Int J Adv Manuf Technol
DOI 10.1007/s00170-012-3913-x

ORIGINAL ARTICLE

A queuing approach for making decisions about order penetration point in multiechelon supply chains

E. Teimoury · M. Modarres · I. G. Khondabi · M. Fathi

Received: 18 May 2010 / Accepted: 10 January 2012

Abstract This study is dedicated to order penetration point (OPP) strategic decision making which is the boundary between make-to-order (MTO) and make-to-stock (MTS) policies. A multiproduct multiechelon production supply chain is considered where the first production stage manufactures semifinished products based on an MTS policy to supply the second production stage which operates on the MTO policy. The producer desires to find the optimal fraction of processing time fulfilled by supplier and optimal semifinished products buffer capacity in OPP. To calculate system performance indexes, the matrix geometric method is employed. Afterward, optimal solutions are obtained by enumeration and direct search techniques. Moreover, the system behavior is analyzed by the numerical example. It is shown that system total cost is a concave function of increasing completed percentage in first production stage. According to the total cost function elements, managers desire to locate OPP where to balance the order fulfillment delay cost, holding cost and the cost of disposing unsuitable items. Finally, the impact of different amounts of storage capacity on OPP and total cost are analyzed. Also, the manner of expected numbers of unsuitable products, semifinished products, and expected order completion delay are analyzed versus various quantities of storage capacity and production rate.

Keywords Queuing system · Supply chain · Logistics · Order penetration point (OPP) · MTS/MTO queue · Matrix geometric method (MGM)

E. Teimoury · M. Fathi (✉)
Department of Industrial Engineering,
Iran University of Science and Technology,
Tehran, Iran
e-mail: mfathi@iust.ac.ir

E. Teimoury
Logistics & Supply Chain Researches & Studies Group,
Institute for Trade Studies & Research,
Tehran, Iran

M. Modarres
Department of Industrial Engineering,
Sharif University of Technology,
Tehran, Iran

I. G. Khondabi
Department of Engineering, Shahed University,
Tehran, Iran

1 Introduction

In terms of supply chain design, customer satisfaction lead time and inventory costs are common problems that keep managers actively encouraged. In a production supply chain system, there are various generators of uncertainties such as demands' arrival times, processing times, transportation lead times, reliability of the machines, and the capability of the operators [1]. Customers consider the response time of the order completion as a service performance measure [2]. According to the new business model of Internet/telephone ordering and quick response time requirement, make-to-order (MTO) business model is growing quickly [3, 4]. On the one hand, make-to-stock (MTS) production system can meet customer orders fast, but confronts inventory risks associated with short product life cycles and unpredictable demands. On the other hand, MTO producers can provide a variety of products and custom orders with lower inventory risks, although usually have longer customer lead times. Moreover, in MTS production, products are stocked in advance, while in MTO production, a product only starts to be produced when an order of demand is received. In some cases, custom products share approximately all the parts of the standard products and can be produced by alternating the existing standard parts with some further works, thus the assembler usually

Published online: 26 February 2012

contemplates embedding MTO processes into the mainstream MTS lines which in turn forms a hybrid production system. Order penetration point (OPP) is a concept which enables the decision makers to make use of a hybrid MTS/MTO system, applying the abovementioned queuing theory. This is considered a suitable way to model uncertainties which affects the OPP.

There are some articles of making decisions on OPP which appeared in the literature with many names such as decoupling point (DP), delay product differentiation (DPD), and product customization postponement. The term DP, in the logistics framework, was first introduced by Sharman [5] where he argued the DP's dependency on a balance between product cost, competitive pressure, and complexity. Adan and Van der Wal [6] analyzed the effect of MTS and MTO production policies on order satisfaction lead times. Arreola-Risa and DeCroix [7] studied a simple queuing environment where customers are served in first-come-first-served order regardless of their classes. Moreover, they provided a closed form formulae for making decision about the production strategy for each customer type. Recently, Yavuz Günalay [8] studied the efficient management of MTS or MTO production–inventory system in a multi-item manufacturing facility. Rajagopalan [9] proposed a mixed-integer nonlinear program production model which optimized (Q, r) (the production lot size and inventory reorder point), parameters of every product's inventory system. A comprehensive literature review on MTS–MTO production systems and revenue management of demand fulfillment can be found in Perona et al. [10] and Quante et al. [11]. The trade-off between aggregation of inventory (or inventory pooling) and the costs of redesigning the production process is studied by Aviv and Federgruen [12] where they do not consider congestion impacts, whereas Gupta and Benjaafar [13] considered the impact of capacity restrictions and congestion. That is, they proposed a common framework to examine MTO, MTS, and DPD systems in which production capability is considered. Furthermore, they analyzed the optimal point of postponement in a multistage queuing system. The DPD issue in manufacturing systems is studied by Jewkes and Alfa [2] in which they decided on where to locate the point of differentiation in a manufacturing system, and also what size of semifinished products inventory storage should be considered. In addition, they presented a model to realize how the degree of DPD affects the trade-off between customer order completion postponement and inventory risks, when both stages of production have nonnegligible time and the production capacity is limited. Also from a different point of view, the concept of order decoupling zone is introduced as an alternate to the DP concept by Wikner and Rudberg [14].

Recently, Ahmadi and Teimouri [15] studied the problem of where to locate the OPP in an auto export supply chain by using dynamic programming. Furthermore, a notable literature review in positioning DPs and studying the positioning of multiple DPs in a supply network can be seen in Sun et al. [16], but their positioning model did not make any decisions about the optimal semifinished buffer size and optimal fraction of processing time fulfilled by the upstream of DP. Jeong [17] developed a dynamic model to simultaneously determine the optimal position of the decoupling point and production–inventory plan in a supply chain. Also, many applications and methods for determining the OPP are surveyed in Olhager [18, 19], Yang and Burns [20], Yang et al. [21], Rudberg and Wikner [22], and Mikkola and Larsen [23].

The presented model tries to find equilibrium customer service levels with inventory costs, such as developed models in the literature. However, presented model differs from the studied articles in several ways. First, a two-stage MTS/MTO production model is used for each product type in a multiproduct, multiechelon supply chain. Second, the considered model gives the optimal transportation mode, optimal semifinished products warehouse capacity, and optimal fraction of processing time fulfilled by the supplier of each product type in an integrated model.

The supply chain which is considered as a basic model in this paper is composed of two production stages. In the first production stage, each product type's supplier supplies semifinished products on an MTS policy for a producer in the second production stage. The second stage producer will customize the products based on an MTO policy. The semifinished products will be completed as a result of specific customer orders. The supposed model obtains the optimal vehicles for the transportation of the completed products to each demand point.

In order to balance the costs of customer order fulfillment delay and inventory costs, each product type producer tries to find the optimal fraction of processing performed by the supplier and its optimal semifinished products buffer storage. The remainder of this paper is organized as follows. The problem description and formulation are reviewed in Sections 2 and 3. Also, the queuing aspect and performance evaluation indexes are studied in Section 3. Besides, the described model is studied with an additional warehouse capacity constraint in Section 4. Section 5 is dedicated to a numerical example. And finally, the study is concluded in Section 6.

2 Problem description and list of symbol

The following notations are used for the mathematical formulation of considered model.

Sets and indices:

i Products type index $i=1, 2,.., L$

m_i Semifinished products buffer storage capacity for product type i index m_i=1, 2,…, S_i
j Vehicles type index j=1, 2,.., J

Decision variables:

θ_i Percentage of completion for product type i in first production stage
S_i Optimal storage capacity of type i semifinished products
x_{ij} 1 if vehicle j is dedicated to logistic process of product type i, otherwise is 0.

