness of the SPTA schedule, we compare the total waiting times of jobs using the SPTA schedule to that of a lower bound. If we consider only the bottleneck facility (the facility with slowest processing rate), use an SPTA schedule with preemption in that facility, and assume that the waiting time of any job in the queue of any other facility is zero, then the total weighted waiting time of jobs at this system will be a lower bound for those in our model. Let Z_{LB} denote the lower bound for the total weighted waiting times of jobs in the system and

$$Z_{SPTA} = \sum_{i=1}^n \{c^d W_i\}$$

Table 5
Arrival probabilities and mean processing times of product types

	δ_l	p_{E_1}	p_{E_2}	p_{S_1}	p_{S_2}	p_{S_3}	p_{S_4}	p_{S_5}	p_{S_6}
$l = 1$	0.2	20	30	10	15	6	30	15	10
$l = 2$	0.3	15	10	25	5	45	15	20	10
$l = 3$	0.15	5	15	10	10	15	20	25	15
$l = 4$	0.25	10	20	30	15	10	5	10	20
$l = 5$	0.1	20	5	5	10	5	10	30	15

Table 6
Comparison of SPTA schedule and lead-time quotation with the lower bound for $K = 3$ and 5 product types

$K = 3$	$n = 10$	$n = 100$	$n = 1000$	$n = 5000$	$K = 5$	$n = 10$	$n = 100$	$n = 1000$	$n = 5000$
Z_{LB}/Z_{SPTA}	0.962	0.813	0.847	0.833		0.859	0.790	0.822	0.816
Z_{SPTA}/Z_{LT}	0.874	0.933	0.952	0.947		0.941	0.922	0.925	0.931
Z_{LB}/Z_{LT}	0.848	0.753	0.802	0.786		0.807	0.735	0.766	0.751

denote the total weighted waiting times with the SPTA-based schedule. The comparison of the total waiting times resulting from the use of our heuristic to that of the lower bound is presented in Table 6 for different numbers of jobs.

Also, to explore the effectiveness of the lead-time quotation portion of Algorithm 2, we present the ratios of the total quoted lead times plus tardiness costs,

$$Z_{LT} = \sum_{i=1}^n \{c^d d_i + c^T (W_i - d_i)^+\}$$

for this case to Z_{SPTA} and Z_{LB} in Table 6 using equal weights for different product types, $c^d = 5$ and $c^T = 7$. Observe that Z_{SPTA} is a lower bound on the objective associated with applying Algorithm 2 with an SPTA-based schedule, and Z_{LB} is a lower bound on applying Algorithm 2 with any schedule. Observe that the difference between Z_{SPTA} and the lower bound is less than 20% and the lead-time quotation algorithm gives results that are less than 7% worse than Z_{SPTA} . We can conclude that the lead-time quotation part of Algorithm 2 is effective for quoting short and reliable lead times when the SPTA schedule is used. However, the difference between the lower bound and Z_{SPTA} is larger than we would like. This is probably because this lower bound considers only the bottleneck facility and allows preemption, and thus may not be very tight. Indeed, developing a tighter lower bound would be useful, and is a problem we leave for future research. As our bound is likely not tight, to develop an understanding of the effectiveness of the SPTA-based scheduling algorithm for our problem, in Table 7, we also compare the total cost in our model using an SPTA version of Algorithm 2 with the total costs obtained when a commonly used schedule, FCFS policy, is employed to schedule the jobs.

