Int. J. Production Economics 126 (2010) 341–349

Contents lists available at ScienceDirect

Int. J. Production Economics

journal homepage: www.elsevier.com/locate/ijpe

Order-fulfillment performance analysis of an assemble-to-order system with unreliable machines

Chunyan Gao [a,b], Houcai Shen [a,b,*], T.C.E. Cheng [a,b]

[a] Department of Management Science and Engineering, Nanjing University, Nanjing, China
[b] Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 1 July 2009
Accepted 16 April 2010
Available online 29 April 2010

Keywords:
Inventory
Queuing
Assemble-to-order
Performance measures

ABSTRACT

We study a multi-item, multiple classes of demand, assemble-to-order system. The inventory of each item is kept at the item level and controlled by the base-stock policy with a finite capacity. Each item is replenished by an independent unreliable machine. Each type of demand arrives according to a Poisson process with an individual rate and requires a subset of the items. When the item requirements of an arriving demand cannot be satisfied entirely, two kinds of stockout may occur, namely total-order-service and partial-order-service. We formulate the system as a queuing network and deduce that it is a quasi-birth–death process. Applying the matrix-geometric solution approach, we derive the exact joint steady-state distribution of on-order inventories, based on which we compute the order-based and item-based the fill rate within a time window and the service level. We present numerical examples to show how system performance varies with changes in system parameters to show the importance of taking machine failures into consideration.

1. Introduction

This paper studies order-based and item-based performance measures of a multi-item, multiple classes of demand, assemble-to-order (ATO) production-inventory system with unreliable machines. The inventory of each item (component) is kept only at the item level and controlled by the base-stock policy. Upon the arrival of a demand, the required components are assembled if necessary to satisfy the demand. Each item is replenished by a dedicated failure-prone machine that processes orders on a first-come-first-served (FCFS) basis. In this paper we use the terms item and component interchangeably.

The ATO system described above is widely used in manufacturing, especially in the electronics manufacturing industry, as a means to quickly respond to market uncertainty by postponing product differentiation to the latest stage of production possible. An example of firms that use the ATO system is Dell, which supplies standard changeable computer components, including main boards, processors, display units, cache memory chips, etc., and assembles the required components into PCs according to customer demands. In China, many firms have implemented the ATO system in their manufacturing facilities, too. The adoption of the ATO system is not confined to manufacturing. For example, the multi-item distribution system commonly adopted in the logistics industry is an ATO system.

Considerable research has been devoted to the modelling and analysis of the ATO system. We only focus on the literature on continuous time models. The reader is referred to Song and Zipkin (2003) and Benjaafar and El Hafsi (2006) for details on research on this topic.

Most of the existing research on the continuous time ATO system focused on the base-stock policy, the FCFS service discipline, but they differed in their model assumptions and solution approaches. These studies can be further classified into two categories based on the component supply process, namely (1) the system with exogenous lead times and (2) the system with endogenous lead times. The literature on exogenous lead times models includes Song (1998, 2002), Song and Yao (2002), Lu et al. (2003, 2005), Lu and Song (2005), Lu (2008), Zhao (2009), and Xiao et al. (2010). Studies on the ATO system with endogenous lead times always assumed that components are made to stock by production facilities with finite capacity. These studies are most relative to ours. But the literature on this subject is relatively scant. Song et al. (1999) studied the multi-component, multi-product ATO system, where each of the components is made to stock and replenished by its dedicated machine that has exponential processing times, and inventory is controlled by the base-stock policy. They modelled the system as a network of $M/M/1/N$ queues and derived some key performance measures from the exact joint stationary distribution of outstanding orders,

* Corresponding author at: Department of Management Science and Engineering, Nanjing University, 22 Hankou Road, Nanjing, Jiangsu, China.
Tel.: +86 25 83686775; fax: +86 25 83597501.
E-mail addresses: lgtgaocy@inet.polyu.edu.hk (C. Gao), hcshen@nju.edu.cn (H. Shen), lgtcheng@inet.polyu.edu.hk (T.C.E. Cheng).

0925-5273/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpe.2010.04.014

which can be determined by modelling the system as a quasi-birth–death (QBD) process. Recently, Hoen et al. (2010) also considered a lost-sales partial order service model. They showed that the order-based performance measures can be calculated through analyzing its decomposed subsystems. Dayanik et al. (2003) re-considered their model. They developed and compared several lower bounds on the outstanding orders in the partial-order-service ATO system. Benjaafar and EI Hafsi (2006) considered a single-product ATO system with multiple customer classes. They obtained some structural properties of the optimal policy. Subsequently, El Hafsi (2009) extended their model to the compound Poisson demand case. The literature on multi-product ATO systems almost always assumed that each component is replenished with the exponentially distributed processing time. Such an assumption makes the system tractable. But in manufacturing, the replenishment processes face uncertainty, such as machine breakdowns. Machines are subject to deterioration with usage and age, and machine reliability becomes a key factor that affects system performance. This uncertainty challenges the assumptions of exponential processing times, because the total processing time may follow a more general distribution. In this paper we consider an ATO system with machine failures as a special case of generally distributed processing time.

A large body of work has studied production-inventory system with unreliable machines. Buzacott and Shanthikumar (1993) and Gershwin (1994) presented various systems with failure-prone machines. Song (2009) extensively reviewed the literature on the manufacturing system with machine failures. In this paper we extend the model of Song et al. (1999) to the more realistic case of the multi-item, multiple classes of demand ATO system with unreliable machines. We develop an inventory-queue model for performance analysis and examine how machine parameters such as the machine failure rate and repair rate affect system performance. We also present some numerical examples to show the importances of taking the machine failures into consideration.

The rest of this paper is organized as follows: We present the model description and assumptions in Section 2. In Section 3 we discuss the total-order-service ATO model. Section 4 studies the partial-order-service model. In Section 5 we give some numerical examples to show how different system parameters affect system performance and generate managerial insights on the ATO system. We conclude the paper in Section 6.

2. Model description

We adopt the basic model assumptions and notation as those in Song et al. (1999). There are J different items with $\Omega=\{1,2,\ldots J\}$ being the set of item indices. Let K be a subset of Ω, i.e., $K\subseteq\Omega$, and let $|K|$ denote the number of elements in K. There are multiple classes of demand, each of which requires a fixed kit (subset) of the items and arrives independently according to a Poisson process. We say that a demand is of type K if it requires only one unit of each item in K and zero unit in $\Omega-K$. Let ψ be the collection of all types of demand. Let λ^K be the arrival rate of type K demand, λ be the overall demand arrival rate, i.e., $\lambda=\sum_{K\in\psi}\lambda^K$, and q^K be the probability that an arriving demand is of type K, i.e., $q^K=\lambda^K/\lambda$. We use $S(i)$ to denote the set of all types of demand containing item i, i.e., $S(i)=\{K: i\in K, K\in\psi\}$, so the demand rate for item i is $\lambda_i=\sum_{K\in S(i)}\lambda^K$.

