Mathematical and Computer Modelling 54 (2011) 2750–2759

Contents lists available at SciVerse ScienceDirect

Mathematical and Computer Modelling

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

Inventory management in a base-stock controlled serial production system with finite storage space

Kuo-Hwa Chang, Yang-Shu Lu *

Department of Industrial and Systems Engineering, Chung Yuan Christian University, 200, Chung Pei Rd., Chung-Li, 32023, Taiwan, ROC

ARTICLE INFO

Article history:
Received 11 June 2010
Received in revised form 28 June 2011
Accepted 28 June 2011

Keywords:
Queuing network
Phase-type distribution
Production planning
Inventory control
Base-stock policy

ABSTRACT

We consider a serial production system controlled by the base-stock policy, in which customer demands must be satisfied immediately or it will be considered lost. Since the exact analysis is impossible for the general system, we present a phase-type approximation for a base-stock controlled serial production system. The numerical results indicate that this approximation provides good estimates for performance measures such as fill rate and mean queue-length distributions of each station. In addition, a cost model is constructed to determine the optimal base-stock level.

1. Introduction

The purpose of supply chain management is to minimize operating costs while maintaining customer satisfaction. Three factors that cause uncertainty are widespread in the supply chain system. They are the difficulty in predicting customer demands, high inventory level, and an unstable manufacturing process (see [1]). Among them, difficulty in predicting customer demands is the main source of uncertainty. Therefore, businesses must maintain high inventory in order to satisfy unpredictable customer demands. Businesses use various production strategies or technological improvements, such as Total Quality Management (TQM), Enterprise Resource Planning (ERP), and 6 Sigma etc., to increase production efficiency. The main issue is to maintain Quality of Service (QoS) while reducing factors that cause uncertainty.

Due to the uncertainty and diversity of customer demands, manufacturers have switched from the traditional Build-to-Forecast (BTF) or Make-to-stock (MTS) strategies to Build-to-Order (BTO) or Make-to-Order (MTO) strategies. Taking the laptop computer assembly as an example, when a laptop computer company receives orders from Internet, it will pass these orders to the assemblers and ask them to have the product shipped immediately. A typical example of a detailed assembly process is shown in Table 1 with the maximum and minimum of operating time needed in each step. The laptop computers are stored in the warehouse after the testing step (the 6th step). In order to achieve 100% QoS, assemblers can prepare a lot of finished laptop computers in the warehouse, however, it is not cost-efficient and is unrealistic. As seen in Table 1, the total assembly time is between 31.88 h and 40.48 h. In stead of preparing a lot of finished computers, it is still possible to adopt a Build-to-order strategy if we start to assemble a new computer with plenty of time ahead. The assembler initially prepares some finished laptop computers in the warehouse, called base stock, and starts to assemble a new laptop computer when an order is received. This new one will used to fulfill the potential orders that arrive later. Another example can be seen in a fast-food restaurant. Customers coming to restaurant order their food and want to have it immediately. Fulfilling orders immediately can be done by preparing a lot of cooked food in the food rack, however, some of it may remain in the rack too long to sustain its original quality. Therefore, only a certain number of newly cooked food will be prepared in the rack and

* Corresponding author. Tel.: +886 3 2654427; fax: +886 3 2654499.
E-mail address: g9402404@cycu.edu.tw (Y.-S. Lu).

0895-7177/$ – see front matter
doi:10.1016/j.mcm.2011.06.063

Table 1
Typical assembly steps of laptop computers.

Station	Step	Min	Max
1	Sales order	0.41	0.62
2	Production planning	1.48	2.22
3	Preparing materials	1.85	2.38
4	Manufacturing	17.66	21.70
5	Testing	3.90	4.76
6	Shipping	6.58	8.08
			Unit time: hour

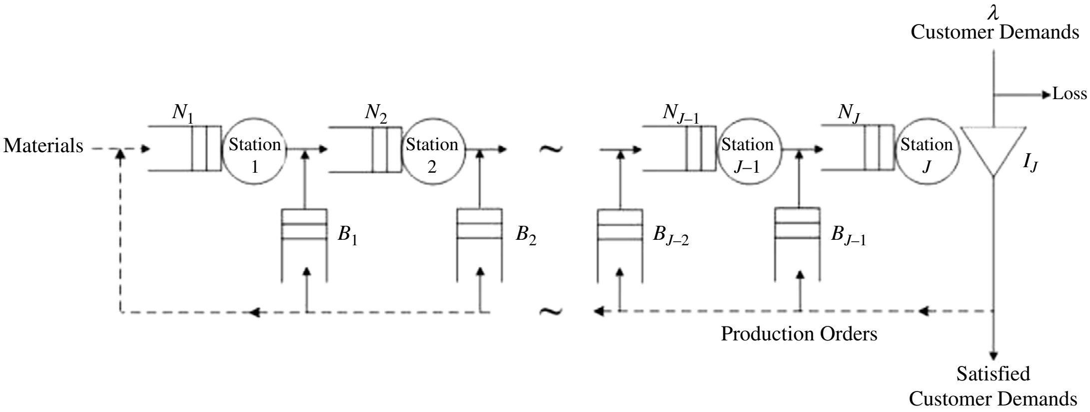

Fig. 1. The serial production system.

new ones will be made if an order is received. To emphasize the immediate need for products in modeling, we can regard the customers in the above systems as no-wait customers that will be lost if they cannot be satisfied immediately.