In Table 7, using $c^T = 12$ and different values for h and c^d , we present the ratios of the total

Table 7
Comparison of combined strategy with pure strategies for multiple product type models

$K = 3$	$h = 8, c^d = 11$	$h = 8, c^d = 10$	$h = 8, c^d = 5$	$h = 4, c^d = 5$	$h = 2, c^d = 5$
$Z_{MTO-MTS}/Z_{MTS}$	0.732	0.732	0.695	0.887	0.936
$Z_{MTO-MTS}/Z_{MTO}$	0.603	0.666	0.851	0.789	0.628
$Z_{SPTA-LTQ}/Z_{FCFS-LTQ}$	0.934	0.943	0.918	0.930	0.956
$K = 5$	$h = 8, c^d = 11$	$h = 8, c^d = 10$	$h = 8, c^d = 5$	$h = 4, c^d = 5$	$h = 2, c^d = 5$
$Z_{MTO-MTS}/Z_{MTS}$	0.762	0.762	0.717	0.869	0.939
$Z_{MTO-MTS}/Z_{MTO}$	0.693	0.724	0.871	0.812	0.687
$Z_{SPTA-LTQ}/Z_{FCFS-LTQ}$	0.861	0.885	0.874	0.891	0.925

Table 8
Effect of demand rate and congestion level in the system

Demand rate	Congestion level	$\frac{Z_{MTO-MTS}}{Z_{MTS}}$	$\frac{Z_{MTO-MTS}}{Z_{MTO}}$	$\frac{Z_{SPTA}}{Z_{FCFS}}$	$\frac{Z_{LB}}{Z_{SPTA}}$	$\frac{Z_{SPTA}}{Z_{LT}}$
1/20	0.99	0.898	0.685	0.852	0.795	0.927
1/25	0.80	0.869	0.812	0.891	0.816	0.921
1/30	0.66	0.878	0.856	0.902	0.834	0.942
1/35	0.57	0.862	0.887	0.917	0.867	0.949
1/40	0.50	0.857	0.912	0.945	0.901	0.963

costs

$$\sum_{i=1}^K \left\{ \left(\sum_{j=1}^N h_{ij} E[I_{ij}] \right) + c_i^d E[d_i] + c_i^T E[W_i - d_i]^+ \right\}$$

found employing our heuristics (that is, using (3.2) to find inventory levels and Algorithm 2 for scheduling and lead-time quotation) to that of the total costs with pure MTO and MTS strategies and the total costs using an FCFS schedule. As we see in Table 7, the costs with the combined strategy is about 20% less on average than the pure strategies. Also, using an SPTA-based schedule leads to about a 10% cost decrease over using an FCFS schedule. Finally, we observe that as c^d decreases, the combined system moves toward an MTO system while as h decreases, the MTS system gives better results.

We also analyze the impact of the congestion level on system performance. We present the results in Table 8 using five different product types with the same supply chain structure and with arrival probabilities and processing times as given in Table 5 using parameters $h = 4$, $c^d = 5$ and $c^T = 7$. We increase the demand rate, and thus the congestion level at each facility, gradually from $\frac{1}{40}$ to $\frac{1}{20}$. The congestion levels in Table 8 denote the

congestion levels at the bottleneck facility (i.e. the most congested facility) calculated using the arrival probabilities and processing times in Table 5 and the demand rates in Table 8. Observe that the MTS system gives better results as congestion increases and the MTO system performs significantly better if the congestion decreases. Also, observe that the SPTA-based schedule as explained in Algorithm 2 performs much better than the FCFS schedule as the congestion increases, although the performance of the SPTA-based schedule diverges from the lower bound as the congestion increases. In addition, in the last column of Table 8, we compare the effectiveness of only the lead-time quotation component of Algorithm 2 by comparing

$$Z_{SPTA} = \sum_{i=1}^n \{c^d W_i\}$$

and

$$Z_{LT} = \sum_{i=1}^n \{c^d d_i + c^T (W_i - d_i)^+\},$$

and see that the lead-time quotation algorithm performs very well even if the congestion level is very high. This is primarily due to the fact that when

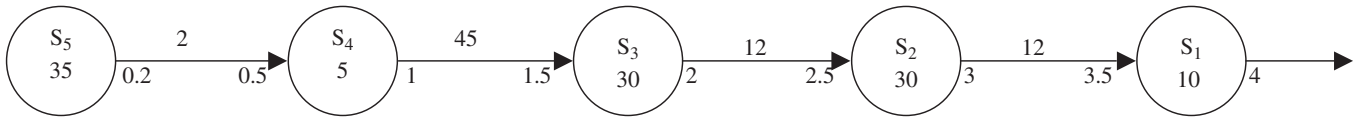


Fig. 4. Structure 1: 4-supplier flow shop model.