For any i, $i\in\Omega$, the inventory of component i is independently controlled by the base-stock policy with the base-stock level s_i and replenished by its corresponding dedicated unreliable machine i. The processing time is exponentially distributed with a rate μ_i and the machine processes the items on the FCFS basis. Economies of scale in replenishment are not considered. We assume that each machine failure is operational dependent, which means that a machine may fail only when it is operating. The up time of machine i follows an exponential distribution with a mean $1/\xi_i$. Machine i is repaired immediately once it breaks down and it takes an exponentially distributed time with a rate r_i to complete the repair. Let $f_{R_i}(x)$ denote the probability density function of the time to repair machine i. Note that when $r_i\to+\infty$, $i=1,\ldots J$, the model reduces to that of Song et al. (1999). After repair, a machine works as good as new. The processing time, up time and repair time of each machine are mutually independent. We ignore the assembling time because it is much shorter than the item processing time.

The demand for item i can be filled immediately from its buffer if possible; otherwise, it joins the backlog queue i with a finite capacity $b_i\le+\infty$ until item i is available. b_i can be interpreted as customers' patience (see Song et al., 1999); $b_i=0$ means backlog is not allowed, while $b_i=+\infty$ means completely backlogging. When a demand arrives requiring more than one item, but finds some of the required items' backlog queues are full while the others' are not, two kinds of stockout may occur as Song et al. (1999)'s model: total-order-service (TOS), where the demand is accepted or rejected as a whole, and partial-order-service (POS), where the demands for the items are treated separately. Further, if we accept a demand but some of the required items cannot be filled immediately from their buffers, then the whole demand is delayed and backlogged. Throughout the paper we use a superscript to denote the demand type and a subscript to denote the item type.

Notation:

F^{Kx}	order-based type K order fill rate with window x. It is equal to the probability of satisfying a type K order within time period x
SL^K	order-based type K service level. It is equal to the probability of accepting a type K order
F_i^x	item-based fill rate of item i. It is the probability of satisfying the requirement for item i immediately from buffer i
SL_i	item-based service level of item i. It is the probability of accepting the requirement for item i
W^K	the waiting time until a type K demand is filled
$OH_i(t)$	on-hand inventory of item i at time t
$B_i(t)$	backorder of item i at time t
$IO_i(t)$	on-order inventory of item i at time t
$M_i(t)$	the state of machine i at time t, $M_i(t)=0$ or 1, where 0 denotes that the machine is down and 1 denotes that the machine is in operation
W_i	the waiting time of filling a requirement for item i

3. Performance analysis of the TOS model

In this section we study the performance measures of the total-order-service ATO model. Without loss of generality, we assume that at the initial time the on-hand inventory of item $i, i\in\Omega$, is s_i. Inventory is controlled by the base-stock policy. From basic inventory theory, we have

$$OH_i(t)=[s_i-IO_i(t)]^+,\quad B_i(t)=[IO_i(t)-s_i]^+, \tag{1}$$

where the marginal distribution function of $IO_i(t)$ can be derived from the joint distribution function of $[(IO_1(t),M_1(t)),(IO_2(t),M_2(t)),\ldots,(IO_J(t),M_J(t))]$. In the following subsections we derive this joint distribution function.

3.1. Stationary distribution

Let $(IO(t),M(t))=[(IO_1(t),M_1(t)),(IO_2(t),M_2(t)),\ldots,(IO_J(t),M_J(t))]$. It is evident that $(IO(t),\ M(t))$ is a $2J$-dimensional positive recurrent, irreducible, aperiodic Markov process with a finite state space $\Theta=\{(n,m)=[(n_1,m_1),(n_2,m_2),\ldots,(n_J,m_J)]:(n_i,m_i)\in\Theta_i,i\in\Omega\}$, where $\Theta_i=\{(0,1)\bigcup\{(n_i,m_i):n_i=1,2,\ldots,N_i,N_i=s_i+b_i,m_i=0,1\}\}$. Therefore there exists a unique stationary distribution. Let $[(IO_1,M_1),\ldots,(IO_J,M_J)]$ be the stationary state and $\pi(n,m)$ be the stationary distribution, i.e.,

$$\pi(n,m)=\Pr[(IO_1=n_1,M_1=m_1),(IO_2=n_2,M_2=m_2),\ldots\ldots,(IO_J=n_J,M_J=m_J)].$$

There are $\Pi_{i=1}^{J}[2N_i+1]$ system states. The system state changes at any instant when a demand arrives, a machine breaks down, a machine is repaired or an order of item is released. Therefore with transition rate λ_K, the state $[(n_1,m_1),(n_2,m_2),\ldots,(n_J,m_J)]$, satisfying $0\le n_i\le N_i-1$ for any $i\in K$, transits to state $(n',m')=[(n_1',m_1'),(n_2',m_2'),\ldots,(n_J',m_J')]$, where $m_i'=m_i$, and

$$n_i'=\begin{cases}n_i+1, & i\in K,\\ n_i & \text{otherwise.}\end{cases}$$

With transition rate μ_i, the state $[(n_1,m_1),\ldots,(n_i,1),\ldots,(n_J,m_J)]$, where $n_i\ge 1$, transits to the state $[(n_1,m_1),\ldots,(n_i-1,1),\ldots,(n_J,m_J)]$. With transition rate ξ_i, the state $[(n_1,m_1),\ldots,\ (n_i,1),\ldots,(n_J,m_J)]$,$n_i\ge 1$, transits to the state $[(n_1,m_1),\ldots,\ (n_i,0),\ldots,(n_J,m_J)]$; and the converse transition rate is r_i. Then we can derive the infinitesimal generator $\tilde{Q}$ in lexicographical order as given by

$$\tilde{Q}=\begin{bmatrix}\tilde{B}_0 & \tilde{C}_0 & & & & \\ \tilde{A}_1 & \tilde{B} & \tilde{C} & & & \\ & \tilde{A} & \tilde{B} & \tilde{C} & & \\ & & \ddots & \ddots & \ddots & \\ & & & & \tilde{A} & \tilde{B}\end{bmatrix},$$

where

$$\tilde{A}_0=\begin{pmatrix}0\\ \mu_1 I\end{pmatrix},\quad \tilde{A}=\begin{pmatrix}0 & 0\\ 0 & \mu_1 I\end{pmatrix},\quad \tilde{C}_0=(0\ C),\quad \tilde{C}=\begin{pmatrix}C & 0\\ 0 & C\end{pmatrix},$$

$$\tilde{B}_0=B_0,\quad \tilde{B}=\begin{pmatrix}B_0-r_1 I & r_1 I\\ \xi_1 I & B_0-(\mu_1+\xi_1)I\end{pmatrix},\quad \tilde{B}_1=\begin{pmatrix}B_1 & r_1 I\\ \xi_1 I & B_2\end{pmatrix}.$$

It is noted that Q has the block-tridiagonal form, where B_i, $i=0,1,2$, and C are all $\Pi_{i=1}^{J-1}(2N_i+1)$- dimensional square matrices, which can be obtained from the above analysis of the state transition rates and I is an identity matrix with order $\Pi_{i=1}^{J-1}(2N_i+1)$. More details on the infinitesimal generator are given in Latouche and Ramaswamy (1999). This special form of the infinitesimal generator indicates that $(IO(t),M(t))$ is a finite QBD process. Let $\pi=(\pi_0,\pi_1,\ldots,\pi_{N_1})$ denote the system stationary distribution of the joint on-order inventories, where π_i denotes the joint probability distribution of the states with level i.