Parameters:

$V(\theta_i)$ The value per unit of semifinished products (dollar/unit)
τ_i Constant fraction of the MTO processing rate for product type i
μ_i Production rate for product type i per each unit
C_{Ui} The cost of disposing an unsuitable item of type i (dollar/unit)
C_{Hi} The holding cost for semifinished products of type i for unit time (dollar/unit)
C_{Wi} The cost of customer order fulfillment delay for each unit of time for product type i (dollar/unit)
C_{Ci} The cost of establishing type i semifinished products storage capacity for each unit of time (dollar/unit)
c_{ij} Transportation cost of finished product type i by vehicle j for each unit of time (dollar/unit)
t_{ij} Transporting time for product type i with vehicle j
Cap_{ij} Capacity of vehicle j for product type i

Expected performance measures:

$E(N_i)$ The expected number of ith type semifinished products in the system
$E(W_i)$ The expected customer order completion delay for product type i—the time from when a customer order enters the system until its product is completed
$E(U_i)$ The expected number of ith type unsuitable products produced per unit time

A multiproduct multiechelon production supply chain is considered. In this system, it is assumed that the demands arrive according to a Poisson process with rate λ. Each customer orders one unit of ith product type with a probability of q_i where $\sum_{i=1}^{L} q_i = 1$ and $\lambda_i=\lambda q_i$, i=1, 2,…, L. The production times of workstations for all product types are assumed to be exponentially distributed with rates μ_i, i=1, 2,…, L where $\sum_{i=1}^{L} \mu_i = \mu$. These assumptions about arrival and service time's distributions return to the fact that the customer arrivals and system service times are memoryless. More specifically, the properties of random arrival and service times related to the future do not depend on any other information from further in the past. Therefore, the interval times between two consecutive arrivals follow an exponential distribution and demands arrival follow a Poisson process. These explanations are true for production time's exponential distributions. It is supposed that the supplier has an infinite source of raw materials and never faces shortage. The producer has to determine the optimal storage capacity of type i semifinished products (S_i, i=1, 2,…, L).

Each product type supplier produces a semifinished product [100% θ_i completed ($0<\theta_i<1$)] to be delivered to the producer. The producer then completes the remaining $1-\theta_i$ fraction according to a particular customer order. It should be noted that the supplier is not necessarily in a different organization from the producer; the "supplier" and "producer" may be two successive stages in a same organization. It is chosen to model θ_i as a continuous variable so that greater insights into the overall relationship between θ_i and the performance of the system can be gained. The assumption also facilitates the computational analysis. Therefore, results are presented as if the producer can implement any values of θ_i. If this is not the case, the model enables the managers to quickly identify the best choice of θ_i among a finite number of feasible alternatives. According to market characteristics studied by Jewkes and Alfa [2], there is a probability of $\phi_i(\theta_i)$ that a semifinished product is not suitable for customization and so $\phi_i(\theta_i)$ is monotonically increasing with θ_i which is a rational assumption. Figure 1 illustrates a diagram depicting proposed model.

It is assumed that there is some logistics time to supply items from producers to demand points. The logistics process is modeled by using queuing notation $M/D^{\mathrm{Cap}}{}_{ij}/\infty$ in a continuous time (as discussed by Purdue and Linton [24] and Kashyap et al. [25]), where M denotes the exponential arrivals of completed products to logistics process which is logical, owing to semifinished products completion time, D represents that each vehicle service time is deterministic, Cap_{ij} is the jth vehicle capacity for product type i which is deterministic, and the vehicles do not transport any product up to the time they become filled to capacity. It is also assumed that infinite vehicles are available to supply the order; this assumption is represented by ∞ in the queuing notation. These transport assumptions hold good for third party logistics which can be applied in practical situations.

3 Problem formulation

The entire explained system in Section 2 can be described by a Markov process with state (n_i, m_i), where n_i is the number of customers in the system waiting for each finished product type i and m_i is the number of type i semifinished products in its semifinished product storage [2]. Therefore,

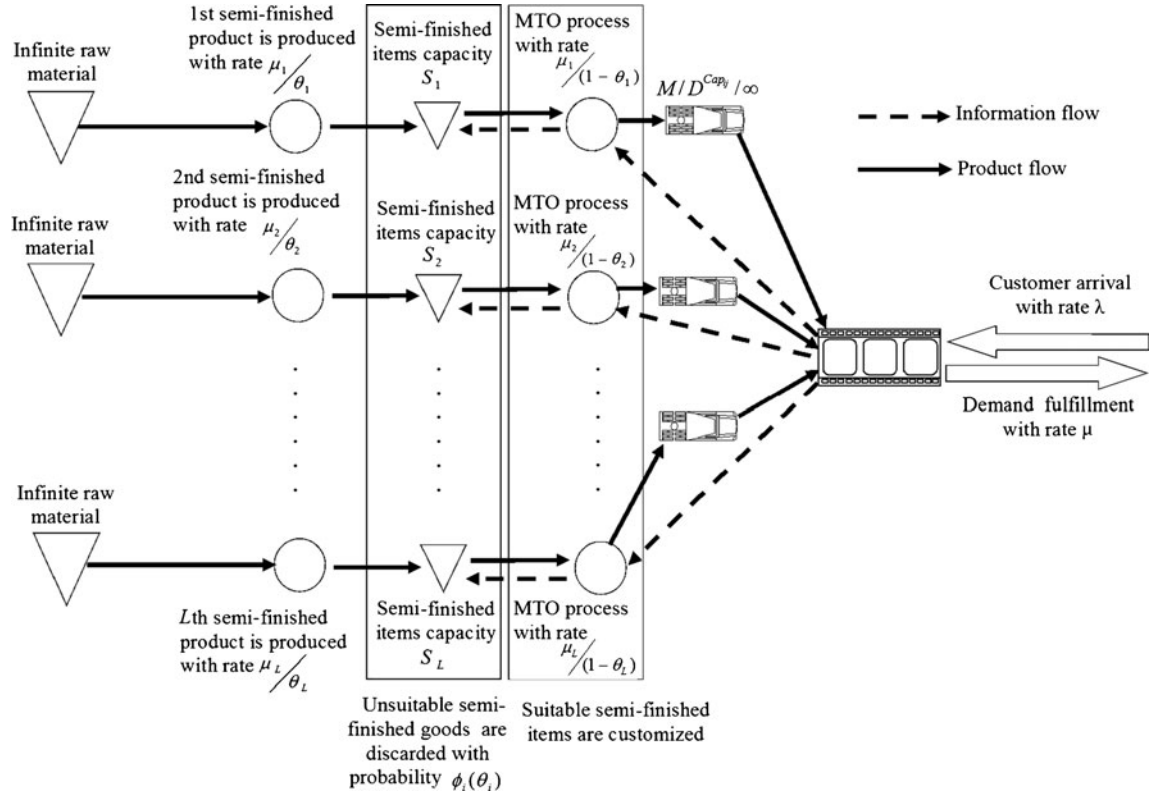


Fig. 1 The multiproduct hybrid MTO/MTS production supply chain system

the state space is denoted by $\Omega = \{n_i \geq 0, 0 \leq m_i \leq S_i\}$, which is depicted in Fig. 2 with transition rates.