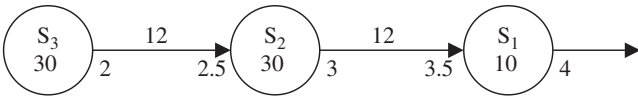


Fig. 5. Structure 2: 2-supplier flow shop model.

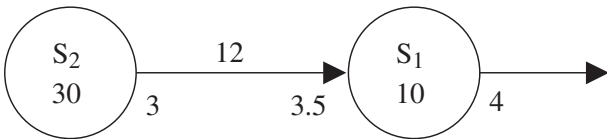


Fig. 6. Structure 3: single supplier, single manufacturer model.

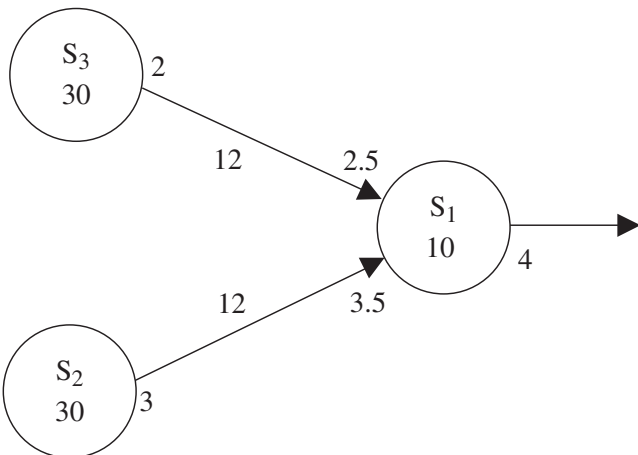


Fig. 7. Structure 4: 2 parallel supplier model.

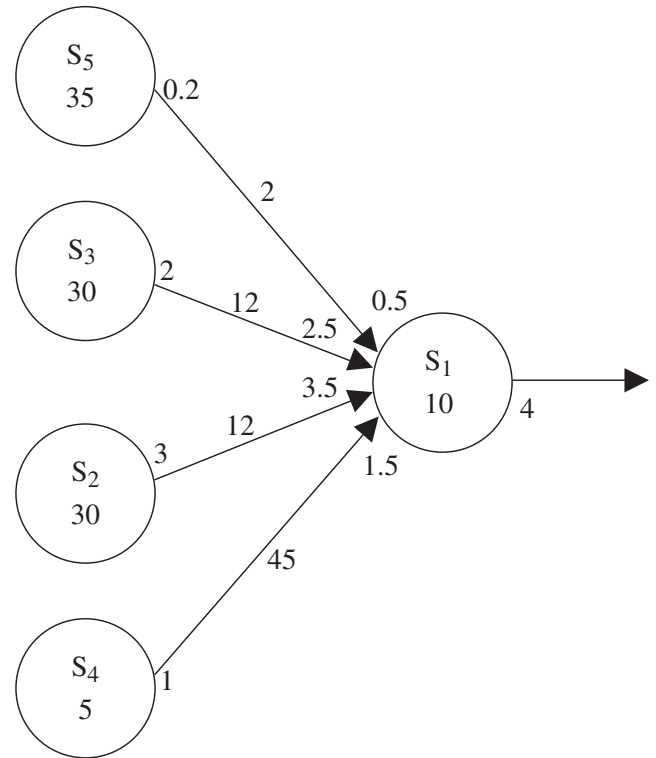


Fig. 8. Structure 5: 4 parallel supplier model.

congestion is high, the inventory levels are also high, so that when an order arrives, it can frequently be satisfied from inventory so there is no error in lead-time quotation.