According to the equilibrium equation $\pi\tilde{Q}=0$, we have

$$\pi_0\tilde{B}_0+\pi_1\tilde{A}_0=0,$$

$$\pi_0\tilde{C}_0+\pi_1\tilde{B}+\pi_2\tilde{A}=0,$$

$$\pi_{k-1}\tilde{C}+\pi_k\tilde{B}+\pi_{k+1}\tilde{A}=0,\quad k=2,3,\ldots,N_1-1,$$

$$\pi_{N_1-1}\tilde{C}+\pi_{N_1}\tilde{B}_1=0.$$

Solving the above equations recursively, we obtain

$$\pi_k=\pi_{k-1}R_k,\quad k=1,2,\ldots,N_1,$$

where

$$R_1=-\tilde{C}_0[\tilde{B}+R_2\tilde{A}]^{-1},$$

$$R_k=-\tilde{C}[\tilde{B}+R_{k+1}\tilde{A}]^{-1},\quad k=N_1-1,\ N_1-2,\ldots,2,$$

$$R_{N_1}=-\tilde{C}\tilde{B}_1^{-1}.$$

π_0 can be obtained by the normalization equation, i.e.,

$$\pi_0 e_1+\sum_{i=1}^{N_1}\pi_i e_2=\pi_0 e_1+\pi_0\sum_{i=1}^{N_1}\prod_{j=1}^{i}R_j e_2=1,$$

where e_i, $i=1,2$, are column vectors of ones.

Note that we solve the $\prod_{i=1}^{J}(2N_i+1)$ equations by the matrix-geometric solution approach, instead of by the direct approach, because it is not only efficient but also effective to do so. To deal with large-scale systems, Song et al. (1999) proposed two approximate methods, which are applicable to our model. One method only focuses on several major demands due to the Pareto phenomenon, while the other method divides the multiple items into several disjoint sets that are either independent or weakly dependent. We propose another method here that does not consider the failure of a machine if its failure rate is much smaller than its production rate.

3.2. Order-based performance measures of the TOS model

The order-based performance measures can be easily obtained by the joint stationary distribution of on-order item inventories. Let

$$(n,m)^K=\{(n_i,m_i):i\in K\},$$

$$C^K=\{(n,m)^K:\ \text{for any } i\in K,n_i<s_i+b_i,m_i=0,1\},$$

$$C^K(L)=\{(n,m)^K:L\subseteq K,\text{ if } i\in L,\text{ then } i \text{ satisfies } s_i\le n_i<s_i+b_i,m_i=0,1;\ \text{else } j\in K-L,\text{ and } j \text{ satisfies } n_j<s_j,m_j=0,1\},$$

$$C^K_{L_0}(L)=\{(n,m)^K\in C^K(L),L_0\subseteq L:\ \text{if } i\in L_0,\text{ then } m_i=0;\ \text{else } i\in L-L_0,m_i=1\},$$

$$\tilde{\pi}_{(n,m)^K}=P\{(n,m)^K|(n,m)^K\in C^K\}=\pi_{(n,m)^K}/\textstyle\sum_{(n,m)^K\in C^K}\pi_{(n,m)^K}.$$

Obviously, $C^K_{L_0}(L)\subseteq C^K(L)\subseteq C^K$ and $\emptyset\in L_0$. $C^K(L)$ is the set of states upon the arrival of a type K demand if it is accepted. Given the requirement for item $i,i\in K$, if $i\in L$, then the item cannot be filled immediately from its buffer and the demand joins the backlog queue; otherwise, if $i\in K-L$, then the requirement for item i can be satisfied immediately.

Therefore we have

$$F^K=P\{(n,m)^K:(n,m)^K\in C^K(\emptyset)\}=P\{(n,m)^K:n_i<s_i,\text{ for any } i\in K\},\tag{2}$$

$$SL^K=P\{(n,m)^K:(n,m)^K\in C^K\}=P\{(n,m)^K:n_i<s_i+b_i,\text{ for any } i\in K\}.\tag{3}$$

In order to calculate the order-based fill rate within a time window, we define $\tilde{S}_i$ as the total time needed to complete the processing of item i, which includes the actual processing time and the machine repair time. It follows that the Laplace–Stieltjes Transform (LST) of $\tilde{S}_i$ is $\tilde{F}_{\tilde{S}_i}(s)=\tilde{F}_{S_i}(s+\xi_i-\xi_i\tilde{F}_{R_i}(s))$ (see Buzacott and Shanthikumar, 1993), where $\tilde{F}_{S_i}(s)$ and $\tilde{F}_{R_i}(s)$ are the LSTs of the processing time and the repair time, respectively. Therefore

we have

$$\tilde{F}_{\tilde{S}_i}(s)=\frac{\mu_i(r_i+s)}{s^2+s(r_i+\xi_i+\mu_i)+r_i\mu_i}. \tag{4}$$

We can obtain the probability distribution function of $f_{\tilde{S}_i}(x)$ from (5) as follows:

$$F^{Kx}=P\{W^K\le x\}=\sum_{(n,m)^K\in C^K(\emptyset)}\tilde{\pi}_{(n,m)^K}+\sum_{L\subseteq K,L\neq\emptyset}\sum_{L_0\subseteq L_{(n,m)^K}}\sum_{\in C^K_{L_0}(L)}P\{W^K \le x|(n,m)^K\}\tilde{\pi}_{(n,m)^K}=\sum_{(n,m)^K\in C^K(\emptyset)}\tilde{\pi}_{(n,m)^K}+\sum_{L\subseteq K,L\neq\emptyset}\sum_{L_0\subseteq L_{(n,m)^K}}\sum_{\in C^K_{L_0}(L)}\left[\prod_{i\in L_0}\int_0^x f_{R_i}(t)*f_{\tilde{S}_i}^{(n_i-s_i+1)}(t)\,dt\cdot\prod_{j\in L-L_0}\int_0^x f_{\tilde{S}_j}^{(n_j-s_j+1)}(t)\,dt\cdot\tilde{\pi}_{(n,m)^K}\right], \tag{5}$$

where $f_{\tilde{S}_j}^{(n)}$ denotes the n-fold convolution of $f_{\tilde{S}_j}$.

3.3. Item-based performance measures of the TOS model

The item-based fill rate and the service level can also be obtained from the joint stationary distribution function of (IO,M) as follows:

$$F_i=\sum_{K\in S(i)}\frac{\lambda^K}{\lambda_i}P\{(n,m)^K: IO_i<s_i, \text{ and for all } j\in K-\{i\}, IO_j<s_j+b_j\}, \tag{6}$$

$$SL_i=\sum_{K\in S(i)}\frac{\lambda^K}{\lambda_i}SL^K=\sum_{K\in S(i)}\frac{\lambda^K}{\lambda_i}P\{(n,m)^K\in C^K\}. \tag{7}$$