In Fig. 2 (a, b) stands for $\frac{\mu(1-\phi)}{\theta}$ and $\frac{\mu}{1-\theta}$ for each product type, respectively. The associated balance equations for the

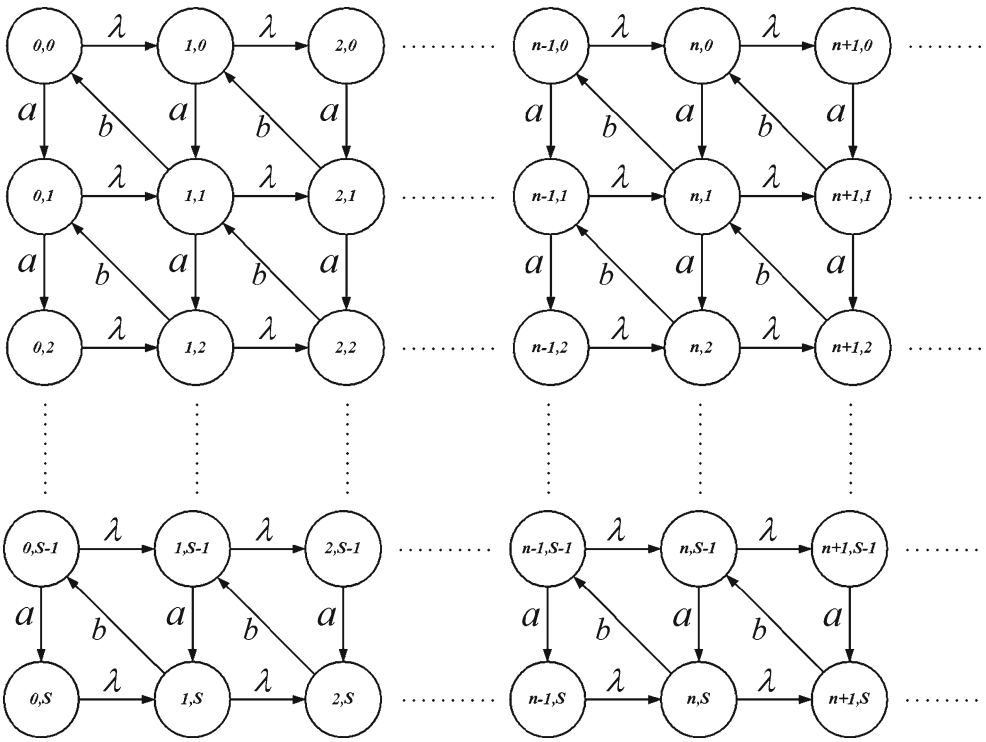


Fig. 2 State transition rates diagram

steady probabilities follow Eq. 1, 2, 3, 4, 5, and 6.

$$\left(\frac{\mu_i(1-\phi_i)}{\theta_i}+\lambda_i\right)P_i(n_i,m_i)=\frac{\mu_i}{1-\theta_i}P_i(n_i+1,m_i+1),\quad n_i=0,m_i=0 \tag{1}$$

$$\left(\frac{\mu_i(1-\phi_i)}{\theta_i}+\lambda_i\right)P_i(n_i,m_i)=\frac{\mu_i(1-\phi_i)}{\theta_i}P_i(n_i,m_i-1)+\frac{\mu_i}{1-\theta_i}P_i(n_i+1,m_i+1),\quad n_i=0,1\le m_i\le S_i-1 \tag{2}$$

$$\frac{\mu_i(1-\phi_i)}{\theta_i}P_i(n_i,m_i-1)=\lambda_i P_i(n_i,m_i),\qquad n_i=0,m_i=S_i \tag{3}$$

$$\left(\frac{\mu_i(1-\phi_i)}{\theta_i}+\lambda_i\right)P_i(n_i,m_i)=\lambda_i P_i(n_i-1,m_i)+\frac{\mu_i}{1-\theta_i}P_i(n_i+1,m_i+1),\quad 1\le n_i,m_i=0 \tag{4}$$

$$\left(\frac{\mu_i(1-\phi_i)}{\theta_i}+\lambda_i+\frac{\mu_i}{1-\theta_i}\right)P_i(n_i,m_i)=\frac{\mu_i(1-\phi_i)}{\theta_i}P_i(n_i,m_i-1)+\frac{\mu_i}{1-\theta_i}P_i(n_i+1,m_i+1)+\lambda_i P_i(n_i-1,m_i),\quad n_i=0,1\le m_i\le S_i-1 \tag{5}$$

$$\left(\lambda_i+\frac{\mu_i}{1-\theta_i}\right)P_i(n_i,m_i)=\frac{\mu_i(1-\phi_i)}{\theta_i}P_i(n_i,m_i-1)+\lambda_i P_i(n_i-1,m_i),\quad 1\le n_i,m_i=S_i \tag{6}$$

There exists the corresponding generator matrix Q_i written in block form (Eq. 7) for the product type i:

$$Q_i=\begin{bmatrix} G_i & A_i & & & \\ C_i & E_i & A_i & & \\ & C_i & E_i & A_i & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \tag{7}$$

Appendix A shows block matrices where A_i, C_i, E_i, and G_i are block matrices with the dimension of $(S_i+1)\times(S_i+1)$. It is notable that A_i giving the rate at which the number of customer orders in the system increases by one, E_i giving the rate at which the number of customer orders in the system either stays at the same level, and C_i giving the rate at which the number of customer orders in the system decreases by one. G_i is the matrix rate at which the customer orders in the system move from zero to one.

Let $F_i=A_i+E_i+C_i$ be a generator matrix with its associated stationary distribution $P_i=[P_{i0},P_{i1},\ldots,P_{iS_i}]$ given as a solution to $P_iF_i=0$, $P_i\mathbf{1}=1$.

$$F_i=\begin{bmatrix} F_{i_{0,0}} & F_{i_{0,1}} & & & \\ F_{i_{1,0}} & F_{i_{1,1}} & F_{i_{1,2}} & & \\ \ddots & \ddots & \ddots & & \\ & F_{i_{S_i-1,S_i-2}} & F_{i_{S_i-1,S_i-1}} & F_{i_{S_i-1,S_i}} & \\ & & F_{i_{S_i,S_i-1}} & F_{i_{S_i,S_i}} & \end{bmatrix} \tag{8}$$

Appendix B illustrates block matrices where $F_{i_{m,m+1}}$, $F_{i_{m,m-1}}$, and $F_{i_{m,m}}$ are $(S_i+1)\times(S_i+1)$. As it is discussed in Neuts [26], the explained Markov chain is stable if $P_iC_i\mathbf{1}>P_iA_i\mathbf{1}$. In order to have a stable system, the producer requires having a service rate that exceeds the arrival rate of customers. In addition, the supply rate of suitable semifinished products to the producer must be more than the customer demands rate.

3.1 Steady-state analysis

The behavior of this supply chain system is studied in a steady state. Let $\Pi_i=[\Pi_{i\mathbf{0}},\Pi_{i\mathbf{1}},\Pi_{i\mathbf{2}},\ldots]$ be the stationary probabilities associated with the Markov chain for each product type so that $\Pi_iQ_i=0$ and $\Pi_i\mathbf{1}=1(i=1,\ 2,\ldots,\ L)$. Due to the matrix geometric theorem [26], equation $\Pi_{i,n+1}=\Pi_{i,n}R_i,\quad n\ge 0$ must be satisfied where R_i is the minimal nonnegative solution to the matrix quadratic equation $A_i+R_iE_i+R_i^2C_i=0$.

It is noteworthy that matrix R_i can be computed very easily using some well-known methods according to Bolch et al. [27]. A simple way to compute R_i is the iterative approach given as $R_i(n+1)=-\left(A_i+R_i(n)^2C_i\right)E_i^{-1}$ until $|R_i(n+1)-R_i(n)|_{nj}<\varepsilon$, with $R_i(0)=0$. The boundary vector Π_{i0} is obtained from $\Pi_{i0}(G_i+R_iC_i)=0$.