Motivated by the above examples, in this study we consider a serial production system (see Fig. 1) with finite storage space. In our system each station, except at final station, adopts MTO production to produce semi-finished products. The final station adopts Make-to-stock (MTS) production and is controlled by base-stock policy. Under base-stock policy, initially there are some finished products in inventory. When a demand arrives, if there are finished products in inventory, this demand will be satisfied immediately. If there are no finished products in inventory, this demand will be lost. A production order will be placed whenever a finished product is taken from inventory or, equivalently, a demand is satisfied. In our system, the percentage of satisfied customer demands, called the fill rate, is considered as the key measure of QoS.

For such a system the first issue is to obtain the optimal base-stock levels at each station. Karl and Stefan [2] determined safety stocks in multi-stage inventory systems with normally distributed demands. The system followed a periodic review base-stock policy. In this paper, they concentrated on situations with stochastic demands and safety stock buffering. They restricted their analysis to high demand items, the stochastic behavior of which can (at least approximately) be described by normal distributions. Duenyas and Patana-Anake [3] considered a multiple-stage tandem production/inventory system producing a single product. They formulated the optimal control problem as a Markov decision process (MDP). Boyaci and Gallego [4] considered a serial base-stock system and studied the problem of minimizing the expected inventory holding costs subject to fill-rate constraints. They also developed bounds on the total system-stock and on base-stock levels of each stage and developed an algorithm to find the optimal inventory policy. Shang and Song [5] studied a serial base-stock inventory model with Poisson demand and a fill-rate constraint. They developed a closed-form approximation for the optimal base-stock levels so that the average inventory-holding cost is minimized while meeting the required fill rate. Chao and Zhou [6] considered the infinite horizon serial inventory system and the optimal echelon base-stock levels for minimizing average cost and discounted cost are obtained in terms of only probability distributions of lead time demands. They also developed bounds and heuristics for optimal inventory control polices.

A serial production system without base stock in which customers will wait in the system for their finished products is relatively easy to analyze. It is actually a classic tandem queuing system, however, the exact analysis on tandem queue system with base stock is usually impossible. Buzacott et al. [7] and Svoronos et al. [8] both developed approximation approaches for analyzing the serial production system with intermediate inventory and each station is controlled by a base-stock policy. Buzacott et al. [7] studied the recursive relations of the waiting times based on sample path analysis. Lee and Zipkin [9] used phase-type approximation mentioned in Svoronos and Zipkin [8] to approximate the number of customers and the waiting times in a system by assuming each station is an M/M/1 queue. By the tests results, their approach provides quite accurate estimates on performance measures. Zipkin [10] considered the same system as in Lee and Zipkin [9] but with feedback customers. Duri et al. [11] showed that these two approaches by Buzacott et al. [7] and Lee and Zipkin [9] are equivalent. They then extended Lee and Zipkin's system by introducing Coxian-2 service times. Wang and Su [12] also extended Lee and Zipkin's system by adding an end station with hyper-exponential service time that can be regarded as a distribution center in a supply-chain network. Based on the above researches, it turns out that the phase-type approximation is also a good way to study the similar systems and can be further applied on other variants.

Our system is similar to the one in Lee and Zipkin [9], however, in our system customers demands must be satisfied immediately otherwise they will be lost. We extend their phase-type approximation to our system to estimate the effective arrival rate of customers and mean queue-length of each queue. Based on this analysis, we numerically determine the optimal base-stock level according to certain cost structure under the requirement on the fill rate. When there are no intermediate inventories, our system can be analyzed as a closed Jackson network, however, our phase-type approximation approach can be applied to the system with intermediate inventory and this analysis can be extend to other system such as the system with multiple class demands.

The remainder of this paper is organized as follows. In Section 2, firstly we introduce a closed Jackson model for the system without intermediate inventories. Then we present our approach based on phase-type approximation. The approximate method is validated and some numerical examples are tested in Section 3. We conclude our study in Section 4.

2. Inventory-queue model

The base-stock controlled serial production system with finite storage space comprises J single server queues as shown in Fig. 1. We assume that customer demands arrive according to a Poisson process with rate λ. The production time of each station is assumed to be exponentially distributed with respective rates $\mu_j,\ j = 1, \ldots, J$. There are inventories of finished semi-products at each station. When a finished product at station J is taken, a production order will be released to station J itself and it needs a semi-product from station $J-1$. Therefore a demand will also be sent to station $J-1$ as a backorder to require a finished semi-product. Therefore, each station will send a request to its precedent station to require a semi-product and eventually a request will be sent to the first station. If the semi-product is available from the precedent station, the backorder then becomes a production order. At each station, there are queues of backorders and production orders and inventories of finished semi-products. For $j = 1, \ldots, J$, let N_j be the number of production orders at station j (waiting or under processing); let B_j be the number of backorders at station j for $j = 1, \ldots, J-1$ and I_J be the number of finished products at station J. S_j denotes the base-stock level for the finished products at station j. Without any confusion, we also use N_j, B_j and I_j to name the corresponding queues, respectively.