Let $\tilde{p}_i^K(n_i,m_i)$ be the conditional probability of the system being in state $IO_i=n_i$, $M_i=m_i$, given a type $K,K\in S(i)$, demand is accepted. It is equal to

$$\tilde{p}_i^K(n_i,m_i)=\frac{P\{(n,m)^K: IO_i=n_i,M_i=m_i, \text{ and for all } j\in K-\{i\}, IO_j<s_j+b_j\}}{P\{(n,m)^K\in C^K\}}, \tag{8}$$

$$F_i^x=P\{W_i\le x\}=\sum_{K\in S(i)}\frac{\lambda^K}{\lambda_i}\left\{\sum_{n_i=0}^{s_i-1}\left[\tilde{p}_i^K(n_i,0)+\tilde{p}_i^K(n_i,1)\right]+\sum_{n_i=s_i}^{s_i+b_i-1}\left[\left(\int_0^x f_{R_i}(t)*f_{\tilde{S}_i}^{(n_i-s_i+1)}(t)\,dt\right)\tilde{p}_i^K(n_i,0)+\left(\int_0^x f_{\tilde{S}_i}^{(n_i-s_i+1)}(t)\,dt\right)\tilde{p}_i^K(n_i,1)\right]\right\}.$$

Therefore the average expected item-based waiting time is

$$E[W_i]=\sum_{K\in S(i)}\frac{\lambda^K}{\lambda_i}\sum_{n_i=s_i}^{s_i+b_i-1}\left\{\frac{\tilde{p}_i^K(n_i,0)}{r_i}+(n_i-s_i+1)\left(\frac{1}{\mu_i}+\frac{\xi_i}{\mu_i r_i}\right)(\tilde{p}_i^K(n_i,0)+\tilde{p}_i^K(n_i,1))\right\}.$$

4. Performance analysis of the POS model

In this section we consider the partial-order-service model. Under this model, a type K demand is no longer accepted or rejected as a whole, but the constituent items are treated separately, whereby the requirements for items whose backlog queues are not full are accepted, while the requirements for items whose backlog queues are full are rejected.

4.1. Order-based performance measures of the POS model

The analysis is similar to that of the TOS model, with differences in matrices B_i, $i=0,1,\ldots,4$, and C of the infinitesimal generator. In other words, the only differences are in rate $\lambda^K,K\in\psi$, and in the way in which state (n,m) enters state (n',m'), where $m_i'=m_i$,

$$n_i'=\begin{cases}n_i+1 & \text{if } i\in K \text{ and } n_i<s_i+b_i,\\ n_i & \text{otherwise.}\end{cases}$$

No changes happen to the other transition rates. Therefore following the same procedure as that we applied to the TOS model, we can derive the joint steady-state distribution π of (IO,M), from which we can obtain the stationary performance measures. Define

$$D^{K(Q)}=\{(n,m)^K,Q\subseteq K:\ \text{if } i\in Q,n_i<s_i+b_i\},\quad \text{where } Q\neq\emptyset,$$

$$D^{K(Q)}(I)=\{(n,m)^K\in D^{K(Q)},I\subseteq Q:\ \text{if } i\in I,s_i\le n_i<s_i+b_i;\ \text{else } i\in Q-I,n_i<s_i\},$$

$$D^{K(Q)}[I(I_0)]=\{(n,m)^K\in D^{K(Q)}(I),I_0\subseteq I:\ \text{if } i\in I_0,m_i=0;\ \text{else } i\in I-I_0,m_i=1\},$$

$$\overline{\pi}_{(n,m)^K}=\pi_{(n,m)^K}/[1-P\{(n,m)^K:\ \text{for all } i\in K,n_i\ge s_i+b_i\}].$$

The order-based fill rate of the POS model is the probability that accepted orders for items can be satisfied immediately, while other orders are rejected and no orders join the backlog queues. Therefore

$$F^K=P\{(n,m)^K:(n,m)^K\in D^{K(K)}(\emptyset)\}, \tag{9}$$

$$SL^K=P\{(n,m)^K:(n,m)^K\in D^{K(K)}\}, \tag{10}$$

$$F^{Kx}=P\{W^K\le x\}=\sum_{Q\subseteq KI\subseteq Q,I\neq\emptyset}\sum_{I_0\subseteq I_{(n,m)^K}}\sum_{\in D^{K(Q)}[I(I_0)]}\left[\prod_{i\in I_0}\int_0^x f_{R_i}(t)*f_{\tilde{S}_i}^{(n_i-s_i+1)}(t)\,dt\cdot\prod_{j\in I-I_0}\int_0^x f_{\tilde{S}_j}^{(n_j-s_j+1)}(t)\,dt\right]\cdot\overline{\pi}_{(n,m)^K}+\sum_{Q\subseteq K}\sum_{(n,m)^K\in D^{K(Q)}(\emptyset)}\overline{\pi}_{(n,m)^K}. \tag{11}$$

4.2. Item-based performance measures of the POS model

The stationary item-based performance measures of item i in the POS model can be determined by the marginal distribution of $IO_i,i\in J$, which can be obtained from the joint stationary distribution of (IO,M). But we present another method. POS means that when a demand arrives, the requirements for the items are treated independently, which are either accepted or rejected. Therefore we formulate the system as an $M/M/1/N_i$ queue with unreliable server i. Let $p_i(n_i,m_i),(n_i,m_i)\in\{(0,1)\bigcup\{(n_i,m_i):1\le n_i\le N_i,m_i=0,1\}\}$ denote the stationary distribution of the system being in state $(IO_i=n_i,\ M_i=m_i)$ and $p_i(n_i)=(p_i(n_i,0),p_i(n_i,1))$, i.e.,

$$p_i(0,1)=T,$$

$$p_i(1)=\left(\frac{\lambda_i\xi_i}{(\lambda_i+r_i)\mu_i},\frac{\lambda_i}{\mu_i}\right)T,$$

$$p_i(n_i)=p_i(1)R_i^{n_i-1},\quad n_i=2,3,\ldots,N_i-1,$$

$$p_i(N_i)=p_i(N_i-1)R_i', \tag{12}$$

where

$$R_i = \begin{pmatrix} \frac{\lambda_i}{\lambda_i + r_i}\left(1+\frac{\xi_i}{\mu_i}\right) & \frac{\lambda_i}{\mu_i} \\ \frac{\lambda_i}{\lambda_i + r_i}\frac{\xi_i}{\mu_i} & \frac{\lambda_i}{\mu_i} \end{pmatrix}, \quad R_i' = \begin{pmatrix} \frac{\lambda_i}{r_i}\left(1+\frac{\xi_i}{\mu_i}\right) & \frac{\lambda_i}{\mu_i} \\ \frac{\lambda_i \xi_i}{r_i \mu_i} & \frac{\lambda_i}{\mu_i} \end{pmatrix}$$