3.2 Performance evaluation indexes

Here, the important performance evaluation indexes of the system can be obtained as described below. Let $E[O_i]$ be the mean number of customers' orders for product type i in the system, including the one being served; $E[W_i]$ be the mean customer order completion delay for product type i; $E[N_i]$ be the mean number of semifinished products in the system for product type i; and $E[U_i]$ be the expected number of unsuitable semifinished products disposed per unit time for product type i, then

$$E[O_i]=\Pi_{i1}(I-R_i)^{-2}\mathbf{1}$$

$E[W_i] = \frac{E(O_i)}{\lambda_i}$ (by applying Little's Law), $E[N_i] = \Pi_{i0}(I - R_i)^{-1} y_i$, where $y_i = [0,1,2,\ldots,S_i]^T$, and $E(U_i) = \frac{(1-\Pr(m_i=S_i))\phi_i\mu_i}{\theta_i}$, where m_i denotes the number of semifinished products storage for each product type.

3.3 Mathematical model

The objective function includes the following costs:

1. Disposing of semifinished products that are not appropriate for customizing the customer orders ($C_{U_i} V(\theta_i) E(U_i)$),
2. Holding semifinished products in buffer storage ($C_{H_i} V(\theta_i) E(N_i)$),
3. Providing storage capacity for the semifinished products ($C_{C_i} S_i$).
4. Customer order fulfillment delay ($\sum_{i=1}^{L}\sum_{j=1}^{J} x_{ij} C_{W_i} \left(\mathrm{Cap}_{ij} \cdot E(W_i) + t_{ij}\right)$), and
5. Transportation cost ($\sum_{i=1}^{L}\sum_{j=1}^{J} c_{ij} \cdot \mathrm{Cap}_{ij} \cdot x_{ij}$).

The mathematical formulation of the model is as follows:

$$\min_{S_i,\theta_i,x_{ij}} \mathrm{TC}\left(S_i,\theta_i,x_{ij}\right) = C_{U_i} V(\theta_i) E(U_i) + C_{H_i} V(\theta_i) E(N_i) + C_{C_i} S_i + \sum_{i=1}^{L}\sum_{j=1}^{J} x_{ij} C_{W_i} \left(\mathrm{Cap}_{ij} \cdot E(W_i) + t_{ij}\right) + \sum_{i=1}^{L}\sum_{j=1}^{J} c_{ij} \cdot \mathrm{Cap}_{ij} \cdot x_{ij} \tag{9}$$

subject to:

$$\sum_{j=1}^{J} x_{ij} = 1 \qquad \forall i \tag{10}$$

$$\tau_i \frac{\mu_i}{(1-\theta_i)} \le \frac{1}{E(W_i)} + \frac{\mathrm{Cap}_{ij}}{t_{ij}} \qquad \forall i \tag{11}$$

$$0 < \theta_i < 1.0 \quad \forall i \tag{12}$$

$$S_i = 1,2,\ldots \quad \forall i \tag{13}$$

$$x_{ij} \in \{0,1\} \qquad \forall i,j \tag{14}$$

The model (Eq. 9) minimizes the total expected cost including the cost of semifinished products that are not consistent with customer's order, expected semifinished products holding cost, the cost of establishing storage capacity for semifinished products, expected cost of delay in customer order completion (which include time of customization and logistics), and transportation costs based on the vehicle type selected. Constraint in Eq. 10 assures that logistics process for each product type accomplishes by exactly one vehicle. Constraint in Eq. 11 represents that a constant value (τ_i) of the MTO processing rate for product type i ($\frac{\mu_i}{(1-\theta_i)}$) must be at its most less than the total customer order completion rate which contains customization and logistics process (service level constraint). Constraints in Eqs. 12, 13, and 14 represent the ranges of the model variables.

In order to solve proposed mathematical model, a direct search heuristic method is used. The values of S_i and θ_i must vary in their allowable variation ranges to find their near optimal values. These various values of storage capacity and completion percentage enable us to calculate system performance measures which are used in mathematical model. The outputs of the represented model are the optimal fractions of the process fulfilled by the supplier for each product type, their optimal sem-finished products buffer storage capacity, and the optimal transportation mode for each product type.

4 Studying model under the warehouse capacity constraint

This section studies a more realistic constraint that can be added to the proposed model (see Fig. 3). According to warehouses physical structure, it is not possible to establish every calculated optimal storage capacity for each product type. This is a logical assumption in operational problems. In this section, specific capacity of K is considered for semifinished product warehouse.

Due to separate calculations of optimal storage capacity for each product type, the storage space constraint cannot be applied in the optimization model. Therefore, if the summation of semifinished product storage related to obtained optimal solution for all types of products satisfies the warehouse capacity constraint, the obtained solutions can be considered as optimal storage capacities. However, if the warehouse capacity constraint has not been satisfied, the developed heuristic solution procedure can be used as follows.

Algorithm

Step 1 Find the optimal values S_i^* and θ_i^* for each product type.

Step 2 Calculate $\sum_{i=1}^{L} S_i^*$ (cumulative storage value for all product types). If $\sum_{i=1}^{L} S_i^*$ is smaller than the predefined capacity constraint for central warehouse (K), solutions in step 1 are acceptable: stop. Otherwise, use step 3.

Step 3 Find $\mathrm{TC}_i\left(S_i^* - 1, \theta_i\left(S_i^* - 1\right)\right) - \mathrm{TC}_i(S_i^*, \theta_i^*)$ for each product type. Set $S_i^* - 1 \rightarrow S_i^*$ (if $S_i^* - 1$ is stable) and $\theta_i(S_i^* - 1) \rightarrow \theta_i^*$ for product type with the lowest $\mathrm{TC}_i\left(S_i^* - 1, \theta_i\left(S_i^* - 1\right)\right) - \mathrm{TC}_i(S_i^*, \theta_i^*)$.

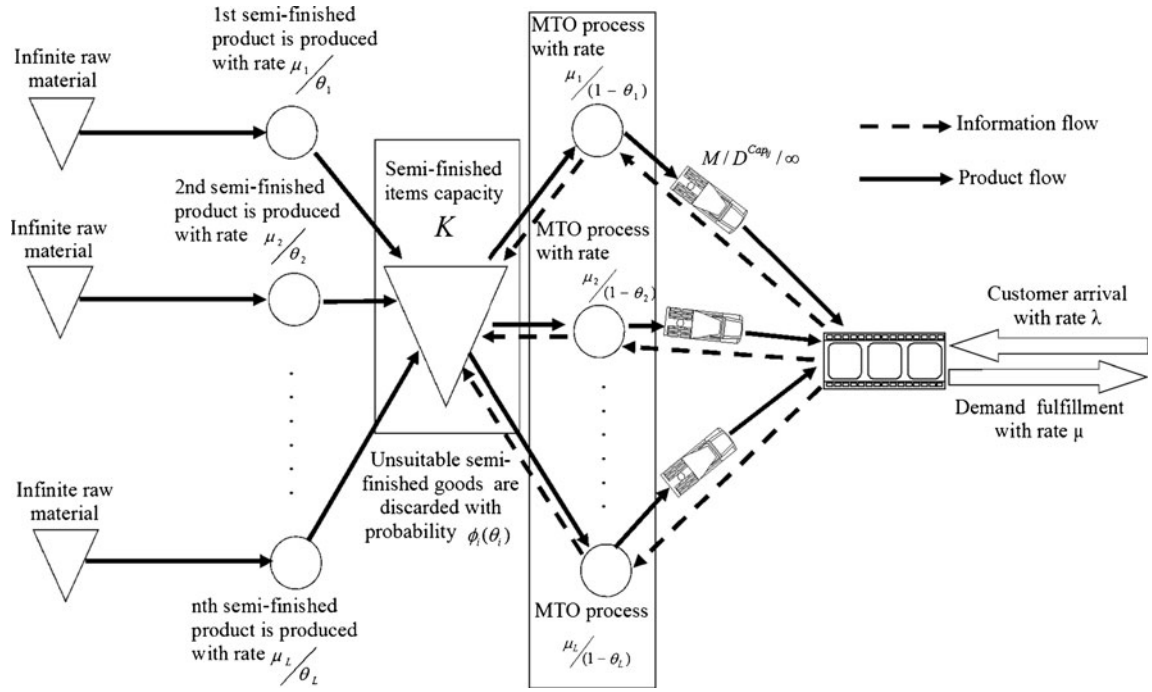


Fig. 3 The multiproduct hybrid MTO/MTS production supply chain with the capacitated warehouse

Step 4 If $\sum_{i=1}^{L} S_i^* \le K$, solutions obtained in step 3 are acceptable: stop. Otherwise, use step 5.