Since production orders can be generated only when one finished product is taken from base-stock, we have

$$\sum_{j=1}^{J} N_j \le S_J, \tag{1}$$

$$B_j = \sum_{i=1}^{j} N_i, \quad j = 1, \ldots, J-1, \tag{2}$$

and

$$I_J = S_J - \sum_{j=1}^{J} N_j. \tag{3}$$

By (2) and (3), we can also obtain $B_{J-1} + N_J = S_J$.

For the system without intermediate inventory, let n_j be the number of production orders at station $j,\ j = 1, \ldots, J$. Let Ω denote the state space and $\Omega = \{(n_1, n_2, \ldots, n_J) | \sum_{j=1}^{J} n_j \le S_J\}$. Since each taken finished product can generate one production order to the system and there are no intermediate inventories, the underlying queuing model with state space Ω is actually a closed Jackson Network with finite population S_J. The joint distribution of queue lengths has a product-form structure with joint probability distribution

$$P(n_1, n_2, \ldots, n_J) = c \cdot \rho_1^{n_1} \rho_2^{n_2} \cdots \rho_J^{n_J}, \tag{4}$$

where $\rho_j = \lambda/\mu_j$ and c is the normalization constant

$$c = \left(\sum_{n_1+n_2+\cdots+n_J=S_J} \rho_1^{n_1} \rho_2^{n_2} \cdots \rho_J^{n_J} \right)^{-1}. \tag{5}$$

Let P_f be the steady-state probability that there are finished products in inventory at station J.

$$P_f = \sum_{n_1+n_2+\cdots+n_J<S_J} P(n_1, n_2, \ldots, n_J). \tag{6}$$

By PASTA property (see [13]), P_f can be regarded as the probability that an arriving customer demand can be satisfied immediately or, equivalently, the fill rate of customer demands. Denote λ_e the effective arrival rate of the customer demands, then $\lambda_e = P_f \lambda$.

For $j = 1, \ldots, J$, let L_{N_j} be the expected value of N_j. For $j = 1, \ldots, J-1$, let L_{B_j} be the expected value of B_j. Let L_{I_J} be the expected value of I_J. Then

$$L_{N_j} = \sum_{n_1+n_2+\cdots+n_J=S_J} n_j P(n_1, n_2, \ldots, n_J). \tag{7}$$

From (2) and (3),

$$L_{B_j} = \sum_{n_1+n_2+\cdots+n_J=S_J} (n_1+n_2+\cdots+n_j) P(n_1, n_2, \ldots, n_J), \tag{8}$$

$$L_{I_J} = \sum_{n_1+n_2+\cdots+n_J<S_J} (S_J - n_1 - n_2 - \cdots - n_J) P(n_1, n_2, \ldots, n_J). \tag{9}$$

The above results from exact analysis will be used to validate our following phase-type approximations for the system without intermediate inventories. The phase-type approximation is more applicable to a complex system then the traditional queuing model analysis. The following analysis can serve as a cornerstone for a future study in a more general system. We briefly introduce the approach in Lee and Zipkin [9], in which customer demands can wait for the finished products if there are no finished products in inventory. Then we present our approach using Lee and Zipkin's system to find the effective arrival rate of our system in which customer demands will be lost if there are no finished products.

We consider a system with intermediate inventories. Let K_j be the number of products needed to produce to fulfill the initial base-stock level at station j. In our system, one taken finished product will generate one request and it will be a backorder waiting for the finished semi-product (in B_j) from a precedent station or a production order is ready for processing (in N_j), therefore, K_j is the sum of those orders in B_j and N_j such that

$$K_1 = N_1 \tag{10}$$

and

$$K_j = B_{j-1} + N_j, \quad j = 2, \ldots, J. \tag{11}$$

From Neuts [14], we know that K_j is of discrete phase-type distribution with some initial distribution π_j and some transition rate matrix $\mathbf{P}_j$. We denote it by $K_j \sim DPH(\pi_j, \mathbf{P}_j)$.

As assumed by Lee and Zipkin [9] (see pp. 939–940), each station is assumed as an independent M/M/1 queue with arrival rate λ and service rate μ_j. The waiting time is of exponential distribution with rate $v_j = (\mu_j - \lambda)$. The $j \times j$ matrix $\mathbf{C}_j$ is defined as follows:

$$\mathbf{C}_j = \begin{bmatrix} -v_1 & v_1 & & & \\ & -v_2 & v_2 & & \\ & & \ddots & \ddots & \\ & & & -v_{j-1} & v_{j-1} \\ & & & & -v_j \end{bmatrix}, \quad j = 1, \ldots, J. \tag{12}$$

Then,

$$\mathbf{P}_j = \lambda(\lambda \mathbf{I} - \mathbf{C}_j) \tag{13}$$

and

$$\pi_j = \gamma_j \mathbf{P}_j. \tag{14}$$

$\mathbf{I}$ denotes an identity matrix and $\mathbf{e}$ denotes a column vector of ones. $\gamma_j,\ j = 2, 3, \ldots, J$, can be obtained recursively by letting

$$\gamma_1 = 1 \tag{15}$$

and

$$\gamma_{j+1} = \left[\gamma_j \mathbf{P}_j^{S_j}, \left(1 - \gamma_j \mathbf{P}_j^{S_j} \mathbf{e}\right)\right]. \tag{16}$$

Note that, $\gamma_{j+1} = [1, 0, \ldots, 0],\ j = 1, \ldots, J-2$ when $S_j = 0,\ j = 1, \ldots, J-1$.