and T can be derived by the normalization equation. Then we have

$$F_i = p_i(0,1) + \sum_{1 \le n_i < s_i} p_i(n_i)e, \tag{13}$$

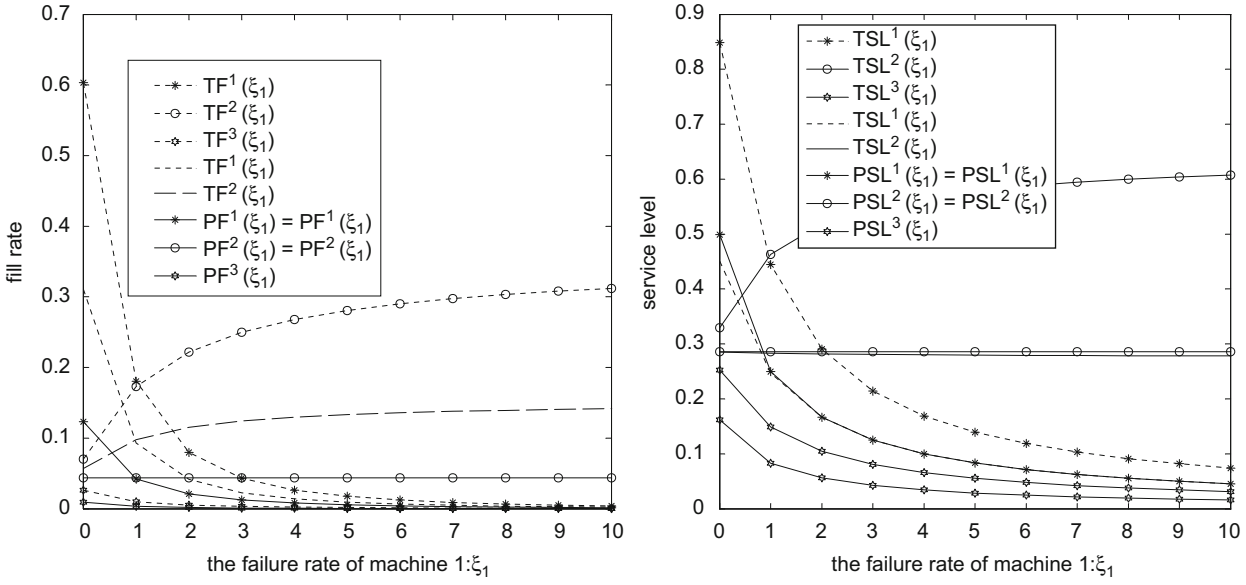

Fig. 1. Performance measures of both TOS and POS versus ξ_1: $\lambda^1 = 2$, $\lambda^2 = 3$, $\lambda^3 = 4$, $\mu_1 = \mu_2 = 3$, $\xi_2 = 0.5$, $r_1 = r_2 = 1$, $s_1 = s_2 = 6$, $b_1 = b_2 = 2$.

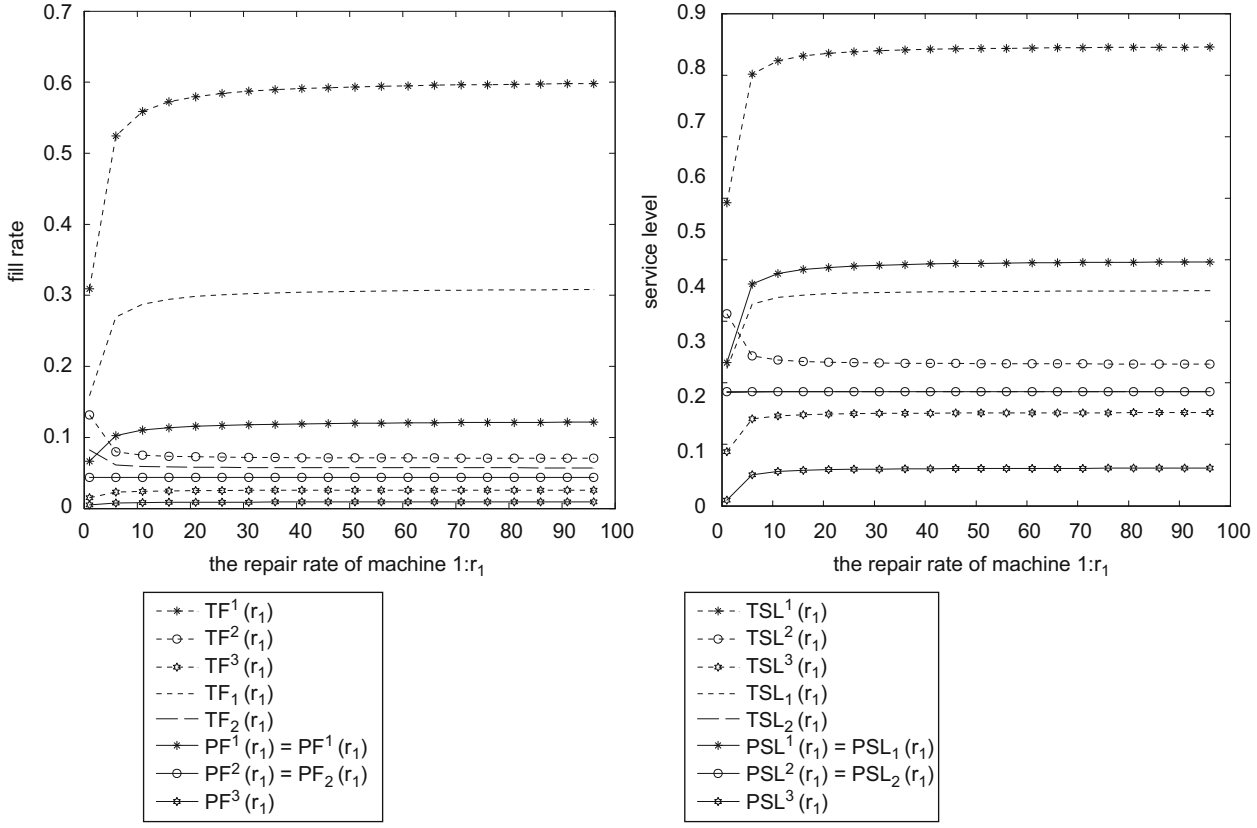

Fig. 2. Performance measures of both TOS and POS versus r_1: $\lambda^1 = 2$, $\lambda^2 = 3$, $\lambda^3 = 4$, $\mu_1 = \mu_2 = 3$, $\xi_1 = \xi_2 = 0.5$, $r_2 = 1$, $s_1 = s_2 = 6$, $b_1 = b_2 = 2$.

$$SL_i = p_i(0,1) + \sum_{1 \le n_i < N_i} p_i(n_i)e, \tag{14}$$

where e is a column vector of ones.

Let $\tilde{p}_i(n_i,m_i)$ be the conditional probability that, given that an order for item i occurs and the order is accepted, the system is in state $(IO_i=n_i, M_i=m_i)$, i.e.,

$$\tilde{p}_i(n_i,m_i) = \frac{p_i(n_i,m_i)}{1-p_i(N_i)e}. \tag{15}$$

It follows that

$$F_i^x = P\{W_i \le x\} = \tilde{p}_i(0,1) + \sum_{n_i=1}^{s_i-1} \tilde{p}_i(n_i)e + \sum_{n_i=s_i}^{N_i-1} \left[\int_0^x f_{R_i}(t) * f_{\tilde{S}_i}^{(n_i-s_i+1)}(t)\,dt \cdot \tilde{p}_i(k,0) + \int_0^x f_{\tilde{S}_i}^{(n_i-s_i+1)}(t)\,dt \cdot \tilde{p}_i(k,1) \right], \tag{16}$$

$$E(W_i) = \sum_{n_i=s_i}^{N_i-1} \left[\frac{1}{r_i} \cdot \tilde{p}_i(k,0) + (n_i-s_i+1)\left(\frac{1}{\mu_i} + \frac{\xi_i}{\mu_i r_i}\right)\tilde{p}_i(k)e \right]. \tag{17}$$

5. Numerical examples

In this section we present some numerical examples to investigate the effects of changes in various system parameters on both order-based and item-based performance measures of the TOS and POS models. We assume that there are two items and three types of demand for the ATO system under study. We use λ^1, λ^2 and λ^3 to denote the arrival rates of demands requiring only item 1, only item 2 and both items, respectively.