Step 5 Go to step 3.

The proposed algorithm is represented schematically in Fig. 4. Although the developed algorithm is so time-consuming due to the enumeration technique used in its steps, it computes a nearly optimal solution with minimum benefit loss.

5 Numerical example

In this section a numerical example is used to show the relation between $\mathrm{TC}(S_i, \theta_i(S_i))$ and variables $\theta_i^*(S_i)$ and S_i, also system analysis is done for a variety of parameters. A production supply chain network containing three product types with three suppliers, three producers, a capacitated warehouse with the capacity of K=7, and three types of transportation vehicles is considered. The following parameter values are considered which are changeless for the

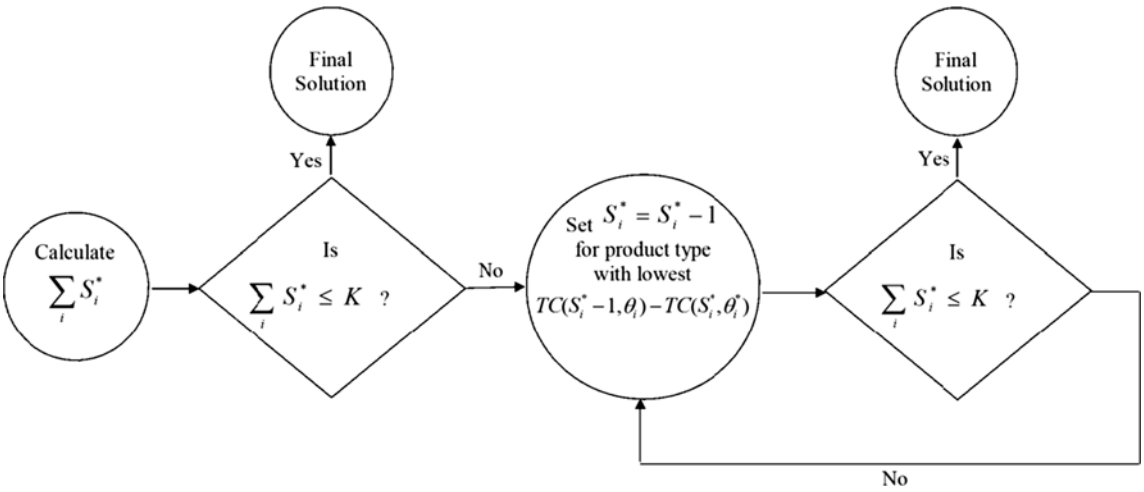


Fig. 4 Heuristic solution procedure

various product types: τ_i=0.05, C_W=1.2, C_H=0.1, and each semifinished product value $V(\theta_i)$ equals to θ_i as assumed by Jewkes and Alfa [2]. Other necessary information about each product type and transportation vehicles characteristics id provided in Table 1 (λ_i is calculated by λq_i for each product type).

As shown in Table 2, the system is not stable for any S_i= 1 due to the supplier disability to provide sufficient suitable semifinished products to satisfy the producers' demands. The optimal values of storage capacities and the process postponement for each product type are shown in bold in Table 2. As it is seen, the semifinished product completed percentage is an ascending function of capacity growth. It is notable that when there is a lower capacity, the semifinished products are highly affected by demand fluctuations and the manufacturer prefers to bear the cost of completion delay for a more customization right and preventing of disposing semifinished products. But as it is seen, by increasing the storage capacity, the affect of demand fluctuations in cost function decreases and the manufacturer prefers to increase the completed percentage in first echelon in order to reduce the product completion delay.

Also there exists a trend in total cost function which is affected by cost parameters. For product type 1 and 3, a reduction in total cost function can be seen which can be explained by the increase of product completed part due to storage growth. This increase in product completed part reduces the completion delay cost which is more than the growth of other four cost parameters. But there exists an increasing trend for total cost function after S_i=3 for product type 1 and 3 and also for all capacities of product type 2, which is reasonable due to structure of cost parameters. It is obvious that holding semifinished products and providing storage capacity are increasing functions of capacity growth which are more effective than other cost parameters and finally increase the total cost function by growing the semifinished product storage. In order to better understand the affect of completion percentage on total cost, the variation of total cost function versus completion percentage of semifinished product type 1 is shown in Fig. 5, it is notable that zero total costs are related to infeasible points.

Table 2 Results of numerical example

Product type	S_i	Optimal vehicle	$\theta_i^*(S_i)$	Total cost
1	1	3	Nonstable	Nonstable
	2		0.26	15.0867
	3		0.29	**15.0300**
	4		0.29	15.2436
	5		0.30	15.6314
	–		–	–
	30		0.30	25.6630
	40		0.30	29.6633
	50		0.30	33.6634
2	1	3	Nonstable	Nonstable
	2		**0.28**	**12.3716**
	3		0.30	12.8561
	4		0.30	13.5357
	–		–	–
	30		0.31	31.7625
	40		0.31	38.7626
	50		0.31	45.7626
3	1	3	Nonstable	Nonstable
	2		0.28	14.0539
	3		**0.31**	**13.9608**
	4		0.32	14.2431
	5		0.33	14.6173
	–		–	–
	30		0.33	24.6454
	40		0.33	28.6457
	50		0.33	32.6458

In this example an optimal transportation vehicle is selected for each product type. It is worth noting that the logistic process does not follow an assignment model, and each vehicle type can be used for more than one product type. The derived results of change trend in Fig. 5 are in accordance with Jewkes and Alfa [2].