Since $K_j \sim DPH(\pi_j, \mathbf{P}_j)$,

$$P(K_j \le k) = 1 - \pi_j \mathbf{P}_j^k \mathbf{e}, \quad k \ge 0, \tag{17}$$

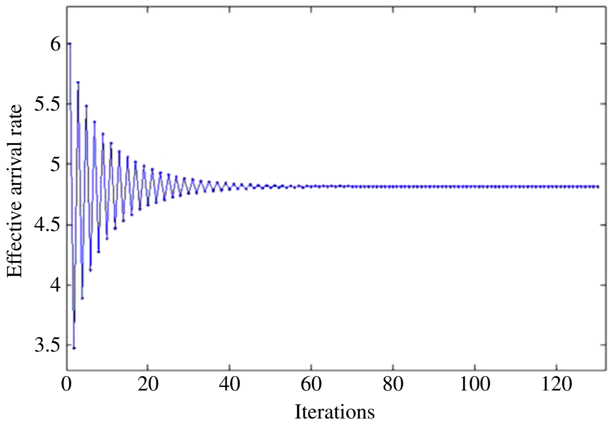

Fig. 2. The convergence of $\lambda_e^{(\ell)}$ ($\lambda = 6$, $\mu_1 = \mu_2 = \mu_3 = 10$ and $S_3 = 5$).

and the expected value of K_j, L_{K_j}, is given by

$$L_{K_j} = \pi_j(\mathbf{I} - \mathbf{P}_j)^{-1}\mathbf{e}. \tag{18}$$

The number of backorders at station j, B_j, has a shifted discrete phase-type distribution and

$$B_j \sim DPH(\pi_j \mathbf{P}_j^{S_j}, \mathbf{P}_j). \tag{19}$$

The expected value of B_j, L_{B_j}, is given by

$$L_{B_j} = \pi_j \mathbf{P}_j^{S_j}(\mathbf{I} - \mathbf{P}_j)^{-1}\mathbf{e}. \tag{20}$$

By Little's Law, we obtain the expected value of N_j at station j,

$$L_{N_j} = \frac{\lambda_e}{\mu_j - \lambda_e}, \quad j = 1, \ldots, J. \tag{21}$$

Finally, we obtain the expected value of I_J by using (1) such that

$$L_{I_J} = S_J - \sum_{j=1}^{J} L_{N_j}. \tag{22}$$

In Lee and Zipkin's system, there are no inventory ($I_J = 0$) if and only if $K_J \geq S_J$. Therefore, a customer demand will be satisfied by a finished product in inventory with probability $1 - P(K_j \geq S_J)$. In our loss system, we use it to approximate the fill rate P_f. The distribution of K_J will depend on the arrival rate of production orders to the first station, which is the effective arrival rates of customer demands λ_e. Let p_0 be the actual probability that there are no inventory in I_J in our system, therefore

$$\lambda_e = \lambda(1 - p_0). \tag{23}$$

The approach is to use $1 - P(K_J \geq S_J)$ obtained from Lee and Zipkin's system to approximate the fill rate and corresponding effective arrival rate of our system and recursively use the newly estimated effective arrival rate as the arrival rate to Lee and Zipkin's system to modify $1 - P(K_J \geq S_J)$ and the effective arrival rate until the estimated effective arrival rates converge.

The following is our approach for obtaining the estimation of λ_e. We denote $\lambda_e^{(\ell)}$ to be the approximated effective arrival rate obtained in the ℓth iteration.

Initial step. Let $\lambda_e^{(0)} = \lambda$.

General step ℓ. Let $\lambda_e^{(\ell-1)}$ be the arrival rate to Lee and Zipkin's system and use Eq. (15) and iterative Eqs. (12)–(14) and (16) to obtain the steady-state probability of no inventory $p_0^{(\ell)}$, where $p_0^{(\ell)} = P(K_J \geq S_J)$ (Eq. (17)). Let $\lambda_e^{(\ell)} = \lambda(1 - p_0^{(\ell-1)})$.

Stopping rule. If $|\lambda_e^{(\ell-1)} - \lambda_e^{(\ell)}| < \varepsilon$, then stop. ε is the tolerance error and $\varepsilon = 10^{-4}$ in this study.

Note that $\lambda_e^{(2\ell)}$, $\ell = 0, 1, 2 \ldots$ decrease and are bound from below by actual effective arrival rate λ_e and $\lambda_e^{(2\ell+1)}$, $\ell = 0, 1, 2 \ldots$ increase and are bound from above by λ_e. Fig. 2 shows the convergence of the effective arrival rates obtained by the approximation in a three-station system.