Fig. 1 shows how performance measures respond to changes in the failure rate ξ_1. From Fig. 1, we see that the order-based fill rate and service level of demand $K, K \in S(i)$, the item-based fill rate and service level of item i decrease with ξ_i, while the order-based fill rate and service level of demand $K, K \notin S(i)$, increase with ξ_i. The item-based fill rate and service level of item $j, j \neq i$, in the TOS model also increase with ξ_i, but they do not change in the POS model. The effect of increasing the repair rate is equivalent to that of decreasing the failure rate of a machine, so the performance measures respond to repair rate changes in a manner opposite to that to failure rate changes. We skip the analysis of repair rate's effects on system performance. The reader may refer to Fig. 2 for details. Particularly, when $r_1 \to \infty$, the performance measures tend to be steady and the influence of repair rate on system performance decreases. Under such a condition, our model reduces to the ATO system with reliable machines. In other words, the machine failure rate and repair rate have a profound effect on system performance.

Figs. 3–5 show how demand arrival rates affect system performance measures. Fig. 3 shows the effects of λ^1 on the performance measures of the TOS model and the POS model, where the other parameters were fixed. From the graphs, we can conclude that the order-based fill rate and service level of demand K, and the item-based fill rate and the service level of item i, where $i \in K$, decrease with λ^K in both models. It is noted that the arrival rate of demand K has no influence on the order-based fill rate and service level of demand M, where $M \notin \{\cup_{i \in K} S(i)\}$, in both models. This means that the order-based fill rate of demand M is

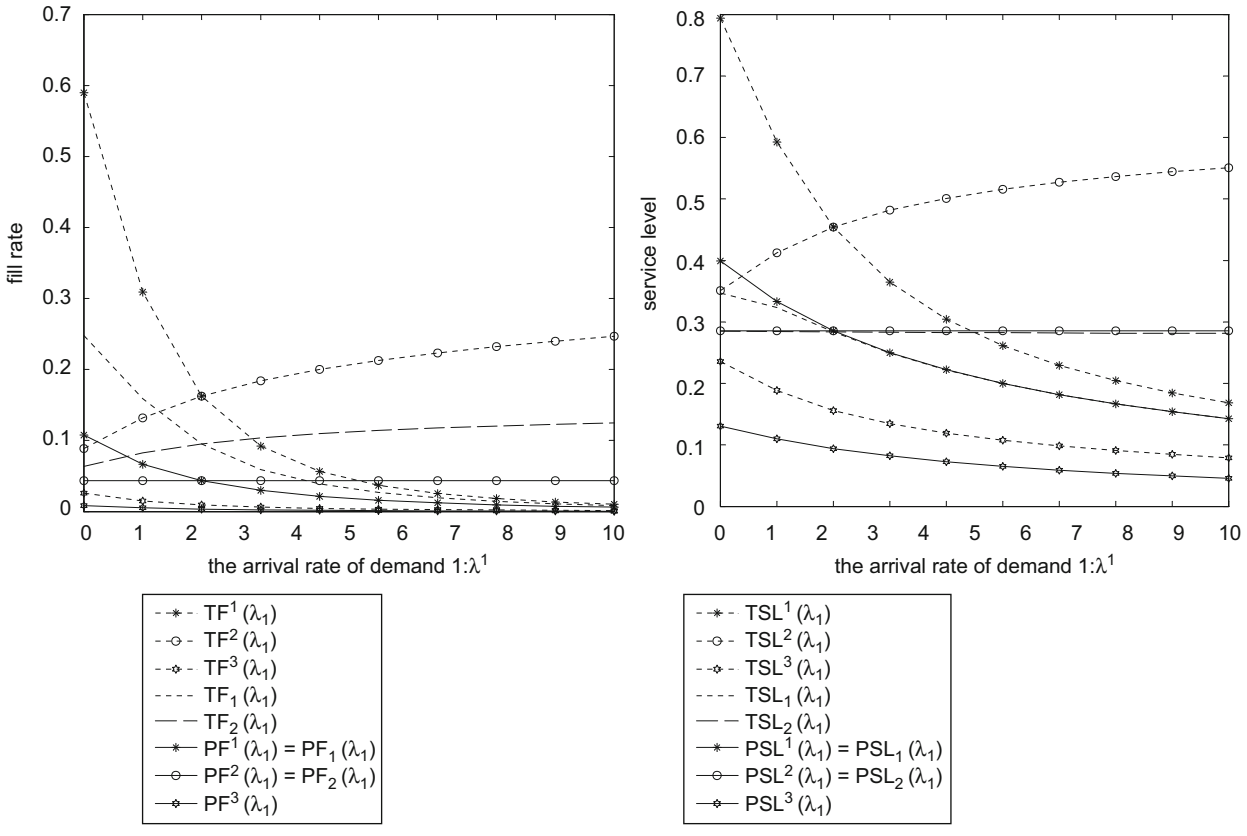

Fig. 3. Performance measures of both TOS and POS versus λ^1: $\lambda^2=3$, $\lambda^3=4$, $\mu_1=\mu_2=3$, $\xi_1=\xi_2=0.5$, $r_1=r_2=1$, $s_1=s_2=6$, $d_1=d_2=2$.