4.4. Effect of supply chain structure on the system

We also study the effect of supply chain structure on the system and on the effectiveness of our heuristics. For this purpose, we consider the supply chain structures shown in Figs. 4–8, moving from serial to parallel facility models. We complete a series of simulations using these supply chain structures. We use the appropriate LP formulation to determine target inventory levels, and then use

Algorithm 2 during the simulation to schedule and quote due dates. In Table 10, we compare the total costs

$$Z = \sum_{i=1}^K \left\{ \left(\sum_{j=1}^N h_{ij} E[U_{ij}] \right) + c_i^d E[d_i] + c_i^T E[W_i - d_i]^+ \right\}$$

when our heuristic is used to the total costs with pure MTO and MTS strategies, and the total costs using an FCFS schedule with a mixed strategy.

Processing times for product type $l = 1$, transshipment times, and unit holding costs, are shown on the graphs. We consider four additional product types with the arrival probabilities and processing times shown in Table 9. The transshipment and unit holding costs for different product types are equal and as shown on the graphs. The order inter-arrival times are taken to be independent and exponentially distributed with mean 40 and unit tardiness cost $c^T = 8$. We consider two possible unit lead-time costs, $c^d = 5$ and 3.3. Each entry in Table 10 is the

average of 10 simulation runs with $n = 5000$ jobs in each run.

As can be seen in Table 10, the combined MTO–MTS approach performs significantly better than pure MTO or MTS approaches as the number of facilities in series increases. Indeed, as the number of suppliers in series increases, the combined model performs better than pure models with the difference increasing to more than 50%. For a serial supply chain, we see that the facilities employ a “all or nothing” strategy. That is, a facility either carries all the inventory required to decouple the upstream facilities from the downstream ones or it carries no inventory at all, and there is thus effectively a boundary line between upstream facilities operating MTS, and downstream facilities operating MTO. Also, if we add a new supplier in series on the upstream side of this boundary line, then that supplier uses an MTS strategy in the optimal solution provided that the inventory holding cost at that supplier is less than the inventory holding cost at the successor of that supplier. For example, the boundary for structure 3 with $c^d = 3.3$ is located between the supplier and the manufacturer—the manufacturer uses an MTO strategy and the supplier uses an MTS strategy. When we add the third facility as in structure 2, we see that facility 3 also uses an MTS strategy in the optimal solution. In addition, we see from structure 1 that facilities 4 and 5 also use MTS strategies since they are also on the upstream side of this boundary.

Table 9
Arrival probabilities and mean processing times of product types

Pr. type	δ_l	p_5	p_4	p_3	p_2	p_1	Pr. type	δ_l	p_5	p_4	p_3	p_2	p_1
$l = 2$	0.3	15	45	20	20	30	$l = 4$	0.25	40	30	15	15	40
$l = 3$	0.15	10	5	25	25	20	$l = 5$	0.1	10	15	5	5	10

Table 10
Effect of supply chain structure on the system

$c^d = 5$	ST 1	ST 2	ST 3	ST 4	ST 5	$c^d = 3.3$	ST 1	ST 2	ST 3	ST 4	ST 5
$Z_{\text{MTS-MTO}}$	0.533	0.766	0.896	1	0.966	0.530	0.741	0.851	0.895	0.814	
$\frac{Z_{\text{MTS}}}{Z_{\text{MTS-MTO}}}$	0.448	0.627	0.731	0.872	0.785	0.627	0.884	0.972	1	0.954	
$\frac{Z_{\text{MTO}}}{Z_{\text{SPTA-LTQ}}}$	0.762	0.957	0.891	0.829	0.802	0.786	0.934	0.818	0.753	0.791	
$\frac{Z_{\text{FCFS-LTQ}}}{Z_{\text{FCFS-LTQ}}}$											