Table 2
Comparisons of exact and approximate effective arrival rates ($\lambda = 3$).

The number of stations	μ_1	μ_2	μ_3	μ_4	S_J	Exact	App.	%Error
2	6.5	6.5	–	–	2	2.250 (1.5)	2.202 (2.5)	−2.13
3	6.5	6.5	6.5	–	2	1.953 (2.5)	1.911 (7.0)	−2.15
4	6.5	6.5	6.5	6.5	2	1.716 (3.0)	1.764 (29)	2.80
2	6.5	6.5	–	–	4	2.784 (1.5)	2.703 (2.5)	−2.91
3	6.5	6.5	6.5	–	4	2.616 (2.5)	2.541 (5.0)	−2.87
4	6.5	6.5	6.5	6.5	4	2.439 (3.0)	2.346 (3.5)	−3.81
2	6.5	6.5	–	–	6	2.937 (1.5)	2.898 (2.5)	−1.33
3	6.5	6.5	6.5	–	6	2.865 (2.5)	2.775 (3)	−3.14
4	6.5	6.5	6.5	6.5	6	2.766 (3.0)	2.667 (2)	−3.58
2	6.5	6.5	–	–	8	2.982 (1.5)	2.967 (2.5)	−0.50
3	6.5	6.5	6.5	–	8	2.955 (2.5)	2.913 (3.0)	−1.42
4	6.5	6.5	6.5	6.5	8	2.910 (3.0)	2.847 (2.0)	−2.16
2	6.5	6.5	–	–	10	2.994 (1.5)	3.000 (1.5)	0.20
3	6.5	6.5	6.5	–	10	2.985 (2.5)	3.000 (1.5)	0.50
4	6.5	6.5	6.5	6.5	10	2.967 (3.0)	2.937 (1.0)	−1.01
2	6.5	6.5	–	–	15	3.000 (1.5)	3.000 (1.5)	0.00
3	6.5	6.5	6.5	–	15	3.000 (2.5)	3.000 (1.0)	0.00
4	6.5	6.5	6.5	6.5	15	3.000 (3.0)	3.000 (1.0)	0.00
2	7.5	7.0	–	–	2	2.340 (1.5)	2.289 (2.0)	−2.18
3	7.5	7.0	6.5	–	2	2.019 (2.5)	1.971 (4.0)	−2.38
4	7.5	7.0	6.5	6.0	2	1.746 (3.0)	1.788 (16.0)	2.41
2	7.5	7.0	–	–	4	2.838 (1.5)	2.775 (2.0)	−2.22
3	7.5	7.0	6.5	–	4	2.670 (2.5)	2.610 (3.0)	−2.25
4	7.5	7.0	6.5	6.0	4	2.469 (3.0)	2.373 (3.0)	−3.89
2	7.5	7.0	–	–	6	2.961 (1.5)	2.937 (1.5)	−0.81
3	7.5	7.0	6.5	–	6	2.895 (2.5)	2.850 (1.5)	−1.55
4	7.5	7.0	6.5	6.0	6	2.787 (3.0)	2.691 (1.5)	−3.44
2	7.5	7.0	–	–	8	2.991 (1.5)	3.000 (1.0)	0.30
3	7.5	7.0	6.5	–	8	2.967 (2.5)	3.000 (1.0)	1.11
4	7.5	7.0	6.5	6.0	8	2.919 (3.0)	2.862 (2.8)	−1.95
2	7.5	7.0	–	–	10	2.997 (1.5)	3.000 (1.0)	0.10
3	7.5	7.0	6.5	–	10	2.991 (2.5)	3.000 (1.0)	0.30
4	7.5	7.0	6.5	6.0	10	2.970 (3.0)	2.946 (1.0)	−0.81
2	7.5	7.0	–	–	15	3.000 (1.5)	3.000 (1.0)	0.00
3	7.5	7.0	6.5	–	15	3.000 (2.5)	3.000 (1.0)	0.00
4	7.5	7.0	6.5	6.0	15	3.000 (3.0)	3.000 (1.0)	0.00

3. Numerical results

In this section, we evaluate the effective arrival rate obtained from our approximation. The results are compared with those obtained from the exact solutions. We first compare both exact results and approximate results on the systems of various numbers of stations with no intermediate inventories. The results are included in Table 2. The column labeled "Exact" contains the results obtained from exact solutions and the column labeled "App." contains the results from our approximation. The corresponding computation times are recorded in the brackets. The percentage of relative errors compared with the exact solutions are shown in column "%Error". As we can see, although our approach underestimates the effective arrival rates, the relative errors for the approximate effective arrival rates are within 5%. The relative errors decrease as the base-stock levels increase and the relative errors increase as the number of stations increase. As for the computation time, both approaches have similar computation times. However, our approach has slightly less computation times on the system with more stations and higher base-stock level. This is because when stock level is high and effective arrival rate is close to the total arrival rate, it needs less recursive loops to have the approximate effective arrival rates converge. On the other hand, the procedure of product form solution depends on the size of the state space determined by the numbers of stations and the base-stock level, therefore, it is intuitive that the computation time of the product form procedure increases as the as numbers of stations and base-stock levels increase and are insensitive to parameters.