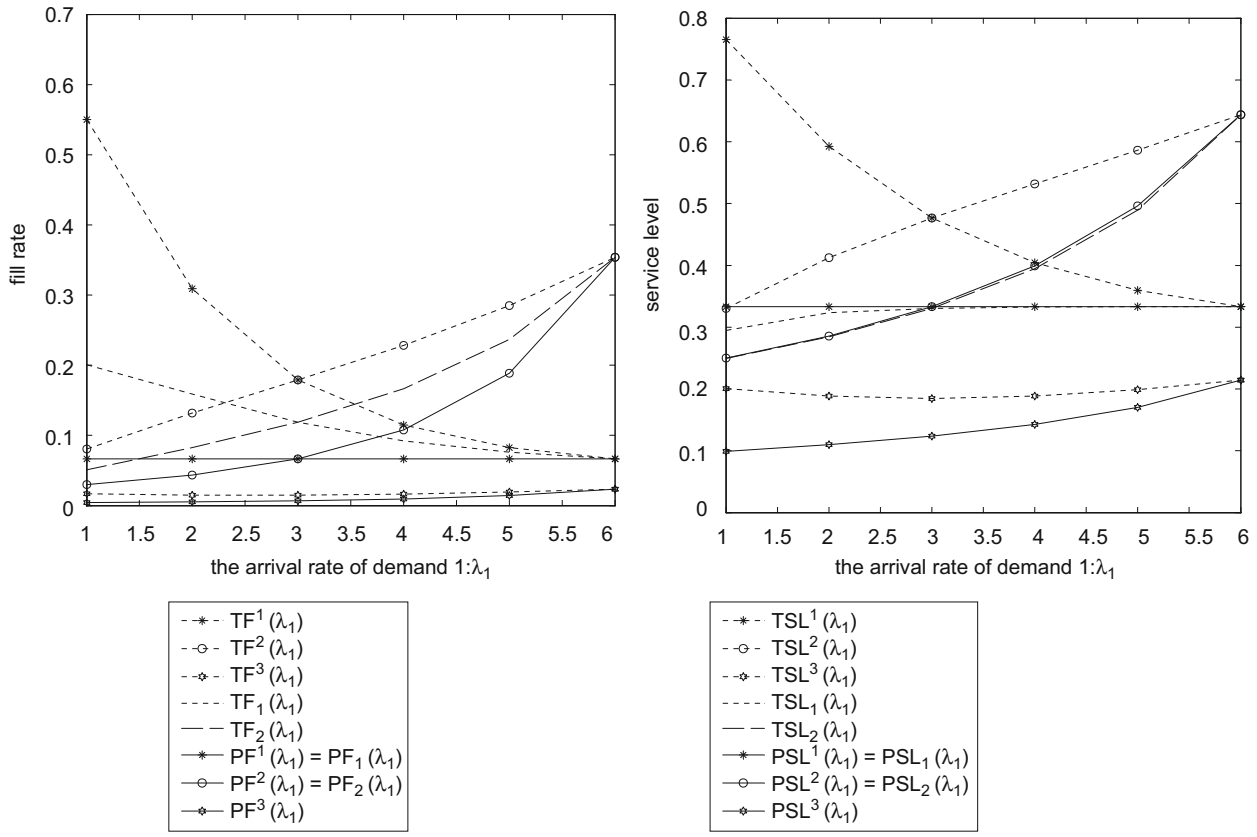

Fig. 4. Performance measures of both TOS and POS versus λ^1 with fixed λ_1: $\lambda_1=\lambda^1+\lambda^3=6$, $\lambda^2=3$, $\mu_1=\mu_2=3$, $\xi_1=\xi_2=0.5$, $r_1=r_2=1$, $s_1=s_2=6$, $d_1=d_2=2$.

only affected by the demand rate of $\bigcup_{i\in M}S(i)$. The item-based fill rate and service level of item j, where $j\notin K$, in the POS model are not affected by λ^K. These results are not surprising. In Fig. 4, we kept the total demand rate for item 1 at a fixed level, i.e., $\lambda_1=\lambda^1+\lambda^3=6$, and investigated the performance measures. First, the order-based fill rate and service level of demand 2 in both models increase with λ^1. This is because λ^3 decreases as λ^1 increases, which could increase the probability of filling demand 2 immediately. Second, the order-based fill rate and service level of item i in the POS model are unchanged since λ_i is fixed. Third, the item-based fill rate and service level of item 2 in the TOS model change without the monotone property because demand 3 requires item 2. Fourth, in the POS model, the item-based fill rate and service level of item i decrease with the total demand rate for item i, i.e., λ_i. In Fig. 5, we changed the arrival rate of a demand that requires multiple items. From the figure we see that the performance measures associated to that required items are affected. Obviously, the order-based fill rate and service level of demand K decrease with λ^K. In particular, the item-based fill rate and service level of item $i, i\in K$, in the POS model decrease with λ^K.

How to allocate inventory buffers with capacity limits to each inventory item in order to improve performance measures and save cost is a great challenge faced by inventory managers. Tables 1 and 2 show the performance measures by fixing $s_1+s_2=12$ in both models. We see that the performance measures fluctuate to some extent. In Table 1, the order-based fill rate of demand 1 reaches the maximum at $s_1=6$, while that of demand 2 reaches the maximum at $s_1=2$. The fluctuations of the order-based fill rate and service level of demand 3 are much smaller because demand 3 requires both two items, so increasing one but decreasing the other balances out their individual influences. In Table 2, the changes in the POS model are not obvious, which may look unreasonable at first sight. But if we look deeper, we see that the reason may be due to machine failure. It is evident that items are consumed more quickly in the POS model, which means the machines must keep operating, which increases the probability of machine breakdown. In conclusion, increasing s_1 while simultaneously increasing the probability of machine breakdown neutralizes base-stock level's influence on system performance. These observations provide good insights to inventory management in that if inventory capacity is limited, we can adjust the base-stock level of each item to maximize system performance.

6. Concluding remarks

We examined a multi-item, multiple classes of demand assemble-to-order system. Each demand arrives according to a Poisson process with a certain rate and requires a subset of the items. Inventories are kept at the item-level and controlled by the base-stock policy with unreliable production facilities. Stockout of the ATO system is divided into two kinds, namely total-order-service and partial-order-service. Under certain assumptions, we formulated the system as a queuing network with the infinitesimal generator being a QBD process. Applying the matrix-geometric solution approach, we derived the exact joint stationary distribution of on-hand inventories effectively and efficiently, based on which we computed some key system performance measures. We then presented numerical examples to show how various system parameters affect system performance and generate useful managerial insights into the ATO system.

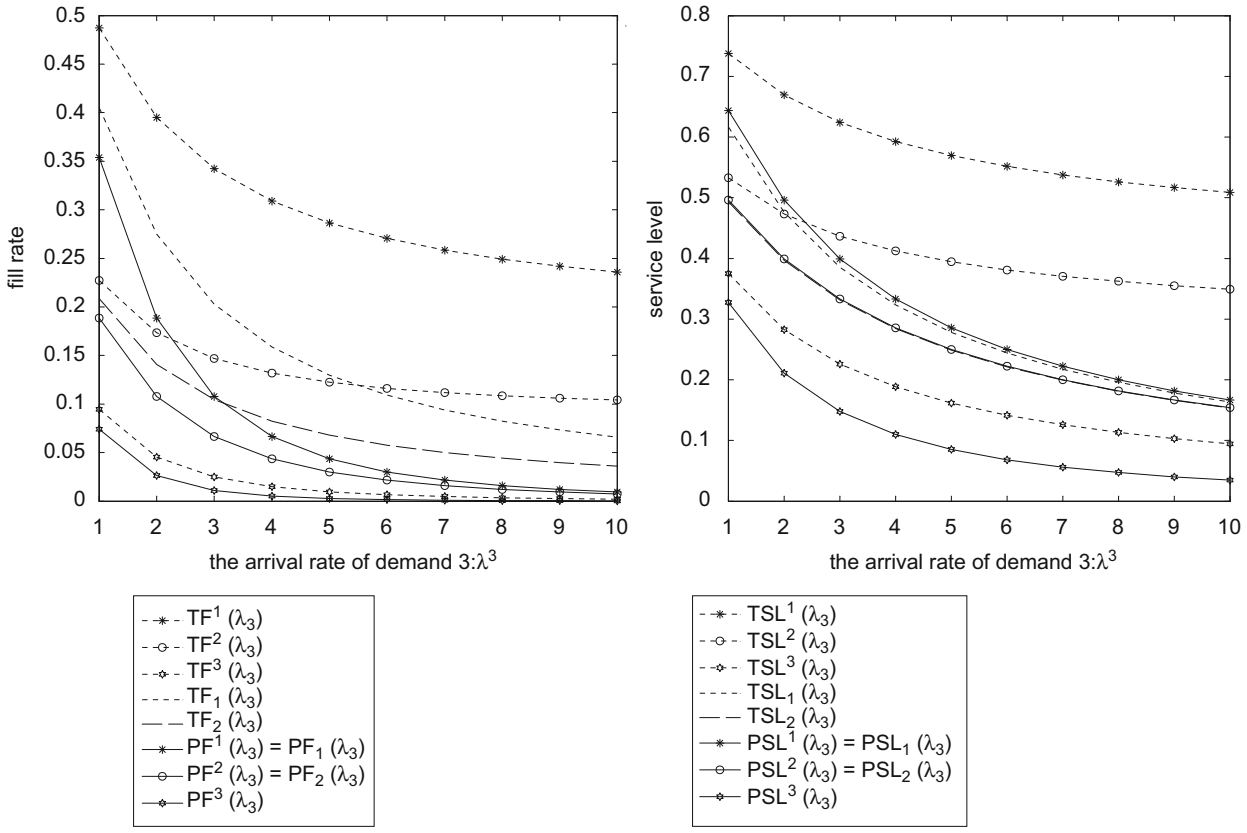

Fig. 5. Performance measures of both TOS and POS versus λ^3: $\lambda^1=2$, $\lambda^2=3$, $\mu_1=\mu_2=3$, $\xi_1=\xi_2=0.5$, $r_1=r_2=1$, $s_1=s_2=6$, $d_1=d_2=2$.