Due to warehouse capacity, the satisfaction condition $\sum_{i=1}^{L} S_i^* \le K$ must be checked and if the storage capacity

Table 1 Parameters' data for numerical example

Product type	λ_i	μ_i	ϕ_i	C_{c_i}	$C_{U\bar{i}}$	Transport time by vehicle j			Transport cost by vehicle j			Transport capacity by vehicle j		
						1	2	3	1	2	3	1	2	3
1	0.7	1	$0.9\theta_1$	0.4	1	10	8	5	0.25	0.26	0.3	5	4	3
2	0.6	1	$0.7\theta_2$	0.7	0.8	10	8	5	0.30	0.32	0.38	3	2	2
3	0.9	1.2	$0.75\theta_3$	0.4	0.7	10	8	5	0.19	0.21	0.28	4	3	3

The computational results are based on the MATLAB 7.1 implementation where the total cost is computed for $0.01 \le \theta_i \le 0.99$ in increments of 0.01 where S_i varies from 1 to 50

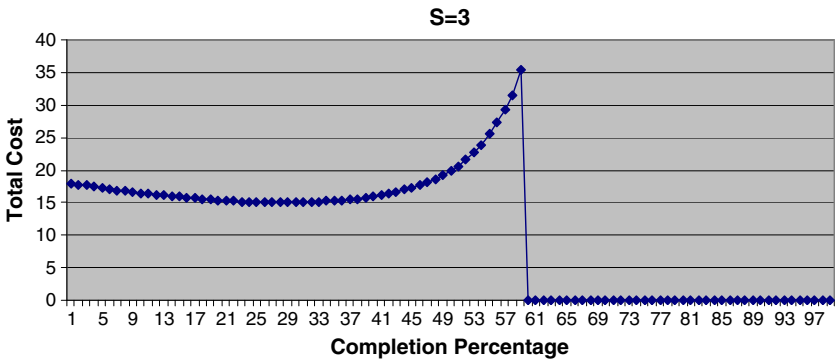


Fig. 5 $TC(S_i, \theta_i(S_i))$ versus $\theta_i^*(S_i)$ for product type 1

constraint does not hold, the developed heuristic solution must be run:

Step 2 $\sum_{i=1}^{L} S_i^* = 3 + 2 + 3 > K = 7$

Step 3 $\left.\begin{array}{l} \mathrm{TC}_1(2,0.26) - TC_1(3,0.29) = 15.08 - 15.03 = 0.05 \\ \mathrm{TC}_2(1,\theta_2(1)) = \text{nonstable} \\ \mathrm{TC}_3(2,0.28) - \mathrm{TC}_1(3,0.31) = 14.05 - 13.96 = 0.09 \end{array}\right\} \Rightarrow S_1^* = 2, \theta_1^* = 0.26$

Step 4 $\sum_{i=1}^{L} S_i^* = 2 + 2 + 3 \leq K = 7$

Now the near optimal solutions with minimum benefit loss are obtained. The storage capacity for first, second, and third type products equals to 2, 2, and 3, respectively. Moreover, the OPP must stand after the 0.26, 0.28, and 0.31 of product completion in each product's supply chain and the transportation vehicle for all products still remains the third one due to the explained reasons. In Section 5 the system manner under different parameter variations must be analyzed. The logical manner of surveyed system characteristics can be considered as a validation for the proposed model and the performed computations.

5.1 Affect of demand λ fluctuations on total cost function and OPP location

According to queuing theory fundamentals, the customer arrival rate must be lower than systems' service rate due to establish system stability; otherwise, the queue length goes infinite. In this example the service rate for product type 1 is equal to 1, so the affect of varying customer arrival rate on the total system cost is studied by the values between 0.2 and 0.9 which increments by 0.1. Changes in total cost function for various values of customer arrival rate are shown in Fig. 6. It is notable that zero total costs are related to infeasible points.

Higher arrival rate enhances the system busy time, and the queue length and customer order fulfillment time are increased, consequently. It is obvious that the total cost of system will be augmented by increasing system busy time, due to the mentioned explanations and the fact that there is no unemployment cost for the system.

In addition to the affect of increasing λ_i on positioning, OPP is remarkable. As it is shown in Fig. 7, increasing λ_i has an increasing effect on optimal value of OPP. The derived results of change trend in Fig. 6 are in accordance with Jewkes and Alfa [2].

Moreover, the place of OPP differentiates with different values of demand rate. It is obvious that the expected number of customers will increase by growing demand, therefore the queue length will raise and customers must wait a longer time to get service. This fact enforces the manufacturer to complete a more percentage of products in the first echelon and satisfies the demand with less delay. This increase in OPP reduces the queue length and completion delay which follows the reduction in the total cost function.

5.2 Affect of production rate μ fluctuations on system performance measures

Increasing production rate has a specific effect on each system performance measures such as increasing service rate and customer satisfaction, but the cost of increasing production rate is a preventive factor of enjoying these advantages. In this section the affect of various production rates on some system performance measures is studied and shown schematically in Fig. 8 (μ stands for production rate).

In Fig. 8a the expected order completion delay is reduced by increasing production rate which is justifiable by queue length. The higher production rate leads to a higher demand satisfaction rate, therefore the queue length would be decreased and this is equal to less completion delay due to Little's laws. The affect of production rate on expected number of unsuitable products is shown in Fig. 8b. The production rate μ is used as a linear coefficient in calculating the expected number of unsuitable products, and without any conceptual explanations, it is expectable to have an increasing manner of $E(U_i)$ by growing μ.

The expected number of semifinished products in the system decreases versus production growth, because the higher rate of production satisfies the customer demands with a higher service rate and a lower value of semifinished

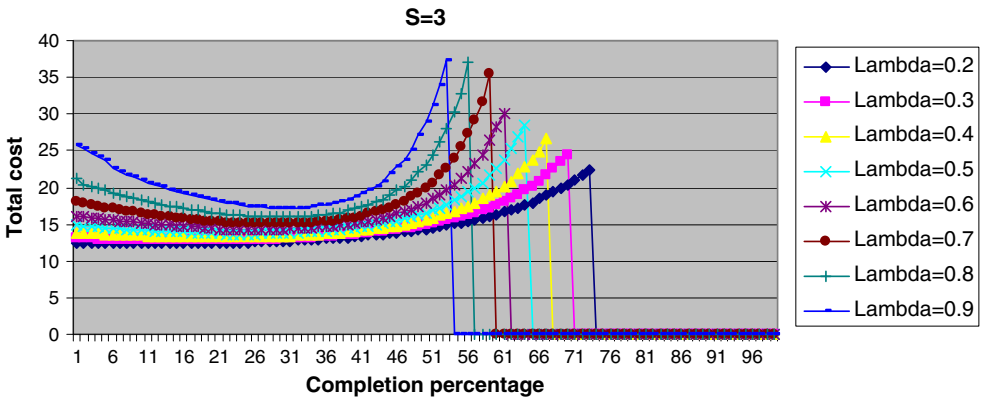


Fig. 6 $TC(S_i, \theta_i(S_i))$ versus $\theta_i^*(S_i)$ and λ_i for product type 1

products would be remained in the system, consequently. This reduction of semifinished products against production rate is depicted in Fig. 8c.

5.3 Affect of various capacity storages on system performance measures

Increasing of expected number of semifinished products is the first thing which is expected due to increasing of storage capacity. But growing the storage capacity values will affect the other system performances which are depicted in Fig. 9 schematically.

As it is discussed, growing storage capacity enhances the expected number of semifinished products which is shown in Fig. 9a. This growing trend is approximately linear which is expected from the initial numerical example results where the minimum costs are related to lower values of capacity storage.

Figure 9b shows the expected number of unsuitable products growth. This trend can be interpreted by referring to the formula of calculating unsuitable products. The probability of storage not being full is used, and it is obvious that growing storage capacity will reduce the probability of a full storage and enhance the probability of having an empty storage. The probability of having empty storage is multiplied to the numerator of unsuitable products calculating formula, so the growth of storage capacity will enhance the expected number of semifinished products which must be discarded.

Figure 9c is dedicated to showing the variations of expected order completion delay for storage capacity. Growing storage capacity will enhance the expected number of semifinished products in the system, and the demands will satisfied with higher rate which reduces the queue length. As it was discussed, lower queue length will logically reduce the completion delay. The derived results of change trend in Fig. 9 are in accordance with Jewkes and Alfa [2].

6 Conclusion

OPP is the boundary between MTO and MTS policies. In this article an optimization model was presented to determine OPP in a multiproduct multiechelon supply chain. The affects of product customization postponement on customer order completion delay and inventory risks were discussed. In order to evaluate performance measures, a simple queuing model and an explicitly matrix geometric method was applied.