We now consider a three-station production system ($J = 3$) to study the influence of different permutations of service rates on the effective arrival rates. We consider two sets of service rates, (5.5, 6.0, 6.5) and (6.5, 7, 7.5), and three to five different base-stock levels (6 to 10). The results of effective arrival rates and queue lengths are shown in Tables 3 and 4, respectively. As we can see from these two tables, for different permutations of service rates the effective arrival rates are still the same under the same base-stock level and so are the expected number of finished products (L_{I_3}). The expected number of production orders at a station is more dependent on the service rate than on the permutation. However, if we locate a faster station in the front and a slower station in the back, then we can reduce the total expected number of backorders.

Table 3
Comparison of exact and approximate effective arrival rates with different permutations of service rates.

μ_1	μ_2	μ_3	S_3	Exact	App.	%Error
6.5	6.0	5.5	6	2.811	2.733	−2.77
			7	2.883	2.790	−3.23
			8	2.928	2.880	−1.64
6.0	5.5	6.5	6	2.811	2.733	−2.77
			7	2.883	2.790	−3.23
			8	2.928	2.880	−1.64
5.5	6.5	6.0	6	2.811	2.733	−2.77
			7	2.883	2.790	−3.23
			8	2.928	2.880	−1.64
7.5	7.0	6.5	6	2.895	2.850	−1.55
			7	2.943	2.907	−1.22
			8	2.967	3.000	1.11
7.0	6.5	7.5	6	2.895	2.850	−1.55
			7	2.943	2.907	−1.22
			8	2.967	3.000	1.11
6.5	7.5	7.0	6	2.895	2.850	−1.55
			7	2.943	2.907	−1.22
			8	2.967	3.000	1.11

Table 4
Comparison of queue-lengths of a three-station system.

μ_1	μ_2	μ_3	S_3	L_{N_1}		L_{N_2}		L_{N_3}	
				Exact	App.	Exact	App.	Exact	App.
6.5	6.0	5.5	6	0.717	0.726	0.817	0.837	0.947	0.988
			7	0.761	0.752	0.873	0.869	1.021	1.030
			8	0.793	0.796	0.914	0.923	1.076	1.100
			9	0.815	0.814	0.943	0.947	1.116	1.130
			10	0.831	0.857	0.963	1.000	1.142	1.200
6.0	5.5	6.5	6	0.817	0.834	0.947	0.985	0.716	0.724
			7	0.873	0.869	1.021	1.030	0.761	0.724
			8	0.914	0.923	1.076	1.100	0.793	0.796
			9	0.943	0.947	1.115	1.130	0.815	0.814
			10	0.963	1.000	1.144	1.200	0.831	0.857
5.5	6.5	6.0	6	0.947	0.988	0.716	0.726	0.816	0.837
			7	1.021	1.030	0.761	0.752	0.873	0.869
			8	1.076	1.100	0.793	0.796	0.914	0.923
			9	1.115	1.130	0.815	0.815	0.943	0.947
			10	1.144	1.200	0.830	0.857	0.963	1.000
μ_1	μ_2	μ_3	S_3	L_{B_1}		L_{B_2}		L_{I_3}	
				Exact	App.	Exact	App.	Exact	App.
6.5	6.0	5.5	6	0.717	0.726	1.534	1.563	3.519	3.449
			7	0.761	0.752	1.634	1.621	4.344	4.349
			8	0.793	0.796	1.707	1.720	5.216	5.180
			9	0.815	0.814	1.759	1.761	6.125	6.109
			10	0.831	0.857	1.794	1.857	7.062	6.943
6.0	5.5	6.5	6	0.816	0.834	1.764	1.820	3.518	3.457
			7	0.873	0.869	1.894	1.899	4.343	4.343
			8	0.914	0.923	1.990	2.040	5.216	5.181
			9	0.943	0.947	2.059	2.077	6.125	6.109
			10	0.963	1.000	2.107	2.200	7.060	6.943
5.5	6.5	6.0	6	0.947	0.988	1.664	1.714	3.518	3.449
			7	1.021	1.030	1.782	1.782	4.343	4.349
			8	1.076	1.100	1.869	1.896	5.216	5.181
			9	1.115	1.130	1.931	1.945	6.126	6.108
			10	1.144	1.200	1.974	2.057	7.062	6.943

In the following, we consider a three-station system in which every station has base stock and production is controlled by base-stock policy. It is the generic system discussed in Buzacott et al. [7] and Svoronos and Zipkin [8]. However, in our system customer demands will be lost if there are no finished products in inventory. There are no exact solutions for this system, therefore we compare our results with the results obtained from simulations. The comparisons are shown in Table 5. The column labeled "Sim". contains the results from simulations. The relative errors are with in 5% except the first case ($\mu_1 = \mu_2 = \mu_3 = 5$ and $S_1 = S_2 = S_3 = 2$).