Table 1
Performance measures of TOS versus s_1 with fixed $s_1+s_2=12$: ($\lambda^1=2$, $\lambda^2=3$, $\lambda^3=4$, $\mu_1=\mu_2=3$, $\xi_1=\xi_2=0.5$, $r_1=r_2=1$, $d_1=d_2=2$).

s_1	s_2	F^1	F^2	F^3	SL^1	SL^2	SL^3	F_1	F_2	SL_1	SL_2
2	10	0.213	0.145	0.011	0.544	0.429	0.178	0.108	0.089	0.300	0.286
3	9	0.253	0.141	0.013	0.563	0.423	0.182	0.130	0.087	0.309	0.285
4	8	0.279	0.138	0.014	0.576	0.419	0.185	0.143	0.086	0.315	0.285
5	7	0.296	0.135	0.015	0.586	0.415	0.187	0.152	0.084	0.320	0.285
6	6	0.309	0.132	0.015	0.593	0.412	0.188	0.159	0.082	0.323	0.284
7	5	0.318	0.128	0.015	0.599	0.409	0.189	0.164	0.080	0.326	0.283
8	4	0.326	0.123	0.015	0.604	0.405	0.189	0.168	0.077	0.327	0.282
9	3	0.333	0.114	0.015	0.608	0.400	0.189	0.171	0.072	0.329	0.279
10	2	0.340	0.098	0.014	0.614	0.391	0.188	0.174	0.062	0.330	0.275

Table 2
Performance measures of POS versus s_1 with fixed $s_1+s_2=12$: ($\lambda^1=2$, $\lambda^2=3$, $\lambda^3=4$, $\mu_1=\mu_2=3$, $\xi_1=\xi_2=0.5$, $r_1=r_2=1$, $d_1=d_2=2$).

s_1	s_2	$F^1=F_1$	$F^2=F_2$	F^3	$SL^1=SL_1$	$SL^2=SL_2$	SL^3
2	10	0.0561	0.0438	0.0046	0.3274	0.2857	0.1081
3	9	0.0624	0.0438	0.0049	0.3307	0.2857	0.1090
4	8	0.0650	0.0437	0.0051	0.3322	0.2857	0.1094
5	7	0.0661	0.0437	0.0051	0.3328	0.2857	0.1096
6	6	0.0665	0.0437	0.0052	0.3331	0.2856	0.1096
7	5	0.0668	0.0435	0.0051	0.3332	0.2855	0.1096
8	4	0.0668	0.0430	0.0051	0.3333	0.2852	0.1095
9	3	0.0669	0.0417	0.0050	0.3333	0.2845	0.1093
10	2	0.0669	0.0382	0.0046	0.3333	0.2826	0.1087

Many interesting problems remain to be explored. Demands are not necessarily serviced on the FCFS basis and they may have different priorities. Does the priority allocation policy outperform the FCFS allocation policy? Does an optimal inventory allocation policy exist? What is the optimal inventory allocation policy? We should also consider how preventive maintenance improves system performance.

Acknowledgments

This work was supported in part by The Hong Kong Polytechnic University under Grant number G-U634, the National Natural Science Foundation of China under Grant number 70831002, and Jiangsu Natural Science Foundation under Grant number BK2008273. The authors thank the anonymous reviewers for helpful comments.

References

Benjaafar, S., EI Hafsi, M., 2006. Production and inventory control of a single product assemble-to-order system with multiple customer classes. Management Science 52, 1896–1912.

Buzacott, J., Shanthikumar, J., 1993. Stochastic Models of Manufacturing Systems. Prentice-Hall, Englewood Cliffs, NJ.

Dayanik, S., Song, J.S., Xu, S.H., 2003. The effectiveness of several performance bounds for capacitated production, partial-to-service, assemble-to-order systems. Manufacturing and Service Operations Management 5, 230–251.

El Hafsi, M., 2009. Optimal integrated production and inventory control of an assemble-to-order system with multiple non-unitary demand classes. European Journal of Operational Research 194, 127–142.

Gershwin, S.B., 1994. Manufacturing Systems Engineering. Prentice-Hall, Englewood Cliffs, NJ.

Hoen, K.M.R., Gullu, R., van Houtum, G.J., Vliegen, I.M.H., 2010. A simple and accurate approximation for the order fill rates in lost-sales assemble-to-order systems. International Journal of Production Economics, in press, doi:10.1016/j.ijpe.2009.12.012.

Latouche, G., Ramaswamy, V., 1999. Introduction to Matrix Analytic Methods in Stochastic Modeling. SIAM, Pennsylvania.

Lu, Y.D., 2008. Performance analysis for assemble-to-order systems with general renewal arrivals and random batch demands. European Journal of Operational Research 185, 635–647.

Lu, Y.D., Song, J.S., 2005. Order-based cost optimization in assemble-to-order systems. Operations Research 53, 151–169.

Lu, Y.D., Song, J.S., Yao, D., 2003. Order fill rate, lead-time variability, and advance demand information in an assemble-to-order system. Operations Research 51, 292–308.

Lu, Y.D., Song, J.S., Yao, D., 2005. Backorder minimization in multiproduct assemble-to-order systems. IIE Transactions 37, 763–774.

Song, D.P., 2009. Production and preventive maintenance control in a stochastic manufacturing system. International Journal of Production Economics 119, 101–111.

Song, J.S., 1998. On the order fill rate in a multi-item, base-stock inventory system. Operations Research 46, 739–743.

Song, J.S., 2002. Order-based backorders and their implications in multi-item inventory systems. Management Science 48, 499–516.

Song, J.S., Xu, S.H., Liu, B., 1999. Order-fulfillment performance measures in an assemble-to-order system with stochastic leadtimes. Operations Research 47, 131–149.

Song, J.S., Yao, D., 2002. Performance analysis and optimization of assemble-to-order systems with random lead times. Operations Research 50, 889–903.

Song, J.S., Zipkin, P., 2003. Supply chain operations: assemble-to-order system. In: de Kok, T., Graves, S. (Eds.), Handbooks in Operations Research and Management Science. North-Holland, Amsterdam, pp. 561–596.

Xiao, Y., Chen, J., Lee, C.Y., 2010. Optimal decisions for assemble-to-order systems with uncertain assembly capacity. International Journal of Production Economics 123, 155–165.

Zhao, Y., 2009. Analysis and evaluation of an assemble-to-order system with batch ordering policy and compound Poisson demand. European Journal of Operational Research 198, 800–809.