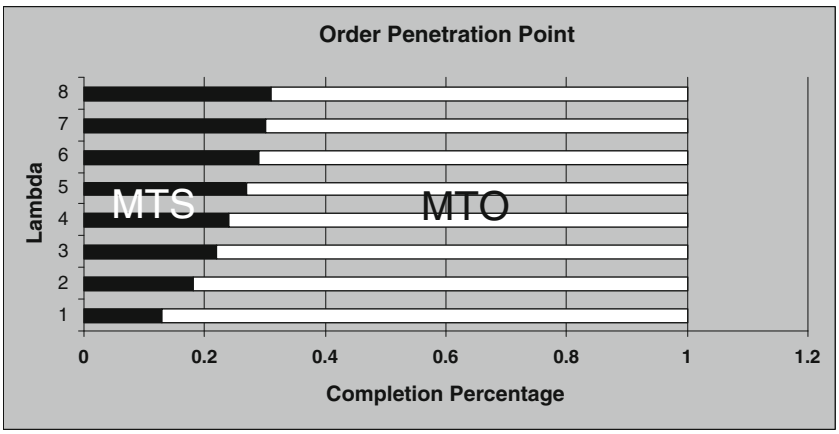
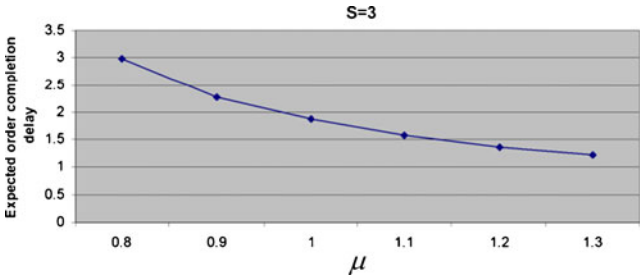
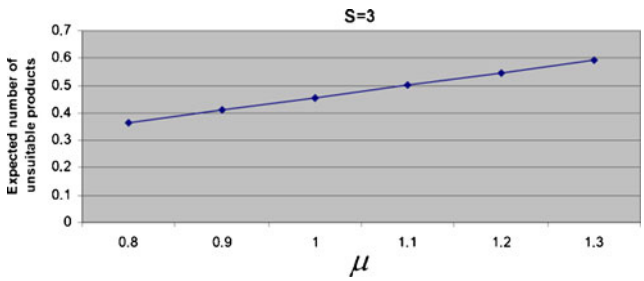


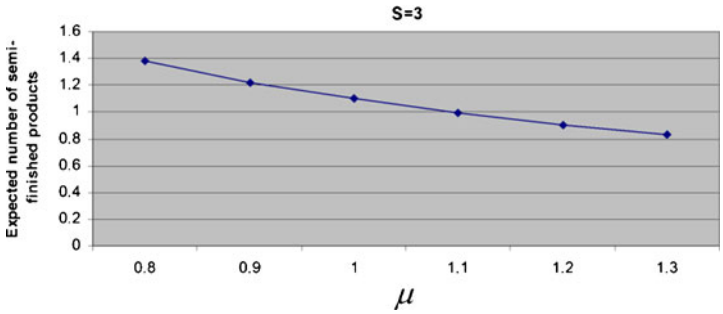
Fig. 7 OPP versus λ_i for product type 1



A. Expected order completion delay versus μ



B. Expected number of unsuitable products versus μ

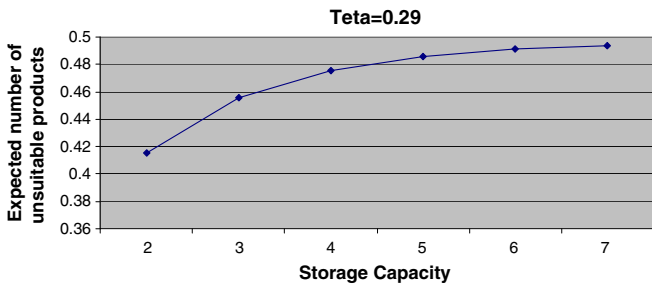


C. Expected number of semi-finished products versus μ

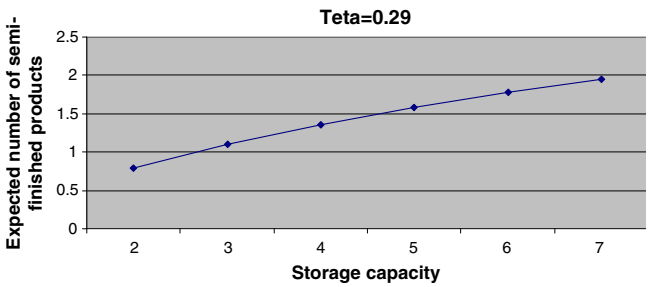
Fig. 8 **a**–**c** System performance measures versus μ

The proposed model aims to obtain the optimal OPP in a supply chain, optimal level of buffer storage capacity, as well as the best finished products transporting vehicle for each product type. The transportation is modeled as a logistic process where each vehicle has a constant capacity and a deterministic delivery time. In addition, the problem with a real warehouse capacity constraint is considered as a development of the main model of the article.

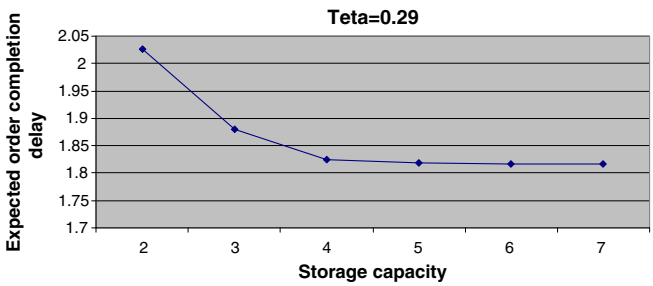
The numerical example and the sensitivity analysis explain the system manner under various chain parameters. It



B. Expected number of unsuitable products versus storage capacity



A. Expected number of semi-finished products versus storage capacity



C. Expected order completion delay versus storage capacity

Fig. 9 **a**–**c** System performance measures versus storage capacity

is shown that the total cost function increases by the increasing demand arrival rate, and it is concave in increasing completed percentage of semifinished products. Furthermore, it is shown that the semifinished product completed percentage is an ascending function of capacity growth by the considered assumptions of considered model. The observation of arrival demand fluctuations showed the increase of optimal OPP due to increasing values of lambda. Moreover, the affect of production rate μ and capacity storage fluctuations show that the expected number of semifinished products is decreasing in μ but increasing in storage capacity, the expected number of unsuitable products is increasing in both μ and storage capacity and the expected order completion delay is decreasing in both μ and storage capacity.

Manufacturers must consider the storage capacity, disposal, and order completion delay costs as the important decision-making parameters. As it is obvious the increasing OPP would increase the storage holding and disposal costs and decrease the completion delay cost. Also, decreasing OPP has an opposite effect on holding, disposal, and completion delay costs. Manufacturers must locate the OPP where the related costs get balanced, and the summation of all considered costs must be minimized. The rates of service and customer arrivals are effective factors which must be considered on choosing OPP, too. Applying the capacity constraint in customers queue, relaxing the assumptions of exponentially distributed arrival and service times, and considering the impatient customers in arrival demands can be as future research possibilities.

Acknowledgment The authors are thankful for constructive comments of the reviewers and the editor that certainly improved the presentation of the paper.