Table 5
Comparison of effective arrival rates on systems with intermediate inventories.

S_1	S_2	S_3	$\mu_1 = \mu_2 = \mu_3 = 5.0$			$\mu_1 = 6.5,\ \mu_2 = 5.5,\ \mu_3 = 5.0$		
			Sim.	App.	%Error	Sim.	App.	%Error
2	2	2	2.367	2.187	−7.60	2.382	2.292	−3.78
6	2	2	2.388	2.289	−4.15	2.397	2.322	−3.13
2	6	2	2.436	2.367	−2.83	2.439	2.364	−3.08
2	2	6	2.886	2.748	−4.78	2.922	2.820	−3.49
6	6	2	2.445	2.364	−3.31	2.439	2.364	−3.08
6	2	6	2.901	2.802	−3.41	2.922	2.865	−1.95
2	6	6	2.937	2.883	−1.84	2.940	2.889	−1.73
6	6	6	2.943	2.886	−1.94	2.946	2.892	−1.83
S_1	S_2	S_3	$\mu_1 = 5.5,\ \mu_2 = 5.0,\ \mu_3 = 6.5$			$\mu_1 = 5.0,\ \mu_2 = 6.5,\ \mu_3 = 5.5$		
			Sim.	App.	%Error	Sim.	App.	%Error
2	2	2	2.514	2.406	−4.30	2.460	2.364	−3.90
6	2	2	2.535	2.436	−3.91	2.478	2.394	−3.39
2	6	2	2.613	2.541	−2.76	2.514	2.394	−4.77
2	2	6	2.937	2.886	−1.74	2.934	2.883	−1.74
6	6	2	2.613	2.544	−2.64	2.508	2.400	−4.31
6	2	6	2.949	2.904	−1.53	2.955	2.910	−1.52
2	6	6	2.976	3.000	0.81	2.967	3.000	1.11
6	6	6	2.979	3.000	0.70	2.976	3.000	0.81

Table 6
Parameters of the example of laptop computer assembly.

Station j	Steps	Exponential (μ_j)	Uniform (Min, Max)
1	Sales order	1.942	(0.41, 0.62)
2	Production planning	0.541	(1.48, 2.22)
3	Preparing materials	0.473	(1.85, 2.38)
4	Manufacturing	0.051	(17.66, 21.70)
5	Testing	0.231	(3.90, 4.76)
6	Shipping	0.136	(6.58, 8.08)

Unit time: hour.

Table 7
Comparison of fill rates.

S_6	Sim.	App.	%Error	S_6	Sim.	App.	%Error
1	0.558	0.513	−8.06	6	0.971	0.953	−1.85
2	0.763	0.723	−5.24	7	0.984	0.971	−1.32
3	0.857	0.823	−3.95	8	0.995	0.988	−0.70
4	0.915	0.887	−3.10	9	1.000	1.000	0.00
5	0.945	0.920	−2.64	10	1.000	1.000	0.00

We also reconsider the example of laptop computer assembly mentioned in Section 1. We estimate fill rate using our approach and compare it with those obtained from simulation. In simulations, we use a uniform distribution to simulate the operating time at each step with two end points equal to the maximum and minimum of the operating time shown in Table 1. The mean of the exponential service time at each station is set to be equal to the median of the operating time. The corresponding parameters are in Table 6. Customer demands arrive to this system 0.03 per hour on average. The comparisons between the results from the simulations and our approach on fill rates are presented in Table 7 and Fig. 3. Our results are close to simulations, although the relative errors for cases with $S_6 = 1$ and 2 are more than 5%. The errors decrease as base-stock levels increase. This implies that if a high fill rate is required and we need to set a higher base-stock level, our model can deliver a relatively accurate approximation. Suppose the required fill rate is 95%, then, from Table 7, the minimum base-stock level should be set to 6.

We further consider an example with a cost structure. The associated costs are described as follows. C_P is the penalty cost for each unsatisfied customer demand; C_{N_j} is the cost for each production order in N_j per unit time, $j = 1, \ldots, J$; C_{B_j} is the cost for each backorder in B_j per unit time, $j = 1, \ldots, J-1$; C_{I_J} is the holding cost for each finished product in I_J per unit time. Then, the total expected cost denoted by TC is defined as

$$\mathrm{TC} = C_P(1 - P_f)\lambda + \sum_{j=1}^{J} C_{N_j} L_{N_j} + \sum_{j=1}^{J-1} C_{B_j} L_{B_j} + C_{I_J} L_{I_J} \quad (24)$$

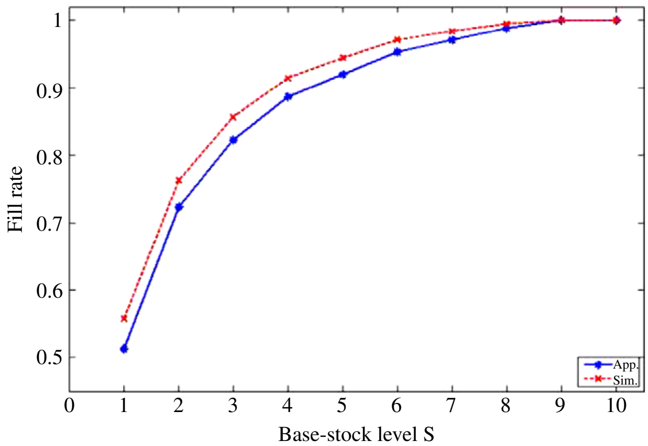

Fig. 3. Comparisons on fill rate.