For the structures with suppliers in parallel (structures 4 and 5), observe that using the appropriate pure strategy (MTS or MTO depending on the parameters) leads to performance quite close to that of the combined strategy. This is because the lead time of a job depends on the maximum lead time of the suppliers in parallel. If there is only one bottleneck supplier, then holding inventory at that single supplier will decrease the lead time for the system until the bottleneck supplier’s lead time is balanced with the lead times of the other suppliers. However, after the lead times are balanced, carrying more inventory at only one of the suppliers does not help at all and we need to hold inventory at every supplier to decrease the lead time for the system, leading to high inventory costs. Thus, rather than hold inventory at each of these suppliers, either holding only finished goods inventory at the manufacturer or holding no inventory at all tends to be more profitable, depending on the parameters. For example, for structure 4, using a pure MTS strategy is optimal when $c^d = 5 > h_{S_1} = 4$ and a pure MTO strategy is optimal when $c^d = 3.3 < h_{S_1} = 4$ since carrying inventory at only one of the suppliers does not help the system.

We also found that Algorithm 2 which is based on SPTA schedule performs about 15% better on average than the algorithm in which we schedule all the jobs according to FCFS at all the facilities and quote lead times accordingly. Recall that the difference in performance appears to depend primarily on the processing times of the product types at different facilities, and appears not to be significantly impacted by supply chain structure.

5. Conclusion

In this paper, we consider stylized models of complex MTO–MTS supply chains in a stochastic,

multi-item environment and designed effective algorithms for inventory placement, job scheduling, and lead-time quotation problems. Through computational analysis, we observe that our heuristics perform very well in many cases and can improve system performance dramatically. We observe that combined MTO–MTS systems perform significantly better than pure MTO or MTS systems, more than 50% better in some cases, and that an SPTA-based algorithm for scheduling the jobs performs much better than the generally used FCFS approach, especially for congested systems. We also observe that the MTS system gives better results for more congested systems and MTO performs better as the congestion level in the supply chain decreases.

We also explore the effect of supply chain structure and several other system parameters on supply chain performance. We see that the combined MTO–MTS approach performs significantly better than pure MTO or MTS approaches as the number of facilities in series increases and as the number of parallel facilities decreases. In serial supply chains, a facility either carries all the inventory required to decouple the upstream facilities from the downstream ones or it carries no inventory at all. Thus, in serial supply chains, the initial upstream stages operate as MTS systems and the later stages operate as MTO systems. However, for supply chains with parallel suppliers, the mixed MTS/MTO strategy has less significant advantage over the appropriate pure strategy because with these system structures, carrying more inventory at only one of the suppliers does not help the system and it is necessary to hold inventory at every supplier to decrease the lead time for the system, leading to high inventory costs. Thus, rather than holding inventory at each of these suppliers, either holding only finished goods inventory at the manufacturer (a pure MTS system) or holding no inventory at all (a pure MTO system) tends to be very effective. We also observe that the effectiveness of SPTA-based scheduling algorithms over the those based on FCFS scheduling approaches depends primarily on the processing times of the product types at different facilities, and appears not to be significantly impacted by supply chain structure. Thus, overall, it is especially important to consider employing a mixed MTO–MTS strategy for serial supply chains, and to employ an SPTA-based scheduling and lead-time quotation algorithm like the one we propose when the system is congested.

Of course, these are stylized models, and real-world supply chains have many complex characteristics that are not captured by these models. Nevertheless, this is to the best of our knowledge, the first study that explores inventory positioning, scheduling and lead-time quotation together in the context of relatively complex supply chains, and we believe that our qualitative insights will apply to more detailed models, and to real-world systems.

In the future, we intend to consider more complex functions of lead time in the objective function. We will also consider systems in which the manufacturer does not have to accept all orders and has the option to reject certain orders, and systems in which the customers might choose not to place an order depending on the quoted lead time. We can also incorporate pricing and capacity decisions into these models, and analyze contracts and gaming strategies for these systems. In each case, our focus will be on developing strategies for system design, and for scheduling and lead-time quotation.