Appendix A

$$G_i = \begin{bmatrix} G_{i_{0,0}} & G_{i_{0,1}} & & \\ & G_{i_{1,1}} & G_{i_{1,2}} & \\ & \ddots & \ddots & \\ & & G_{i_{S_i-1,S_i-1}} & G_{i_{S_i-1,S_i}} \\ & & & G_{i_{S_i,S_i}} \end{bmatrix}_{(S_i+1)\times(S_i+1)}$$

$$G_{i_{m,m}} = \begin{cases} -\left(\lambda_i + \frac{\mu_i(1-\phi_i)}{\theta_i}\right) & 1 \le i \le L, \quad 0 \le m \le S_i - 1 \\ -\lambda_i & 1 \le i \le L, \quad m = S_i \end{cases}$$

$$G_{i_{m,m+1}} = \frac{\mu_i(1-\phi_i)}{\theta_i} \qquad 1 \le i \le L, \quad 0 \le m \le S_i - 1 \tag{A.1}$$

$$E_i = \begin{bmatrix} E_{i_{0,0}} & E_{i_{0,1}} & & \\ & E_{i_{1,1}} & E_{i_{1,2}} & \\ & \ddots & \ddots & \\ & & E_{i_{S_i-1,S_i-1}} & E_{i_{S_i-1,S_i}} \\ & & & E_{i_{S_i,S_i}} \end{bmatrix}_{(S_i+1)\times(S_i+1)}$$

$$E_{i_{m,m}} = \begin{cases} -\left(\lambda_i + \frac{\mu_i(1-\phi_i)}{\theta_i}\right) & 1 \le i \le L, \quad m = 0 \\ -\left(\lambda_i + \frac{\mu_i(1-\phi_i)}{\theta_i} + \frac{\mu_i}{1-\theta_i}\right) & 1 \le i \le L, \quad 1 \le m \le S_i - 1 \\ -\left(\lambda_i + \frac{\mu_i}{1-\theta_i}\right) & 1 \le i \le L, \quad m = S_i \end{cases}$$

$$E_{i_{m,m+1}} = \frac{\mu_i(1-\phi_i)}{\theta_i} \qquad 1 \le i \le L, \quad 0 \le m \le S_i - 1 \tag{A.2}$$

$$C_i = \begin{bmatrix} 0 & 0 \\ I\frac{\mu_i}{1-\theta_i} & 0 \end{bmatrix}_{(S_i+1)\times(S_i+1)} \tag{A.3}$$

$$A_i = [I\lambda_i]_{(S_i+1)\times(S_i+1)} \tag{A.4}$$

Appendix B

$$F_{i_{m,m}} = \begin{cases} -\left(\frac{\mu_i(1-\phi_i)}{\theta_i}\right) & 1 \le i \le L, \quad m = 0 \\ -\left(\frac{\mu_i(1-\phi_i)}{\theta_i} + \frac{\mu_i}{1-\theta_i}\right) & 1 \le i \le L, \quad 1 \le m \le S_i - 1 \\ -\left(\frac{\mu_i}{1-\theta_i}\right) & 1 \le i \le L, \quad m = S_i \end{cases} \tag{B.1}$$

$$F_{i_{m,m+1}} = \frac{\mu_i(1-\phi_i)}{\theta_i} \qquad 1 \le i \le L, \quad 0 \le m \le S_i - 1 \tag{B.2}$$

$$F_{i_{m,m-1}} = \frac{\mu_i}{1-\theta_i} \qquad 1 \le i \le L, \quad 1 \le m \le S_i \tag{B.3}$$

References

1. Chopra S, Meindl P (2003) Supply chain management: strategy, planning, and operations. Prentice Hall, 2nd edn, 592 pages, ISBN-13: 978-0131010284
2. Jewkes EM, Alfa AS (2009) A queuing model of delayed product differentiation. Eur J Oper Res 199:734–743
3. Hajfathaliha A, Teimoury E, khondabi IG, Fathi M (2011) Using queuing approach for locating the order penetration point in a two-echelon supply chain with customer loss. Int J Bus Res Manag 6(1) ISSN 1833-3850;E-ISSN 1833-8119
4. Teimoury E, Modarres M, Kazeruni Monfared A, Fathi M (2011) Price, delivery time, and capacity decisions in an M/M/1 make-to-order/service system with segmented market. Int J Adv Manuf Technol. doi:10.1007/s00170-011-3261-2
5. Sharman G (1984) The rediscovery of logistics. Harv Bus Rev 62 (5):71–79
6. Adan IJBF, Van der Wal J (1998) Combining make to order and make to stock. OR-Spektrum 20(2):73–81
7. Arreola-Risa A, DeCroix GA (1998) Make-to-order versus make-to-order in a production-inventory system with general production times. IIE Trans 30(8):705–713
8. Günalay Y (2010) Efficient management of production-inventory system in a multi-item manufacturing facility: MTS vs MTO. Int J Adv Manuf Technol 54:1179–1186. doi:10.1007/s00170-010-2984
9. Rajagopalan S (2002) Make-to-order or make-to-stock: model and application. Manag Sci 48(2):241–256
10. Perona M, Saccani N, Zanoni S (2009) Combining make-to-order and make-to-stock inventory policies: an empirical application to a manufacturing SME. Prod Plan Control 20(7):559–575. doi:10.1080/09537280903034271
11. Quante R, Meyr H, Fleischmann M (2009) Revenue management and demand fulfillment: matching applications, models, and software. OR Spectr 31(1):31–62. doi:10.1007/s00291-008-0125-8
12. Aviv Y, Federgruen A (2001) Design for postponement: a comprehensive characterization of its benefits under unknown demand distributions. Oper Res 49(4):578–598
13. Gupta D, Benjaafar S (2004) Make-to-order, make-to-stock, or delay product differentiation? A common framework for modeling and analysis. IIE Trans 36:529–546
14. Wikner J, Rudberg M (2005) Introducing a customer order decoupling zone in logistics decision making. Int J Logist Res Appl 8 (3):211–224
15. Ahmadi M, Teimouri E (2008) Determining the order penetration point in auto export supply chain by the use of dynamic programming. J Appl Sci 8(18):3214–3220
16. Sun XY, Jib P, Sun LY, Wang YL (2008) Positioning multiple decoupling points in a supply network. Int J Prod Econ 113:943–956
17. Jeong IJ (2011) A dynamic model for the optimization of decoupling point and production planning in a supply chain. Int J Prod Econ 131(2):561–567
18. Olhager J (2003) Strategic positioning of the order penetration point. Int J Prod Econ 85:319–329
19. Olhager J (2010) The role of the customer order decoupling point in production and supply chain management. Comput Ind 61 (9):863–868
20. Yang B, Burns ND (2003) Implications of postponement for the supply chain. Int J Prod Res 41(9):2075–2090
21. Yang B, Burns ND, Backhouse CJ (2004) Postponement: a review and an integrated framework. Int J Oper Prod Manag 24(5):468–487
22. Rudberg M, Wikner J (2004) Mass customization in terms of the customer order decoupling point. Prod Plan Control 15(4):445–458
23. Mikkola JH, Larsen TS (2004) Supply chain integration: implications for mass customization, modularization and postponement strategies. Prod Plan Control 15(4):352–361
24. Purdue P, Linton D (1981) An infinite-server queue subject to an extraneous phase process and related models. J Appl Probab 18:236–244
25. Kashyap BRK, Liu L, Templeton JGC (1990) On the $GI^X/G/\infty$ system. J Appl Probab 27:671–683
26. Neuts MF (1981) Matrix-geometric solutions in stochastic models: an algorithmic approach. Johns Hopkins University Press, Baltimore
27. Bolch G, Greiner S, De Meer H, Trivedi KS (2006) Queueing networks and Markov chains: modeling and performance evaluation with computer science. Wiley, Hoboken, New Jersey