Table 8
Fill rates with corresponding base-stock levels.

S_3	P_f	S_3	P_f
1	0.513	6	0.957
2	0.695	7	1.000
3	0.807	8	1.000
4	0.879	9	1.000
5	0.918	10	1.000

Table 9
TC on various base-stock levels.

S_3	TC	S_3	TC	S_3	TC
1	28.703	6	17.065	11	22.802
2	20.807	7*	16.845	12	24.301
3	18.871	8	18.322	13	25.801
4	17.328	9	19.811	14	27.300
5	17.155	10	21.305	15	28.800

We need to find an optimal base-stock level with minimum total cost subject to a required minimum fill rate. Let *SL* be the minimum fill rate (service level) required, then the optimization problem is

$$\begin{aligned}&\text{Min TC}\\&\text{s.t.}\quad P_f \geq SL.\end{aligned}$$

We consider a three-station production system as an example. Let $\lambda = 3$, $\mu_1 = 7$, $\mu_2 = 8$, and $\mu_3 = 7$. Suppose that the fill rate of customer demands is required to be at least 90%. The corresponding results on various base-stock levels are presented in Table 8. In order to satisfy the requirement, the base-stock level must be set at least to 5. As we can see from the table, P_f converges to 1 as S_3 gets larger. It is because the large base-stock level leads to more finished products in inventory and more undergoing production of new products and ensures that an arriving customer demand will have a greater chance to obtain a finished product from the inventory.

Let $C_P = \$50$, $C_{N_1} = C_{N_2} = C_{N_3} = \$2.5$, $C_{B_1} = C_{B_2} = \$2$ and $C_{I_3} = \$1.5$. The results for the total expected costs on various base-stock levels are shown in Table 9. Since the requirement of the minimum fill rate is 90%, we consider only the base-stock level greater than or equal to 5. Therefore, the optimal base-stock level is 7 as indicated by an asterisk in Table 9.

4. Conclusions

In this study, we consider a practical serial production system that can be seen in laptop computer assembly. The main contribution of this study is that the fill rate can be estimated by using phase-type approximation. We validate our results by comparing with the exact solutions for the system without intermediate inventories and with simulations for general systems. Our approximation provides an accurate estimation. The study here using phase-type approximation can also be a cornerstone to further research on multiple-product and multiple-demand production systems.

References

[1] H.L. Lee, C. Billington, Material management in decentralized supply chains, Operations Research 41 (5) (1993) 835–847.
[2] I. Karl, M. Stefan, Safety stocks in multi-stage inventory systems under different service measures, European Journal of Operational Research 106 (1) (1998) 57–73.
[3] I. Duenyas, P. Patana-Anake, Base-stock control for single-product tandem make-to-stock systems, IIE Transactions 30 (1998) 31–39.
[4] T. Boyaci, G. Gallego, Serial production/distribution systems under service constraints, Manufacturing & Service Operations Management 3 (2001) 43–50.
[5] K.H. Shang, J.S. Song, A closed-form approximation for serial inventory systems and its application to system design, Manufacturing & Service Operations Management 8 (4) (2006) 394–406.
[6] X.L. Chao, S.X. Zhou, Probabilistic solution and bounds for serial inventory systems with discounted and average costs, Naval Research Logistics 54 (2007) 623–631.
[7] J.A. Buzacott, S.M. Price, J.G. Shanthikumar, Service level in multistage MRP and base stock controlled production systems, in: G. Fandel, T. Gulledge, A. Jones (Eds.), New Directions for Operations Research in a Manufacturing Systems, Springer, New York, 1992, pp. 445–464.
[8] A. Svoronos, P. Zipkin, Evaluation of one-for-one replenishment policies for multi-echelon inventory systems, Management Science 37 (1) (1991) 68–83.
[9] Y.J. Lee, P. Zipkin, Tandem queues with planned inventories, Operations Research 40 (5) (1992) 936–947.
[10] P. Zipkin, Processing network with inventories: tandem queues with feedback, European Journal of Operational Research 80 (2) (1995) 344–349.
[11] C. Duri, Y. Frein, M.D. Mascolo, Performance evaluation and design of base stock systems, European Journal of Operational Research 127 (1) (2000) 172–188.
[12] F.F. Wang, C.T. Su, Performance evaluation of a multi-echelon production, transportation and distribution system: a matrix analytical approach, European Journal of Operational Research 176 (2) (2007) 1066–1083.
[13] R.W. Wolff, Poisson arrivals see time averages, Operations Research 30 (2) (1982) 223–231.
[14] M. Neuts, Matrix-Geometric Solutions in Stochastic Models, Johns Hopkins University Press, Baltimore, 1981.