References

- Axsater, S., 1993. Continuous review policies for multi-level inventory systems with stochastic demand. In: Graves, S.C., Rinnoy Kan, A.H., Zipkin, P.H. (Eds.), *Handbooks in Operation Research and Management Science*, vol. 4. *Logistics of Production and Inventory*. North-Holland Publishing Company, Amsterdam, The Netherlands (Chapter 4).
- Conway, R., Maxwell, W., Miller, L., 1967. *Theory of Scheduling*. Addison-Wesley Publishing Company, Reading, MA.
- Diks, E.B., de Kok, A.G., Lagodimos, A.G., 1996. Multi-echelon systems: A service measure perspective. *European Journal of Operational Research* 95, 241–263.
- Ettl, M., Feigin, G.E., Lin, G.Y., Yao, D.D., 2000. A supply network model with base-stock control and service requirements. *Operations Research* 48.
- Federgruen, A., 1993. Centralized planning models for multi-echelon inventory systems under uncertainty. In: Graves, S.C., Rinnooy Kan, A.H., Zipkin, P.H. (Eds.), *Handbooks in Operations Research and Management Science*, vol. 4. *Logistics of Production and Inventory*. North-Holland Publishing Company, Amsterdam, The Netherlands (Chapter 3).
- Glasserman, P., Tayur, S., 1995. Sensitivity analysis for base-stock levels in multi-echelon production-inventory systems. *Management Science* 41, 263–281.
- Graves, S.C., Willems, S.P., 1996. Strategic safety stock placement in supply chains. In: *Proceedings of the 1996 MSOM Conference*, Hanover, NH.
- Graves, S.C., Willems, S.P., 2000. Optimizing strategic safety stock placement in supply chains. *Manufacturing and Service Operations Management* 2, 68–83.

- Graves, S.C., Willems, S.P., 2003. Erratum: Optimizing strategic safety stock placement in supply chains. *Manufacturing and Service Operations Management* 5, 176–177.
- Inderfurth, K., 1991. Safety stock optimization in multi-stage inventory systems. *International Journal of Production Economics* 24, 103–113.
- Inderfurth, K., 1993. Valuation of leadtime reduction in multi-stage production systems. In: Fandel, G., Gullledge, T., Jones, A. (Eds.), *Operation Research in Production Planning and Inventory Control*. Springer, Berlin, Germany, pp. 413–427.
- Inderfurth, K., 1994. Safety stocks in multistage, divergent inventory systems: A survey. *International Journal of Production Economics* 35, 321–329.
- Inderfurth, K., Minner, S., 1998. Safety stocks in multi-stage inventory systems under different service measures. *European Journal of Operational Research* 106, 57–73.
- Kaminsky, P., Kaya, O., 2008a. Scheduling and due date quotation in a MTO supply chain, submitted for publication.
- Kaminsky, P., Kaya, O., 2008b. An analysis of a combined make-to-order/make-to-stock system. *IIE Transactions*, to appear.
- Kaminsky, P., Simchi-Levi, D., 1998. Probabilistic analysis and practical algorithms for the flow shop weighted completion time problem. *Operations Research* 46, 872–882.
- Kaminsky, P., Simchi-Levi, D., 2001. Probabilistic analysis of an on-line algorithm for the single machine completion time problem with release dates. *Operations Research Letters* 29, 141–148.
- Lee, H.L., Billington, C., 1993. Material management in decentralized supply chains. *Operations Research* 41, 835–847.
- Magnanti, T.L., Shen, Z.M., Shu, J., Simchi-Levi, D., Teo, C.P., 2006. Inventory placement in acyclic supply chain networks. *Operations Research Letters* 34, 228–238.
- Minner, S., 1997. Dynamic programming algorithms for multi-stage safety stock optimization. *OR Spektrum* 19, 261–271.
- Simpson, K.F., 1958. In-process inventories. *Operations Research* 6, 863–873.
- Wallstrom, B., 1980. On the M/G/1 queue with several classes of customers having different service time distributions. Report 1-19, Lund Institute of Technology.
- Xia, C., Shanthikumar, G., Glynn, P., 2000. On the asymptotic optimality of the SPT rule for the flow shop average completion time problem. *Operations Research* 48, 615–622.